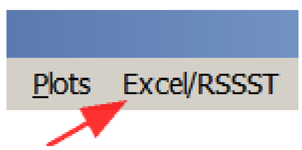
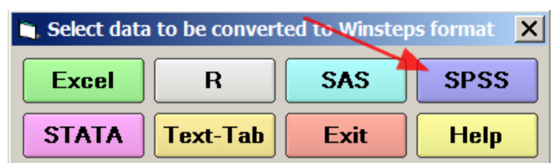
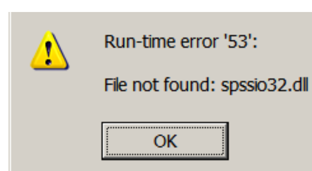
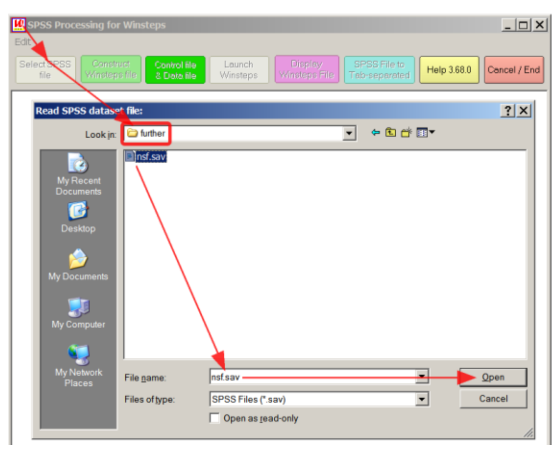
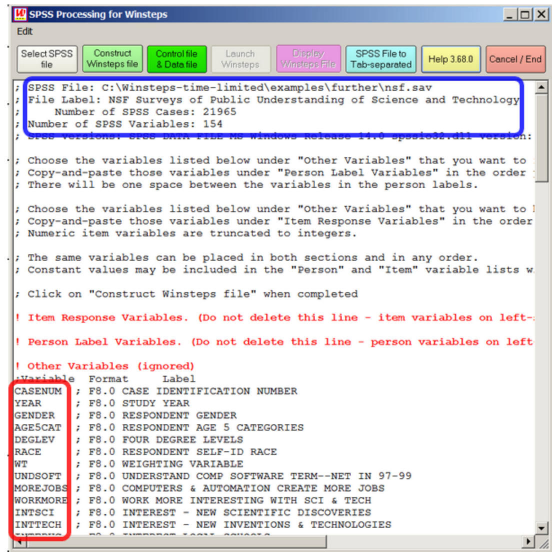
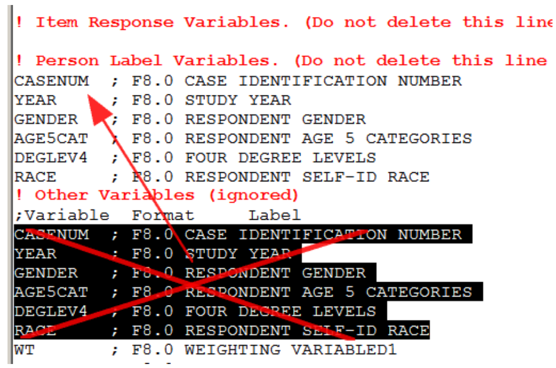
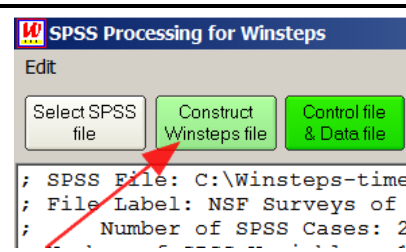
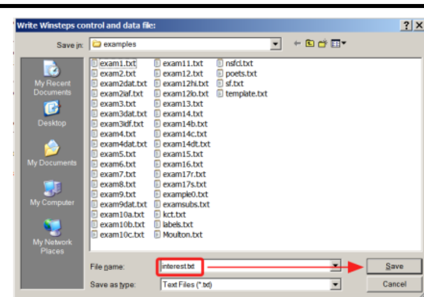
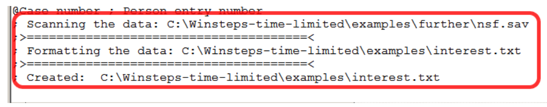
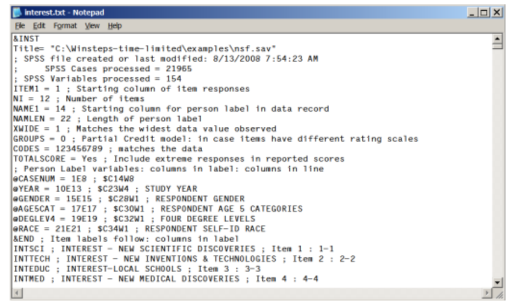
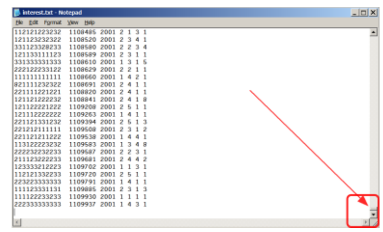
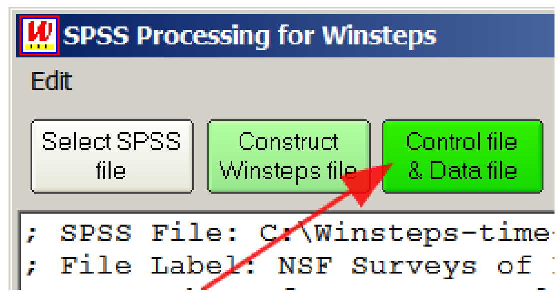
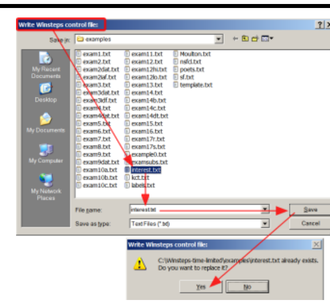
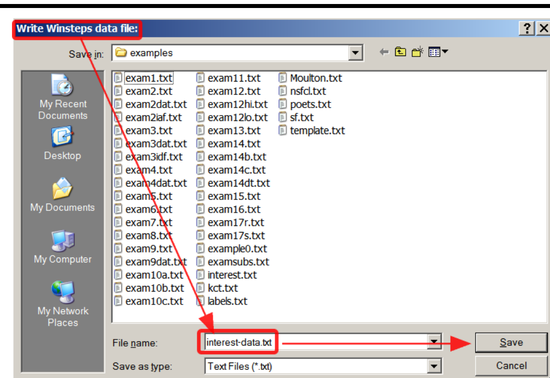
#	**Practical Rasch Measurement - Further Topics : www.winsteps.com** **Mike Linacre, instructor - July 2011**	
1.	**Tutorial 2. Polytomous analysis:** ***This is the most difficult week, but it is very practical!*** • Importing data from SPSS, SAS, Stata, Excel • Rescoring data • Polytomous models • Polytomous estimation • Average measures and category thresholds If you don't know the meaning of a word, then please look at the "Glossary" Lesson.	
2.	**A. Data from SPSS (.sav, not .spv) (SAS, Stata, Excel, and more)**	
3.	We are going to perform an exploratory analysis of the *"National Science Foundation Surveys of Public Understanding of Science and Technology, 1979-2001"* We will extract the actual, real data from the original SPSS file. *No faking here!* This will be an adventure!	
4.	Launch *Winsteps*	
5.	On the Winsteps menu bar, Click on "Excel/RSSST" Click on "SPSS"	 
6.	*If you see this error message,* *Browse to "c:\Winsteps"* *Rename "spssio32.hld" to "spssio32.dll"* *then try again ....*	
7.	"SPSS Processing for Winsteps" window displays. ***SAS, Stata and Excel** follow this same procedure.* *Please try this with your own data. Any questions or problems? - Winsteps Help* Click on "Select SPSS file" "Read SPSS dataset file" displays *The SPSS dataset files have suffix .sav (If suffixes don't display for you, see Lesson 1 Appendix 2.)* Double-click on *further* folder Double-click on nsf.sav	

8.	Your “SPSS” window should now display: Summary details of the SPSS file display (in the blue box). There are 21,965 cases (persons). The list of SPSS variables displays (in the red box). This SPSS data file contains 154 of the original 278 National-Science-Foundation Survey variables. There are 213 variables. Some are demographic, such as “YEAR”, the year the survey was conducted. Most are survey items, such as “INTSCI”, respondent’s interest in scientific discoveries.	

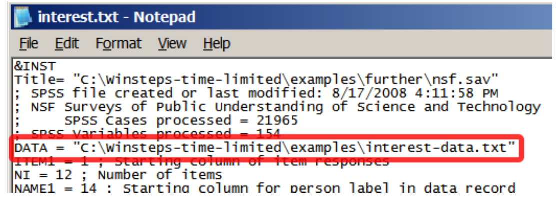
9.	Let’s select some demographic variables to become the Winsteps person labels.
	Cut-and-paste from “! Other Variables”
	`CASENUM   ; F8.0 CASE IDENTIFICATION NUMBER`
	`YEAR      ; F8.0 STUDY YEAR`
	`GENDER    ; F8.0 RESPONDENT GENDER`
	`AGE5CAT   ; F8.0 RESPONDENT AGE 5 CATEGORIES`
	`DEGLEV4   ; F8.0 FOUR DEGREE LEVELS`
	`RACE      ; F8.0 RESPONDENT SELF-ID RACE`
	
	Paste under “! Person Label Variables”
		

10.	Let’s select some item response variables: Copy-and-paste 6 “INTEREST” variables and 6 “INFORMED” variables from “! Other Variables” Paste under “! Item Response Variables”	
11.	Now we will extract the SPSS data into a Winsteps Control and data file: Click on “Construct Winsteps file”	
12.	Save the new Winsteps Control and Data file. I am calling mine: “interest.txt” in folder “examples”	

13.	The bottom of the SPSS window shows the stages of constructing the Winsteps file.	
14.	The Winsteps file displays in NotePad. You can see that it has the familiar format: Winsteps control variables at the top &END Item labels END NAMES Data	
15.	Scroll down to the bottom of the NotePad file. You can see that it is a long file, which we need to edit. So it will be easier, and less error-prone, if we split the Winsteps control variables and the data into separate files.	
16.	Close the NotePad window	
17.	In the SPSS window, click on "Control file & Data file"	
18.	First, the file for the control variables: "Write Winsteps control file:" box Click on "interest.txt" Click on "Save" Replace? Click on "Save"	
19.	Then the file for the data: "Write Winsteps data file:" File name: "interest-data.txt" Save	

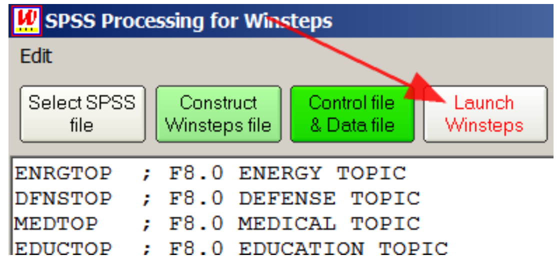
20. The control file displays in NotePad:
It includes the DATA= variable for the data file

Remove the **;** before **ISGROUPS=0**



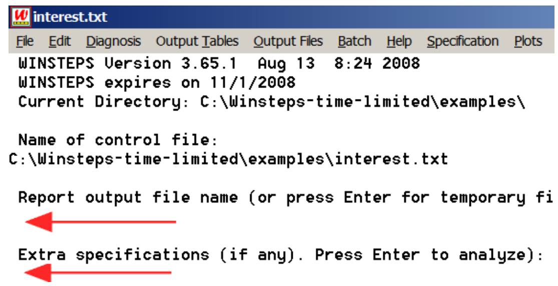
21. Let's see what happens when Winsteps gets to work on this!
In the Winsteps SPSS window,
click on "Launch Winsteps"

This is exciting! I wonder what will happen ...

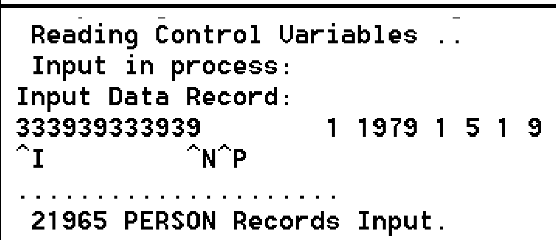


22. Winsteps launches:
"Report output file name..."
Press Enter
"Extra specifications ...
Press Enter

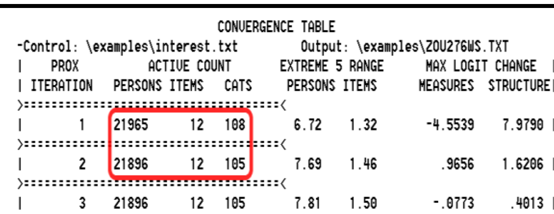
Winsteps starts processing ....



23. The first data record is shown. It looks somewhat unusual: **3's and 9's**, but ^I and ^N tell us that they are the item responses. ^P tells us where the person label starts.
There is a line of "....................." = 21 "."
Each "." means 1,000 cases is being processed. And, as expected, there are 21,965 person records.



24. The **Convergence Table** also has some curious features: 21965 persons, good! 12 items - yes, we selected 12 items. 108 CATS (rating scale CATegorieS). For dichotomous data we saw only 2 CATS. These data are polytomous. PROX iterations give us initial estimates for the JMLE estimation.



25. Then 21965 reduces to 21896. Winsteps has discovered that 69 persons had extreme scores. They do not provide useful information for comparing item difficulties, but they will be included in the reports. 108 categories reduces to 105 categories. 3 extreme categories (top or bottom rating-scale-categories) contained no observations and so were dropped from the analysis.

26. This dataset contains surveys collected over 22 years.
Green box: “Checking connectivity” confirms that the dataset is one cohesive unit.
If every person responds to every item on your test. Then "connectivity" is always perfect. The data are one unit.
But our data may be from a computer-adaptive test, or it may be several datasets from different tests combined into one analysis. Every person did not respond to every item. Now we need to check that there is overlap between the persons and the items. If there is overlap, then all the measures can be compared, so that there is "one cohesive unit".
But perhaps all the boys responded to items 1-10, and all the girls responded to items 11-20, but there is no overlap. Then the "boys" analysis is a separate unit from the "girls" analysis. Winsteps will warn us when this “disconnection” has happened. When this happens, the measures for the boys cannot be compared to the measures for the girls, and the difficulties of items 1-10 cannot be compared to the difficulties for items 11-20.

```
Checking connectivity ...
>=====================================<
|Control: \examples\interest.txt        Output: \examples\ZOU276WS.TXT
|    JMLE     MAX SCORE   MAX LOGIT    LEAST CONVERGED      CATEGORY STRUCTURE|
| ITERATION   RESIDUAL*    CHANGE    PERSON ITEM     CAT   RESIDUAL    CHANGE |
>=====================================<
|       1     12166.25       .7484   6107      4*      0   -4393.23  -1.5016|
>=====================================<
......|
|      20       145.64      -.0053   4840      4*      0     -42.65   -.0041|
>=====================================<
|      21       131.49      -.0047   4840      4*      0     -40.33   -.0040|
--------------------------------------------------------------------------------
 Calculating Fit Statistics
>=====================================<
Standardized Residuals N(0,1) Mean: .00 S.D.: 1.02
```

To see what a disconnected analysis looks like, analyze Examsubs.txt with Winsteps

27. *Red box:* The JMLE estimation process converges (stops) when the maximum logit change is too small to see in the output, less than .005 logits. This is much more exact than we need for our current explorations.

*If your **computer is slow,** then press **Ctrl+F** to stop the estimation process, and then skip to the next stage. Your results will be almost the same as mine.*

28. *Blue box:* Do you remember the “standardized residuals” in the spreadsheet at the end of Lesson 1? An assumption of JMLE estimation is that these are sampled from a unit-normal distribution, N(0,1) - *if “N(0,1)” looks strange to you, then please refer to Lesson 1 Appendix 7.* We have (0, 1.02) - reasonable values, very close to (0,1)

29. Beneath the Convergence Table, summary statistics for this analysis are shown.
Red box: The person “separation” is 2, corresponding to a person “test” reliability (like Cronbach Alpha, KR-20) of 0.8. This indicates that this test can distinguish between high and low performers (2 performance levels) in the sample.
Blue box: The item separation is 91 (huge), with reliability 1.00 (perfect). With this large person sample, the item difficulties are estimated exceedingly precisely.

```
C:\Winsteps-time-limited\examples\nsf.sav
--------------------------------------------------------------------------------
| PERSONS   21965 INPUT   21965 MEASURED              INFIT         OUTFIT    |
|           SCORE     COUNT     MEASURE   ERROR     IMNSQ  ZSTD  OMNSQ   ZSTD|
| MEAN       30.6      12.0       -1.31     .35       .65   -.3    .92     .0|
| S.D.       14.1        .0         .89     .18       .54    .8    .87     .9|
| REAL RMSE    .40  ADJ.SD      .80  SEPARATION  2.01  PERSON RELIABILITY  .80|
|-------------------------------------------------------------------------------|
| ITEMS      12 INPUT      12 MEASURED                INFIT         OUTFIT    |
| MEAN   55933.9   21965.0         .00     .01       .82  -2.3   1.05    3.6|
| S.D.   17999.0        .0         .80     .00       .33   8.0    .22    6.3|
| REAL RMSE    .01  ADJ.SD      .80  SEPARATION 90.80  ITEM   RELIABILITY 1.00|
--------------------------------------------------------------------------------
```

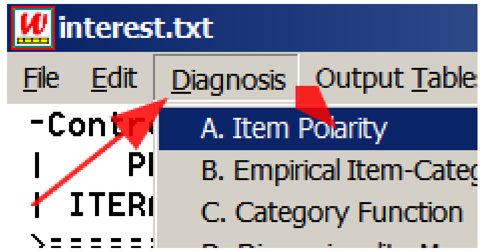
30. If the person separation and reliability are too low:
1. Increase the test length, the number of items in the test.
2. Increase the ability range of the sample being tested.
3. Increase the number of response categories per item.

If item separation and item reliability are too low:
1. Increase the person sample-size. *Test more people!*

31. *For more about “Separation” and “Reliability”, see* Appendix 1. Reliability and Separation Statistics.

32. **B. Diagnosis Menu A. Item Polarity**

33. *Oh, no! Not again ....* ***Yes, every time ....***
If the items do are not all pointing in the same direction, then the latent trait is not going anywhere.
Winsteps Analysis Window
Winsteps menu bar
Click on "Diagnosis"
Click on "A. Item Polarity"



34. Table 26.1 gives us valuable insights into the data.
Green box: the sample size is so large, that the reported measure standard errors are effectively zero (extremely high precision). This is like measuring your own weight to the nearest gram. It is true as an instantaneous weight, but for practical purposes, we know that the measure will rarely be exactly this value again.

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	DISPLACE	ITEM
3	36243	21965	.73	.01	1.27	9.9	1.44	9.9	.26	.44	47.1	53.5	-.01	INTE
9	41750	21965	1.00	.01	1.25	9.9	1.34	9.9	.29	.47	48.4	52.8	-.01	INFE
1	36856	21965	1.10	.01	1.10	5.0	1.12	9.9	.38	.43	58.3	57.2	-.01	INTS
2	37764	21965	.56	.01	1.09	4.3	1.13	9.9	.38	.43	59.4	58.1	-.01	INTT
8	49002	21965	.19	.01	1.05	2.5	1.13	9.9	.42	.46	55.9	58.9	-.01	INFT
7	47628	21965	.80	.01	1.01	.6	1.08	6.5	.45	.46	61.3	59.5	-.01	INFS
5	48159	21965	-.16	.01	.70	-9.9	1.05	3.1	.48	.50	51.3	51.2	-.01	INTS
11	54176	21965	-.34	.01	.65	-9.9	.96	-2.2	.49	.50	47.8	53.9	-.01	INFS
6	72503	21965	-.83	.00	.42	-9.9	.94	-1.1	.67	.65	46.2	44.7	.00	INTD
12	79914	21965	-1.00	.00	.43	-9.9	.75	-6.2	.69	.66	47.0	43.9	.00	INFD
4	78976	21965	-.91	.00	.43	-9.9	1.07	.7	.70	.67	57.7	54.5	.00	INTM
10	88236	21965	-1.13	.00	.43	-9.9	.56	-6.7	.72	.68	52.4	48.7	.00	INFM
MEAN	55933.9	21965	.00	.01	.82	-2.3	1.05	3.6			52.7	53.1		
S.D.	17999.0	.0	.80	.00	.33	8.0	.22	6.3			5.3	5.0		

35. *Orange boxes: For more about ZSTD, see* Appendix 2. Computing INFIT and OUTFIT "ZSTD" Fit Statistics. The extreme standardized statistics reflect the huge sample size. We are absolutely certain that these data do not fit the Rasch model (perfectly), the null hypothesis of "these data fit the Rasch model" can definitely be rejected. But the mean-squares are not excessive, so these data support measurement. The size of the departure in the data from the Rasch model is not overwhelming. It is a small departure, but we are certain that it didn't happen due to the randomness predicted by the Rasch model.
Red box: The point-measure correlations are all positive. Good! All the items seem to be pointing in the same direction.
Pink box: The observed correlations differ somewhat from their prediction, but the Infit and Outfit mean-squares tell us this is not a serious cause of concern.
Blue box: The "displacement" warns us that the estimates of some item difficulties may be .01 logits from the exact JMLE values. This is a tiny amount, but if we are making decisions based on the tiny S.E.s then we need to eliminate the displacement. We can do this by tightening the convergence criteria: LCONV=, RCONV=, CONVERGE=.

36. But, before taking any action, look at Table 26.3.
Red box: the scores on the item are 1,2,3 7,8. We expect a rating scale to go 1,2,3,4,5,6,7,8. *Why is there a gap?*
Blue boxes: The top category (8) has a lower average measure (-.13) than category 7, and the top category also has a huge misfit (outfit mean-square = 3.1). *What is wrong?*
Orange box: It would be helpful if each option had a description to help us interpret these data.

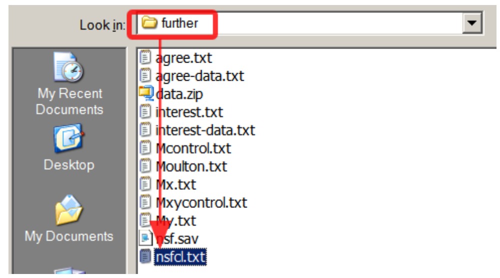
ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	DATA %	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM
3	1	1	11128	51	-1.53	.01	1.2	-.23	INTEDUC
	2	2	7533	34	-1.21	.01	1.6	.09	
	3	3	3275	15	-.89	.01	1.4	.20	
	7	7	8	0	.56	.34	1.0	.04	
	8	8	21	0	-.13*	.27	3.1	.04	

37. C. Category Label File CLFILE=

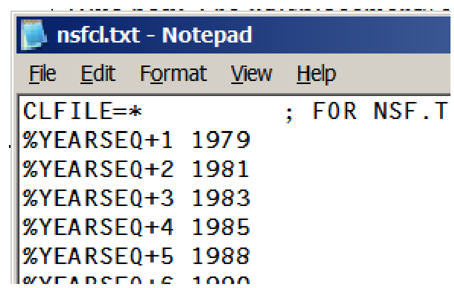
38. Yes, each option does have a description in the NSF documentation *usnsf2001-science.pdf*
We can now see what 1,2,3,7,8,9 mean for this item. Let's add this information to the Winsteps output.

```
INTEDUC     INTEREST-LOCAL SCHOOLS
            Value       Label
             1.00       Very Interested
             2.00       Moderately Interested
             3.00       Not Interested
             7.00       WON'T SAY
             8.00       DON'T KNOW
             9.00       NOT ASKED THAT YEAR
```

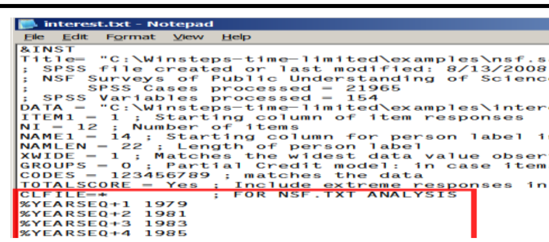
39. There are about 50 pages of details in the NSF document, please extract it from www.winsteps.com/a/further-data.zip into a folder called c:\Winsteps\further
Double-click on: nsfcl.txt



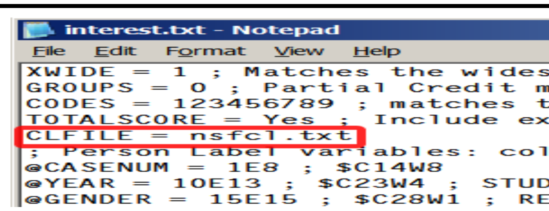
40. nsfcl.txt display in NotePad. It is a text file:
CLFILE=* starts a "Category Label" list. This is a Winsteps variable. "=*" states that a list of values follows.
%YEARSEQ means "this is for the item labeled *YEARSEQ"*
+1 means "this is for the category coded *1"*
1979 means "the label for this category is *1979"*
nsfcl.txt contains 885 category labels like the first one.
The list ends with "*".



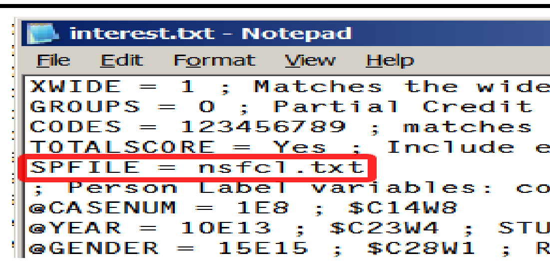
41. There are three ways we can incorporate this information into a Winsteps control file:
either: *(don't do this now!)* 1. We can copy-and-paste all 887 lines of this file into the Winsteps control file. CLFILE= can go anywhere in the control file before &END. But this is awkward.



42. or: *(don't do this now!)* 2. We can comment out the first and last lines in nsfcl.txt
*;CLFILE=**
*;**
Then put this control variable in the Winsteps file:
CLFILE = nsfcl.txt

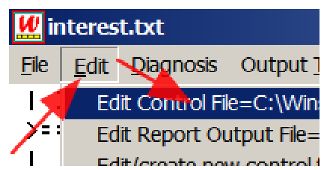
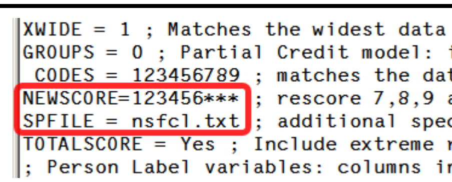
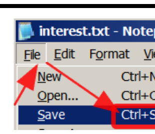
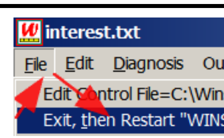
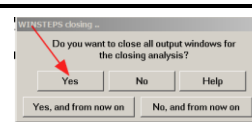


43. or: ***(Yes! Do this now!)*** 3. Make no changes to nsfcl.txt
Then put this control variable in the Winsteps file:
SPFILE = c:\Winsteps\further\nsfcl.txt
This tells Winsteps that nsfcl.txt is an additional specification-file containing more Winsteps control variables. When Winsteps read nsfcl.txt, then it will discover the control variable CLFILE=*



44.	D. Valid codes in the data: CODES= and NEWSCORE=	
45.	Did you notice what 7, 8, 9 mean for this item? "Won't say", "Don't know", "Not asked".	NSF state: Missing value codes are typically 7, 8, 9,
46.	Currently: "CODES = 123456789" - this means "process the data as observations on a 9-category rating scale in which each ascending number represents a higher qualitative level of the latent trait. But 7,8,9 are not on the intended latent trait of "Scientific Interest and Information" **Winsteps has been analyzing the wrong latent trait!** With the current scoring, higher categories mean "No meaningful response". So the reported latent trait is "Meaninglessness of the responses". The higher the person raw score, the more meaningless are the person's responses!	
47.	We need to tell Winsteps to ignore data which does not contribute to measurement of the intended latent trait. One way is to omit the unproductive data codes:	`CODES = 123456`

48.	But we may want summary statistics for each of 7,8,9. So another approach is to recode 7,8,9 into missing data. The 7,8,9 in CODES= will cause them to be reported in the Option/Distractor Table 26.3. But the non-numeric values in NEWSCORE= ("*"), will cause 7,8,9 to be treated as missing data, and ignored for estimation purposes.	`CODES = 123456789`
		`NEWSCORE = 123456***`
		The code and its scoring align vertically

49.	Now to edit our Winsteps control file, Interest.txt: Winsteps Analysis Window Click on Edit menu Click on Edit control file	
50.	Interest.txt displays in a NotePad window Edit into the file, at a convenient location before &END: NEWSCORE=123456*** SPFILE = nsfcl.txt	
51.	Save the NotePad file. Ctrl+S works well! We have revised the Control file. We want to analyze it.	
52.	Winsteps Analysis window: Click on File menu Click on "Exit and Restart ..."	
53.	Close all windows from the previous analysis of Interest.txt. There is nothing we want to look at later.	

54. Run Winsteps in the usual way. *Yes, you know what to do!* Look at the Convergence Table.
As before, we start with 21965 persons. It reduces to 21836. This time 129 persons are discovered to have extreme scores.
And the number of active rating-scale categories (across all items) started at 72 (= 6 categories for 12 items). It has dropped to 42 because 30 categories were unused. (Remember that earlier it was 108).

```
|     PROX         ACTIVE COUNT        E>
| ITERATION   PERSONS ITEMS   CATS      F
>=======================================<
|         1   21965      12    72
>=======================================<
|         2   21842      12    42
>=======================================<
|         3   21836      12    42
\---------------------------------------/
```

55. The person separation has dropped from 2 to 1.5, and the reliability from .80 to .70. We expected this, because: "fewer categories → lower precision → lower reliability"

```
 ---------------------------------------------
 ASURED               INFIT        OUTFIT    |
 SURE    ERROR     IMNSQ  ZSTD  OMNSQ   ZSTD|
 1.39      .63      1.01    -.1   1.02    -.1|
 1.32      .36       .54    1.3    .61    1.3|
   SEPARATION  1.52  PERSON RELIABILITY   .70|
```

56. Onwards to Diagnosis Menu: A. Item Polarity. Table 26.1.
Red box: The point-measure correlations are all nicely positive. So the scoring of our items is oriented with our new latent trait.
Blue box: "G" refers to item-grouping. "0" means "this item is in a group by itself." This reminds us that the control file contains "GROUPS=0" or "ISGROUPS=0" (they mean the same thing), specifying the Rasch "Partial Credit Model"

```
---------------------------------------------------------
|PT-MEASURE |EXACT MATCH|                              |
|CORR.  EXP.| OBS%  EXP%| ITEM                    G    |
+-----------+-----------+------------------------------|
|  .39   .55| 50.3  58.6| INTEDUC                 0    |
|  .42   .55| 50.8  56.8| INFEDUC                 0    |
|  .46   .46| 72.5  72.2| INTMED                  0    |
|  .46   .52| 56.8  59.4| INTDFNS                 0    |
|  .48   .53| 57.0  58.5| INFDFNS                 0    |
|  .53   .51| 64.3  63.6| INFMED                  0    |
|  .58   .54| 59.0  57.5| INTSPACE                0    |
|  .59   .52| 66.1  62.9| INFTECH                 0    |
|  .59   .54| 65.5  62.2| INTTECH                 0    |
|  .59   .52| 64.9  61.5| INFSPACE                0    |
|  .60   .54| 64.8  61.2| INTSCI                  0    |
|  .62   .52| 67.9  63.7| INFSCI                  0    |
+-----------+-----------+------------------------------|
```

57. Scroll down to Table 26.3. Amazing progress!
Blue box: 7,8 are scored "***", which means "ignored".
Notice that only categories 7, 8, 1, 2, 3 are used for this item. Categories 4, 5, 6, 9 are not used. They were dropped from the CATS count.
Red box: The average measures advance with the meaningful categories, 1, 2, 3. Good!
Orange box: Outfit mean-squares for the categories. The data are noisy, but not excessively (OUTF MNSQ<2.0).

```
---------------------------------------------------------------
|ENTRY  DATA  SCORE |   DATA    | AVERAGE  S.E.  OUTF PTMEA|         |
|NUMBER CODE  VALUE |  COUNT  % | MEASURE  MEAN  MNSQ CORR.| ITEM    |
|-------------------+-----------+--------------------------+---------|
|    3  7       *** |     8   0*|   -1.75  1.24        .00 |INTEDUC  | 7 WON'T SAY
|       8       *** |    21   0*|    -.74   .44        .01 |         | 8 DON'T KNOW
|       1         1 | 11128  51 |   -1.85   .01   1.3 -.33 |         | 1 Very Interested
|       2         2 |  7533  34 |   -1.25   .01   1.6  .09 |         | 2 Moderately Interested
|       3         3 |  3275  15 |    -.30   .03   1.7  .34 |         | 3 Not Interested
```

58. *Purple box:* A correlation is computed for each category by scoring that category "1", and all the other categories "0". The lowest category, 1, has a negative correlation. The middle category, 2, a zero correlation. The top category, 3, a positive correlation. Categories, 7, 8, which are not expected to relate to the new latent trait, have correlation zero. Exactly right! *Please ask if you don't understand why ...*
Yellow box: labels are shown for the categories. Oops! Look closely at the wording. Category 1 is "Very Interested". The data are coded backwards! We expect "more score → more interest", but NSF have coded the data "more score → less interest". When we try to explain our findings, **this reverse scoring is going to confuse everyone.** *We must reverse the scoring!*

59. Table 26.3. Next item: 9. INFEDUC
Red box: Category 4 with 1 observation and no label.
Diagnosis: ***Data entry error at NSF!***
Comment: 40+ years of experience with computers have demonstrated to me that even the most carefully screened data files can contain garbage. So this is no surprise.

```
    9  7       *** |    11   0*|   -1.56  1.21        .00 |INFEDUC  | 7 WON'T SAY
       9       *** |     1   0*|    -.54              .00 |         | 9 NOT ASKED THAT YEAR
       8       *** |    33   0*|    -.22   .44        .03 |         | 8 DON'T KNOW
       1         1 |  7265  33 |   -2.09   .01   1.2 -.35 |         | 1 Very Well Informed
       2         2 |  9831  45 |   -1.35   .01   1.4  .04 |         | 2 Moderately Informed
       3         3 |  4823  22 |    -.52   .02   1.4  .35 |         | 3 Poorly Informed
       4         4 |     1   0 |    -.09          .1  .01 |         | 4
```

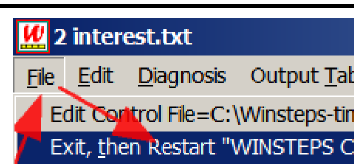
60.	Look at the category labels in Table 26.3. There are two rating scales: *Informed* and *Interested.* Our audience will be puzzled if we squash the two rating scales together. But our audience will be overwhelmed if we try to communicate a different version of *"Informed"* for each of its 6 items, and a different version of *"Interested"* for each of its 6 items. As Albert Einstein said, *"Science should be as simple as possible, but no simpler."* So we need to estimate two rating scales (one for each cluster of 6 items), not one for all 12 items, and not 12 rating scales, one for each individual item..	1 Very Well Informed 2 Moderately Informed 3 Poorly Informed 1 Very Interested 2 Moderately Interested 3 Not Interested

61.	**E. Complex data recoding: IREFER=, IVALUE= and ISGROUPS=**	

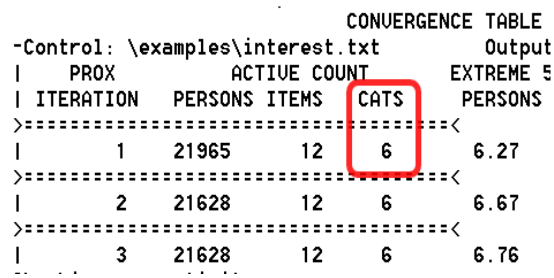
62.	Our first attempt at recoding the data using NEWSCORE= was productive, but now we see that it is not enough.
	We can see that in these data, only 1,2,3 are valid. And we want to reverse them so that “more score → more interest”
		This will do the job:
		`CODES = 123456789`
		`NEWSCORE = 321******`

63.	But we may discover we need to recode the two rating scales differently, so we will prepare for that: Items 1-6 are “InTerest” items, let’s call them T-type items. Items 7-12 are “InFormation” items, let’s call them F-type items.	For data rescoring: IREFER = TTTTTTFFFFFF Each item is one character. Items 1-6 are T-type “inTerest” items. Items 7-12 are F-type “inFormation” items.
64.	Now for the rescoring (replacing NEWSCORE=): IVALUET= means “rescore T-type items” IVALUEF= means “rescore F-type items” We have rescored the categories of the items. The reports should make better sense.	`CODES = 123456789` ; original data `IVALUET=321******` ; rescored `IVALUEF=321******` ; rescored We are reversing the scoring of 1, 2, 3 “1” is rescored “3” for T and F items.
65.	The data were analyzed with the “Partial Credit Model” (ISGROUPS=0 or GROUPS=0), where each item defines its own rating scale. This allows items to have different numbers of categories.	ISGROUPS=0 Each of our 12 items defines its own rating scale.
66.	But we want to specify that the *Interest items share the same rating scale,* and the *Information items share another rating scale.* This will simplify communication, but otherwise there is usually not much difference in the analyses between Partial Credit and Grouped rating scales.	ISGROUPS = TTTTTTFFFFFF The first 6 items share the same rating scale, “T”. The second 6 items share the same rating scale, “F”. We use the same letters as IREFER= to avoid mistakes.

67.	Edit these changes into our control file, Interest.txt
	
	I have commented out the earlier instructions with “;” in case they are needed again.
	
	Save the revised control file (Ctrl+S)
		`XWIDE = 1 ; Matches the widest data va`
		
		`; comment: GROUPS = 0 ; Partial Credit`
		`ISGROUPS=TTTTTTFFFFFF`
		`IREFER = TTTTTTFFFFFF`
		`CODES = 123456789 ; matches the data`
		`IVALUET= 321******`
		`IVALUEF= 321******`
		`; comment: NEWSCORE=123456*** ; rescor`
		
		`SPFILE = nsfcl.txt ; additional specif`

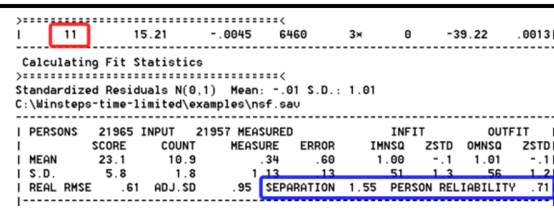
68.	Winsteps Analysis window: Click on File menu Click on “Exit and Restart ....” Run the interest.txt analysis again ...	

69. *Convergence Table:*
CATS 6: six categories. 3 categories for each of our two rating scales. Excellent! Yes, we should be able to explain 6 categories. We had absolutely no hope of explaining the 108 original categories , or even 36 categories (3 for each of 12 items).
Cleaning up our data by removing the data-entry errors has caused 337 persons to have extreme scores, and so to provide no information about relative item difficulties. *But we have lost garbage, not meaning!*



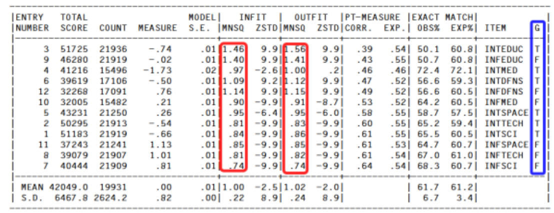
```
                                        CONVERGENCE TABLE
-Control: \examples\interest.txt                 Output
|    PROX             ACTIVE COUNT           EXTREME 5
| ITERATION    PERSONS ITEMS     CATS          PERSONS
>========================================<
|         1    21965       12        6          6.27
>========================================<
|         2    21628       12        6          6.67
>========================================<
|         3    21628       12        6          6.76
```

70. The data are better organized, so Winsteps performs fewer iterations to convergence, and the person reliability has improved slightly .71 (from .70)



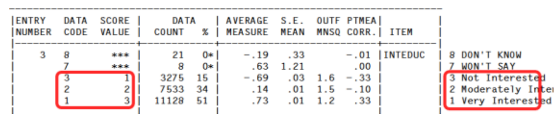
```
>=====================================<
|    11       15.21      -.0045    6460      3*      0     -39.22    .0013|
--------------------------------------------------------------------------
 Calculating Fit Statistics
>=====================================<
Standardized Residuals N(0,1)  Mean: -.01 S.D.: 1.01
C:\Winsteps-time-limited\examples\nsf.sav
--------------------------------------------------------------------------
| PERSONS   21965 INPUT   21957 MEASURED            INFIT         OUTFIT  |
|           SCORE     COUNT      MEASURE   ERROR    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN       23.1      10.9          .34     .60     1.00    -.1  1.01    -.1|
| S.D.        5.8       1.8         1.13     .13      .51    1.3   .56    1.2|
| REAL RMSE   .61   ADJ.SD       .95  SEPARATION  1.55  PERSON RELIABILITY  .71|
|-------------------------------------------------------------------------|
```

71. Click on “Diagnosis” menu, “A. Item Polarity”.
Table 26.1:
Red box: Outfit and Infit mean-squares are improving by becoming nearer to 1.0.
Blue box: In the “Groups” column, G, the items are identified by their rating-scale groups, T or F.



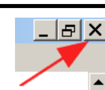
```
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL                    MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|          |
|NUMBER  SCORE   COUNT   MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM    G|
|--------------------------------------+----------+----------+-----------+-----------+----------|
|     3  51725   21936    -.74     .01|1.46   9.9|1.56   9.9|  .39   .54| 50.1  60.8| INTEDUC  T|
|     9  46280   21919    -.02     .01|1.40   9.9|1.41   9.9|  .43   .55| 50.7  60.8| INFEDUC  F|
|     4  41216   15496   -1.73     .02| .97  -2.6|1.00    .2|  .46   .46| 72.4  72.1| INTMED   T|
|     6  39619   17106    -.50     .01|1.09   9.2|1.12   9.9|  .47   .52| 56.6  59.3| INTDFNS  T|
|    12  32268   17091     .76     .01|1.14   9.9|1.15   9.9|  .49   .52| 56.6  60.5| INFDFNS  F|
|    10  32005   15482     .21     .01| .90  -9.9| .91  -8.7|  .53   .52| 64.2  60.5| INFMED   F|
|     5  43231   21250     .26     .01| .95  -6.4| .95  -6.0|  .58   .55| 58.7  57.5| INTSPACE T|
|     2  50295   21913    -.54     .01| .81  -9.9| .83  -9.9|  .60   .55| 65.2  59.4| INTTECH  T|
|     1  51183   21919    -.66     .01| .84  -9.9| .86  -9.9|  .61   .55| 65.5  60.5| INTSCI   T|
|    11  37243   21241    1.13     .01| .85  -9.9| .85  -9.9|  .61   .53| 64.7  60.7| INFSPACE F|
|     8  39079   21907    1.01     .01| .81  -9.9| .82  -9.9|  .61   .54| 67.0  61.0| INFTECH  F|
|     7  40444   21909     .81     .01| .74  -9.9| .74  -9.9|  .64   .54| 68.3  60.7| INFSCI   F|
|--------------------------------------+----------+----------+-----------+-----------+----------|
| MEAN 42049.0   19931     .00     .01|1.00  -2.5|1.02  -2.0|           | 61.7  61.2|          |
| S.D.  6467.8  2624.2     .82     .00| .22   8.9| .24   8.9|           |  6.7   3.4|          |
-------------------------------------------------------------------------------------------
```

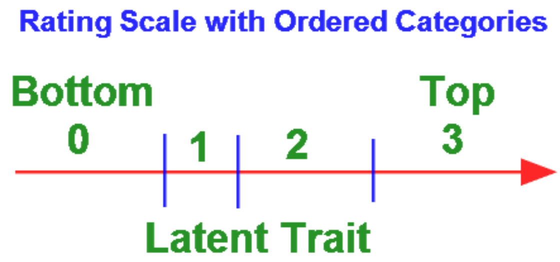
72. Table 26.3: *Excellent!*
Red box: data codes are scored so that “more category → more of the variable”
Data entry errors are now rescored as missing data “***”.



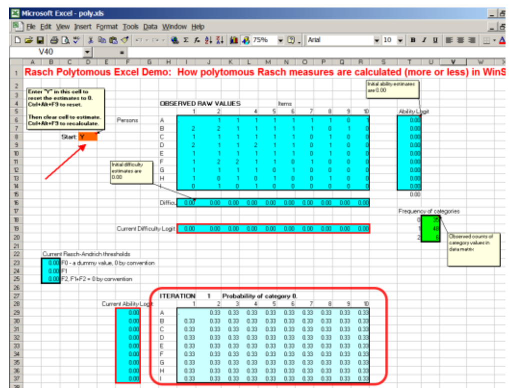
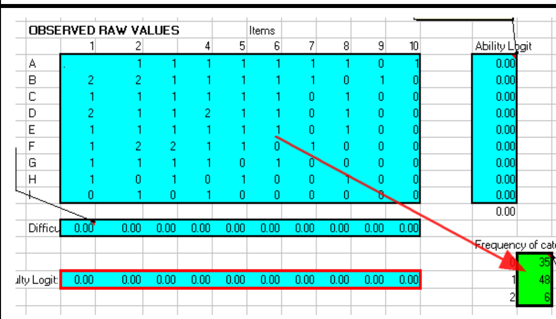
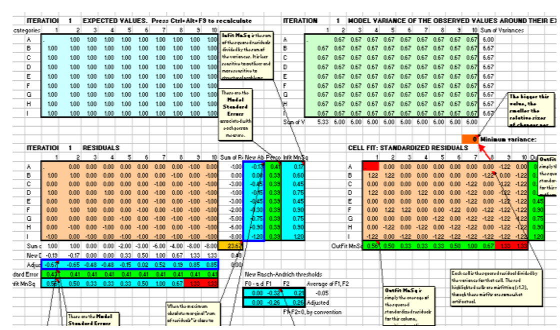
```
------------------------------------------------------------------
|ENTRY   DATA  SCORE |    DATA    | AVERAGE  S.E.  OUTF PTMEA|          |
|NUMBER  CODE  VALUE |  COUNT   % | MEASURE  MEAN  MNSQ CORR.| ITEM     |
|--------------------+------------+--------------------------+----------|
|     3   8      *** |     21   0*|    -.19   .33        -.01 |INTEDUC   | 8 DON'T KNOW
|         7      *** |      8   0*|     .63  1.21         .00 |          | 7 WON'T SAY
|         3        1 |   3275  15 |    -.69   .03   1.6  -.33 |          | 3 Not Interested
|         2        2 |   7533  34 |     .14   .01   1.5  -.10 |          | 2 Moderately Inte
|         1        3 |  11128  51 |     .73   .01   1.2   .33 |          | 1 Very Interested
```

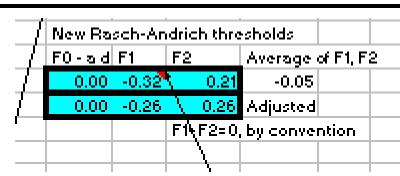
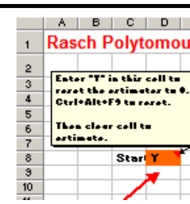
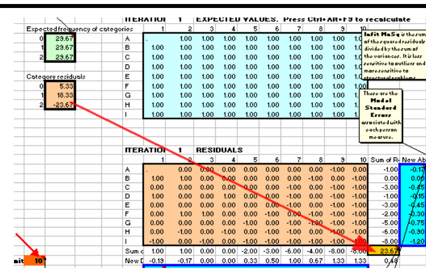
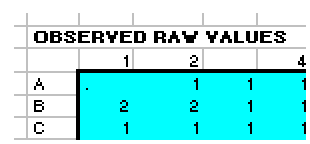
73. We will return to this analysis.
But, for now, close all windows

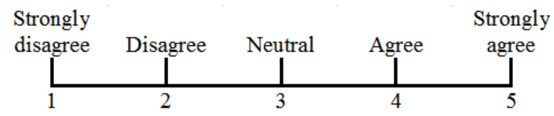
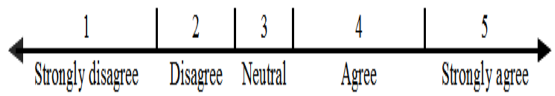
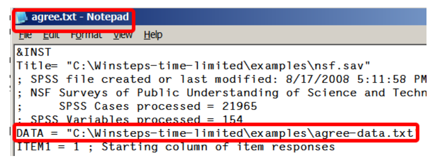
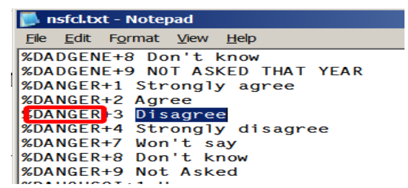
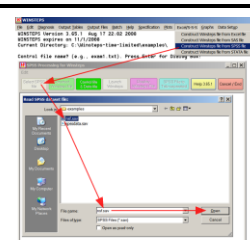
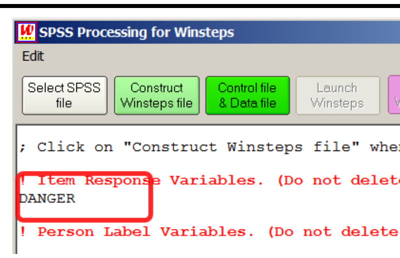
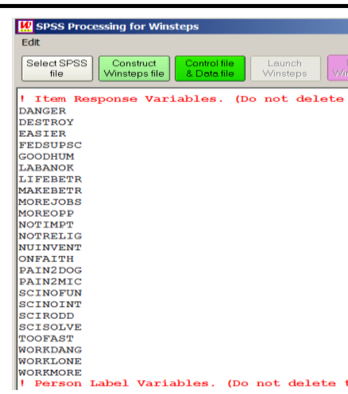


74.	**F. Polytomous Rasch Models**	
75.	Early on, Georg Rasch conceptualized a rating scale as a many dichotomies, each with its own dimension, but we find it more productive to think of a rating scale as the division of the latent trait into ordered categories, qualitatively advancing along the latent trait. These are called “polytomies”. “Polytomous” comes from the Greek words “Poly” (meaning “many”) and “tomos” (meaning “division” or “slice”). Our word “atom” comes from the Greek “a” (meaning “no”) and “tomos”, because the ancient Greeks thought that atoms could not be split.	 The bottom and top categories reach to infinity, and so are always infinitely wide.
76.	Rasch polytomous analysis has developed from two profound insights. First, Erling Andersen perceived that Rasch measurement is based on counts of qualitatively ordered observations. Then, David Andrich perceived that the relationship between adjacent polytomous categories has the form of a Rasch dichotomy.	“Counts are the sufficient statistics for Rasch measures” - Andersen “The fundamental relationship is the log-odds of adjacent categories” - Andrich
77.	The Andrich “Rating Scale Model” (RSM), from which have developed many variants, is composed of 3 parameters: B_n = Ability of person *n* D_i = Difficulty of item *i* F_j = The “Rasch-Andrich threshold”. The “step” difficulty of **observing** category *j* relative to category *j-1*.	$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = B_n - D_i - F_j$ where P_{nij} is the probability of observing category *j* for person *n* on item *i*. *F_j is also called “step calibration”*
78.	Imagine a situation in which all the persons are equally able, and all the items are equally difficult, and they match, so that $B_n = D_i$, then an estimate of F_j is	$F_j \approx \log\left(\frac{\text{count of observations in } j-1}{\text{count of observations in } j}\right)$
79.	By changing the subscripts, we obtain the Partial Credit Model (PCM) of Geoff Masters, and the Grouped model we are using for the NSF data. F_{ij} = rating scale is specific to the item = Partial Credit F_{jg} = rating scale is specific to the group of items = Group	$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = B_n - D_i - F_{ij}$ $\log\left(\frac{P_{nigj}}{P_{nig(j-1)}}\right) = B_n - D_{ig} - F_{jg}$
80.	So where do “sufficient statistics” come into this? Person raw score → Person ability estimate Item raw score (or p-value) → Item difficulty estimate Count of observations in a category (category frequency) → Rasch-Andrich threshold.	The “father of modern statistics”, Ronald A. Fisher, perceived that a “sufficient statistic” contains all the information in the data from which to estimate the value of a parameter.

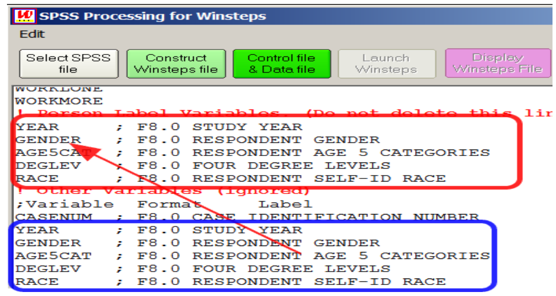
81.	The polytomous Rasch models include the *"Andrich" Rating Scale model (RSM):* all the items share the same rating scale. *"Masters" Partial Credit model (PCM):* each item has its own rating scale. *Grouped Rating Scale model:* groups of items share the same rating scale. *Binomial Trials (Bernoulli) model:* this is an RSM with preset values of F_j. *Poisson model:* this is an RSM with preset values of F_j and an infinite number of categories. *Success and Failure models:* These are incremental dichotomous models, implemented in Winsteps, but not recommended. They seem to be ideal for various processes, but they have proved too fragile when the data depart from strict adherence to those processes, which empirical data always do. And there are many more models in what Jürgen Rost calls "the growing family of Rasch models".

82.	**G. Polytomous Rasch Estimation** **Study this closely** if you want to understand how Winsteps estimates the parameters of polytomous models. **Glance through this** if your focus is less mathematical, but do answer the question at #Error! Reference source not found.	
83.	Estimating the measures for a polytomous model is more complex than a dichotomous because we have to consider the rating-scale categories. Let's think about a rating-scale with 3 categories: 0, 1, 2. The model equations for the category probabilities become:	$P_0 = P_0$ $P_1 = \exp(B_n - D_i - F_1)\ P_0$ $P_2 = \exp(B_n - D_i - F_2)\ P_1$ $= \exp(B_n - D_i - F_2) \exp(B_n - D_i - F_1)\ P_0$ $= \exp(2(B_n-D_i) - (F_1+F_2))\ P_0$
84.	Three categories are all there are so $P_0 + P_1 + P_2 = 1$, which enables us to give an explicit equation for P_0, and similarly P_1 and P_2.	$P_0 = 1 / (1 + \exp(B_n - D_i - F_1)$ $+ \exp(2(B_n-D_i) - (F_1+F_2)))$ P_1 and P_2 follow.
85.	Look at http://www.winsteps.com/furthercourse2/poly.xls also at *c:\Winsteps\further\poly.xls* This is an estimation spreadsheet for 3 category items. In the top rectangle is the raw data: 0, 1, 2 and missing. In this example, the initial values of all the Rasch parameters (persons, items, R-A thresholds) are 0.0. The probabilities for each category for each of the three original data-points are shown in 3 rectangles. The probabilities for category 0 are shown here. Initially every category probability is 0.33.	
86.	The frequency-counts for each of the categories in the raw data are shown in the green box. These are the sufficient statistics for the R-A threshold estimates. We are estimating measures for the Andrich Rating Scale model, so we count the frequency of categories (C_0, C_1, C_2) over the entire data set. Any misfit to the Rasch model is ignored at this stage.	
87.	The expected values, E, for each observation are: $E = 0*P_0 + 1*P_1 + 2*P_2$ The variance values, W, for each observation are: $W = 0*0*P_0 + 1*1*P_1 + 2*2*P_2 - E*E$ Person ability and item difficulty estimation follows the same process we used for dichotomous data. The standard errors, Infit mean-squares and Outfit mean-squares are computed in exactly the same way as before.	

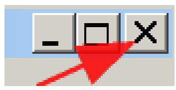
88.	But now we must estimate the Rasch-Andrich thresholds: F_1 and F_2. Here is the logit-linear Andrich Rating Scale Model: We can rewrite this in terms of F_j. When estimating F_j, we can think of B_n and D_i as constants. P_{nij} / $P_{ni(j-1)}$ corresponds to the frequencies of the categories in the data.	$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = B_n - D_i - F_j$ rewritten: $F_j = B_n - D_i - \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right)$
89.	This suggests an estimation equation for F_j. A better estimate of F_j is one for which the ratio of the accumulated category probabilities for each pair of adjacent categories (across all the persons and items) more closely matches the ratio of the observed frequencies of those categories.	$F_j' = F_j + \log\left(\frac{\sum P_{nij}}{\sum P_{ni(j-1)}}\right) - \log\left(\frac{C_j}{C_{(j-1)}}\right)$ where $\sum$ is summed across all persons and items for the category
90.	Each time a new set of $\{F_j\}$ estimates is produced, they are all adjusted by the same amount so that their sum is zero. This establishes that ***"the difficulty of the items is defined to be at the location on the latent variable where the probabilities of its top and bottom categories are equal."*** This is also the location of the average of the R-A thresholds for the item.	$\Sigma(F_j)=0$ constrains the R-A thresholds and defines the item difficulty locations.
91.	The R-A threshold estimation equations are computed at the bottom-center of the worksheet. They compare the observed frequency of adjacent categories with the expected frequencies.	
92.	You can now perform your own polytomous estimation in the worksheet by removing the "Y" in the orange cell. Then pressing Ctrl+Alt+F9 for each iteration through the data.	
93.	The convergence cell now includes the category residuals between the observed and expected frequencies of the categories in the data.	
94.	Does this work for you? If so, please try altering the 0, 1, 2 data. Type "Y" in the orange cell, then press "enter" to reset. Remove the "Y" and press Enter, then see what happens.	

95.	**H. Items with Rating Scales**	
96.	Conventional statistical analysis assumes that rating-scale category numbers are linear measures defining a finite latent trait. Each rating scale category is conceptualized to be a point.	
97.	In Rasch analysis, each rating scale category is conceptualized to be an interval (zone) on an infinite latent variable. The extreme categories correspond to intervals extending to infinity. For the purposes of estimating Rasch measures, all the categories are assumed to correspond to ascending qualitative levels of the latent trait.	 Notice that category widths are unequal.
98.	In your "c:\Winsteps\further" folder you will find Winsteps control file: "**agree.txt**" and its data file "agree-data.txt". It contains most of the "agreement" items in the NSF data file. Here is how I constructed it *(do this if you want to).* *Otherwise go to #**105***	 agree.txt - Notepad File Edit Format View Help &INST Title= "C:\Winsteps-time-limited\examples\nsf.sav" ; SPSS file created or last modified: 8/17/2008 5:11:58 PM ; NSF Surveys of Public Understanding of Science and Techn ; SPSS Cases processed = 21965 ; SPSS Variables processed = 154 DATA = "C:\Winsteps-time-limited\examples\agree-data.txt ITEM1 = 1 ; Starting column of item responses
99.	*1. File nsfcl.txt has all the option/distractor labels.* *I opened it and found 24 items with "disagree" in them.* *The first one is DANGER.*	 nsfcl.txt - Notepad File Edit Format View Help %DADGENE+8 Don't know %DADGENE+9 NOT ASKED THAT YEAR %DANGER+1 Strongly agree %DANGER+2 Agree %DANGER+3 Disagree %DANGER+4 Strongly disagree %DANGER+7 Won't say %DANGER+8 Don't know %DANGER+9 Not Asked
100.	*2. Launch Winsteps.* *Then SPSS interface* *Then, in "examples", nsf.sav*	
101.	*3. Copy-and-paste the item name, "DANGER" from nsfcl.txt to the SPSS window.* *This is exactly the same as copying the item name from further down the SPSS window*	 SPSS Processing for Winsteps Edit Select SPSS file Construct Winsteps file Control file & Data file Launch Winsteps ; Click on "Construct Winsteps file" wher ! Item Response Variables. (Do not delete DANGER ! Person Label Variables. (Do not delete
102.	*4. Copy over 23 other "disagree" item names from nsfcl.txt to the SPSS window* *I excluded LUCKYNUM from this selection*	 SPSS Processing for Winsteps Edit Select SPSS file Construct Winsteps file Control file & Data file Launch Winsteps ! Item Response Variables. (Do not delete DANGER DESTROY EASIER FEDSUPSC GOODHUM LABANOK LIFEBETR MAKEBETR MOREJOBS MOREOPP NOTIMPT NOTRELIG NUINVENT ONFAITH PAIN2DOG PAIN2MIC SCINOFUN SCINOINT SCIRODD SCISOLVE TOOFAST WORKDANG WORKLONE WORKMORE ! Person Label Variables. (Do not delete

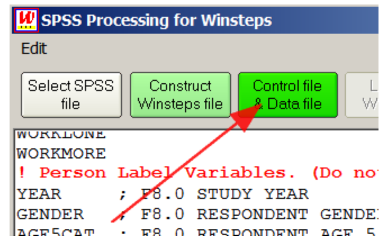
103. *5. Copy the person demographics from the SPSS variable list*



104. *6. Control and data file:*
Control file: ***my-agree.txt***
Data file: ***my-agree-data.txt***

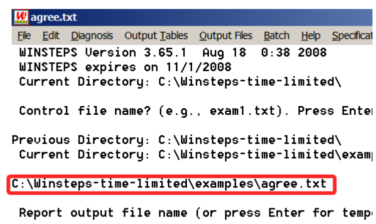
Close the SPSS window: 

My-agree.txt will not include some extra control-instructions in agree.txt

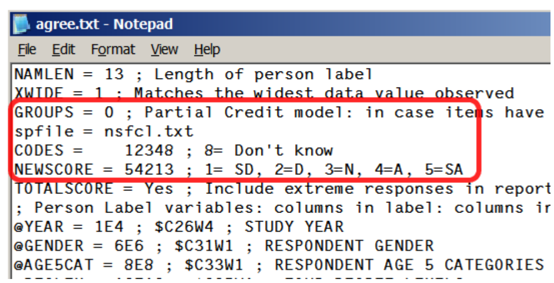


Mike: redo this analysis. The output is incorrect

105. Launch Winsteps
Winsteps menu bar
Click on "File"
Click on "Open file"
In "further", click on "agree.txt"

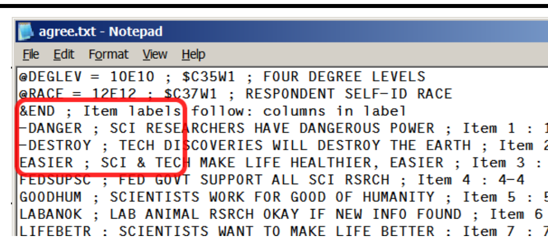


106. Winsteps menu bar
Click on "Edit Control File"
"agree.txt" displays in a NotePad window:
Groups=0 - all 24 items have the same rating scale, but we want to verify it operates the same way, so have deliberately left this as the "Partial Credit" model where each item is specified to have its own rating scale.



107. spfile = nfscl.txt - this has all the categories labels
`CODES =` **1**234**8** - according to nfscl.txt, these are the valid codes. 8 = "Don't know"
`NEWSCORE =` **5**42**1**3 - recoding to make
Blue numbers: "Strongly agree": Data = 1, recoded = 5
Red numbers: "Don't know": Data = 8, recoded = 3 = Neutral ← *This is my guess!*

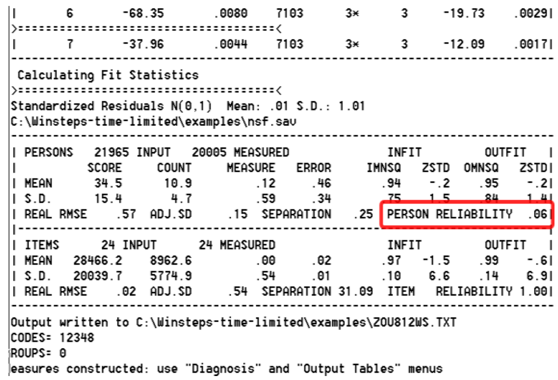
108. After &END, in the item labels, some items are "in favor of science", and some items are "against science". I put "-" in front of item labels whose wording seems to be against science, but I did not change the data for them.



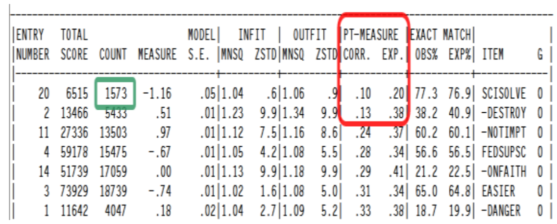
109. Wrong numbers??

Winsteps analysis window:
No extra instructions ....
Run the analysis ...

Person reliability .06 (very low) - Have me measured anything more than random noise?
Observed person measure S.D. = 0.59 logits.
The average standard error of the person measures is around 0.5 logits. (Arithmetic average = 0.46 logits, statistical average = RMSE = 0.57 logits)



110. **Diagnosis menu: A. Item Polarity**
All the correlations are positive. Good!
Red box: But "SCISOLVE" and "DESTROY" report low point-measure correlations (.10, .13), noticeably below their expected values (.20, .38).
Green box: "SCISOLVE" has a low response count (1573)

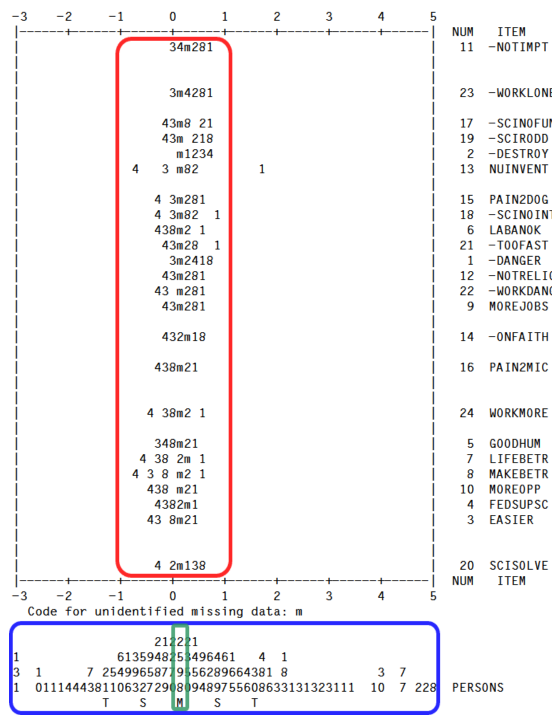


111. **Diagnosis menu: B. Empirical Item-Category Measures**

Red box: the average measures for each data-code for each item are shown. These are squashed together on each line. They are ordered backwards (4-3-2-1) because the original data was scored backwards.

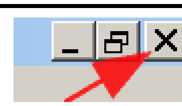
Blue box: this is the person distribution. It is wide, but
Green box in the Blue box: the distribution is very central. There are 2,598 persons located at about the mean, "M", of the person distribution. The standard deviation (distance between "S" and "M") of the person distribution is around 0.6 logits, only slightly larger than the average standard error.

We usually expect to see a distribution with a single peak that looks like a normal distribution. But we like to see it more spread out (bigger standard deviation, bigger observed variance, less central). The bigger the observed variance (= S.D.^{2}) in the measures (or scores), then the bigger the reliability:
Reliability = "observed variance - error variance" / "observed variance".
In this example, the observed variance is small, so that the reliability is small.

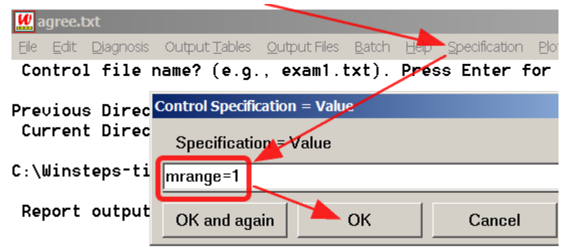


"m" in Table 2.5 and Table 2.15 indicates the average measure for missing data. If we don't see an **"m"**, then the data are "complete" (there are no missing data).

112. Close this NotePad window

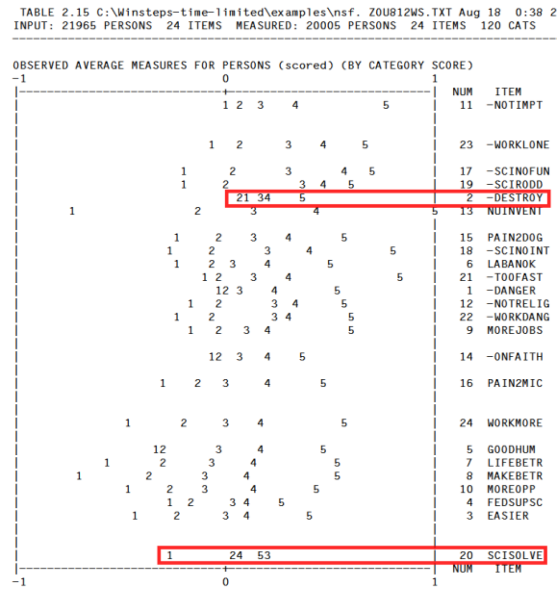


113. Let us widen the red box:
Winsteps menu bar
Click on "Specification"
Type into the specification box: "mrange=1"
This sets the range of Table 2 at 1 logit each side of the local origin.
Click on "OK"

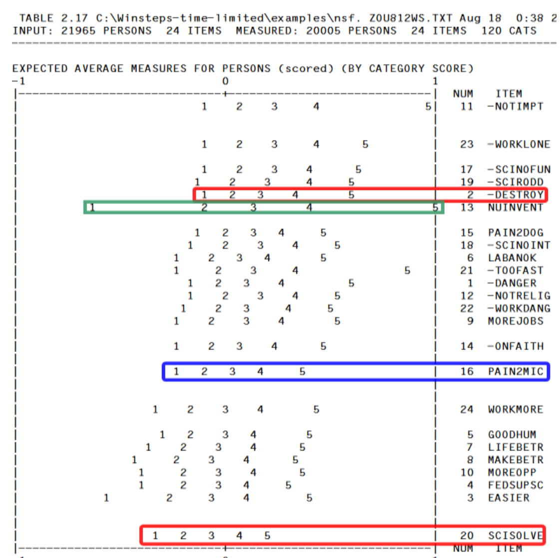


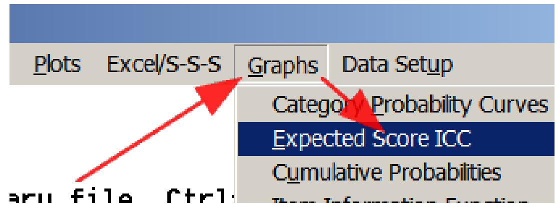
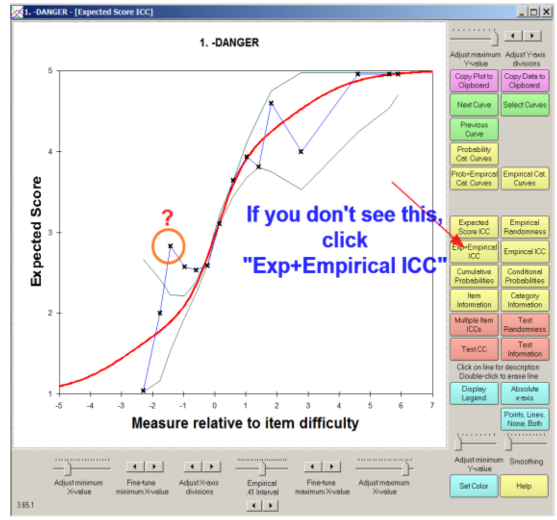
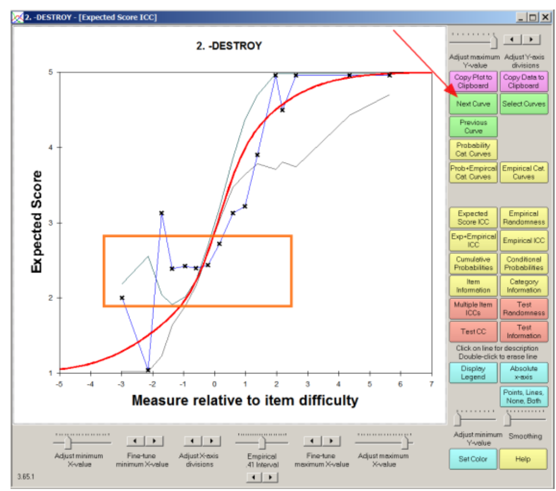
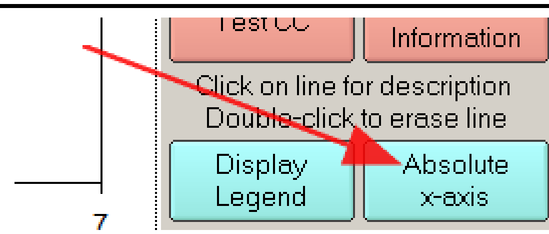
114. **Output Tables menu: 2. Measure Forms (All)**

Scroll down to **Table 2.15.** This is between Table 2.5 and Table 2.7.
Table 2.15: "Observed Average Measures For Persons (Scored) (By Category Score)"
There are many pictures telling similar stories. We choose the one that matches what we want.
In this picture, our "NEWSCORE=" values are displayed.
Red boxes: We expect the category scores to progress "1-2-3-4-5" up the latent variable for every item. This is true for all items except "-DESTROY" and "SCISOLVE" which we saw in Diagnosis-A (Table 26) had low correlations.
In #107, I guessed that "Don't know" (coded 8 in the original data) would act like a neutral category (3 on the Likert scale), and now that looks like a good guess!



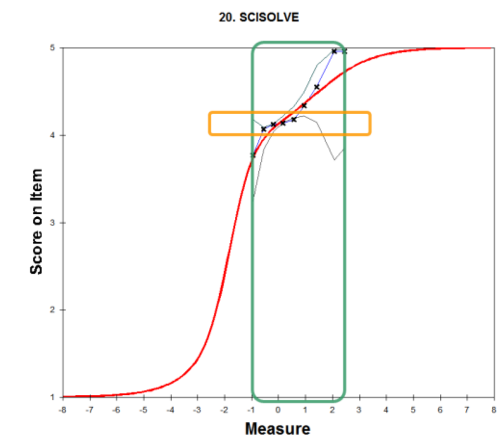
115. If everything had worked perfectly, what would Table 2.15 look like?
Scroll down to **Table 2.17: "Expected Average Measures For Persons (Scored) (By Category Score)"**
This is what Table 2.15 would look like if the data fit the Rasch model perfectly. The wider and narrower spread of 1-2-3-4-5 is because we are using the Partial Credit model GROUPS=0.
Green box: NUINVENT is about the same in both subtables 2.17 (expected) and subtable 2.15 (observed).
The widest spread is NUINVENT
Red boxes: This shows that 1-2-3-4-5 should have been widely spread for -DESTROY and SCISOLVE.
Blue box: The narrower spreads are items such as PAIN2MIC.
Let's look more closely at what these spreads mean ...



116.	**I. Item Characteristic Curves (ICCs)**	
117.	Winsteps Analysis window Winsteps menu bar Click on Graphs Click on “Expected Score ICC”	
118.	The ICCs for Item 1, “-Danger” display. *Red curve:* This is the “model ICC” or “expected ICC”. It is what the ICC would look like if the data exactly fit the Rasch model. *Blue line with x’s:* These are the “empirical ICC”. Each “x” summarizes the responses (y-axis) by the persons (x-axis) near the x-axis location of the “x”. The blue lines join the x’s to guide the eye. *Grey-green lines:* These are the 95% confidence bands around the expected ICC. x’s outside these lines are unexpected. *Orange circle:* This is the most unexpected cluster of responses to the Danger item. But it may be merely due chance.	
119.	Click on “Next Curve” Item 2, “-Destroy” display. This is one of our suspect items. *Orange box:* Now we can see where the problem is. It is in the area where we expect responses (y-axis) from 1. “Strongly disagree” to 3. “Neutral”. The empirical ICC (blue line) is almost flat (horizontal). The correlation between responses to the item and the abilities/attitudes of the sample (x-axis) is almost 0 in this region. This item is not useful for measurement at the lower end.	
120.	Click on “Absolute x-axis” this will enable us to see the curves move right-and-left with increasing-and-decreasing difficulty	

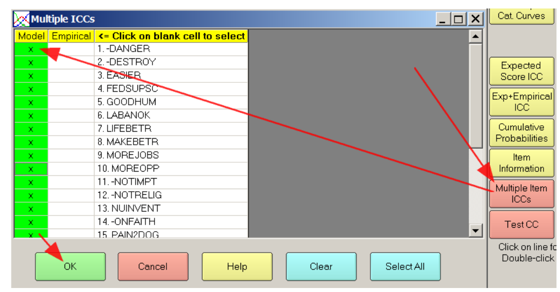
121. Click “Next Curve” and take a look at each curve. Do you see some patterns? Most empirical (blue) curves are close to their model (red) curves.

Item 20, “SCISOLVE”.
Green box: The range of person measures on the x-axis is narrower than for the other items. In Table 26, #110, we noticed that fewer people responded to this item.
Orange box: They were also much more homogeneous in their overall responses. The empirical (blue) curve is too flat. Here is another item which is not helping to construct measurement.



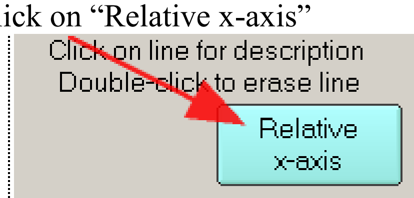
122. The test designers undoubtedly intended that all their agreement items should share the same rating scale structure. Did they?

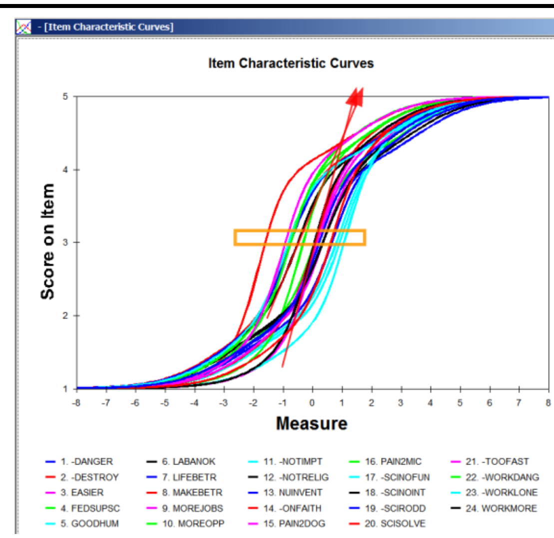
Click on “Multiple Item ICCs”
Click on all the boxes in the “Model” column.
Scroll down to click on all 24 items.
Then click on “OK”



123. You can see all 24 ICCs superimposed.
Orange box: There is too much going on in this box. It looks a mess, because the items have different difficulties.
Red arrows: but we can see some items have different slopes.

Click on “Relative x-axis”





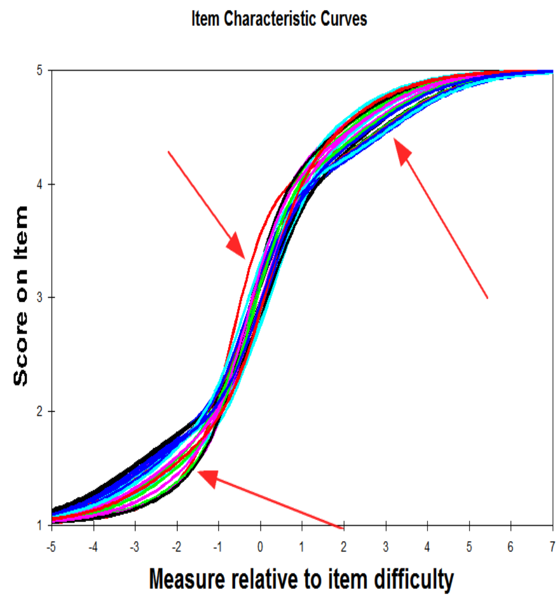
124. Now all the model ICCs are plotted relative to their item difficulties. We can see that there are outlying items (red arrows), but the general pattern is the same across all items.

The steeper curves are for items which are more discriminating between high and low performers. But if a curve is too steep, then the item is acting like a switch, no longer providing useful measurement information for comparing persons of different ability levels. In this situation, there is always a decision:

1. Model each item to have its own rating scale, the "Partial Credit model" - so obtaining more exact measures (we hope), but measures which are more influenced by **accidents in the data**,

or

2. Model the items to share the same rating scale, the "Andrich Rating Scale model", so greatly simplifying communication, and obtaining measures which are more robust against **accidents in the data**.



Every item difficulty is located at "0"

125. **Accidents in the data?** What are we talking about? Look back at **Diagnosis A. Table 26.3**, the Distractor/Option Table. The first item listed is "SCISOLVE".

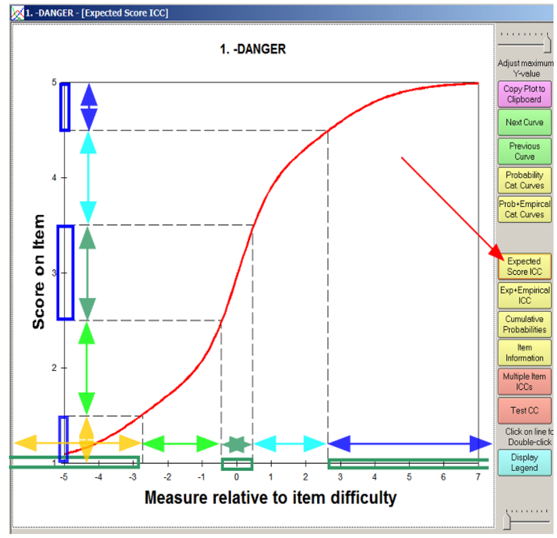
Red box: Do you see the very small counts for the lowest three categories? Only "4" for the lowest category? This is meager data on which to base decisions about the ICCs for this item. If the item is modeled to have its own rating scale (partial-credit model), a data-entry error, or misunderstandings by a few respondents about the item or the rating scale, could influence the shape of this ICC considerably.

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM	
20	4	1	4	0	-.26	.34	1.0	-.04	SCISOLVE	4 Strongly Disagree
	3	2	30	2	.05	.09	1.2	-.01		3 Disagree
	8	3	17	1	.20	.08	1.2	.03		8 DON'T KNOW
	2	4	1210	77	.06*	.01	1.0	-.12		2 Agree
	1	5	312	20	.18*	.03	1.0	.12		1 Strongly agree
	MISSING ***		20392	93*	.13	.01		.02		

126. **J. Rasch-Half-Point Thresholds**

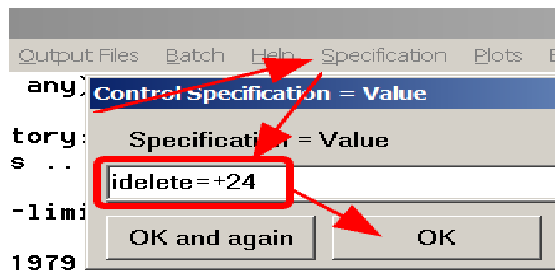
127. In the Graphs window,
Click on "**Expected Score ICC**"
The model ICC displays with - - - - lines to indicate the half-point thresholds and zones. These answer the questions:
1. "What is the average rating for people at this location on the latent variable?"
2. "For a sample with an observed average rating on the item, what is their expected average ability measure?"

Middle Blue box (y-axis). This is the zone from an expected score of on the item of 2.5 to 3.5 Each score zone corresponds to an average expected rating 0.5 score points above and below the category value.
Middle Green box (x-axis): The range of measures corresponding to the middle blue box is from -0.5 to +0.5 logits.
Top and bottom blue boxes: the extreme categories (4.5 to 5, 1.0 to 1.5) correspond to infinitely wide measure-ranges on the x-axis *(left and right green boxes).*



Ends of green boxes are the Rasch-Half-Point thresholds.

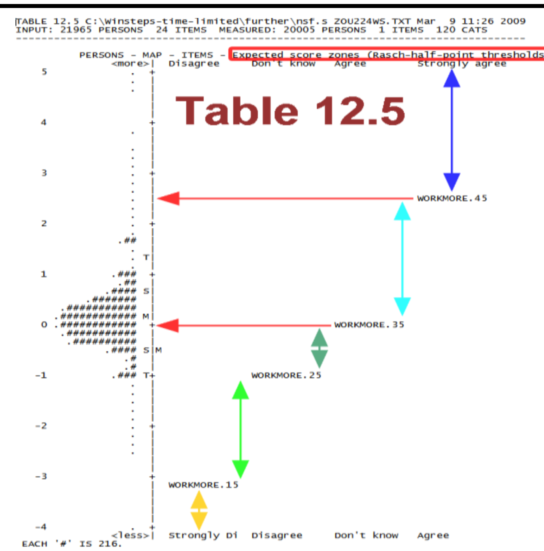
128. Let's look at output based on Rasch-Half-Point thresholds. To make them easier to see, we will look at them for only one item:
Winsteps menu bar
Click on "Specification"
Type into the box: idelete=+24
Click on OK
This temporarily deletes all items except item 24.

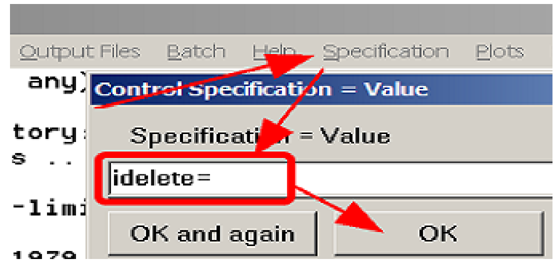


129. Winsteps menu bar
Click on Output Tables
Click on Table 12

Scroll down to **Table 12.5.**

The expected score zones are shown for Item 24, "WORKMORE".
Orange arrow is "Strongly disagree" (below W.15)
Light green arrow is "Disagree" (W.15 to W.25)
Dark green arrow is "Don't know" (W.25 to W.35)
Light blue arrow is "Agree" (W.35 to W.45)
Dark blue arrow is "Strongly agree" (above W.55)
These arrows are the same in #**127**



130.	Winsteps menu bar
	Click on “Specification”
	Type into the box: idelete=
	Click on OK
	This restores all the temporarily deleted all items
	`idelete=`
	`CURRENTLY REPORTABLE ITEMS = 24`
		

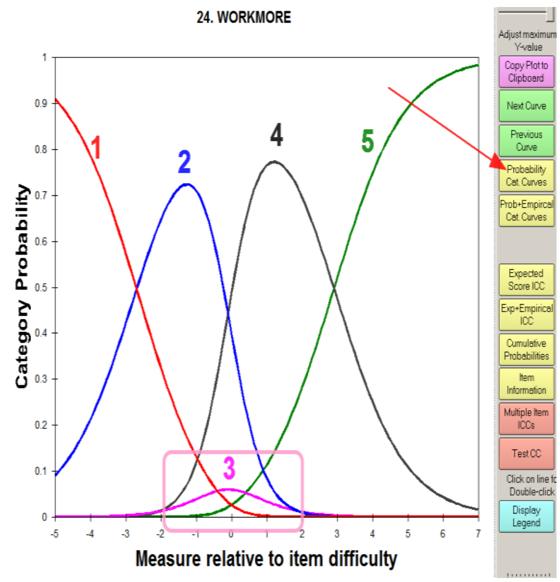
131. **K. Modal *(pronounced: Mode-al)* Probability Curves**

132. *Question: "At what category on the rating scale is the person functioning?"*
Answer: The modal probability curves!

Click on "Probability Cat.(egory) Curves" for any of our 24 items, and you will see the same story. I'm looking at item 24, WORKMORE.

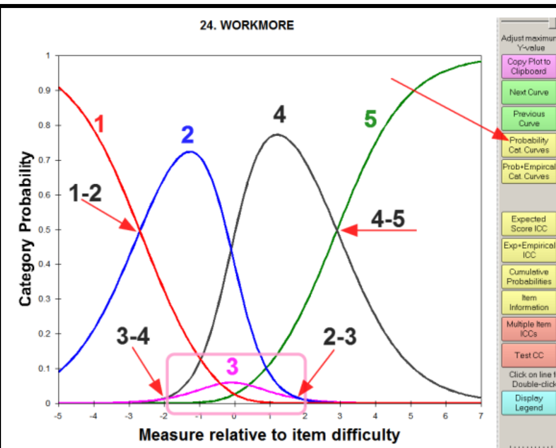
The "modal" perspective on category boundaries on the latent variable is to identify them with the intersections of the category probability curves. This simplifies inferences about which category is most likely to be observed for any item at any point along the latent variable.

Inferential simplicity → Probability curves
→ Category boundaries



133. For item 24, four of the five categories are "modal". They are the most likely of the categories to be observed at some location on the latent variable. They look like a "range of hills". So, as we progress along the latent variable, we are most likely to observe "1" for the lowest performers, "2" for the low performers, "4" for the high performers, and "5" for the highest performers.
But we are never likely to observe "3", which is our "Don't know" category.

134. For NSF, the low frequency of "Don't Know" was good news! They want people to express a clear opinion. But for us it presents a problem.
Red Arrows: These probability curves are drawn to conform with the Rasch model. The Rasch model parameters, the Rasch-Andrich thresholds {Fj}, are the intersections between adjacent categories. For categories 1-2 and 4-5 they are nicely position. But for categories 2-3 and 3-4 they appear to be in the wrong places. 3-4 is to the left of 2-3. The thresholds are "disordered".
Disordered thresholds are a source of considerable contention in the Rasch literature. *What to do about them?*



135. David Andrich is adamant. *"Disordered thresholds are a violation of the principles underlying the Rasch model. They must be eliminated!"* He perceives the category-intervals on the latent variable to correspond to the modal intervals of the categories. But category 3 is never modal. It does not have an interval on the latent variable, so it must be removed.

136. What can we do with Category 3 for Item 24? Here are its statistics from Table 26.3. There are 4 options:
A. Combine (collapse) category 3 with category 2.
B. Combine (collapse) category 3 with category 4.
C. Make category 3 missing data.
D. Keep category 3 unchanged.

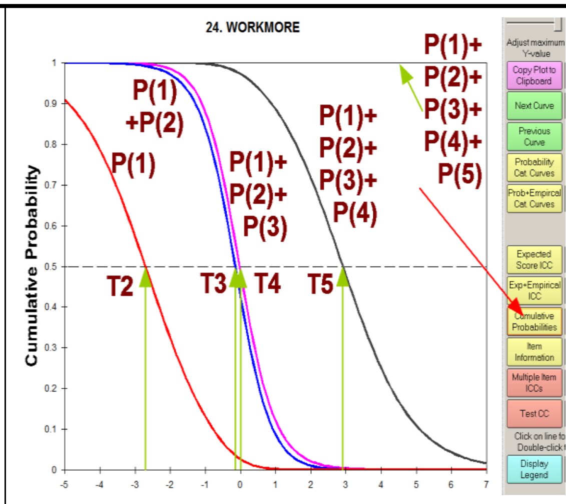
ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM		
24	4	1	142	1	-.47	.05	1.0	-.13	WORKMORE	4	Strongly Disagree
	3	2	2528	22	-.21	.01	.9	-.35		3	Disagree
	8	3	540	5	.00	.02	1.0	-.05		8	Don't know
	2	4	7407	65	.18	.00	.9	.21		2	Agree
	1	5	821	7	.58	.03	1.0	.28		1	Strongly agree

137.	Look at the "Count" columns. Category 3 has 540 observations, 5% of the data. We would rather not lose that much expensive data if we can avoid it. If we add collapse categories 3 and 4 (7407 observations) then, in the probability curve picture, the peak for "4" will become even higher. We would rather collapse 2 and 3 (2528 observtions), so that the peak for 2 is about the same height as the peak for 4. That will also give a more even spread of category counts: 3068-7040, not 2528-7947. So, one vote for 2+3.	SCORE VALUE | DATA COUNT 1 | 142 2 | 2528 3 | 540 4 | 7407 5 | 821
138.	Look at the "Average Measure" column, the average attitude of the people who chose category 3 was .00 logits, this is slightly closer to the attitude for category 4, .18 logits. So it makes sense to collapse 3 with 4. One vote for 3+4.	AVERAGE MEASURE -.47 -.21 .00 .18 .58
139.	Look at the "Outfit Mean-square" column for Item 24: Red box: There is nothing to choose between them. No vote, but the nice fit of 1.0 suggests that perhaps we should keep this category! Blue box: For Item 21, we notice that category 3 is slightly overfitting (0.9), but category 4 is slightly unpredictably noisy (1.2). We would like to smooth out the misfit across the categories, so combining categories 3 and 4 would probably make the combined fit 1.1.	Item 24: OUTF MNSQ 1.0 .9 1.0 .9 1.0 Item 21: OUTF MNSQ 1.1 1.0 .9 1.2 1.4
140.	Look at the category descriptions: Does it make better sense to combine "Don't know" with "Agree" or with "Disagree"? Politicians like to think that "Don't knows" will vote for them when the election comes, but they would be foolish to think that way as they are campaigning for voters to vote for them. I don't know about "Don't Know". One vote for "Missing Data".	Strongly Disagree Disagree Don't know Agree Strongly agree
141.	We've had a vote for every option! So the safest path is to make "Don't know" into missing data. According to the original coding of 8= "Don't know", it is likely that was also NSF's choice. We could easily change agree.txt for that choice: `CODES     = 1234 ;` 8 is not listed, so is treated as missing data. `NEWSCORE = 4321` would do it quickly.	

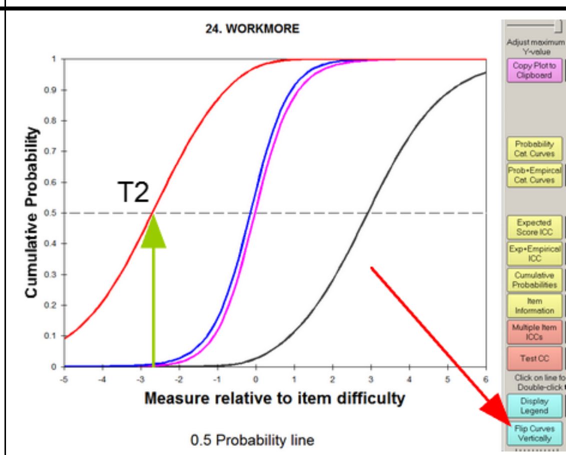
142.	***This is important:*** "F_j = ... difficulty of **observing** ....." F_j is **not** the **difficulty of category *j*** relative to category *j-1*. Many published papers contain crucial errors in interpreting their rating-scale findings. A category may be **more difficult to perform** (so qualitatively higher on the latent variable), **but easier to observe** (so having a lower F_j). Those papers misinterpret F_j as the substantive "category difficulty" and so mis-report that an easily-observable higher category indicates *less* of the latent variable than a more-difficult-to-observe lower category. *Example:* This is true of many transitional states. For instance, "0=can't drive a car", "1=learning to drive a car", "2=can drive a car". I can easily observe which of my current friends "can drive" a car and which "can't drive" a car, but I can't recall observing any of those friends "learning to drive a car", but all those that now drive must have gone through that stage. "2=Can drive" is more *difficult to perform* than "1=Learning to drive", but "2=Can drive" is *easier to observe* than "1=Learning to drive".	
143.	*Example:* We have a latent trait of *"people in a building"*. At night there are few people in the building. During the day there are many people in the building. But there are crucial occupancy numbers to do with fire and security regulations. So the categories are:	**Building Occupancy Rating Scale** 0 = 0-99 people in the building 1=100 2=101-999 3=1000 4=1001-upwards.
144.	If we count the number of people in the building at 5 minute intervals, then categories 0, 2, and 4 will be much easier to observe than categories 1 and 3. On some days, we may not obtain a rating of 1 or 3, even though the low and high categories for those days are 0 and 4.	Occupancy Rating Scale (Frequency) 0 100 1 1 2 140 3 2 4 45
145.	The $\{F_j\}$ will not ascend smoothly with the category numbers. The Rasch-Andrich thresholds will be *disordered.*	F1 ≈ log(frequ(0)/frequ(1)) = log (100/1) = 4.6 F2 ≈ log(frequ(1)/frequ(2)) = log (1/140) = -4.9 F3 ≈ log(frequ(2)/frequ(3)) = log (140/2) = 4.2 F4≈ log(frequ(3)/frequ(4)) = log (2/45) = -3.1
146.	We can force the R-A thresholds to have ascending values (to be ordered) by collapsing categories, 0+1, 2+3, 4	log(freq(0+1)/freq(2+3)) = log (101/142) = -0.3 log(freq(2+3)/frequ(4)) = log (142/45) = 1.1
147.	but Building Managers would complain that the qualitative advances up the rating scale of "Building Occupancy" are important to them, not the threshold-parameter values of our Rasch rating scale!	

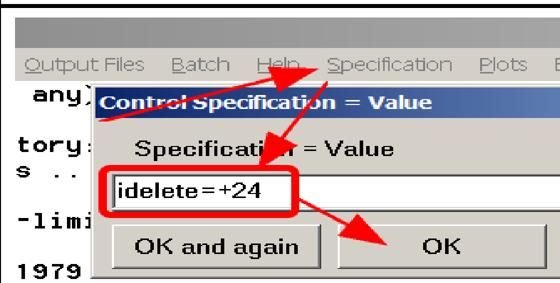
<table>
<tr><th>148.</th><th colspan="2">L. 50% Cumulative Probability Curves: Rasch-Thurstone Thresholds</th></tr>
<tr><td>149.</td><td>Winsteps Graph Screen:
Click on “Cumulative Probability Curves” to see the location of the Rasch-Thurstone thresholds where the .5 probability line crosses the cumulative probability curves.

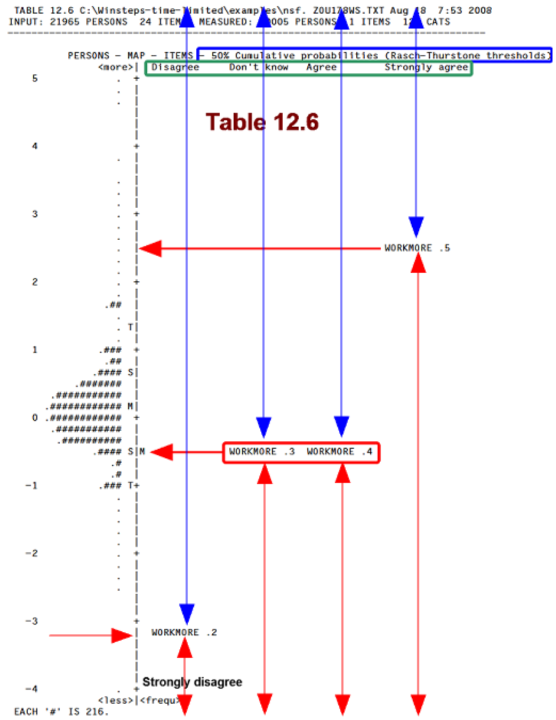
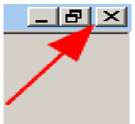
The red line is the probability of observing “category 1 or below”.

At each .5 point (T2, T3, T4, T5), a person with a measure corresponding to the green arrow has a 50% chance of being observed in a category below 2 (or 3 or 4 or 5) and a 50% chance of being observed in a category at or above 2 (or 3 or 4 or 5).</td><td>
</td></tr>
<tr><td>150.</td><td>You may find these curves easier to understand if you “flip” them … (red arrow)

The red line is now the probability of observing “category 2 or above”.

These answer the question: “What performance level has the person reached on the item?”
A person with measure “T2” has a probability of responding “50% in category 1, 50% in category 2 or above”.</td><td>
</td></tr>
<tr><td>151.</td><td>Let’s look at output based on 50% cumulative probabilities (Rasch-Thurstone thresholds).

Winsteps menu bar
Click on “Specification”
Type into the box: idelete=+24
This temporarily deletes all items except item 24.</td><td>
</td></tr>
</table>

152.	Winsteps menu bar Click on Output Tables Click on Table 12 Scroll down to Table 12.6. The 50% cumulative probability thresholds are shown for Item 24, “WORKMORE”. At the bottom left is “Strongly disagree”. At the top right is “Strongly Agree”. Below WORKMORE.2 is “Strongly Disagree”. Above .2 are all the other categories. Below WORKMORE.3 is “Agree” or “Strongly Disagree”. Above .3 are all the other categories. Below WORKMORE.4 is “Don’t Know”, “Agree” or “Strongly Disagree”. Above .4 are “Agree” and “Strongly Agree” The distance between .3 and .4 is so small they are on the same line in the Figure. Below WORKMORE.5 is “Agree”, “Don’t Know”, “Agree” or “Strongly Disagree”. Above .5 is “Strongly Agree”	Rasch-Thurstone Thresholds 
153.	That’s the end of the Lesson. Well done!	

1.

Appendix 1. Reliability and Separation Statistics

2. "What is the difference between *good reliability and bad reliability*?"
In both Classical Test Theory (CTT) and Rasch theory, "Reliability" reports the reproducibility of the scores or measures, *not* their accuracy or quality. In Winsteps there is a "person sample" reliability. This is equivalent to the "test" reliability of CTT. Winsteps also reports an "item" reliability. CTT does not report this.

3. Charles Spearman originated the concept of reliability in 1904. In 1910, he defined it to be the ratio we now express as: Reliability = True Variance / Observed Variance. Kuder-Richardson KR-20, Cronbach Alpha, split-halves, etc. are all estimates of this ratio. They are estimates because we can't know the "true" variance, we must infer it in some way.

4. **What happens when we measure with error?**

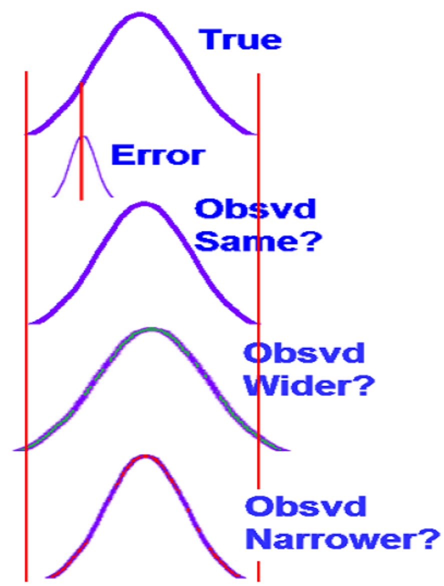
Imagine we have the "true" distribution of the measures. Each is exact. Then we measure them. We can't measure exactly precisely. Our measurements will have measurement error. These are the measurements we observe. What will be the distribution of our observed measures?

Option 1. The observed distribution will be the same as the true distribution: some measures will be bigger, some smaller. Overall, the measurement errors cancel out.

Option 2. The observed distribution will be wider than the true distribution. The measurement errors will tend to make the measures spread out.

Option 3. The observed distribution will be narrower than the true distribution. The measurement errors will tend to make the measures more central. There will be "regression toward the mean".

Think carefully: Is it Option 1, 2 or 3?



5. *Answer:* Let's imagine that all the true measures are all exactly the same. Then the measurement errors will spread them out. The observed distribution will be wider than the true distribution. As we widen the true distribution, the observed distribution also widens. So *Option 2.* is the correct answer.

6. Here is the fundamental relationship when measurement errors are independent of the measures themselves (as we usually conceptualize them to be). It is an example of Ronald Fisher's "Analysis of Variance":

Observed Variance =
True Variance + Error Variance

7. Reliability = True Variance / Observed Variance
Reliability = (Observed Variance - Error Variance) / Observed Variance

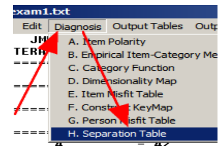
8. So now let's proceed to compute the Rasch-Measure-based Reliability for the current samples of persons and items

9. Look at Table 17 (or any person measure or item measure Table).
There is a column labeled "Measure". The variance of this column is the "Observed variance". It is the columns standard deviation squared.

The column labeled "Model S.E." or "Real S.E." quantifies the error variance for each person or item. In this example. the S.E. for child 32, "Tracie", is 1.30. So her error variance is 1.30^2 = 1.69. We can do this for each child.
The "error variance" we need for the item Reliability equation is the average of the error variances across all the items.
You can do this computation with Excel, if you like, but Winsteps has done it for you!

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REAL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	KID	
24	11	14	3.73	.94	.95	.1	.39	-.1	.69	85.7	88.8	Mike	M
32	11	14	3.73	1.30	1.94	1.7	.88	.4	.61	71.4	88.8	Tracie	F
7	10	14	2.85	.95	.36	-1.5	.15	-.5	.81	100.0	89.1	Blaise	M
11	9	14	1.94	.98	.52	-1.0	.18	-.4	.85	92.9	88.7	David	M
28	9	14	1.94	.98	.65	-.6	.22	-.4	.84	92.9	88.7	Rod	M
33	9	14	1.94	1.29	1.74	1.4	.71	.2	.77	78.6	88.7	Walter	M
35	9	14	1.94	1.24	1.60	1.2	.67	.2	.78	78.6	88.7	Zula	F
3	8	14	.92	1.06	.99	.2	.32	-.2	.86	85.7	91.1	Audrey	F
17	8	14	.92	1.06	.39	-1.0	.13	-.6	.89	100.0	91.1	James	M
20	8	14	.92	1.06	.99	.2	.32	-.2	.86	85.7	91.1	Kim	F
29	8	14	.92	1.06	.39	-1.0	.13	-.6	.89	100.0	91.1	Ron	M
34	8	14	.92	1.06	.39	-1.0	.13	-.6	.89	100.0	91.1	William	M
2	7	14	-.26	1.11	.18	-1.3	.08	-.7	.92	100.0	93.2	Anne	F
8	7	14	-.26	1.11	.18	-1.3	.08	-.7	.92	100.0	93.2	Brenda	F
12	7	14	-.26	1.55	1.94	1.2	1.19	.6	.82	85.7	93.2	Don	M
13	7	14	-.26	1.11	.18	-1.3	.08	-.7	.92	100.0	93.2	Dorothy	F
16	7	14	-.26	1.11	.18	-1.3	.08	-.7	.92	100.0	93.2	Helen	F
18	7	14	-.26	1.11	.18	-1.3	.08	-.7	.92	100.0	93.2	Janet	F
19	7	14	-.26	2.26	4.12	2.5	5.28	2.0	.65	71.4	93.2	Joe	M
22	7	14	-.26	1.11	.18	-1.3	.08	-.7	.92	100.0	93.2	Lisa	F
23	7	14	-.26	1.11	.18	-1.3	.08	-.7	.92	100.0	93.2	Martha	F
25	7	14	-.26	2.25	4.08	2.5	6.07	2.2	.65	71.4	93.2	Pete	M
27	7	14	-.26	1.36	1.49	.8	.51	.0	.85	85.7	93.2	Rick	M
31	7	14	-.26	1.36	1.49	.8	.51	.0	.85	85.7	93.2	Thomas	M
6	6	14	-1.37	1.00	.43	-.9	.16	-.5	.89	92.9	90.4	Betty	F
14	6	14	-1.37	1.07	1.16	.5	.97	.5	.83	92.9	90.4	Elsie	F
30	6	14	-1.37	1.00	.43	-.9	.16	-.5	.89	92.9	90.4	Susan	F
5	5	14	-2.23	.93	1.10	.4	.51	.0	.81	78.6	87.1	Bert	M
9	5	14	-2.23	.89	1.02	.2	.81	.3	.80	92.9	87.1	Britton	F
1	4	14	-2.94	.82	.61	-1.2	.29	-.2	.80	92.9	84.6	Adam	M
26	4	14	-2.94	.82	.61	-1.2	.29	-.2	.80	92.9	84.6	Richard	M
4	3	14	-3.61	.90	1.22	.8	.86	.4	.69	85.7	82.5	Barbara	F
10	3	14	-3.61	.82	.78	-.7	.36	-.1	.74	85.7	82.5	Carol	F
21	2	14	-4.32	.91	1.09	.4	.46	.0	.67	78.6	85.7	Linda	F
15	0	14	-6.62	1.85	MINIMUM ESTIMATED MEASURE							Frank	M

10. On the Winsteps menu bar,
Click on "Diagnosis"
Click on "H. Separation Table"



11. Let's investigate the second Table of numbers:
SUMMARY OF 35 MEASURED (EXTREME AND NON-EXTREME) KIDS
This Table corresponds to Cronbach-Alpha. Indeed, if Cronbach-Alpha is estimable, its value is below the Table:
CRONBACH ALPHA (KR-20) KID RAW SCORE RELIABILITY = .75

This Table summarizes the person distribution. The mean (average) person measure is -.37 logits. The (observed) Person S.D. is 2.22 logits. So the observed variance = 2.22^2 = 4.93.
The square-root of the average error variance is the RMSE = "root-mean-square-error". There is one RMSE for the "Real SE" = 1.21, and a smaller one for the "Model SE" = 1.05. The "true" RMSE is somewhere between. So the "model" error variance 1.05^2 = 1.10.

In the Table, "Adj. SD" means "Adjusted for Error" standard deviation, which is generally called the "True S.D."

SUMMARY OF 35 MEASURED (EXTREME AND NON-EXTREME) KIDS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	6.7	14.0	-.37	1.03				
S.D.	2.4	.0	2.22	.17				
MAX.	11.0	14.0	3.73	1.85				
MIN.	.0	14.0	-6.62	.82				

REAL RMSE 1.21 ADJ.SD 1.86 SEPARATION 1.55 KID RELIABILITY .70
MODEL RMSE 1.05 ADJ.SD 1.96 SEPARATION 1.87 KID RELIABILITY .78
S.E. OF KID MEAN = .38

KID RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) KID RAW SCORE RELIABILITY = .75

"True" Variance = "Adjusted for error variance"

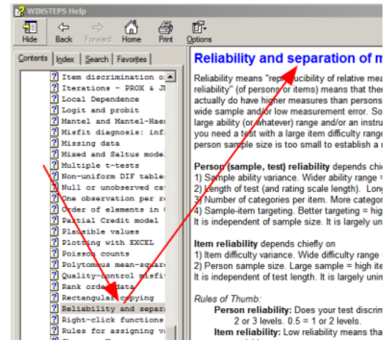
"Model Reliability" =
(S.D. ² - Model RMSE²) / S.D. ²
= $(2.22^2 - 1.05^2) / 2.22^2$ = 0.78

12. Here is a useful Table showing how the average, RMSE, standard error, the True S.D., the Observed S.D. and the Reliability relate to each other. It is from the Winsteps Help "Special Topic", "Reliability". This Table is very important to the understanding of the reproducibility (=Reliability) of measures. Please look at ....

Error RMSE	True SD	True Variance	Observed Variance	Signal-to-Noise Ratio	Separation = True SD / RMSE	Reliability = True Variance / Observed Variance
1	0	0	1	0	0	0
1	1	1	2	1	1	0.5
1	2	4	5	2	2	0.8
1	3	9	10	3	3	0.9
1	4	16	17	4	4	0.94

13. Winsteps menu bar
Click on Help
Click on Contents
Click on Special Topics
Click on Reliability

Read the Reliability topic. Notice particularly that 0.5 is the minimum meaningful reliability, and that 0.8 is the lowest reliability for serious decision-making.



14. Of course, "High Reliability" does not mean "good quality"! A Reliability coefficient is sample-dependent. A "Test" doesn't have a reliability. All we have is the reliability for this sample on this test for this test administration.

Since Reliability coefficients have a ceiling of 1.0, they become insensitive when measurement error is small. As the standard error decreases, the separation increases, but the reliability squeezes toward its maximum value of 1.0. That is why Ben Wright devised the "Separation Coefficient".

Separation = True SD / RMSE	Reliability = True Variance / Observed Variance
0	0
1	0.5
2	0.8
3	0.9
4	0.94

Separation = True "Adjusted" S.D. / RMSE

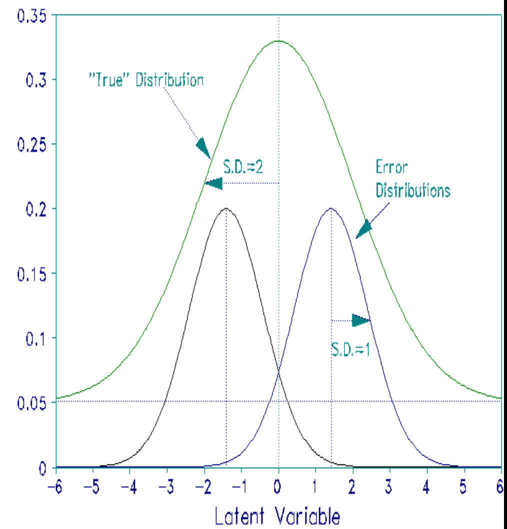
15.

True S.D	**Standard Error** RMSE	**Separation** = True S.D. / RMSE	**True Variance** = True S.D.²	**Observed Variance** = True Variance + RMSE²	**Reliability** = True Variance / Observed Variance
1	100.00	.01	1	10001	0.00
1	1.00	1	1	2.00	0.50
1	0.50	2	1	1.25	0.80
1	0.33	3	1	1.11	0.90
1	0.25	4	1	1.06	0.94
1	0.20	5	1	1.04	0.96
1	0.17	6	1	1.03	0.97
1	0.14	7	1	1.02	0.98
1	0.12	8	1	1.01	0.98
1	0.11	9	1	1.01	0.99
1	0.10	10	1	1.01	0.99

Notice how, as the standard error decreases, the separation increases, but the reliability squeezes toward its maximum value of 1.0

16. *The Person Reliability reports how reproducible is the person measure order of this sample of persons for this set of items*
So how can we increase the "Test" Reliability? For Winsteps, this is how can we increase the "person sample" reliability?
1. Increase the observed standard deviation by testing a wider ability range
2. Increase the person measurement precision so that we decrease the average person S.E. - we do this most effectively by increasing the number of items on the Test.

Increasing person sample size will not increase person reliability unless the extra persons have a wider ability range.

17.	In Rasch situations, we also have an item reliability. This reports *how reproducible is the item difficulty order for this set of items for this sample persons.* Since we don't usually want to change the set of items, the solution to low item reliability is a bigger person sample.	***If the item reliability is low, you need a bigger sample!***
18.	Here is the picture from http://www.rasch.org/rmt/rmt94n.htm showing how a reliability of 0.8 really works. The upper green line shows the conceptual "normal" distribution of a sample with standard deviation of "2", as if we could measure each person perfectly precisely without any measurement error. Now let's measure a sub-sample of persons, all of whose "true" measures are at -1.5. We would expect them to be spread out in a bell-shaped distribution whose standard deviation is the standard error of measurement. Let's say that the S.E. is 1. This is the left-hand lower curve. Now let's do the same thing for a sub-sample of persons, all of whose "true" measures are at +1.5. This is the right-hand lower curve.	
19.	In the Figure above, notice what happens when we add the two lower curves. Their sum approximates the top The entire true person distribution can be explained by two "true" levels of performance, a high performance and a low performance, measured with error. So what is the reliability here? Reliability = True Variance / (True Variance + Error Variance) = True S.D.^{2}/ (True S.D.^{2} + S.E.^{2}) = 2^2 / (2^2 + 1^2) = 0.8 So a reliability of 0.8 is necessary for to reliably distinguish between higher performers and low performers. Or perhaps high-medium-low, if the decisions are regarding the extreme tails of the observed distribution.	
20.	*Reliability rules-of-thumb:* 1. If the Item Reliability is less than 0.8, you need a bigger sample. 2. If the Person Reliability is less than 0.8, you need more items in your test. 3. Use "Real Reliability" (worst case) when doing exploratory analyses, "Model Reliability" (best case) when your analysis is as good as it can be. 4. Use "Non-Extreme Reliability" when doing exploratory analysis, use "Extreme+Non-Extreme Reliability" when reporting. 5. High item reliability does ***not*** compensate for low person reliability.	

1.

Appendix 2. Computing INFIT and OUTFIT "ZSTD" Fit Statistics

2.

Mean-square statistics indicate the size of the misfit, but statisticians are usually more concerned with the improbability of the misfit, its "significance".
So corresponding to each mean-square there is a ZSTD statistic showing the probability of the mean-square as a unit-normal deviate (again, see Lesson 1 Appendix 7 if you don't know about these).
The ZSTD is the probability associated with the null hypothesis: "These data fit the Rasch model". In conventional statistics, when $p<.05$, i.e., ZSTD is more extreme than ±1.96, then there is "statistical significance", and the null hypothesis is rejected.

ZSTD = "the probability of a mean-square, standardized like a z-statistic, which has a N(0,1) distribution"

Wilson-Hilferty transformation:

$q^2 = 2/\text{d.f.}$,
where d.f. ≈ MnSq divisor

$$\text{ZSTD} = (\text{MnSq}^{1/3} - 1)(3/q) + (q/3)$$

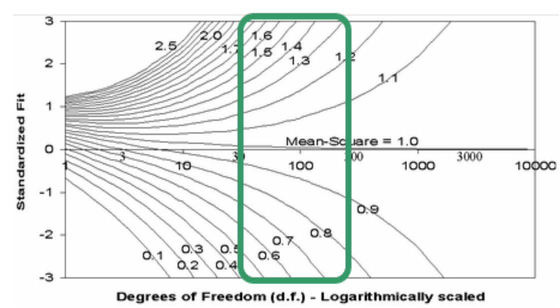
3.

ZSTD means "Standardized like a Z-score", i.e., as a unit-normal deviate. So we are looking for values of 2 or more to indicate statistically significant model misfit.

```
------------------------
.|    INFIT   |  OUTFIT   |
 |MNSQ   ZSTD |MNSQ   ZSTD|
-+------------+-----------+
.|4.08    2.5|6.07    2.2|.
```

4.

The relationship between significance (ZSTD) and size (MnSq) is controlled by the degrees of freedom (d.f.). See the plot in Winsteps Help "Misfit Diagnosis .." or http://www.winsteps.com/winman/diagnosingmisfit.htm
We can see that if the d.f. (x-axis) are too small (less than 30) even huge misfit is statistically insignificant, but if the d.f. are too large (greater than 300), then substantively trivial misfit is statistically significant.
Notice that mean-squares greater than 1, noisy underfit, are reported with positive ZSTD, but mean-squares less than 1, muted overfit, are reported with negative ZSTD.



5.

When sample sizes become huge, then all misfit becomes statistically significant (red boxes). Here the sample sizes are in the thousands. Even the substantively trivial mean-square of 1.12 is reported as statistically significant.

```
+--------------------------------------------------------------
|ENTRY    RAW                       MODEL|   INFIT   |  OUTFIT  |
|NUMBER  SCORE   COUNT   MEASURE   S.E.  |MNSQ  ZSTD|MNSQ  ZSTD|
|----------------------------------------+----------+----------+
|     1   4000   14000    -3.06     .03| .61  -9.9  .29  -9.9
|     2   7000   14000     -.27     .04| .18  -9.9  .08  -9.9
|     3   8000   14000      .98     .03|1.03   1.3| .32  -9.9
|     4   3000   14000    -3.73     .03|1.23   9.9  .90  -1.6|
|     5   5000   14000    -2.34     .03|1.12   7.9  .51  -9.9|
```