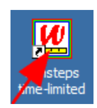
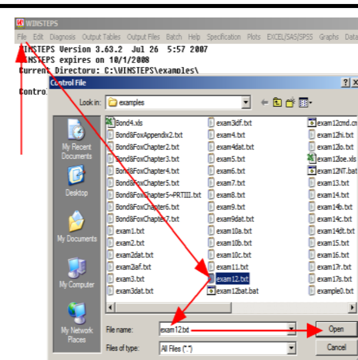
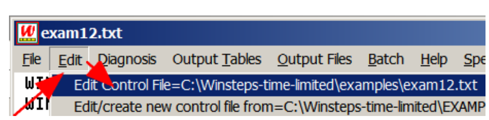
#	Winsteps Rasch Tutorial 3 Mike Linacre, Instructor – June 2012	
1.	**Tutorial 3.** ***More good stuff!*** • Partial Credit Model • Category Description • Standard errors and Reliability • Anchoring This lesson builds on Lessons 1 and 2, so please go back and review when you need to. If you run into difficulties or want to talk about what you are learning, **please post to the Discussion Forum**. http://www.winsteps.com/forum ***What this Course is really about:*** We are learning a lot of technical words and details, but this Course is really about a more powerful way of thinking about the world around us. The Course imagines the world to be composed of perfect equal-interval latent variables. These express the meaning in everything that is around us. We can reconstruct those perfect variables from imperfect data. Those perfect variables give us great insights into why things are the way they are. They gives us the security that comes from knowing what will probably happen. They even gives us power to change the future. All this sounds impossible, far out of the reach of mere humans. But "You have to believe the impossible" (Howard Head, inventor of Head skis and Prince tennis rackets). This Course is about the impossible.	
2.	**A. Rasch-Masters Partial Credit Model**	
3.	In 1982, Geoff Masters (now the Chief Executive of the Australian Council for Educational Research) took rating scale analysis a step further than the Andrich Model. Geoff was investigating multiple-choice questions and the fact that some distractors (incorrect options) are closer to correct than others. Shouldn't the examinee obtain "partial credit" for choosing a partially correct answer? We expect the partial-correctness structure to be different for different items, so Geoff constructed a version of the Rasch rating scale model where the rating scale (partial-credit scale) is specific to each item. *Masters GN 1982. A Rasch model for partial credit scoring. Psychometrika 47 149-174*	
4.	The **Rasch-Masters Partial Credit Model** specifies the probability, P_{nij}, that person *n* of ability measure B_n is observed in category *j* of a rating scale specific to item *i* of difficulty measure (or calibration) D_i as opposed to the probability $P_{ni(j-1)}$ of being observed in category *(j-1)* of a rating scale with categories *j=0,m* *Reminder ...* the "item difficulty" is the Rasch measure of the item. the "person ability" is the Rasch measure of the person. and we can continue to other areas ... the "rater severity" is the Rasch measure of the rater the "task challenge" is the Rasch measure of the task	$\log_e(P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_{ij}$ or $\log_e(P_{nij} / P_{ni(j-1)}) = B_n - D_{ij}$

5.	The rating scale structure $\{F_{ij}\}$ is now specific to item *i.* It works exactly like the rating scale structure for the Rasch-Andrich model. But there is a conceptual difference. We can think about the item difficulty and then impose the rating scale structure on it, $\{D_i + F_{ij}\}$, or we can think about the combination, $\{D_{ij}\}$. Mathematically they are the same thing. It is usually more straightforward to conceptualize and communicate the item difficulty separately from the rating scale structure, so we will use the $\{D_i + F_{ij}\}$ notation. The $\{F_{ij}\}$ are the "Rasch-Andrich thresholds" even when the model is not the Rasch-Andrich model. They are the points of equal probability of adjacent categories. The item difficulty D_i is the point where the top and bottom categories are equally probable. Suppose that we have two four-category items. Item 1 has the rating-scale: "Never, sometimes, often, always", and Item 2 has the rating-scale: "None, some, a lot, all". We score the categories of both items 1,2,3,4. And suppose that both items have the same overall difficulty. In a Partial Credit analysis, these two items will have different Rasch-Andrich thresholds $\{F_{ij}\}$. **This means that partial credit items with the same number of categories, and the same total raw "marginal" score, taken by the same people, can have different difficulties if the pattern of category usage differs between the items.** In Winsteps, the Partial Credit model is specified with ISGROUPS=0 (same as GROUPS=0)

6.	**B. Rating Scale Model for Item Groups**	
7.	Many assessments, observations instruments and surveys are composed of subsets of items which share the same rating scale. For instance, Items 1 to 10 could be "agreement" items. Items 11-20 could be "frequency" items (never, sometimes, often, always), Items 21-30 could be "quantity" items (none, a few, a lot, all).	
8.	The **Rasch-Grouped Rating Scale Model** specifies the probability, P_{nij}, that person *n* of ability B_n is observed in category *j* of a rating scale specific to a group of items, *g*, applied to item *i* of difficulty D_i as opposed to the probability $P_{ni(j-1)}$ of being observed in category *(j-1)*	$\log_e(P_{nij} / P_{ni(j-1)}) = B_n - D_{gi} - F_{gj}$
9.	Notice the subscript "g". This specifies the group of items to which item i belongs, and also identifies the rating scale structure that belongs to the group. In Winsteps, the "grouped scale" model is specified with ISGROUPS=AAAABBBA... where AAAA means "items 1-4 share scale A", BBB means "items 5-7 share scale B". Item 8 also shares scale A, etc.	
10.	Now for some algebra for polytomous Rasch models. This parallels the algebra for the dichotomous Rasch models we saw earlier: X_{ni} = scored observation E_{ni} = expected value of the observation R_{ni} = residual W_{ni} = variance of the observation around its expectation = the statistical information in the observation	$E_{ni} = \sum_{j=0}^{m} jP_{nij}$ $R_{ni} = X_{ni} - E_{ni}$ $W_{ni} = \sum_{j=0}^{m} \left[(j - E_{ni})^2 P_{nij}\right]$
11.	Let's try the Partial Credit model with a dataset to see what happens. Launch Winsteps	
12.	Our Control file is "exam12.txt" On the Winsteps Menu bar, click on "File" in the file dialog box, click on "exam12.txt" click on "Open".	
13.	Let's take a look at exam12.txt - it has unusual features. On the Winsteps Menu bar, Click on "Edit" Click on "Edit Control File"	

14. Scrutinize this Control file. You should understand most of it. If there is something you don't, look at Winsteps Help or ask on the Discussion Board.

This is part of the Functional Independence Measure®. There are NI=13 items of physical functioning, listed between &END and END NAMES.

There is a CODES= 7-category rating scale, with category names given in a sub-list after CLFILE=*.

@TIMEPOINT= identifies a demographic code in the person label. This will be used in later analyses.

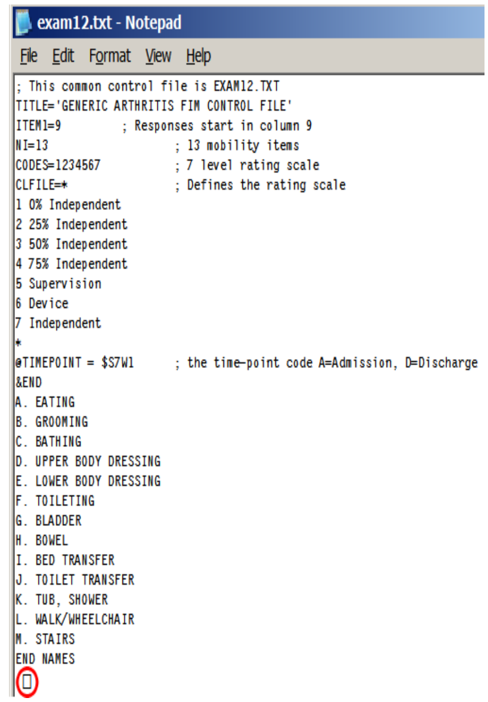
There is no NAME1=, so NAME1=1 is assumed.

There is no NAMELENGTH=, so the person label is assumed to end just before ITEM1= or at the end of the data line (whichever comes first).

There is no data after END NAMES.

You may see a □ *(in the red circle). It is the old MS-DOS end-of-file marker. If you see it, ignore it. Winsteps does!*

We can specify that the DATA is in a separate file with the control instruction: DATA=, but there is none here. We could edit the control file, but not this time ....



15. Winsteps Analysis window:

Report Output ...? Click Enter

Extra Specifications ...? Type in (or copy-and-paste):

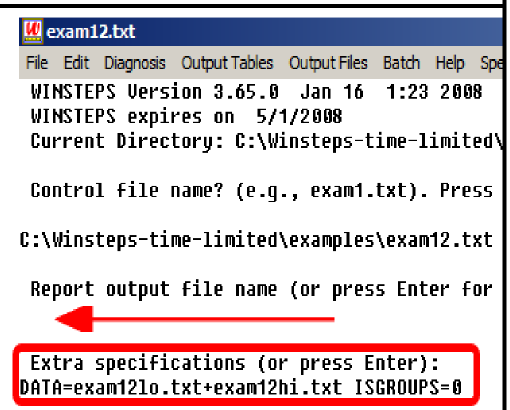
DATA=exam12lo.txt+exam12hi.txt ISGROUPS=0

(Be sure the only space is between .txt and ISGROUPS)

Press **Enter**

The analysis completes ....

We have told Winsteps to access two data files. It will read exam12lo.txt first, and then exam12hi.txt, effectively "stacking" them. Each data file has the same format. Exam12lo was collected at admission to medical rehabilitation. Exam12hi was collected at discharge.



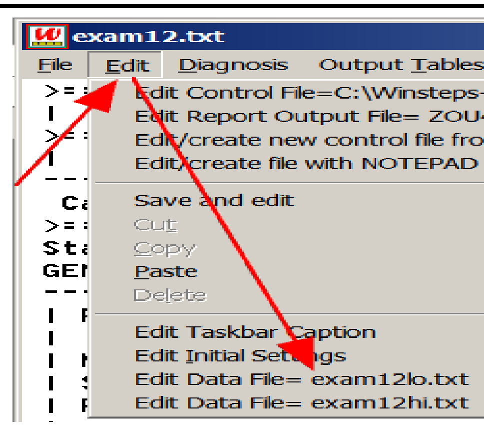
16. Let's look at one of the data files

Winsteps Analysis window:

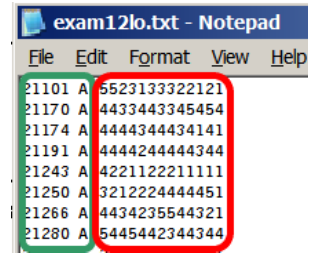
Click on "Edit"

Click on "Edit Data File= exam12lo.txt"

Almost all Winsteps input and output files for the current analysis are shown on the Edit menu list, and are editable. Explore the other choices on this menu. If you see anything that makes you curious (which you will!), talk about it on the **Discussion Board.** If you create a Window for a Table, and then close it, you can access it again using this Edit menu.

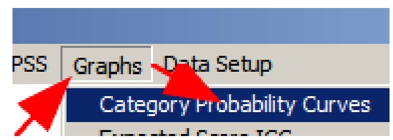


17. Here is the Data File edit window. The person (patient) labels are in the green box. The number is the patient identified. The letter, "A" or "D" is the time-point, Admission or Discharge. The item responses are in the red box. They would look exactly the same if they had been in the control file after END NAMES (or END LABELS, which means the same thing).



18. Back to the Winsteps Analysis window ...
We also specified ISGROUPS=0
This says "Each item is an item-structure group of 1 item, defining its own rating scale structure". In other words, the **Partial Credit Model.**

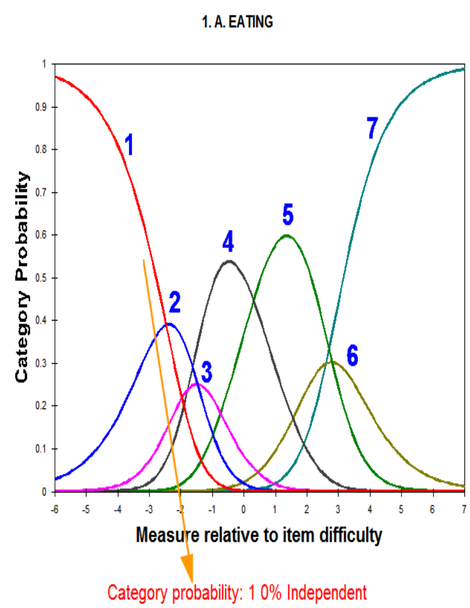
19. What has the Partial Credit model done for us?
Winsteps Menu bar
Click on "Graphs"
Click on "Category Probability Curves"



20. Look at the probability curves for the first item "Eating". These are drawn according to the Partial Credit model for this item. I've numbered them. The peaks of the curves for the 7 categories are always in ascending order along the latent variable, but the cross-over points between the curve for one category and curves of its neighboring categories can be reversed ("disordered"). You can see this for category 3. Its cross-over point with the curve for category 4 is to the left of the cross-over point for category 2. As with the Andrich Rating Scale model, the cross-over, equal-probability points ("thresholds") are the parameters of the Partial Credit model.

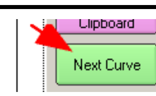
If you want to know which curve is for which category, click on the curve. Its description appears below the plot. I've clicked on the red line of the first category. Its name is "0% Independent", as specified by CLFILE= in the control file, or click on the "Legend" button.

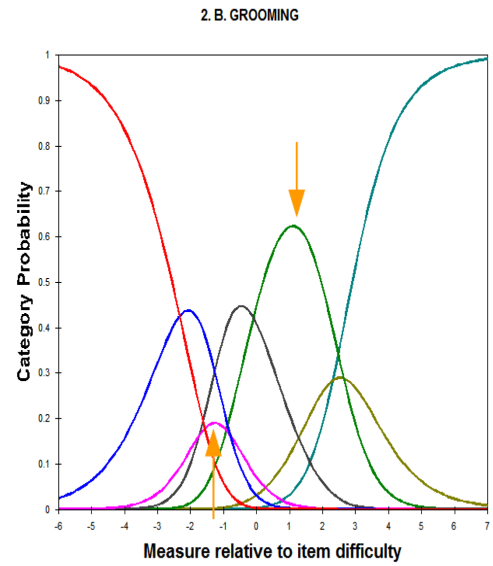
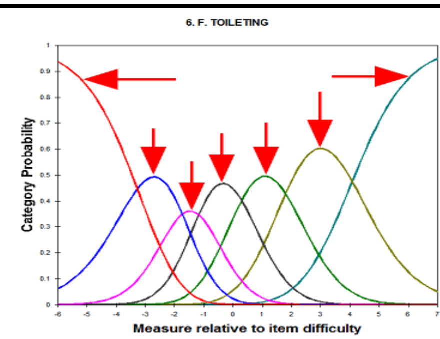
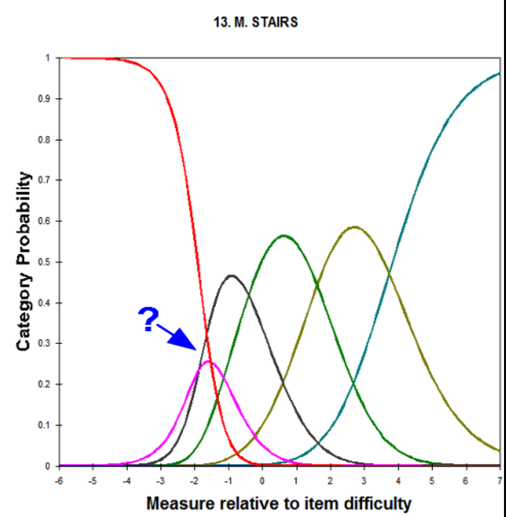
We like this plot to look like a range of hills (a series of distinct peaks). But in this graph, categories 3 and 6 are not distinct peaks. They are never "most probable". This could be a result of the category definitions, but it could also indicate something unusual about this sample of patients.

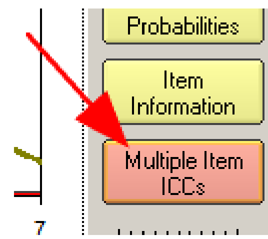
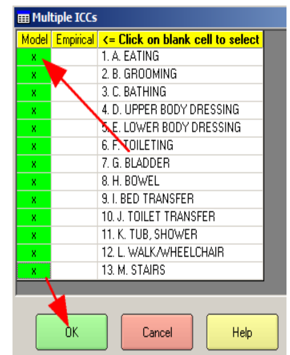
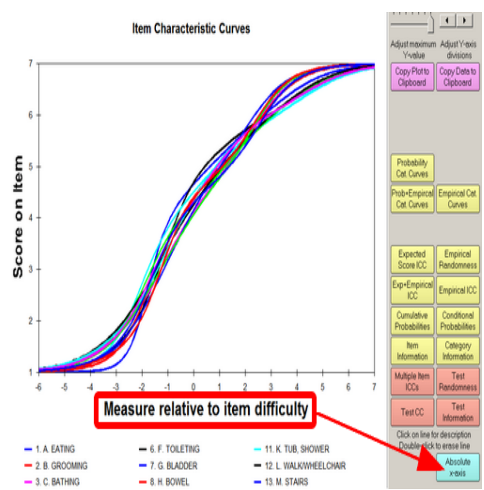


A range of hills:

21. On the Graphs Window
Click on "Next Curve"



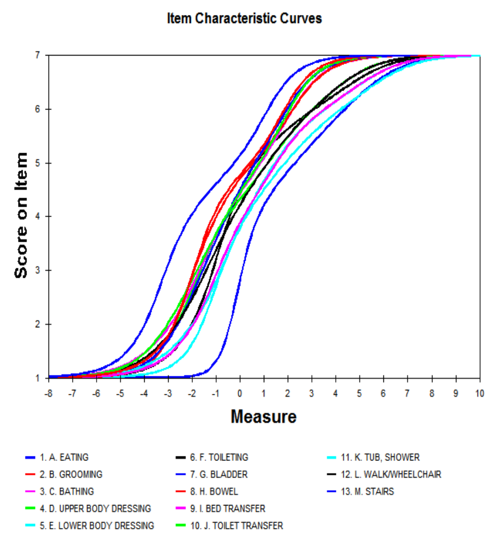
22.	Notice that the category curves for Item 2 "Grooming" look somewhat different from those for Item 1 which were just looking at. If this had been the Andrich Model, the set of curves for item 2 would have looked exactly the same as the set of curves for item 1. We are allowing each item to define its own category-probability structure. This means that the data should fit each item slightly better than for an Andrich rating scale, but it also means that we have lost some of the generality of meaning of the rating scale categories. We can't say "Category 1 functions in this way ...". We have to say "Category 1 for item 1 functions in this way ...." Each partial credit structure is estimated with less data, so the stability of the thresholds is less with the Partial Credit model than with the Rating Scale model. In general, 10 observations per category are required for stable rating-scale-structure "threshold" parameter estimates. "Stable" means "robust against accidents in the data, such as occasional data entry errors, idiosyncratic patients, etc."	
23.	Click "Next Curve" down to Item 6. The shows the category probability curves we like to see for the Partial Credit Model and the Andrich Rating Scale Model. Each category in turn is the modal (most probable) category at some point on the latent variable. This is what we like to see. "A range of hills" (arrowed in red). "Most probable" means "more probable than any other one category", i.e., the modal category, not necessarily "more probable than all other categories combined", i.e., not necessarily the majority category.	
24.	Click "Next Curve" all the way down to Item 13. Stairs. Look at each graph. Do you see many variations on the same pattern? Each item has its own features. A practical question we need to ask ourselves is "Are they different enough to merit a separate rating scale structure for each item, along with all the additional complexity and explanation that goes along with that?" Here are the curves for Item 13. Stairs. **Do you notice anything obviously different?** There is no blue curve. *[If you see a blue curve, you have missed entering ISGROUPS=0. Please go back to #15]* Category 2 has disappeared! These are real data. Category 2 on Item 13 was not observed for this sample of patients.	
25.	We have to decide: Is this Category 2 a "**structural zero**", a category that cannot be observed? Or is category 2 an "**incidental or sampling zero**", a category that exists, but hasn't been observed in this data set. Since this dataset is small, we opt for "sampling zero", so no action is necessary. If it had been a "structural zero", then we would need to re-conceptualize, and rescore, the rating scale for this item.	

26.	The Category Curves are like looking at the Partial Credit structure under a microscope, small differences can look huge. The Category Probability Curves combine into the Model "Expected" ICC which forms the basis for estimating the person measures. So let's look at the model ICCs. On the Winsteps Graphs Window Click on "Multiple Item ICCs"	
27.	Let's go crazy and look at all the Model ICCs at the same time. Position your mouse pointer on the first cell in the "Model" column, and left-click. The cell turns green with an "x" in it. That item is selected. If you click on the wrong cell and it turns green, click on it again to deselect it. Click "OK" when all the cells have been selected.	
28.	If you see "**Measure relative to item difficulty**", Click on the "**Absolute Axis**" button Note: "Relative axis" and "Absolute axis". Let's imagine you are standing on the top of the mountain and I am standing at the foot of the mountain. Then someone says "What is the height to the top of your head?" If we both measure our heights from the soles of our feet, then we are measuring "Relative" to our feet (relative to item difficulty). Our heights will be about the same. If we both measure our heights from sea level, then we are measuring in an "Absolute" way. This is "relative to the latent trait". Our heights will be considerably different. We measure "Relative to item difficulty", when we want to focus on the characteristics of the item. We measure "Absolutely, relative to the latent trait", when we want to place the item in the context of the entire test.	

29. We have great example of modern art! Each colored line is an item ICC. They are position left-to-right according to the item difficulty. Many curves are similar, effectively parallel. The blue curve at the left is 1. Eating. It is an easier item, but its shape resembles those of most of the more difficult items.

The blue line to the right is 13. Stairs. We know it has an unobserved category. This causes the sharp turn, almost a vertical jump, at the lower end of the ICC.

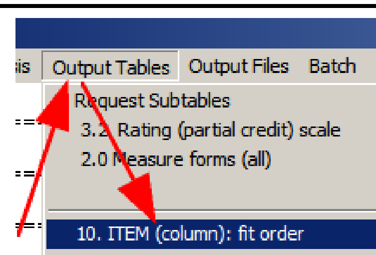
But, in general, these curves are so similar that the improved fit of having a different "model ICC" for each item is not worth the extra communication load of explaining to users of the FIM the niceties of a separate set of threshold estimates for each item - the Partial Credit model.



30. *At any point along the ICC, the steeper it is:*
1. the more discriminating it is between high and low performers.
2. the more statistical information (information in the observation about the person parameter value that is to be estimated) that there is, then the smaller the standard error of person measures in that neighborhood.

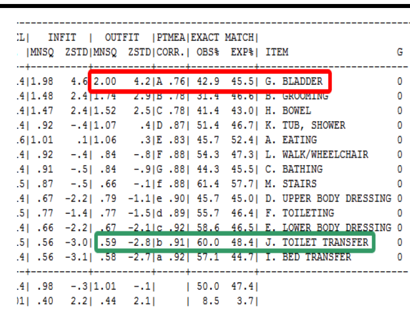
But, if an ICC is steeper at some locations, it must be flatter at other locations, so partial-credit ICCs for the same number categories always contain the same total statistical information.

31. But, even at its best, the Partial Credit Model is still a conceptual ideal, and the data can still be "hornery critters" (wild horned animals, horned creatures) that refuse to be tamed. Has that happened here?
Let's see which item badly underfit ...
Click on Winsteps menu bar
Click on Output Tables
Click on 10. ITEM (column): fit order



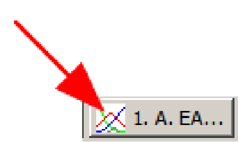
32. Here in Table 10.1, the items in fit order, Item 7. "Bladder" is badly underfitting, the noise is starting to drown out the music. Item 10, "Toilet Transfer" is over-fitting. It is somewhat too predictable. Its music is only at a about half-volume.

What do these mean in terms of the empirical (data-derived) ICC?

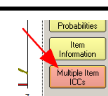


```
-------------------------------------------------------------------------+
L|   INFIT   |  OUTFIT   |PTMEA|EXACT MATCH|                              |
 |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| OBS%  EXP%| ITEM                      G |
-+----------+----------+-----+-----------+------------------------------|
.4|1.98   4.6|2.00   4.2|A .76| 42.9  45.5| G. BLADDER                0 |
.4|1.48   2.4|1.74   2.9|B .78| 31.4  46.6| B. GROOMING               0 |
.4|1.47   2.4|1.52   2.5|C .78| 41.4  43.0| H. BOWEL                  0 |
.4| .92   -.4|1.07    .4|D .87| 51.4  46.7| K. TUB, SHOWER            0 |
.6|1.01    .1|1.06    .3|E .83| 45.7  52.4| A. EATING                 0 |
.4| .92   -.4| .84   -.8|F .88| 54.3  47.3| L. WALK/WHEELCHAIR        0 |
.4| .91   -.5| .84   -.9|G .88| 44.3  45.5| C. BATHING                0 |
.5| .87   -.5| .66   -.1|f .88| 61.4  57.7| M. STAIRS                 0 |
.4| .67  -2.2| .79  -1.1|e .90| 45.7  45.0| D. UPPER BODY DRESSING    0 |
.5| .77  -1.4| .77  -1.5|d .89| 55.7  46.4| F. TOILETING              0 |
.4| .66  -2.2| .67  -2.1|c .92| 58.6  46.5| E. LOWER BODY DRESSING    0 |
.5| .56  -3.0| .59  -2.8|b .91| 60.0  48.4| J. TOILET TRANSFER        0 |
.4| .56  -3.1| .58  -2.7|a .92| 57.1  44.7| I. BED TRANSFER           0 |
-+----------+----------+-----+-----------+------------------------------|
.4| .98   -.3|1.01   -.1|     | 50.0  47.4|                              |
)1| .40   2.2| .44   2.1|     |  8.5   3.7|                              |
-------------------------------------------------------------------------+
```

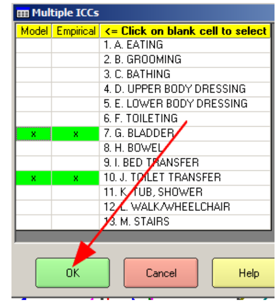
33. Click on the Graphs icon on your Windows Task bar,
or Winsteps Menu Bar
Click on "Graphs"
Click on "Category Probability Curves"



34. Click on "Multiple Item ICCs"
Yes, we are here again ....



35. We want the model and empirical ICCs for items 7 and 10.

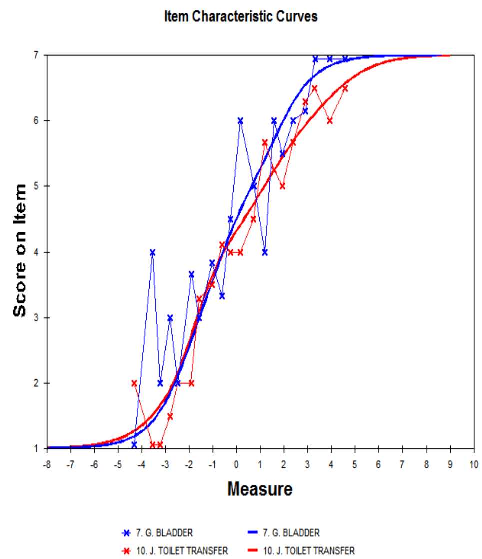


36. Can you see what is going on? There are 4 lines on this plot. The blue lines are for the bad item, "7. Bladder". The red lines are for the good item "10. Toilet Transfer".

The heavy smooth lines are the model ICCs. They differ slightly because we are using the Partial Credit model.

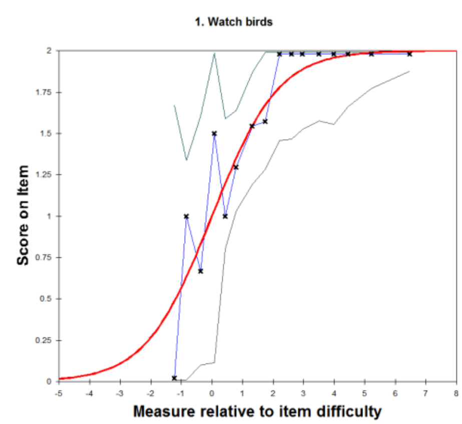
The jagged lines are the empirical ICCs summarizing the data in each interval on the latent variable (x-axis). Notice how much more the blue jagged under-fitting, unpredictable line, departs from its blue smooth line than the red jagged over-predictable line does from its red smooth line. This is what misfit and accuracy is all about. When the jagged line is too far away we cry "These data are too unpredictable!". When the jagged line is too close in, we cry "These data aren't telling me anything new, why did we bother to ask this item?"

We see that partial credit items have ICCs with different slope and different fit even though they have the same number of categories.



37. ***Practical Challenge:*** *Now it's your turn!*
1. Look at Table 10 - it is on your Windows Task bar.
2. Choose one good fitting item - mean-squares near 1.0.
3. Graphs Menu: use the "Exp+Empirical ICC" box to select its expected (model) and empirical (data) ICCs.
4. "Copy" the curves and paste into a Word (or similar) document.
5. Is the empirical line more jagged than you expected? They were for me the first time I saw them!

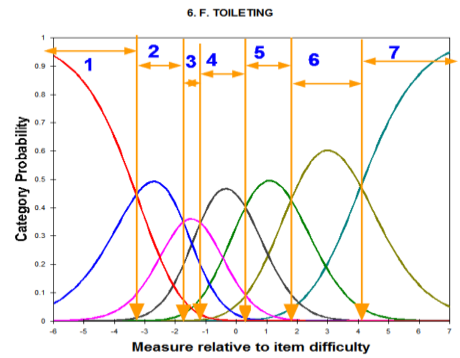
The Rasch model is probabilistic. It predicts randomness in the data. We like to think of the world would be a better place if it were deterministic. But no, **controlled randomness** makes for a better world. There are some examples at *"Stochastic Resonance"* - http://www.rasch.org/rmt/rmt54k.htm

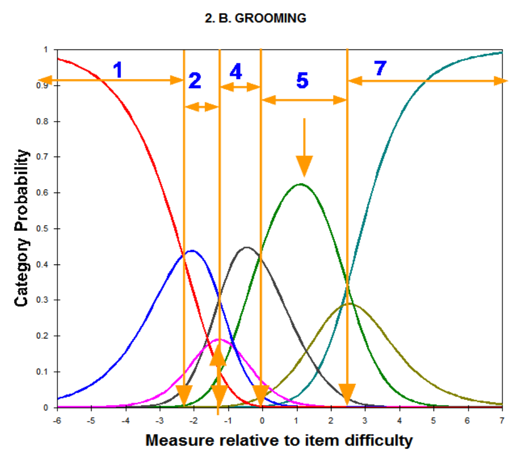
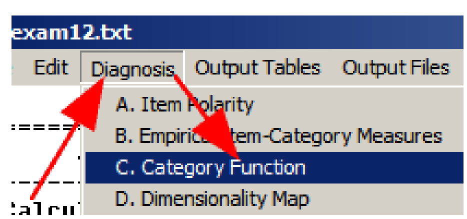


Graph for mean-square = 0.5

38.

<table>
<tr><td>39.</td><td colspan="2">C. Category Description</td></tr>
<tr><td>40.</td><td colspan="2">When talking about the functioning of a rating scale, we need to describe what interval on the latent variable corresponds to what category. There are three approaches in common use:
1. Modal - the most probable category defines the interval.
2. Median - the categories above and below define the interval.
3. Mean - the average category value defines the interval.</td></tr>
<tr><td>41.</td><td colspan="2">Modal - “Rasch-Andrich thresholds” (usually)</td></tr>
<tr><td>42.</td><td>We are continuing the analysis started at #15

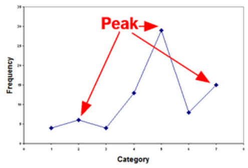
The most probable category, the “mode of the category distribution at each point on the latent variable”, defines the interval. This Figure shows how this works. Notice that “Most probable category” does not mean “More probable than all other categories combined.” For instance, in the “3” interval, you can see that “2 or 4” are more probable to be observed than 3, because the sum of their probabilities is always more than the probability of observing 3. Probability(2+4) > probability (3) everywhere.</td><td>
</td></tr>
<tr><td>43.</td><td>“Disordered Rasch-Andrich Thresholds”
The “Modal” explanation has another pitfall. Remember the “Grooming” item? The modal categories are 1,2,4,5,7. Two categories, 3 and 6, are never modal, never the most probable category. So they have no “modal” interval on the latent variable.

David Andrich, the most ardent proponent of the “modal” interpretation of rating scale functioning, declares this situation to be a violation of the Rasch model, so that the rating scale becomes undefined. From this perspective, the categories must be combined (collapsed). Category 3 must be combined with Category 2 or 4, and Category 6 must be combined with Category 5 or 7.</td><td>
</td></tr>
<tr><td>44.</td><td>This combination can be easily done in Winsteps. Let’s do it. First we need to decide which categories to combine (collapse).
Click on the Winsteps Menu Bar.
Click on “Diagnosis”
Click on “C. Category Function”</td><td>
</td></tr>
</table>

45. Table 3.2 displays in an Edit window. Scroll down to Table 3.3, the details of the Grooming item.

Red box: Notice the “Observed Count” column. In statistics, we like to see a smooth distribution of counts with one peak. These counts are:

4, 6, 4, 13, 20, 8, 15 → three peaks!



Green box: Notice also the large category mean-square statistics for category 4 (2,63, 5.89 where 1.0 is expected). Something is unexpected about the behavior associated with this category. We need to investigate it further.

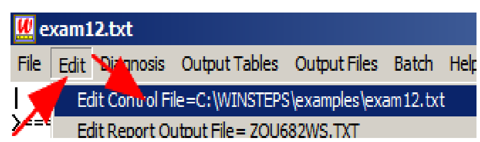
Item 2: Grooming

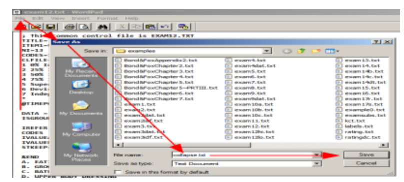
```
SUMMARY OF CATEGORY STRUCTURE.  Model="R"
FOR GROUPING "0" ITEM NUMBER: 2  B. GROOMING

ITEM DIFFICULTY MEASURE OF -.68 ADDED TO MEASURES
-------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||S
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||C
|-------------------+------------+------------++
|  1   1      4   6 | -2.86 -3.01|  1.11   .95||
|  2   2      6   9 | -2.32 -2.24|  1.18  1.03||
|  3   3      4   6 | -2.04 -1.57|   .38   .23||
|  4   4     13  19 |   .24  -.88|  2.63  5.89||
|  5   5     20  29 |  -.26*  .31|  1.48   .98||
|  6   6      8  11 |  2.22  1.71|   .74   .79||
|  7   7     15  21 |  2.44  2.79|  1.74  1.71||
-------------------------------------------------------
```

46. Let’s do a “thought experiment” - a favorite activity of Albert Einstein.
We could combine categories counts to make a smoother distribution. Here is a possibility:
1+2, 3+4, 5, 6+7 = 10, 17, 20, 23

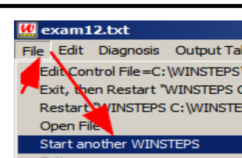
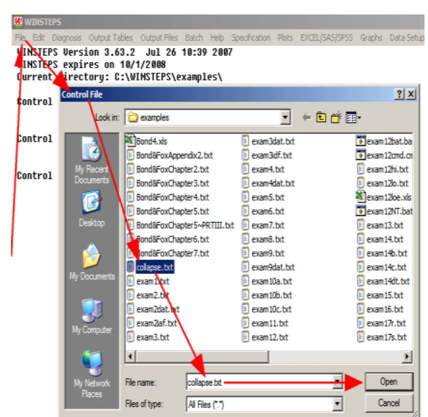
Blue box: Look at the “Observed Averages”. These are the average measures of the persons rated in each category. We notice the very low observed average for category 5, -0.26. This contradicts our theory that “higher measure ➔ higher category”. To eliminate this problem, we could combine category 5 with category 4 or category 6. So some other possibilities are:
A. 1, 2+3, 4+5, 6+7 = 4, 10, 33, 23 - No, almost half the observations (33) are in one category.
B. 1, 2+3, 4, 5+6, 7 = 4, 10, 13, 28, 15
C. 1+2, 3+4, 5+6, 7 = 10, 17, 28, 15
B. and C. both look good. What do you think?
The observed averages for 2 and 3 are close together, and 4 is halfway between 3 and 6, so I am choosing option B.
But, we would also want to consider what the categories mean. Which combination makes more sense?

47. Now to combine the categories ..... 1, 2+3, 4, 5+6, 7
Click on the Winsteps Menu Bar
Click on “Edit”
Click on “Edit Control File”
The exam12.txt Control file displays ...

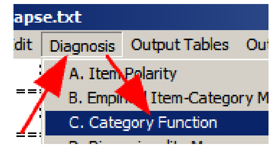


48.	Please type (or copy-and-paste) into the exam12.txt control file (before &END) the red and green type. The control variables previously entered at "Extra Specifications" are shown in the green type: In red type are the instructions to collapse the categories for "Item 2. Grooming". In IREFER= each item has a letter, starting with item 1. All items are coded "A", except item 2 which is coded "B". IVALUEA= rescores A items, IVALUEB= rescores B's.	DATA = exam12lo.txt+exam12hi.txt ; the data files ISGROUPS = 0 ; the Partial Credit model ; we want to rescore item 2: IREFER = ABAAAAAAAAAAA ; Items for rescoring CODES =1234567 ; 7 level rating scale IVALUEA=1234567 ; no rescoring at present ; for item 2: 1, 2+3, 4, 5+6, 7 IVALUEB=1334557 ; collapsing categories for item 2 ; we want to eliminate "structural zero" category 2 STKEEP=no ; squeeze out unobserved categories &END
49.	STKEEP= NO tells Winsteps that missing numbers in IVALUEB= are structural zeroes, and so do not correspond to qualitative levels of the variable. Missing category numbers are squeezed out when the categories are renumbered for analysis. ISGROUPS=0, each item has its own rating-scale thresholds, after the rescoring is done.	STKEEP=NO ; recount the categories IVALUEB = 1334557 is the same as IVALUEB = 1224667 they both mean = 1223445
50.		
51.	"Save as" this edited control file as "**Collapse.txt**" As usual, "Text format" is what we want. We want to leave "exam12.txt" unaltered for when use it later	

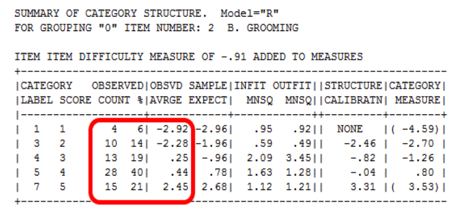
	If we wanted the "A" items to share the same rating scale, and the "B" item(s) to have a different scale, then we would specify:
	ISGROUPS= ABAAAAAAAAA instead of ISGROUPS=0
	Can you decipher how to rescore reversed items?→
		For reversed items:
		`IREFER  = ..x.x..`
		`CODES   = 1234567`
		`IVALUEx = 7654321`

52.	Winsteps menu bar Click on "File" menu Click on "Start another Winsteps" We can have many copies of Winsteps running at the same time	
53.	Winsteps launches .... On the Winsteps Menu bar ... Click on "File" Click on "Open File" In the Control File dialog box, Click on "Collapse.txt" Click on "Open" In the Winsteps Window: Report Output? Click enter Extra Specifications? Click enter The analysis is performed	

54. After the analysis completes,
Winsteps menu bar:
Click on “Diagnosis”
Click on “C. Category Function”

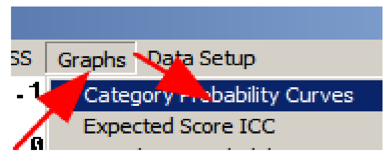


55. Table 3.2 displays.
Scroll down to Table 3.3.
Look at the Observed Count column. *Is it what we expected?*
Look at the Observed Average column. *Is it what we expected?*
How about the Category Label and Score columns?
If I’ve lost you - please ask on the Discussion Board ...



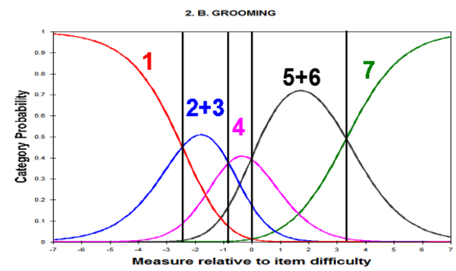
```
SUMMARY OF CATEGORY STRUCTURE.  Model="R"
FOR GROUPING "0" ITEM NUMBER: 2  B. GROOMING

ITEM ITEM DIFFICULTY MEASURE OF -.91 ADDED TO MEASURES
+------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN| MEASURE|
|-------------------+------------+------------++---------+--------+
|  1   1      4   6| -2.92 -2.96|   .95   .92||  NONE   |( -4.59)| 1
|  3   2     10  14| -2.28 -1.96|   .59   .49||   -2.46 |  -2.70 | 2
|  4   3     13  19|   .25  -.96|  2.09  3.45||    -.82 |  -1.26 | 4
|  5   4     28  40|   .44   .78|  1.63  1.28||    -.04 |    .80 | 5
|  7   5     15  21|  2.45  2.68|  1.12  1.21||    3.31 |(  3.53)| 7
+------------------------------------------------------------------
```

56. Now to the Category Probability Curves:
Winsteps menu bar
Click on Graphs
Click on Category Probability Curves



57. On the Graphs window,
Click on the Next Curve button
The revised curves for 2. Grooming display.

Is this what you expected to see? ***Do these curves make sense for a “Modal” explanation of the category intervals on the latent variable?***

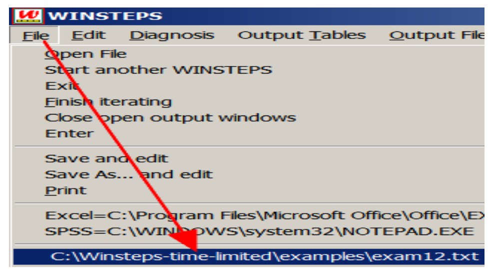


58. Close all windows

59. **D. Median - “Rasch-Thurstone thresholds”**

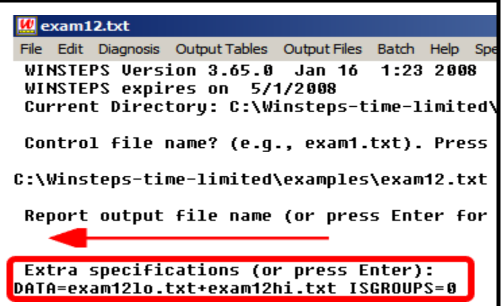
60. **50% Cumulative Probabilities define the intervals**. The “modal” viewpoint considered the rating scale functioning by means of one category at a time. But sometimes our focus is on an accumulation of categories. For instance, in the FIM, categories 1, 2, 3, 4 require physical assistance. Categories 5, 6, 7 do not. So we might want to define one category interval boundary in terms of 4 or below *versus* 5 or above. We can do the same with 1 vs. 2+3+4+5+6+7, then 1+2 vs. 3+4+5+6+7, etc. These cumulative-category interval boundaries are called Rasch-Thurstone thresholds. Psychometrician Leon L. Thurstone used them for his ground-breaking analyses in the 1920s, but he used a normal-ogive model. Rasch used a logistic model, so we call them Rasch-Thurstone thresholds to avoid confusion with Thurstone’s own approach.
They are “median” (middle) thresholds in the sense that the threshold is the location where the probability of being observed in any category below the threshold is the same as the probability of being observed in any category above it.

61. Launch Winsteps
“Quick start” exam12.txt using the file list

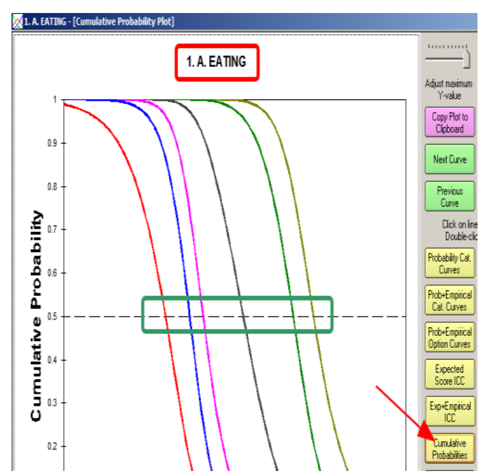


62. Same as before:

Winsteps Analysis window:
Report Output ...? Click Enter
Extra Specifications ...? Type in (or copy-and-paste):
DATA=exam12lo.txt+exam12hi.txt ISGROUPS=0
Press **Enter**
The analysis completes ....



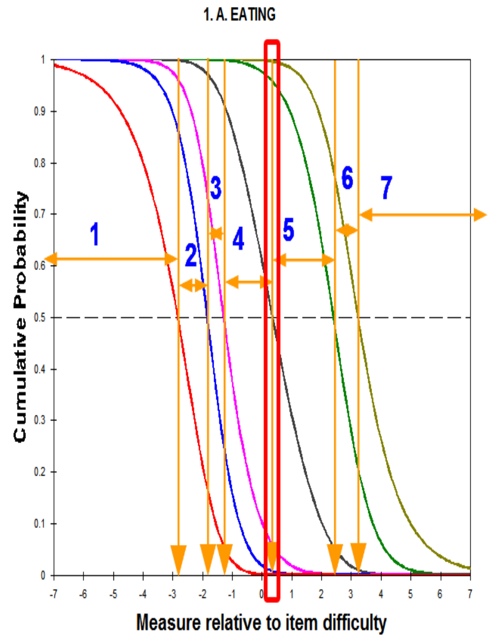
63. In the Graph window,
Click on “Cumulative Probabilities”
The accumulated probability curves display.
The left-hand red probability curve is the familiar curve for category “1”.
The next blue probability curve is the sum of the probabilities for categories 1 and 2.
The next pink curves is the sum of the probabilities for categories 1+2+3.
Then the **black** curve is 1+2+3+4. The green curve is 1+2+3+4+5.
The khaki curve is 1+2+3+4+5+6.
There are 7 categories, so the last curve would be 1+2+3+4+5+6+7. *Where is it?*
What line is the sum of the probabilities of all 7 categories?
Please ask on the Discussion Board if you can’t locate it.



If your curves are upside down, click on “Flip Curves Vertically”

64. The crucial points in this picture are the points where the curves cross the .5 probability line. These are the Rasch-Thurstone thresholds. I've boxed in **red** the vertical arrow corresponding to the medically important 1+2+3+4 vs. 5+6+7 decision point for this item.

According to this conceptualization, the crucial decision point for 1 vs. 2+3+4+5+6+7 is the left-hand orange arrow. So category "1" is to the left of the arrow. The decision point for 1+2 vs. 3+4+5+6+7 is the second orange arrow. So category "2" is between the first two arrows. *And so on ...* Notice that every category has an interval on the 0.5 line. The intervals are always in category order.



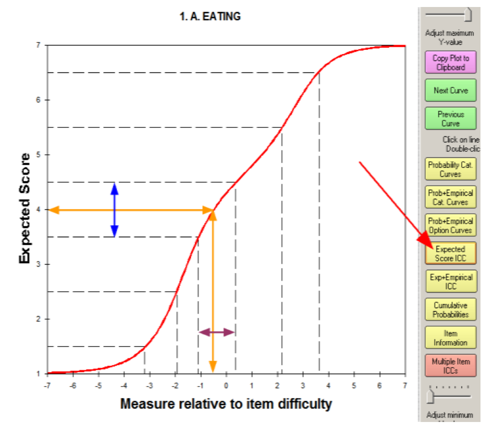
65. It is confusing to most non-specialist audiences to report every category of a rating scale. It is often useful to report a decision-point on the item. This does not require a re-analysis, but rather serious thought about how best to communicate findings.

The Australian Council for Educational Research (ACER) issues reports to millions of parents about their children. ACER looks at the cumulative probability picture, and choose the point on the latent variable where probable overall failure (the lower categories, 1+2+3+4) on the item meets probable overall success (the higher categories, 5+6+7). The decision about which categories indicate failure, and which categories indicate success, is decided by content experts. This "cumulative probability" or "median" transition point (red box) is the "difficulty" of the item *for the report to the parents.* It is not the same as the "difficulty" of the item for Rasch estimation.

66. **E. Mean - “Rasch-Half-Scorepoint” Thresholds**

67. **The average score on the item defines the category**. We may be more interested in predicting the expected score of a person on an item, or in predicting (or interpreting) the average performance of a sample on an item. For these we use terms like “George is a low 3” or “Anna is a high 4”. For a sample, their average performance on the item is 3.24. How are we to display and interpret these?

68. On Graphs menu
Click on **“Expected Score ICC”**
I’m looking at Item “1 A. Eating”
In this plot, the y-axis is the average expected score on the item, according to the Rasch model. It ranges from 1 to 7. The x-axis is the latent variable.
At each point on the x-axis, we ask “what is the average of the ratings we expect people at this point on variable to score?” We compute that expectation and plot it. The result is the red line.
From this picture we can make predictions: At what measure on the latent variable is the average score on the item “4”? For the answer follow the orange arrows in my adjacent Figure ...



69. Click on “ Probability Category Curves”, and ask the same question: What measure best corresponds to an observation of “4”?
Do you get the same or a different answer?
It should be the same. This is an important property of Rasch polytomous models: the peak of a category’s probability curve is at the measure where the expected score on the item is the category value.

Category probabilities and scores on the items: here is how they work. First, let's simplify the situation:

Imagine people advancing up the latent variable from very low to very high. The rating scale goes from 1 to 7 for the item we are thinking about.

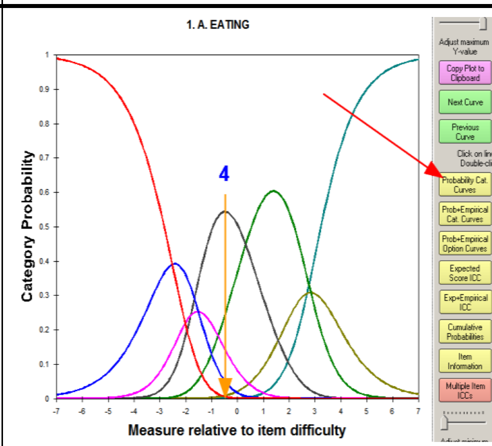
A very low performer: we expect that person to be in category 1 (the bottom category) and to score 1 on the item. OK?

A very high: we expect that person to be in category 7 (the top category) and to score 7 on the item. OK?

Along the way, we expect a middle performer to be in category 4 (the middle category) and to score 4 on the item. Good so far?

And we expect a lowish performer to be in category 3 and to score 3 on the item. Good so far?

But now imagine 1000 people all with the same measure, but between the category 3 performer and the category 4 performer. Let's say say about 1/4 of the way up from the category 3 performer and the category 4 performer. What do we expect to



There are three ideas:

"Category" value: a Likert scale has category values 1,2,3,4,5.

"Response category": the category of our response: 1 or 2 or 3 or 4 or 5

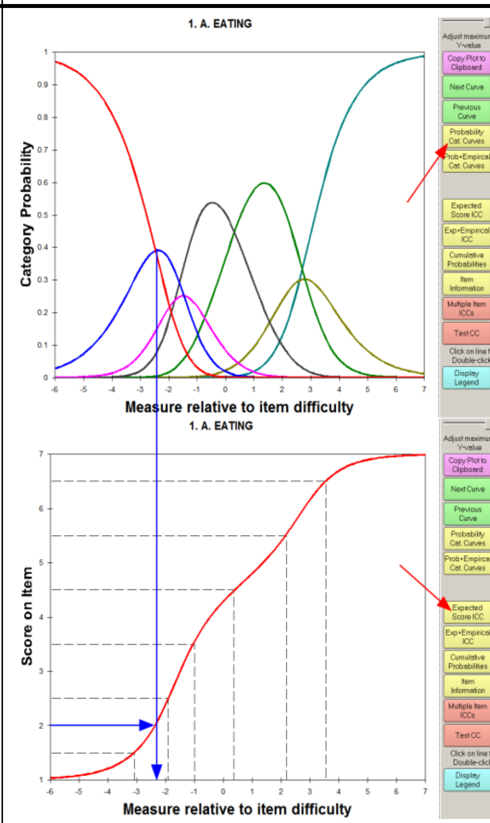
"Our expected response": the sum of the category values and the probability that we would respond in the categories.

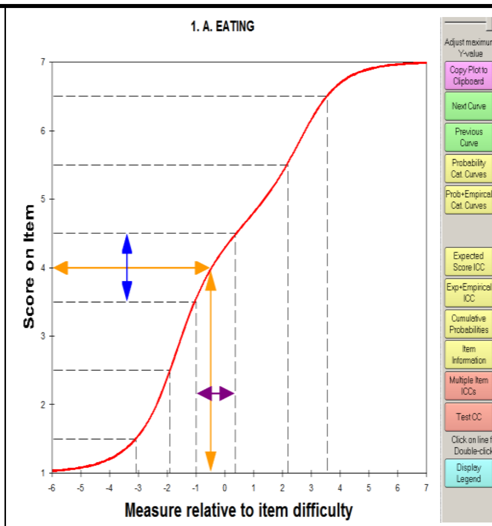
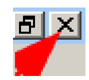
For instance, I could have

.1 probability of responding in category 1 = .1*1 = .1

.3 probability of responding in category 2 = .3*2 = .6

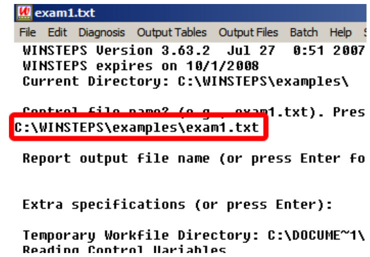
.4 probability of responding in

	observe? Those folks are close to the category 3 performer, so most of them will be observed in category 3. But we are moving into category 4, so some of the 1000 persons will be observed in category 4. The average score on the item for these people will be about: 3*0.75 + 4*0.25 = 3.25. So, overall, for one of these 1000 people, we expect to observe category 3 (the most probable category for these 1000 people), but the expected score on the item by this person is 3.25. The Rasch model is somewhat more complicated than this example, because any category can be observed for any person. But the principle is the same: for a person exactly at "category 3", category 3 has its highest probability of being observed, and that person's expected score on the item is 3.0. As we move from category 3 to category 4, the probability of observing category 3 decreases. The probability of observing category 4 increases. And the expected score moves smoothly from 3.0 to 4.0.	category 3 = .4*3 = 1.2 .15 probability of responding in category 4 = .15*4 = .6 .05 probability of responding in category 5 =.05*5 = .25 My "expected" response (= average response) = .1+.6+1.2+.6+.25 = 2.85 As we advance from strongly disagree (1) to neutral (3) to strongly agree (5), our expected ratings go 1.0, 1.1, 1.2, 1.3 .... 2.9, 3.0, 3.1, ... 4.9, 5.0 **If you don't follow this, ask about it *on the Discussion Forum ....***
70.	When our expected rating (2.0) is the same as a category value ***(2), then the probability that we would respond in the category "2" is the highest.*** An expected score of 2.0 on "Eating" corresponds to a logit difficulty of -2.4 (relative to the difficulty of "Eating"). This corresponds to the highest probability of observing a "2".	

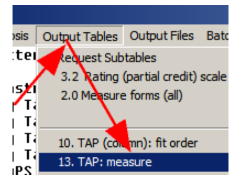
71.	We could say that the interval corresponding to category 4 contains any average expected score on the item from 3.5 to 4.5. These would round to 4. Those intervals are indicated by the --------- in the Figure. I find this "mean" or "average" intervals easier to explain to non-technical audiences than the Modal or Median intervals. Using this Figure, we can say things like "The average rating of 1,000 people at this point on the latent variable is around ..." or "the transition from 3 to 4 (i.e., for people who average 3.5) is at about ...." This Figure has only one curved line. In my experience, audiences become confused when they try to understand several curved lines at once. ***Communication is our big challenge - so choose the approach that best communicates your message to your target audience.***	
72.	Close all Windows	
73.		

74.	**F. Precision, Accuracy, Standard Errors**
75.	We now come to a highly technical, much misunderstood topic in measurement. Ask yourself: "What is the difference between *precision* and *accuracy*?"
76.	*Precision* indicates how closely a measure can be replicated. In shooting arrows at a target, high precision is when the arrows are closely grouped together. Precision is quantified by "standard errors", S.E. This is an internal standard. The measures are conceptually compared with each other.
77.	*Accuracy* indicates how closely a measure conforms to an external target. In shooting arrows at a target, high accuracy is when the arrow hits the bull's-eye. In Rasch measurement, the external standard is ideal fit to the Rasch model, and accuracy is quantified by *fit statistics.*
78.	*Standard Errors* - S.E.s - Every Rasch measure has a precision, a standard error. No measurement of anything is ever measured perfectly precisely. Raw scores, such as "19 out of 20", are often treated as though they are perfectly precise, but they are not. The approximate standard error of a raw score is straight-forward to compute (for 19 out of 20, it is roughly ±1.0 score points), but the computation is rarely done and almost never reported. In Rasch measurement, the S.E.s are always available. The square of a Standard Error is called the Error Variance.
79.	*Estimating the "Model" S.E.* - Earlier we saw that the model variance of an observation around its expectation is W_{ni}. This is also the statistical information in the observation. The Rasch-model-based "Model S.E." of a measure is 1 / square-root (statistical information in the measure) $Model\ S.E.(B_n) = 1 \Big/ \sqrt{\sum_{i=1}^{L} W_{ni}}$ $Model\ S.E.(D_i) = 1 \Big/ \sqrt{\sum_{n=1}^{N} W_{ni}}$
80.	This formula tells us how to improve precision = **reduce the Standard Error**: *Desired result: how to achieve it* 1. For smaller **ability** S.E.s → **increase the test length, L** (number of items taken) 2. For smaller **difficulty** S.E.s → **increase the sample size, N** (number of persons taking the item) 3. For smaller ability S.E.s and difficulty S.E.s → increase the number of categories in the rating scale 4. For smaller ability S.E.s and difficulty S.E. → improve the test targeting = decrease $\lvert B_n-D_i\rvert$
81.	*Estimating the "Real" S.E.* - The Model S.E. is the best-case, optimistic S.E., computed on the basis that the data fit the Rasch model. It is the smallest the S.E. can be. The "Real" S.E. is the worst-case, pessimistic S.E., computed on the basis that unpredictable misfit in the data contradicts the Rasch model. Real S.E. = Model S.E. * Max (1.0, √ INFIT MnSq) http://www.rasch.org/rmt/rmt92n.htm To report the Real S.E., specify: REALSE = YES
82.	The "true" S.E. lies somewhere between the Model S.E. and the Real S.E.. We eliminate unpredictable misfit as we clean up the data. By the end of our analyses, we can usually conclude that the remaining misfit is that predicted by the Rasch model. Accordingly, during our analyses we pay attention to reducing the Real S.E., but when we get to the final report, it is the Model S.E. that is more relevant.

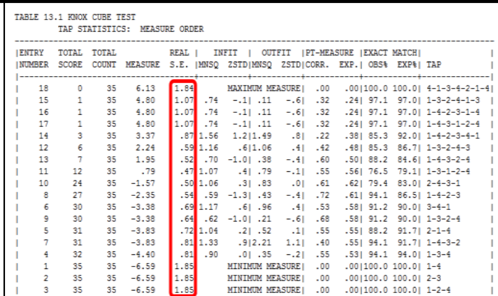
83. Let’s take a look at some Standard Errors.
Launch Winsteps.
Analyze Exam1.txt - the Knox Cube test



84. Let’s look again at the items in measure order:
Click on Winsteps menu bar
Click on Output Tables
Click on 13. TAP: measure

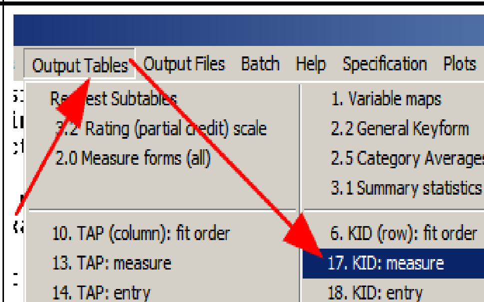


85. Table 13.1 shows the pattern of S.E.s we usually see.
All the children took all the same dichotomous items, so why do the S.E.s differ?
Look at the 4 reasons above. The differences in S.E. must be because of the *person-item targeting.*
The S.E.s of the items are smallest closest to where most of the children are targeted (probability of success, p≈0.5). Not p≈1.0 (item 1-3) nor p≈0.0 (item 18).
Do you understand that? If not, ***Discussion Board ....***



```
TABLE 13.1 KNOX CUBE TEST
        TAP STATISTICS:  MEASURE ORDER
-----------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           REAL |   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|              |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| TAP          |
|------------------------------------+----------+----------+-----------+-----------+--------------|
|    18      0     35    6.13   1.84       MAXIMUM MEASURE|  .00   .00|100.0 100.0| 4-1-3-4-2-1-4|
|    15      1     35    4.80   1.07  .74   -.1| .11   -.6|  .32   .24| 97.1  97.0| 1-3-2-4-1-3  |
|    16      1     35    4.80   1.07  .74   -.1| .11   -.6|  .32   .24| 97.1  97.0| 1-4-2-3-1-4  |
|    17      1     35    4.80   1.07  .74   -.1| .11   -.6|  .32   .24| 97.1  97.0| 1-4-3-1-2-4  |
|    14      3     35    3.37    .87 1.56   1.2|1.49    .8|  .22   .38| 85.3  92.0| 1-4-2-3-4-1  |
|    12      6     35    2.24    .59 1.16    .6|1.06    .4|  .42   .48| 85.3  86.7| 1-3-2-4-3    |
|    13      7     35    1.95    .52  .70  -1.0| .38   -.4|  .60   .50| 88.2  84.6| 1-4-3-2-4    |
|    11     12     35     .79    .47 1.07    .4| .79   -.1|  .55   .56| 76.5  79.1| 1-3-1-2-4    |
|    10     24     35   -1.57    .50 1.06    .3| .83    .0|  .61   .62| 79.4  83.0| 2-4-3-1      |
|     8     27     35   -2.35    .54  .59  -1.3| .43   -.4|  .72   .61| 94.1  86.5| 1-4-2-3      |
|     6     30     35   -3.38    .69 1.17    .6| .96    .4|  .53   .58| 91.2  90.0| 3-4-1        |
|     9     30     35   -3.38    .64  .62  -1.0| .21   -.6|  .68   .58| 91.2  90.0| 1-3-2-4      |
|     5     31     35   -3.83    .72 1.04    .2| .52    .1|  .55   .55| 88.2  91.7| 2-1-4        |
|     7     31     35   -3.83    .81 1.33    .9|2.21   1.1|  .40   .55| 94.1  91.7| 1-4-3-2      |
|     4     32     35   -4.40    .81  .90    .0| .35   -.2|  .55   .53| 94.1  94.0| 1-3-4        |
|     1     35     35   -6.59   1.85       MINIMUM MEASURE|  .00   .00|100.0 100.0| 1-4          |
|     2     35     35   -6.59   1.85       MINIMUM MEASURE|  .00   .00|100.0 100.0| 2-3          |
|     3     35     35   -6.59   1.85       MINIMUM MEASURE|  .00   .00|100.0 100.0| 1-2-4        |
```

86. Let’s look at the children in measure order:
Click on Winsteps menu bar
Click on Output Tables
Click on 17. KID: measure



87. In Table 17.1, we can see the Rasch logit measures for the children, along with their "Model S.E."

Notice that the second child has a standard error of .94 logits. We assume that the imprecision is distributed according to the Normal Distribution. Lesson 2 Appendix 1 talks about the normal distribution.
This means that the precision of measurement is such that we are about 68% sure that the reported measure is within .94 logits of the exact errorless "true" value (if we could discover it).

In #85 and #87, look at measures around 1.94. You can see that the S.E.s (0.98) of the children in Table 17 are bigger than the S.E.s (0.52) of the items in Table 13 with the same measures. This is because there are more children (35) per item (so smaller item S.E.) than there are items (18) per child (so larger child S.E.)

Notice that the S.E.s are biggest in the center of the child distribution. This is unusual. Why is this happening here?
Because there are no central items to target on the children - as we will see shortly.

Reminder: Do you recall the "Pathway" Bubble-chart in Lesson 1? The radius of the bubble corresponds to the standard error.

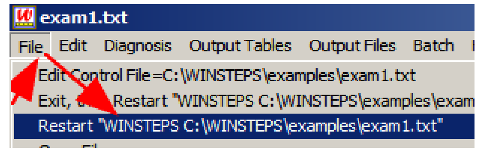
TABLE 17.1 KNOX CUBE TEST
KID STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL 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%	KID	
24	14	18	3.73	.94	.95	.1	.39	-.1	.69	.69	85.7	88.8	Mike	M
32	14	18	3.73	.94	1.94	1.7	.88	.4	.61	.69	71.4	88.8	Tracie	F
7	13	18	2.85	.95	.36	-1.5	.15	-.5	.81	.76	100.0	89.1	Blaise	M
11	12	18	1.94	.98	.52	-1.0	.18	-.4	.85	.81	92.9	88.7	David	M
28	12	18	1.94	.98	.65	-.6	.22	-.4	.84	.81	92.9	88.7	Rod	M
33	12	18	1.94	.98	1.74	1.4	.71	.2	.77	.81	78.6	88.7	Walter	M
35	12	18	1.94	.98	1.60	1.2	.67	.2	.78	.81	78.6	88.7	Zula	F
3	11	18	.92	1.06	.99	.2	.32	-.2	.86	.85	85.7	91.1	Audrey	F
17	11	18	.92	1.06	.39	-1.0	.13	-.6	.89	.85	100.0	91.1	James	M
20	11	18	.92	1.06	.99	.2	.32	-.2	.86	.85	85.7	91.1	Kim	F
29	11	18	.92	1.06	.39	-1.0	.13	-.6	.89	.85	100.0	91.1	Ron	M
34	11	18	.92	1.06	.39	-1.0	.13	-.6	.89	.85	100.0	91.1	William	M
2	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Anne	F
8	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Brenda	F
12	10	18	-.26	1.11	1.94	1.2	1.19	.6	.82	.87	85.7	93.2	Don	M
13	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Dorothy	F
16	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Helen	F
18	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Janet	F
19	10	18	-.26	1.11	4.12	2.5	5.28	2.0	.65	.87	71.4	93.2	Joe	M
22	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Lisa	F
23	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Martha	F
25	10	18	-.26	1.11	4.08	2.5	6.07	2.2	.65	.87	71.4	93.2	Pete	M
27	10	18	-.26	1.11	1.49	.8	.51	.0	.85	.87	85.7	93.2	Rick	M
31	10	18	-.26	1.11	1.49	.8	.51	.0	.85	.87	85.7	93.2	Thomas	M
6	9	18	-1.37	1.00	.43	-.9	.16	-.5	.89	.85	92.9	90.4	Betty	F
14	9	18	-1.37	1.00	1.16	.5	.97	.5	.83	.85	92.9	90.4	Elsie	F
30	9	18	-1.37	1.00	.43	-.9	.16	-.5	.89	.85	92.9	90.4	Susan	F
5	8	18	-2.23	.88	1.10	.4	.51	.0	.81	.81	78.6	87.1	Bert	M
9	8	18	-2.23	.88	1.02	.2	.81	.3	.80	.81	92.9	87.1	Britton	F
1	7	18	-2.94	.82	.61	-1.2	.29	-.2	.80	.76	92.9	84.6	Adam	M
26	7	18	-2.94	.82	.61	-1.2	.29	-.2	.80	.76	92.9	84.6	Richard	M
4	6	18	-3.61	.82	1.22	.8	.86	.4	.69	.72	85.7	82.5	Barbara	F
10	6	18	-3.61	.82	.78	-.7	.36	-.1	.74	.72	85.7	82.5	Carol	F
21	5	18	-4.32	.88	1.09	.4	.46	.0	.67	.68	78.6	85.7	Linda	F
15	3	18	-6.62	1.85		MINIMUM MEASURE			.61	.61	100.0	100.0	Frank	M

88. Though we use statistics to compute the standard errors, they are really an aspect of measurement.
Imagine we measure your height with a tape measure. It might be 195 centimeters. But what is your "true" height? 195.1 cms? 194.8 cms? We don't know, but we do know that your true height is very unlikely to be exactly 195.000 cms. There is a measurement error. So we might report your height as 195.0±0.5 cms.
What does 195.0±0.5 cms mean? It means close to 195.0, and unlikely to be outside 194.5 to 195.5.
It is exactly the same the same with logits: 3.73±0.94 logits means "the true logit measure is close to the observed measure of 3.73 and unlikely to be outside the range 3.73-0.94 to 3.73+0.94".
In fact, the true measure is likely to be inside that range 68% of the time, and outside that range only 32% of the time.

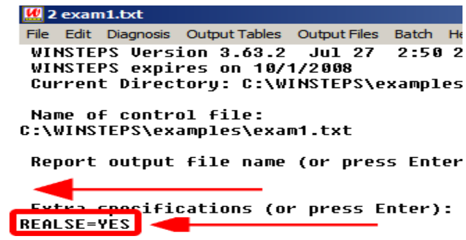
89.	**G. Real Standard Errors**

90. Let's look at the Real Standard Errors - we'll need to run another analysis to compute them:
Winsteps menu bar
Click on "Restart Winsteps"

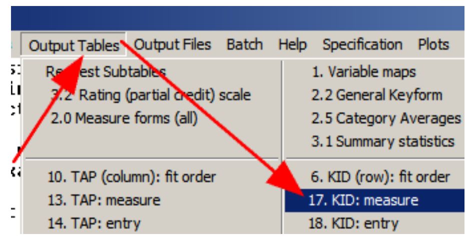


91. Report Output? Click Enter
Extra Specification?
REALSE=Yes
Click Enter

This instructs Winsteps to compute the "Real" standard errors.



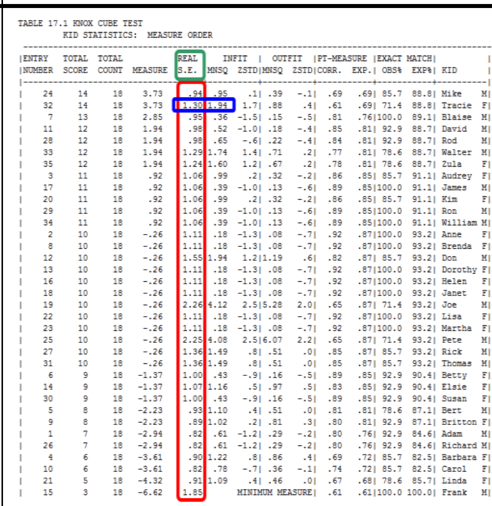
92. Let's look again at the children in measure order:
Click on Winsteps menu bar
Click on Output Tables
Click on 17. KID: measure



93. In Table 17.1, you can see the Real Standard error.

Look at the second child. The Real S.E. is 1.30 logits. Earlier we saw that the Model S.E. is 0.94 logits.
It is the large INFIT Mean-square of 1.94 that causes this difference. The large mean-square indicates that there is unpredictable noise in the data. The Real S.E. interprets this to contradict the Rasch model, and so lowers the reported precision (increases the S.E.)
Real S.E. = Model S.E. * Max(1.0, √ (INFIT MnSq))
so

Model S.E. * √(INFIT MnSq) =
0.94 * √ 1.94 ≈ 1.30 = Real S.E.



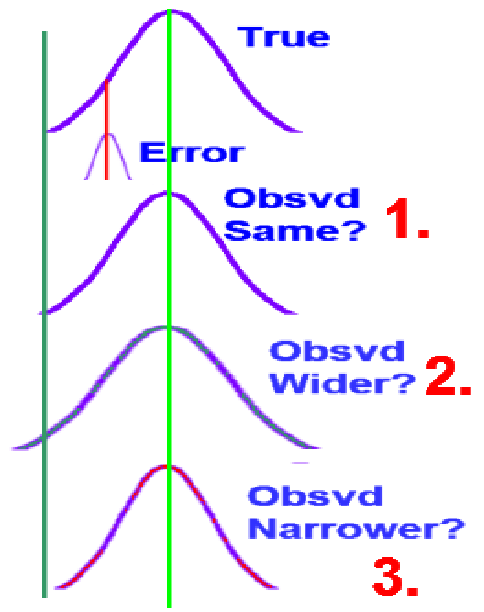
```
TABLE 17.1 KNOX CUBE TEST
       KID STATISTICS:  MEASURE ORDER
 --------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL          REAL     INFIT   |  OUTFIT  |PT-MEASURE |EXACT MATCH|          |
|NUMBER  SCORE  COUNT  MEASURE S.E. MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| KID      |
|---------------------------------------------+----------+-----------+-----------+----------|
|    24     14     18    3.73   .94  .95    .1| .39   -.1|  .69   .69| 85.7  88.8| Mike    M|
|    32     14     18    3.73  1.30 1.94   1.7| .88    .4|  .61   .69| 71.4  88.8| Tracie  F|
|     7     13     18    2.85   .95  .36  -1.5| .15   -.5|  .81   .76|100.0  89.1| Blaise  M|
|    11     12     18    1.94   .98  .52  -1.0| .18   -.4|  .85   .81| 92.9  88.7| David   M|
|    28     12     18    1.94   .98  .65   -.6| .22   -.4|  .84   .81| 92.9  88.7| Rod     M|
|    33     12     18    1.94  1.29 1.74   1.4| .71    .2|  .77   .81| 78.6  88.7| Walter  M|
|    35     12     18    1.94  1.24 1.60   1.2| .67    .2|  .78   .81| 78.6  88.7| Zula    F|
|     3     11     18     .92  1.06  .99    .2| .32   -.2|  .86   .85| 85.7  91.1| Audrey  F|
|    17     11     18     .92  1.06  .39  -1.0| .13   -.6|  .89   .85|100.0  91.1| James   M|
|    20     11     18     .92  1.06  .99    .2| .32   -.2|  .86   .85| 85.7  91.1| Kim     F|
|    29     11     18     .92  1.06  .39  -1.0| .13   -.6|  .89   .85|100.0  91.1| Ron     M|
|    34     11     18     .92  1.06  .39  -1.0| .13   -.6|  .89   .85|100.0  91.1| William M|
|     2     10     18    -.26  1.11  .18  -1.3| .08   -.7|  .92   .87|100.0  93.2| Anne    F|
|     8     10     18    -.26  1.11  .18  -1.3| .08   -.7|  .92   .87|100.0  93.2| Brenda  F|
|    12     10     18    -.26  1.55 1.94   1.2|1.19    .6|  .82   .87| 85.7  93.2| Don     M|
|    13     10     18    -.26  1.11  .18  -1.3| .08   -.7|  .92   .87|100.0  93.2| Dorothy F|
|    16     10     18    -.26  1.11  .18  -1.3| .08   -.7|  .92   .87|100.0  93.2| Helen   F|
|    18     10     18    -.26  1.11  .18  -1.3| .08   -.7|  .92   .87|100.0  93.2| Janet   F|
|    19     10     18    -.26  2.26 4.12   2.5|5.28   2.0|  .65   .87| 71.4  93.2| Joe     M|
|    22     10     18    -.26  1.11  .18  -1.3| .08   -.7|  .92   .87|100.0  93.2| Lisa    F|
|    23     10     18    -.26  1.11  .18  -1.3| .08   -.7|  .92   .87|100.0  93.2| Martha  F|
|    25     10     18    -.26  2.25 4.08   2.5|6.07   2.2|  .65   .87| 71.4  93.2| Pete    M|
|    27     10     18    -.26  1.36 1.49    .8| .51    .0|  .85   .87| 85.7  93.2| Rick    M|
|    31     10     18    -.26  1.36 1.49    .8| .51    .0|  .85   .87| 85.7  93.2| Thomas  M|
|     6      9     18   -1.37  1.00  .43   -.9| .16   -.5|  .89   .85| 92.9  90.4| Betty   F|
|    14      9     18   -1.37  1.07 1.16    .5| .97    .5|  .83   .85| 92.9  90.4| Elsie   F|
|    30      9     18   -1.37  1.00  .43   -.9| .16   -.5|  .89   .85| 92.9  90.4| Susan   F|
|     5      8     18   -2.23   .93 1.10    .4| .51    .0|  .81   .81| 78.6  87.1| Bert    M|
|     9      8     18   -2.23   .89 1.02    .2| .81    .3|  .80   .81| 92.9  87.1| Britton F|
|     1      7     18   -2.94   .82  .61  -1.2| .29   -.2|  .80   .76| 92.9  84.6| Adam    M|
|    26      7     18   -2.94   .82  .61  -1.2| .29   -.2|  .80   .76| 92.9  84.6| Richard M|
|     4      6     18   -3.61   .90 1.22    .8| .86    .4|  .69   .72| 85.7  82.5| Barbara F|
|    10      6     18   -3.61   .82  .78   -.7| .36   -.1|  .74   .72| 85.7  82.5| Carol   F|
|    21      5     18   -4.32   .91 1.09    .4| .46    .0|  .67   .68| 78.6  85.7| Linda   F|
|    15      3     18   -6.62  1.85     MINIMUM MEASURE|  .61   .61|100.0 100.0| Frank   M|
```

94. **H. Reliability and Separation Statistics**

95. "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.

96. 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.

97. What happens when we measure with error?
Imagine we have the "true" distribution of the measures. Each is exactly precise. 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 errors cancel out.
Option 2. The observed distribution will be wider than the true distribution. The errors will tend to make the measures appear more diverse.
Option 3. The observed distribution will be narrower than the true distribution. The errors will tend to be more central. There will be "regression toward the mean".
Think carefully: Is it Option 1, 2 or 3?



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

99. 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

100 Reliability = True Variance / Observed Variance
Reliability = (Observed Variance - Error Variance) / Observed Variance

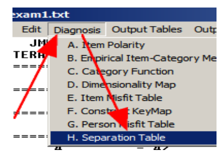
101 So now let's proceed to compute the Rasch-Measure-based Reliability for the current samples of persons and items

102 Look again 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	TOTAL SCORE	TOTAL COUNT	MEASURE	REAL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	KID	
24	14	18	3.73	.94	.95	.1	.39	-.1	.69	.69	85.7	88.8	Mike	M
32	14	18	3.73	1.30	1.94	1.7	.88	.4	.61	.69	71.4	88.8	Tracie	F
7	13	18	2.85	.95	.36	-1.5	.15	-.5	.81	.76	100.0	89.1	Blaise	M
11	12	18	1.94	.98	.52	-1.0	.18	-.4	.85	.81	92.9	88.7	David	M
28	12	18	1.94	.98	.65	-.6	.22	-.4	.84	.81	92.9	88.7	Rod	M
33	12	18	1.94	1.29	1.74	1.4	.71	.2	.77	.81	78.6	88.7	Walter	M
35	12	18	1.94	1.24	1.60	1.2	.67	.2	.78	.81	78.6	88.7	Zula	F
3	11	18	.92	1.06	.99	.2	.32	-.2	.86	.85	85.7	91.1	Audrey	F
17	11	18	.92	1.06	.39	-1.0	.13	-.6	.89	.85	100.0	91.1	James	M
20	11	18	.92	1.06	.99	.2	.32	-.2	.86	.85	85.7	91.1	Kim	F
29	11	18	.92	1.06	.39	-1.0	.13	-.6	.89	.85	100.0	91.1	Ron	M
34	11	18	.92	1.06	.39	-1.0	.13	-.6	.89	.85	100.0	91.1	William	M
2	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Anne	F
8	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Brenda	F
12	10	18	-.26	1.55	1.94	1.2	1.19	.6	.82	.87	85.7	93.2	Don	M
13	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Dorothy	F
16	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Helen	F
18	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Janet	F
19	10	18	-.26	2.26	4.12	2.5	5.28	2.0	.65	.87	71.4	93.2	Joe	M
22	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Lisa	F
23	10	18	-.26	1.11	.18	-1.3	.08	-.7	.92	.87	100.0	93.2	Martha	F
25	10	18	-.26	2.25	4.08	2.5	6.07	2.2	.65	.87	71.4	93.2	Pete	M
27	10	18	-.26	1.36	1.49	.8	.51	.0	.85	.87	85.7	93.2	Rick	M
31	10	18	-.26	1.36	1.49	.8	.51	.0	.85	.87	85.7	93.2	Thomas	M
6	9	18	-1.37	1.00	.43	-.9	.16	-.5	.89	.85	92.9	90.4	Betty	F
14	9	18	-1.37	1.07	1.16	.5	.97	.5	.83	.85	92.9	90.4	Elsie	F
30	9	18	-1.37	1.00	.43	-.9	.16	-.5	.89	.85	92.9	90.4	Susan	F
5	8	18	-2.23	.93	1.10	.4	.51	.0	.81	.81	78.6	87.1	Bert	M
9	8	18	-2.23	.89	1.02	.2	.81	.3	.80	.81	92.9	87.1	Britton	F
1	7	18	-2.94	.82	.61	-1.2	.29	-.2	.80	.76	92.9	84.6	Adam	M
26	7	18	-2.94	.82	.61	-1.2	.29	-.2	.80	.76	92.9	84.6	Richard	M
4	6	18	-3.61	.90	1.22	.8	.86	.4	.69	.72	85.7	82.5	Barbara	F
10	6	18	-3.61	.82	.78	-.7	.36	-.1	.74	.72	85.7	82.5	Carol	F
21	5	18	-4.32	.91	1.09	.4	.46	.0	.67	.68	78.6	85.7	Linda	F
15	3	18	-6.62	1.85	MINIMUM MEASURE				.61	.61	100.0	100.0	Frank	M
MEAN	9.7	18.0	-.37	1.16	.99	-.2	.68	-.1			89.9	90.0		
S.D.	2.4	.0	2.22	.34	.94	1.2	1.29	.7			9.2	3.2		

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



104 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) KID

	TOTAL SCORE	COUNT	MEASURE	REAL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	9.7	18.0	-.37	1.16				
S.D.	2.4	.0	2.22	.34				
MAX.	14.0	18.0	3.73	2.26				
MIN.	3.0	18.0	-6.62	.82	.18	-1.5	.08	-.7

REAL RMSE 1.21 TRUE SD 1.86 SEPARATION 1.55 KID RELIABILITY .70
MODEL RMSE 1.05 TRUE 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 "TEST" RELIABILITY = .75

"True" Variance = "Adjusted for error variance"

"Model Reliability" =
(S.D. ^{2} - Model RMSE^2) / S.D. ^{2}
= (2.22^2-1.05^2) / 2.22^2 = 0.78

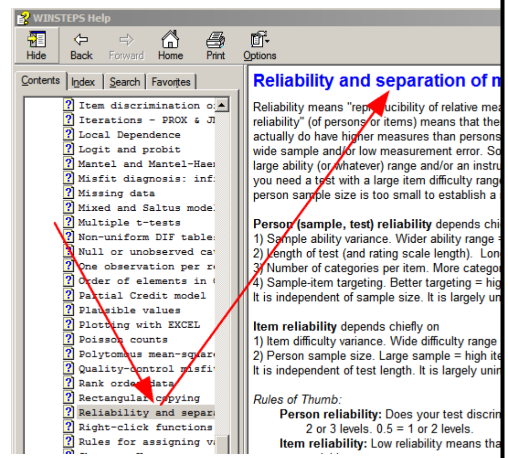
105 Here is a useful Table showing how the average, RMSE, standard error, the True S.D., the Observed S.D. and the Reliability relate to each other. It is from the Winsteps Help "Special Topic", "Reliability". This Table is very important to the understanding of the reproducibility (=Reliability) of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

106 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.



107 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. That is why Ben Wright devised the "Separation Coefficient".

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

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

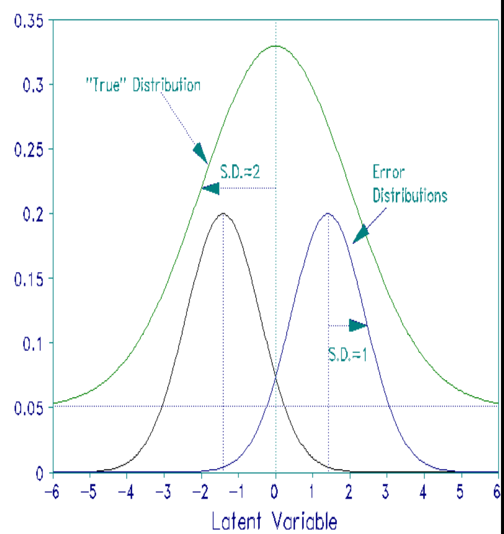
108

True S.D	**Standard Error** RMSE	**Separation** = True S.D. / RMSE	**True Variance** = True S.D.^{2}	**Observed Variance** = True Variance + RMSE^2	**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

<table>
<tr>
<td>109</td>
<td>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. Most effective: Increase the person measurement precision (decrease the average person S.E.) by increasing the number of items on the Test.
2. Effective: Increase the observed standard deviation by testing a wider ability range.
3. Less effective: Improve the targeting of the items on the sample</td>
<td>Increasing person sample size will not increase person reliability unless the extra persons have a wider ability range.</td>
</tr>
<tr>
<td>110</td>
<td>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.</td>
<td>If the item reliability is low, you need a bigger sample!</td>
</tr>
<tr>
<td>111</td>
<td>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 “true” distribution of a sample with standard deviation of “2”, as if we could measure each person perfectly precisely without any measurement error. The x-axis of this curve is drawn above the usual x-axis so that we can see it clearly.

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.</td>
<td>
</td>
</tr>
<tr>
<td>112</td>
<td colspan="2">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)
$= \text{True S.D.}^2 / (\text{True S.D.}^2 + \text{S.E.}^2) = 2^2 / (2^2 + 1^2) = 0.8$
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.</td>
</tr>
</table>

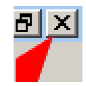
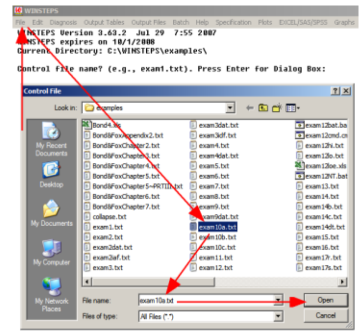
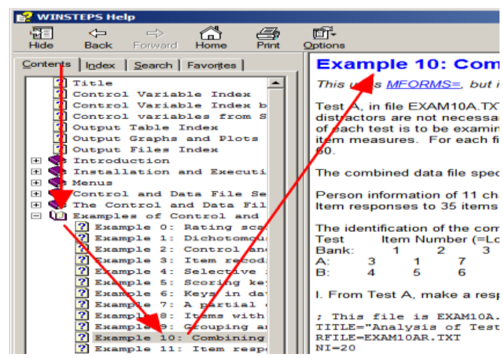
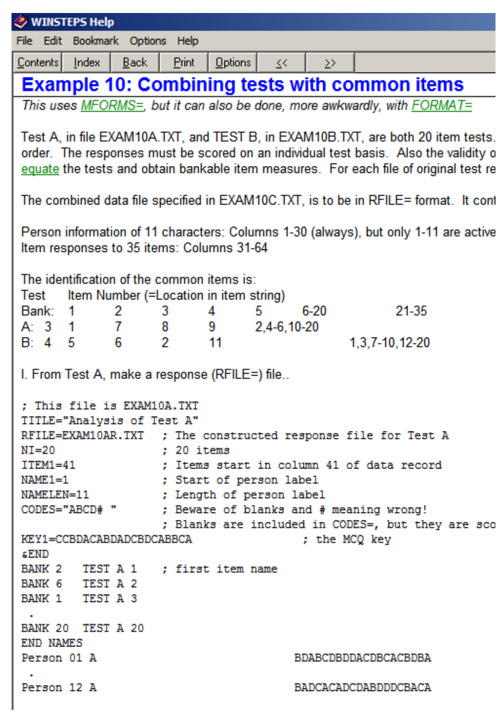
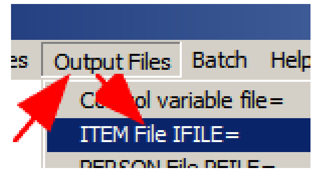
<table>
<tr><td>113.</td><td colspan="2">Reliability rules-of-thumb:
1. If the Item Reliability is less than 0.8, you need a bigger sample.
2. If the Person Reliability is less than 0.8, you need more items in your test.
3. Use “Real Reliability” (worst case) when doing exploratory analyses, “Model Reliability” (best case) when your analysis is as good as it can be.
4. Use “Non-Extreme Reliability” (excludes extreme scores) when doing exploratory analysis, use “Extreme+Non-Extreme Reliability” (includes extreme scores) when reporting.
5. High item reliability does not compensate for low person reliability.</td></tr>
<tr><td>114.</td><td>Close all Winsteps windows</td><td></td></tr>
<tr><td>115.</td><td>Optional Experiment: Analyze Example0.txt, note down the separation and reliability.
Then analyze Example0.txt again.
At the Extra Specification prompt: IDELETE=23
Note down the person separation and person reliability from Table 3.1. Usually “more items → more separation.” do you see that omitting the worst item has increased the separation.
The worst item was doing more harm than good!
Also something has happened to one person. Table 18 tells us who that is.</td><td>This shows where to look in your Example0.txt analysis. It is not the answer!

Table 3.1 for Exam1.txt
<pre>
   SUMMARY OF   34 MEASURED (NON-EXTREME) KIDS
+---+
| RAW   MODEL   INFIT   OUTFIT |
SCORE   COUNT   MEASURE   ERROR   MNSQ   ZSTD   MNSQ   ZSTD
MEAN   6.9   14.0   -.20   1.03   1.03   -.3   .73   -.3
S.D.   2.1   .0   2.07   .11   1.01   1.2   1.45   .5
MAX.   11.0   14.0   3.89   1.15   4.43   2.3   6.86   1.3
MIN.   2.0   14.0   -4.48   .82   .17   -1.6   .08   -.8
REAL RMSE   1.23 TRUE SD   1.66 SEPARATION 1.35 KID   RELIABILITY .65
MODEL RMSE   1.03 TRUE SD   1.79 SEPARATION 1.73 KID   RELIABILITY .75
S.E. OF KID MEAN = .36
+---+
</pre></td></tr>
<tr><td>116.</td><td></td><td></td></tr>
</table>

<table>
<tr><td>117.</td><td colspan="2">I. Output Files and Anchoring</td></tr>
<tr><td>118.</td><td>Launch Winsteps</td><td></td></tr>
<tr><td>119.</td><td>Control file name?
Click on “File” menu
Click on “Open File”
In the File dialog box,
Click on Exam10a.txt
Click on Open
Report Output? Press Enter
Extra Specifications? Press Enter
The analysis is performed ...</td><td></td></tr>
<tr><td>120.</td><td>Let’s find out about this Example ...
Winsteps Menu bar
Click on Help
Click on Contents
Click on Examples
Click on Example 10</td><td></td></tr>
<tr><td>121.</td><td>The Help file tells us that we are looking at two tests, Exam10a.txt and Exam10b.txt with common (shared) items. They are MCQ tests, but the principles are the same for all types of test.
Notice that KEY1= provides the MCQ scoring key for the items.
We are going to do something somewhat different from the procedure in the Help file. We will analyze Exam10a.txt
Write out the item difficulties with IFILE=exam10aif.txt
Edit the item difficulty file, exam10aif.txt, to match Exam10b.txt
Then use the item difficulties from Exam10a.txt in exam10aif.txt to anchor (fix) the item difficulties in the analysis of Exam10b.txt.
The anchored item difficulties will link the tests
This will put the person measures from the Exam10b.txt analysis in the same measurement frame of reference as the person measures from the Exam10a.txt
Notice the entry numbers of the equivalent items:
<table>
<tr><td>Test</td><td colspan="5">Item Number (=Location in item string)</td></tr>
<tr><td>Bank:</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>A:</td><td>3</td><td>1</td><td>7</td><td>8</td><td>9</td></tr>
<tr><td>B:</td><td>4</td><td>5</td><td>6</td><td>2</td><td>11</td></tr>
</table></td><td>
</td></tr>
</table>

122. Let's imagine Test A has been reported, so that we need to report Test B in the Test A frame-of-reference, so that measures on Test B are directly comparable with those on Test A.
We have analyzed Exam10a.txt and estimated the measures.
Let's write out the item difficulties to a file:
Winsteps menu bar
Click on Output Files
Click on ITEM File IFILE=

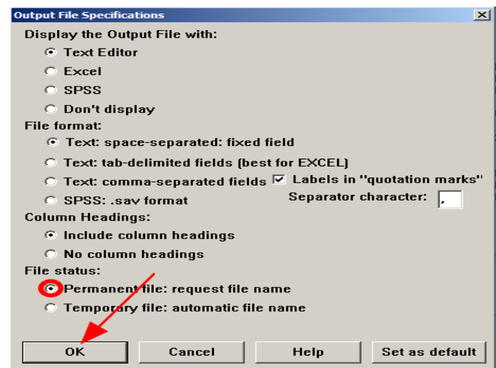


123. We are going to write to disk the item difficulty measures and the other statistics we have seen in the item measure Tables.

So Winsteps needs to know in what format you want them to be.

We want the default Text-file format, except that we want a Permanent file, so we can read the file in our next analysis

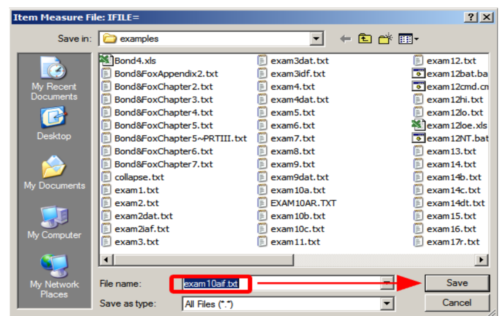
Click on "Permanent file"
Click on "OK"



124. Now Winsteps needs to know the name of the permanent file:

File name: **exam10aif.txt**
Press Enter

I put code letters such as "if" in the file names to remind myself that this is in item file.



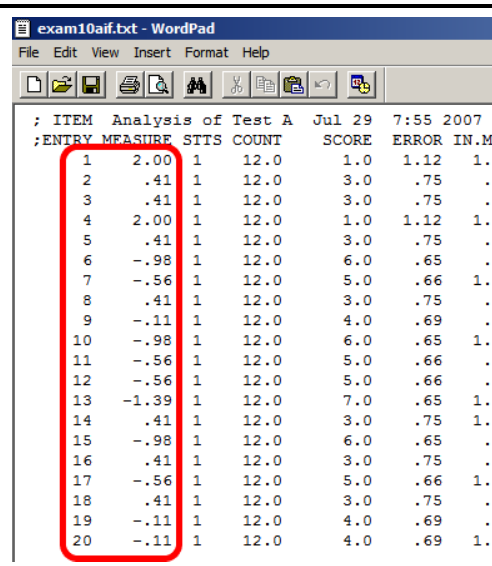
125. **The item file displays.**

The important columns for us are the first two columns:
The item entry number
The item difficulty measure

We want to use these numbers to anchor the item difficulties for the second analysis.

Remember the equivalence:

Test	Item Number (=Location in item string)				
Bank:	**1**	**2**	**3**	**4**	**5**
A:	**3**	**1**	**7**	**8**	**9**
B:	**4**	**5**	**6**	**2**	**11**

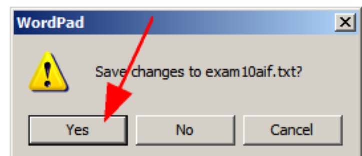
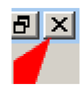
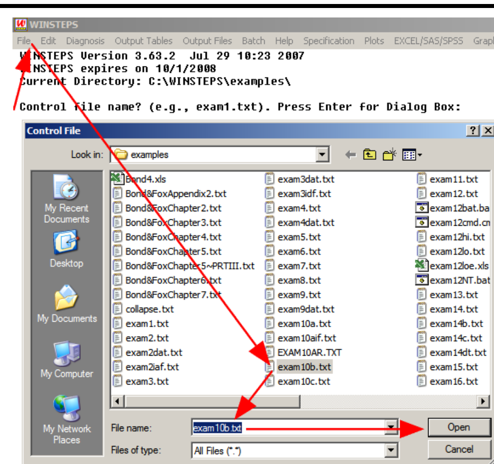


; ITEM Analysis of Test A Jul 29 7:55 2007

;ENTRY	MEASURE	STTS	COUNT	SCORE	ERROR	IN.MS
1	2.00	1	12.0	1.0	1.12	1.4
2	.41	1	12.0	3.0	.75	.5
3	.41	1	12.0	3.0	.75	.6
4	2.00	1	12.0	1.0	1.12	1.5
5	.41	1	12.0	3.0	.75	.9
6	-.98	1	12.0	6.0	.65	.9
7	-.56	1	12.0	5.0	.66	1.0
8	.41	1	12.0	3.0	.75	.5
9	-.11	1	12.0	4.0	.69	.9
10	-.98	1	12.0	6.0	.65	1.0
11	-.56	1	12.0	5.0	.66	.7
12	-.56	1	12.0	5.0	.66	.7
13	-1.39	1	12.0	7.0	.65	1.2
14	.41	1	12.0	3.0	.75	1.1
15	-.98	1	12.0	6.0	.65	.8
16	.41	1	12.0	3.0	.75	.7
17	-.56	1	12.0	5.0	.66	1.8
18	.41	1	12.0	3.0	.75	.5
19	-.11	1	12.0	4.0	.69	.6
20	-.11	1	12.0	4.0	.69	1.6

126.	So let's edit out all the items except 3,1,7,8,9 = 1,3,7,8,9 (both item orders mean the same thing to Winsteps)
	
	We only need the entry numbers and the measures, so I've removed all the other numbers. They would have been ignored.
	exam10aif.txt should look like this →
	Or you could put ; in front of them to make sure they are regarded as comments.
		This is the edited version of
		exam10aif.txt
		`; ITEM  Analysis of Test A  Jul 29`
		`;ENTRY MEASURE`
		`     1    2.00`
		`     3     .41`
		`     7    -.56`
		`     8     .41`
		`     9    -.11  ; this is a comment`

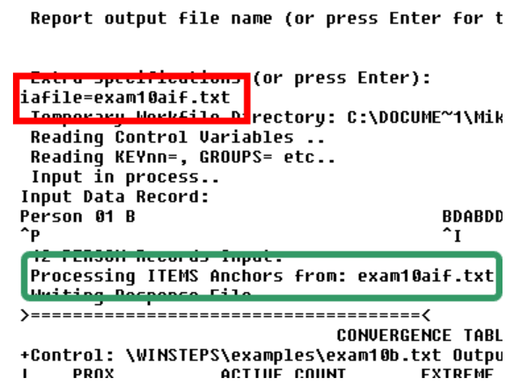
127.	Now **change** the Entry numbers to the "B" numbers.
	We don't need the Test A numbers any more, so I've made the A numbers into comments in case I've made a mistake lining up the numbers.
	exam10aif.txt should look like this →
	
	(The order of the items in the file does not matter.)
		This is the final version of
		exam10aif.txt
		`;    B   Measure`
		`     5     2.00 ; A 1`
		`     4      .41 ; A 3`
		`     6     -.56 ; A 7`
		`     2      .41 ; A 8`
		`     11     -.11 ; A 9`

128.	Save the file (click the diskette icon) We definitely want to save the changes we have made to exam10aif.txt! *Please make no changes to the Winsteps control file.*	
129.	Close all Winsteps windows	
130.	Launch Winsteps	
131.	Control file name? Click on "File" menu Click on "Open File" In the File dialog box, Click on **Exam10b.txt** Click on Open	

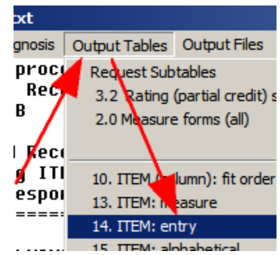
132. Report Output? Press Enter

Extra Specifications? - We want the anchor item difficulties:
`iafile=exam10aif.txt`
Press Enter
The analysis is performed ...

Notice the green box: “Processing ITEMS Anchors ....”
We know that Winsteps has read in our input file of item difficulties.

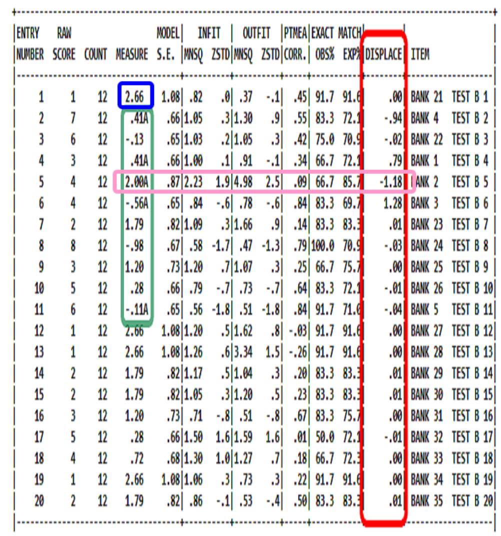


133. Let’s see what has happened to the item difficulty measures.
Winsteps menu bar
Click on Output Tables
Click on “14. ITEM: entry”



134. Table 14, the items in Entry Order, displays.
Green box: The anchored values are indicated by A in the Measure column. The other item difficulties, and the person measures, are estimated relative to those anchored values.
Red box: a new column “Displacement”. This shows how different the measures would be if they weren’t anchored. For instance,
Pink box: if item 5 was unanchored, we would expect its reported difficulty to be (green box) 2.00A + (red box) -1.18 = 0.82 logits.
Item 5 is less difficult (0.82 logits) for these people than it was for the original people (2.00 logits).

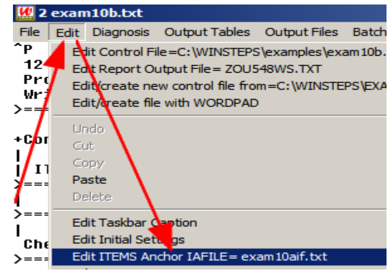
Blue box: notice also that unanchored (no A) Item 1 has a measure of 2.66 logits and a displacement of .00.



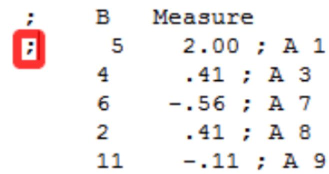
ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	DISPLACE	ITEM
1	1	12	2.66	1.08	.82	.0	.37	-.1	.45	91.7	91.6	.00	BANK 21 TEST B 1
2	7	12	.41A	.66	1.05	.3	1.30	.9	.55	83.3	72.1	-.94	BANK 4 TEST B 2
3	6	12	-.13	.65	1.03	.2	1.05	.3	.42	75.0	70.9	-.02	BANK 22 TEST B 3
4	3	12	.41A	.66	1.00	.1	.91	-.1	.34	66.7	72.1	.79	BANK 1 TEST B 4
5	4	12	2.00A	.87	2.23	1.9	4.98	2.5	.09	66.7	85.7	-1.18	BANK 2 TEST B 5
6	4	12	-.56A	.65	.84	-.6	.78	-.6	.84	83.3	69.7	1.28	BANK 3 TEST B 6
7	2	12	1.79	.82	1.09	.3	1.66	.9	.14	83.3	83.3	.01	BANK 23 TEST B 7
8	8	12	-.98	.67	.58	-1.7	.47	-1.3	.79	100.0	70.9	-.03	BANK 24 TEST B 8
9	3	12	1.20	.73	1.20	.7	1.07	.3	.25	66.7	75.7	.00	BANK 25 TEST B 9
10	5	12	.28	.66	.79	-.7	.73	-.7	.64	83.3	72.1	-.01	BANK 26 TEST B 10
11	6	12	-.11A	.65	.56	-1.8	.51	-1.8	.84	91.7	71.6	-.04	BANK 5 TEST B 11
12	1	12	2.66	1.08	1.20	.5	1.62	.8	-.03	91.7	91.6	.00	BANK 27 TEST B 12
13	1	12	2.66	1.08	1.26	.6	3.34	1.5	-.26	91.7	91.6	.00	BANK 28 TEST B 13
14	2	12	1.79	.82	1.17	.5	1.04	.3	.20	83.3	83.3	.01	BANK 29 TEST B 14
15	2	12	1.79	.82	1.05	.3	1.20	.5	.23	83.3	83.3	.01	BANK 30 TEST B 15
16	3	12	1.20	.73	.71	-.8	.51	-.8	.67	83.3	75.7	.00	BANK 31 TEST B 16
17	5	12	.28	.66	1.50	1.6	1.59	1.6	.01	50.0	72.1	-.01	BANK 32 TEST B 17
18	4	12	.72	.68	1.30	1.0	1.27	.7	.18	66.7	72.3	.00	BANK 33 TEST B 18
19	1	12	2.66	1.08	1.06	.3	.73	.3	.22	91.7	91.6	.00	BANK 34 TEST B 19
20	2	12	1.79	.82	.86	-.1	.53	-.4	.50	83.3	83.3	.01	BANK 35 TEST B 20

135. "Displacement" indicates the difference between the observed and the expected raw scores. There are always displacement values (because the estimation is never perfect) but this column only appears if there are some displacements big enough to merit our attention.
If “displacement” is reported in an **unanchored** analysis, and we are concerned that values like 0.01 or 0.03 are too big, then we need to set the estimation criteria more tightly using LCONV= and RCONV=. This causes Winsteps to perform more iterations.
Some high-stakes examinations are analyzed so that the biggest displacement is 0.0001. This is so that a lawyer cannot say "If you had run Winsteps longer, my client's measure would have been estimated just above the pass-fail point."

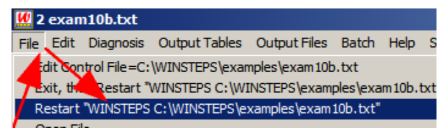
136. Let’s try this!
Winsteps menu bar
Edit menu
Edit ITEMS Anchor IAFILE=



137. Comment out item 5
“Save” the file - click the Diskette icon
Close the NotePad window

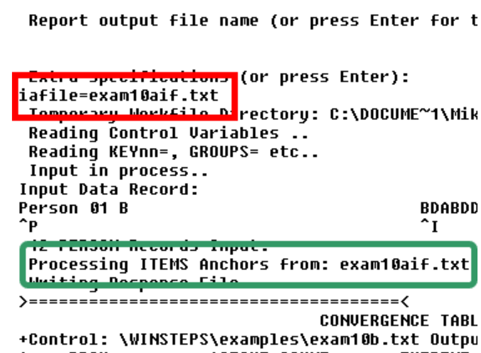


138. Winsteps menu bar
Click on File
Click on Restart Winsteps

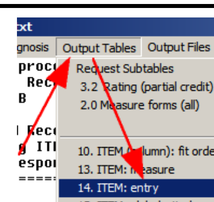


139. Report Output? Press Enter
Extra Specifications? **IAFILE=exam10aif.txt**
Press Enter
The analysis is performed ...

Green box: “Processing ITEMS Anchors ....”
Winsteps has read in our input file of item difficulties.
If we were processing polytomous items, we would also need to anchor the rating scale structures with a SAFILE=, see Winsteps Help.

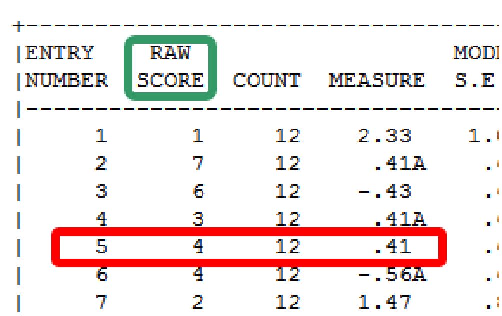


140. Let’s see what has happened to the item difficulty measures this time.
Winsteps menu bar
Click on Output Tables
Click on “14. ITEM: entry”



141. In Table 14, look what is the difficulty of item 5.
Oops! It is .41 logits, but we expected 0.82 logits, about 0.41 logits difference. What has gone wrong? *Nothing!*
Compare the measures for Item 1. Earlier it was 2.66 logits. Now it is 2.33. The entire *frame of reference* has moved .33 logits. Unanchoring item 5 and reanalyzing has caused all the measures of the unanchored items to change, as well as all the person measures. This explains the 0.41 logits.

Green box: This table was produced with TOTALSCORE=No, so it shows “RAW SCORE” instead of “TOTAL SCORE”. The “RAW SCORE” omits extreme persons with minimum possible and maxim possible scores.



<table>
<tr><td>142.</td><td>The computation of “Displacement” assumes that all the other items and all the persons keep their same measures. So we see that Displacement is a useful indication of what an unanchored measure would be, but it does not provide the precise value.

When the Displacement is much smaller than the S.E. then it has no statistical meaning. Its size is smaller than the expected randomness due to the imprecision of the measures.

If you see a Displacement value that is big enough to be alarming, but you are not using anchor values, then the estimation has not run to convergence. Specify smaller values for LCONV= and RCONV=, see Winsteps Help.</td><td>---------------------------
MODEL|| |
MEASURE S.E. ||DISPLACE|
---------------++--------+
2.66 1.08|| .00|
.41A .66|| -.94|
-.13 .65|| -.02|
.41A .66|| .79|
2.00A .87|| -1.18|</td></tr>
<tr><td>143.</td><td>Close all windows</td><td></td></tr>
<tr><td>144.</td><td></td><td></td></tr>
<tr><td>145.</td><td colspan="2">Supplemental Readings</td></tr>
<tr><td>146.</td><td>Bond & Fox Chapter 7: The Partial Credit Model
Bond & Fox Chapter 3: Reliability
Rating Scale Analysis Chapter 7: Fear of Crime
Rating Scale Analysis Chapter 5: Verifying Variables</td><td></td></tr>
</table>