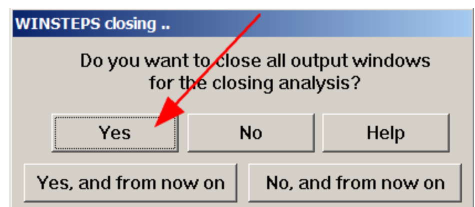
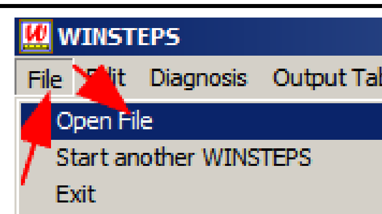
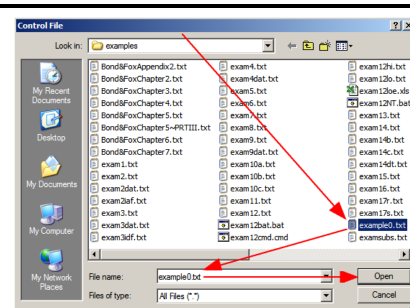
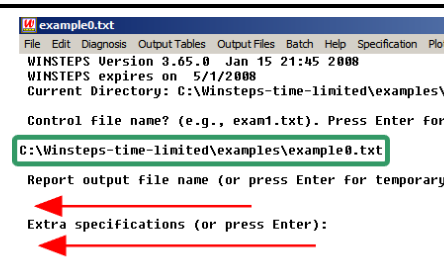
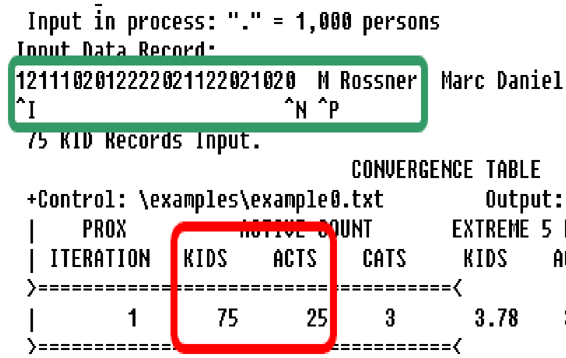
#	Winsteps Rasch Tutorial 2 Mike Linacre, instructor – June 2012	
1.	**Tutorial 2. Fit analysis and Measurement Models** ***Welcome back!*** • Rasch-Andrich Rating Scale Model • Quality-control fit statistics • Scalograms This lesson builds on Lesson 1, so please go back and review when you need to. 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**	
2.	**A. Liking for Science - the control and data file**	
3.	Let's start with rating scales .... Double-click on the Winsteps short-cut on your desktop	
4.	If you see the "Welcome" dialog, please Click on **Don't ask again** Click on **No** You can access this function from **Data Setup** on the Winsteps menu bar.	
5.	Click on **File** Click on **Open File**	
6.	Click on **example0.txt** Click on **Open**	
7.	Let's accept the usual defaults .... Report output file name .... Press **Enter** Extra specifications .... Press **Enter**	

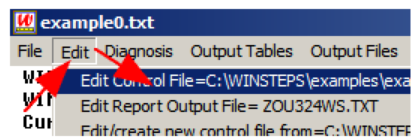
8.

The analysis commences.
Scroll back to the top of the report ...
Notice "Input Data Record". This shows the first data record that was processed (in the green box). We can see that it consists of 0's, 1's and 2's and also a name "Rossner". The data is a 3-category rating scale.
^I means "the item responses start here" (ITEM1=)
^N means "the item responses end here" (NI=)
^P means "the person label starts here" (NAME1=)
The first character of the person label, "M" is a gender indicator. M=Male, F=Female.
All this looks correct. This is the first place to look when an analysis seems to go wrong.
In the red box we see the analysis was of 75 persons ("KIDS" short for "Children") and 25 items ("ACTS" short for "Activities").



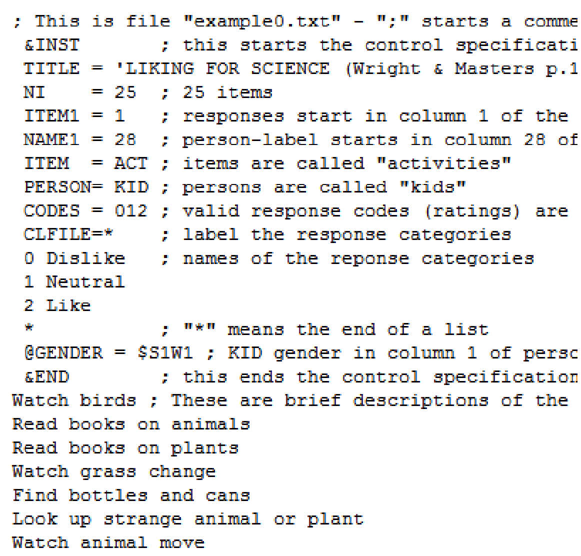
9.

Let's look at the control and data file for this analysis
Click on **Edit** menu
Click on **Edit Control File = ..**

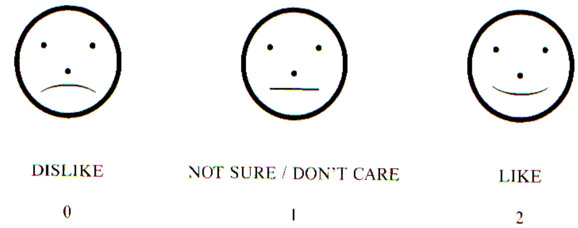


10.

The control and data file displays in a NotePad window.
If this is ragged, see Appendix 5 of Lesson 1.
We can edit this in the Data Setup window, but you will soon discover that it is quicker and easier to edit the control directly. In the control file,
; starts a comment. Anything to the right of ; is ignored.
&INST is optional. It is only here for compatible with very early version of the software. Winsteps is backward compatible with control and data files 20+ years old - unusual in this age of fast-changing computer systems!
The control instructions are "variable = value". They can be UPPER or lower or MixeD case. Spaces before and after = don't matter, same with the order of the variables .
&END is between the control variables and the first item label "Watch Birds"
We must specify:
NI = Number of items
ITEM1= first column of item responses
Usually also:
NAME1 = first column of person label
CODES = valid codes in the item responses



```
; This is file "example0.txt" - ";" starts a comme
 &INST       ; this starts the control specificati
 TITLE = 'LIKING FOR SCIENCE (Wright & Masters p.1
 NI    = 25  ; 25 items
 ITEM1 = 1   ; responses start in column 1 of the
 NAME1 = 28  ; person-label starts in column 28 of
 ITEM  = ACT ; items are called "activities"
 PERSON= KID ; persons are called "kids"
 CODES = 012 ; valid response codes (ratings) are
 CLFILE=*    ; label the response categories
 0 Dislike   ; names of the reponse categories
 1 Neutral
 2 Like
 *           ; "*" means the end of a list
 @GENDER = $S1W1 ; KID gender in column 1 of persc
 &END        ; this ends the control specification
Watch birds ; These are brief descriptions of the
Read books on animals
Read books on plants
Watch grass change
Find bottles and cans
Look up strange animal or plant
Watch animal move
```

11.	This is the “Liking for Science” data from the book “Rating Scale Analysis” (Wright & Masters). 75 children visiting a Science Museum were asked their opinion of 25 science-related activities. They responded using a pictorial rating scale of 3 faces. The meanings of the faces and the scoring were added later ... The museum’s experts decided that the values 0, 1, 2 are qualitatively-advancing levels of **liking for science.** CODES = 012 specifies that these are the values of the observations in the data file. Winsteps sees these three values and automatically detects that the data are on a three-category rating scale.	

12.	The definitions of the 3 categories of the rating scale are given in CLFILE=*. (CLFILE = Category Label File). This is a sub-list of 3 categories. Sublists start with =* and end with *.	`CLFILE=*     ; label the re`
		`0 Dislike   ; names of the`
		`1 Neutral`
		`2 Like`
		`*           ; "*" means the`

13.	Notice also @GENDER. @ means that this is a user-defined control variable. $S1W1 means this references the part of the person label that starts in column 1 of the label (S1) and is one column wide (W1). This is where M or F is in the person label. @Gender : I am going to define some columns in the labels as the "Gender" codes. = S1 : "starting in column 1 of the label, 1 column wide" = S1W2 : "starting in column 1 of the label, 2 columns wide" Winsteps discovers whether the "Gender" is in the item or person label by how it is used. For instance: "PSUBTOTAL = @Gender" means "Subtotals based on the Gender code in the person labels" for Table 28. If gender was in the item label starting in column 1 of the label, 1 column wide (i.e. @gender=$S1W1), I could write the command "ISUBTOTAL=@Gender" meaning subtotals based on the gender code in the items label. So it is the P or the I in the subtotal command that implies where gender is found (or whatever it is you are stipulating) and @gender is just stating the starting column (of the person or item label) and how wide.	`@GENDER = $S1W1 ; KID gende`

<table>
<tr><td>14.</td><td>There are 25 item labels (one per item) then
END NAMES (or END LABELS, they mean the same thing).
The data lines follow. This time the item responses are to the left and person labels to the right. This is usually the simplest layout, because then the item number corresponds to the column number in the data file.</td><td><pre>Watch a rat
Find out what flowers live on
Talk w/friends about plants
END NAMES ;this follows the item names: - the
1211102012222021122021020 M Rossner, Marc Da
2222222222222222222222222 M Rossner, Lawrence
2211011012222122122121111 M Rossner, Toby G.
1010010122122111122021111 M Rossner, Michael</pre></td></tr>
<tr><td>15.</td><td>The last child, Linus Pauling, is in the last line of the data file. There is no need to tell Winsteps how many persons, rows, cases, subjects there are. Winsteps will read to end-of-file and count them up itself.</td><td><pre>2201222001220022220222201 M Stoller, Dave
1001000100011010021200201 M Jackson, Solomon
2012010122210201102202022 M Sandberg, Ryne
2220022002222222222022012 M Patriarca, Ray
1200110112122221022020010 M Pauling, Linus</pre></td></tr>
</table>

16. **B. Liking for Science - first steps in the analysis**

17. Here's a task for you. The most important form of validity for a test is **"Construct Validity".** For us, it evaluates the question **"Are we measuring what we intended to measure?"**

The Museum wants to measure "Liking for Science-related Activities". Look at this list of 25 items.

Which items do you think the children liked most?
Note down the numbers of three items.

Which item do you think the children liked least?
Note down the numbers of three items.

This is the start of our own "construct map". If we were more involved in this, we would note down a list of all the items in order of difficulty as we imagine it. This would be our complete "construct map". This "map" is like a road map from New York to Chicago. We see where we start out, at the lowest category of the easiest-to-like item, and where we are going to, the highest category of the hardest-to-like item. And there are all the other item locations in-between.

The construct map reflects our **Construct Theory.** We will compare our map with what the analysis tells us. **This is the way that we will learn the most from our analysis.** We expect that the data will mainly support our theory. But we also expect the data to contain contradictions to our construct theory. Sometimes these contradictions will improve our theory, sometimes they will raise questions about the quality of the data ….

```
 1 WATCH BIRDS
 2 READ BOOKS ON ANIMALS
 3 READ BOOKS ON PLANTS
 4 WATCH GRASS CHANGE
 5 FIND BOTTLES AND CANS
 6 LOOK UP STRANGE ANIMAL OR PLANT
 7 WATCH ANIMAL MOVE
 8 LOOK IN SIDEWALK CRACKS
 9 LEARN WEED NAMES
10 LISTEN TO BIRD SING
11 FIND WHERE ANIMAL LIVES
12 GO TO MUSEUM
13 GROW GARDEN
14 LOOK AT PICTURES OF PLANTS
15 READ ANIMAL STORIES
16 MAKE A MAP
17 WATCH WHAT ANIMALS EAT
18 GO ON PICNIC
19 GO TO ZOO
20 WATCH BUGS
21 WATCH BIRD MAKE NEST
22 FIND OUT WHAT ANIMALS EAT
23 WATCH A RAT
24 FIND OUT WHAT FLOWERS LIVE ON
25 TALK W/FRIENDS ABOUT PLANTS
```

18. Now back to the analysis .....
In the Windows task bar,
Click on the Winsteps analysis example0.txt

19. *Pretend that this is a new rectangular text dataset (looking like **example0dat.txt**) that we haven't seen before. Let's imagine that we have been asked to help analyze these data. Here's how I would go about it.*

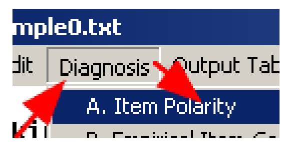
1. Construct the Winsteps Control and Data file - we did this in Lesson 1.
2. Start the analysis and check that
   start of item responses: ITEM1= (^I)
   end of item responses: NI=(^N)
   start of person name/label: NAME1=(^P)
   align correctly with the data record. It's the position of ^ that is important.

```
Input Data Record:
1211102012222021122021020  M Rossner  Marc Daniel
^I                       ^N ^P
```

20. Before worrying about the summary statistics on the analysis screen we need to look at more basic aspects of the analysis.

```
+-----------------------------------------------------------------------------+
| KIDS      75 INPUT      75 MEASURED                INFIT         OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      31.4      25.0         .90     .41       .99    -.2   1.08     .1|
| S.D.       8.4        .0        1.22     .11       .50    1.6   1.04    1.9|
| REAL RMSE    .43  ADJ.SD    1.14  SEPARATION  2.67  KID    RELIABILITY  .88|
|-----------------------------------------------------------------------------|
| ACTS      25 INPUT      25 MEASURED                INFIT         OUTFIT     |
| MEAN      93.0      74.0         .00     .25      1.02    -.2   1.08     .0|
| S.D.      30.9        .0        1.41     .08       .45    2.3    .87    2.8|
| REAL RMSE    .26  ADJ.SD    1.38  SEPARATION  5.32  ACT    RELIABILITY  .97|
+-----------------------------------------------------------------------------+
```

21. On the Winsteps Analysis Window menu bar,
click on **Diagnosis** menu
click on **A. Item Polarity**



22. The "item polarity" Table displays in NotePad. It is identified as Table 26 in the top left corner.
If the Table looks wrapped or ragged, see Lesson1 Appendix 5.

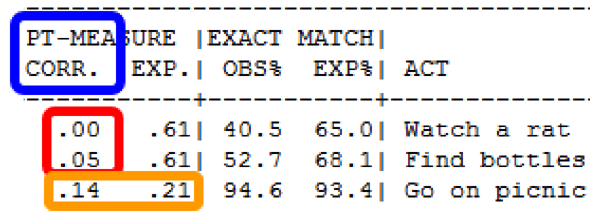
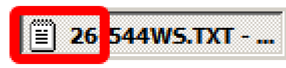
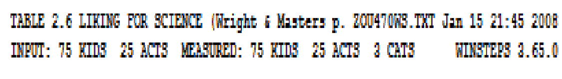
```
TABLE 26.1 LIKING FOR SCIE
INPUT: 75 KIDS   25 ACTS   M
--------------------------
```

23. Table 26.1 displays. This shows the items ordered by **point-measure correlation**. This answers the question: ***Do the responses to this item align with the abilities of the persons?*** A fundamental concept in Rasch measurement is that:
higher person measures ➔ higher ratings on items
higher ratings on items ➔ higher person measures
The point-measure correlations (PT-MEASURE) report the extent to which this is true for each item. **We want to see noticeably positive correlations** (green box).
Negative and close-to-zero correlations (red box) sound alarm bells. Small positive correlations (orange box) may need further investigation.

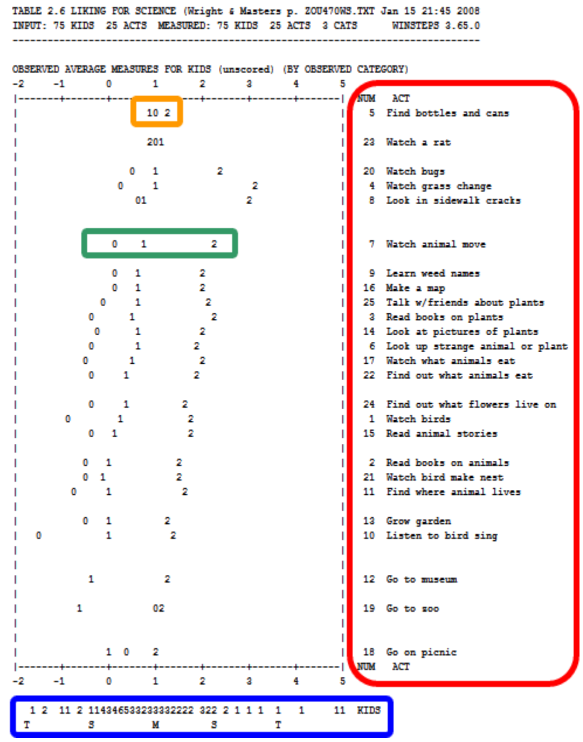
```
---------------------------------------------------
PT-MEASURE |EXACT MATCH|
CORR.  EXP.| OBS%  EXP%| ACT
-----------+-----------+---------------------------
 .00   .61| 40.5  65.0| Watch a rat
 .05   .61| 52.7  68.1| Find bottles
 .14   .21| 94.6  93.4| Go on picnic

 .66   .55| 66.2  57.4| Talk w/frienc
 .70   .53| 73.0  58.7| Watch what ar
 .72   .55| 73.0  57.7| Read books or
-----------+-----------+---------------------------
```

24. No correlations are negative – Good! **Negative correlations usually indicate that the responses to the item contradict the direction of the latent variable.** If there had been negative correlations we would want to check for reversed item wording. If we find it, we will want to rescore the item. Winsteps has facilities for doing this, as we will discover later.

A possible reversed-meaning item that might have appeared in this survey:
Yell at animals
A child who likes to shout at animals will probably not like the other activities. This item would have a negative correlation.

25.	Zero and low positive correlations. We don’t have any negative correlations. Good! But we do have a zero and two small positive correlations - Oops! Let’s make a mental note of these items for investigation later …. .14 is “Go on Picnic”, a very easy-to-like item. The EXP. (expected correlation) shows what the correlation would be, .21, if the data matched the Rasch model. .14 is close to .21, so it looks OK.	
26.	Let’s continue our investigation. Don’t close Table 26. Leave it on the Windows Task bar. The “26” means “Table 26”. Following 26- is a random number to identify the file.	
27.	On the Winsteps Analysis Window menu bar, click on **Diagnosis** menu click on **B. Empirical Item-Category Measures**	
28.	Table 2.6 displays in a NotePad window. Do you notice that the Diagnosis A. was Table 26? The Diagnosis menu displays selected Tables and Sub-Tables from the 34 Tables produced by Winsteps. If this Table is wrapped or ragged, see Appendix 3.	

<table>
<tr>
<td>29.</td>
<td>Table 2.6 is a picture that is packed with meaning.
Red box: these are the items. They are ordered vertically according to their difficulty or challenge according to the data. At the bottom of the red box we see “Go on Picnic”. The children are telling us that this bottom item is the easiest item to like.
The vertical spacing approximates the items’ placement on the linear Rasch dimension, so that going from “Go on Picnic” to “Go to Museum’ is approximately the same advance in item difficulty as going from “Go to Museum” to “Find where animals live”.
The most challenging item, at the top, is “Find bottles and cans”. The children are telling us that this item is most difficult to like. (This survey was conducted when empty bottles and cans were considered to be worthless trash, not recyclable resources.)
Blue box: distribution of the children on the variable. Here our sample ranges across the operational range of the instrument.
Green box: the category numbers are positioned at the average measures of the children in this sample who chose each of them (the empirical average measures).. Here the positions of the category numbers agree strongly with our theory that “higher category ↔ more of the latent variable”.
Orange box: here the empirical average measures for all three categories are close together and disordered: 1-0-2. This item is not agreeing with the other items in defining the latent variable. We must investigate this item.</td>
<td>

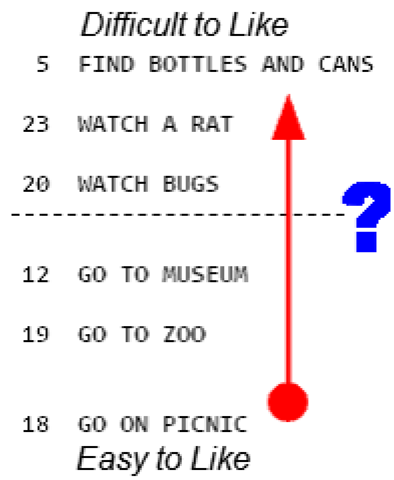


</td>
</tr>
<tr>
<td>30.</td>
<td>Glance down at the Windows Task Bar. You will see that Table 2 is shown.</td>
<td>

</td>
</tr>
</table>

31. **Construct validity:** does this vertical hierarchy make sense? The vertical item hierarchy tells us what "more" and "less" of the latent variable means. **The item hierarchy defines what we are measuring.** Is it what we intend to measure? Look down the item hierarchy in the red box. What is its message? ***Is the Museum measuring what they intended to measure?*** Is there anything here that might surprise science-museum administrators?

Practical Challenge: Imagine you are describing this latent variable (this construct) to the Museum administrators. **Pick out 4 items** which show how children progress from low interest to high interest:
(item for disinterested children) →
(item for slight interest) →
(item for moderate interest) →
(item for enthusiastic children)
Also, look back at your selection of the three most and least likable items**. How did your choices compare with the children's?**

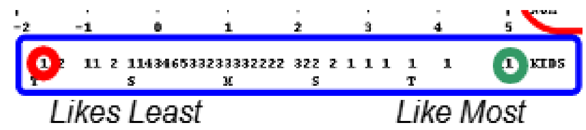
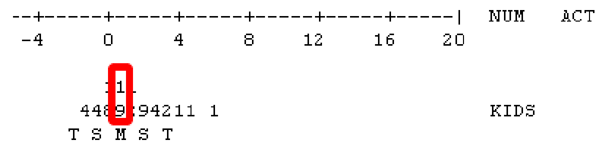
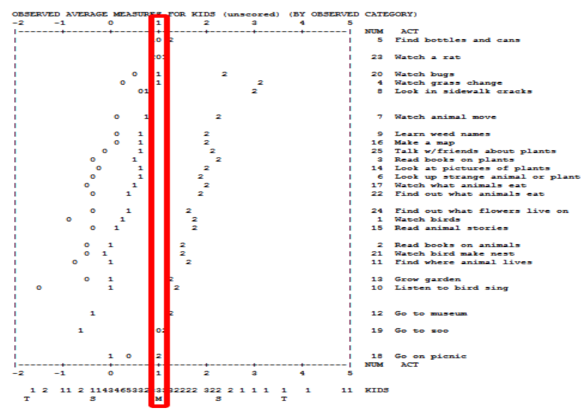


Note: I recently visited a large cultural Museum. It was clear from the guided tour that the Museum's curators thought that *"older = more interesting"*. So we saw lots of repetitive old stuff. But perhaps *"surprising, beautiful = more interesting"*. What do you think? Let's hope the cultural Museum conducts a survey like this one.

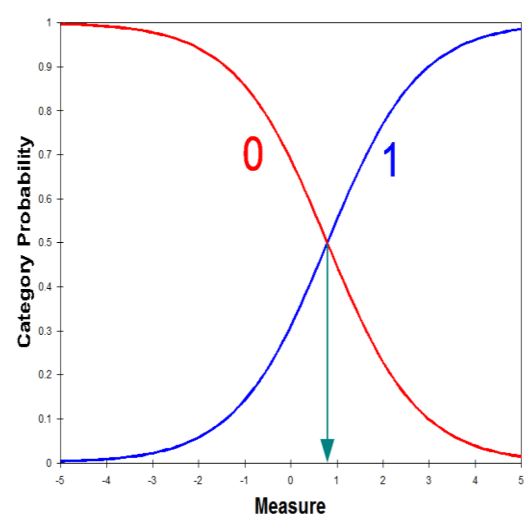
32. **Predictive validity:** do the children's measures match our expectations?
"Predictive" means "predicts future performance" - and when we get that information we will use it. But often that information comes too late or never. So, in statistics, we "predict" things that have already happened, or the current data set. If we are interested in "predictive validity", we must know something about the members of our sample. For instance:
Is a child excelling at science at school? If so, we expect that child to score highly on our survey.
Is a child younger? If so, we expect that child to have a low score on our survey.
So the "predictive validity" of an instrument is revealed by the person ability hierarchy in the same way that the "construct validity" is revealed by the item difficulty hierarchy.

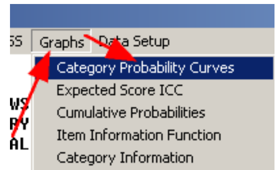
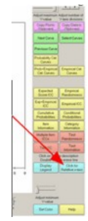
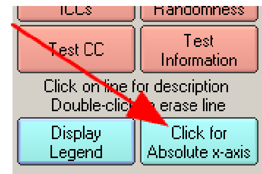
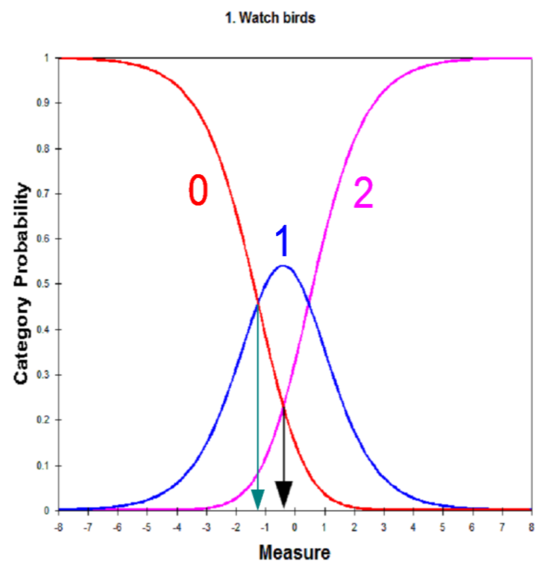
In Rasch analysis the term "ability" is used generically to refer to what me are measuring about the persons, and the term "difficulty" is used generically to refer to what we are measuring about the items.

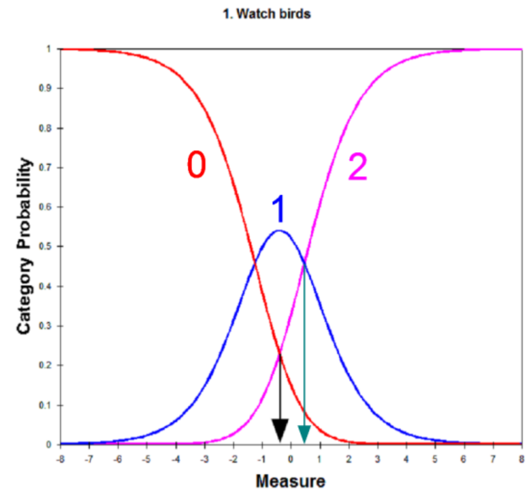
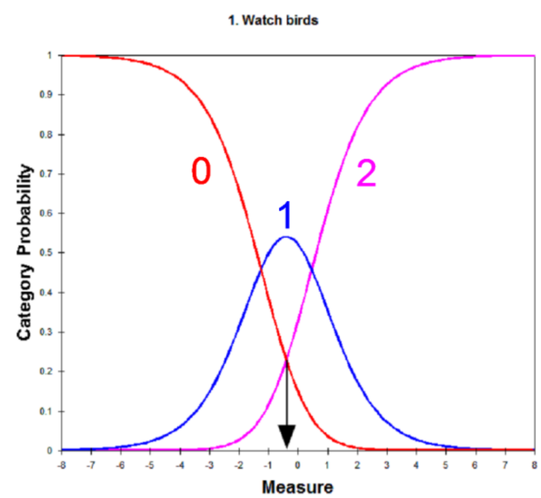
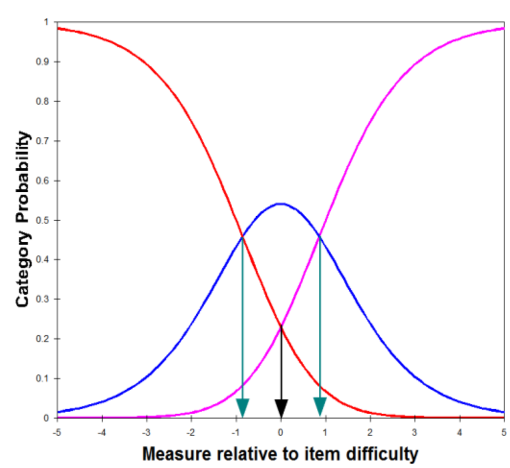
When we report Rasch results for a non-specialist audience, we change the terms to match the construct (latent variable). For instance, for the "Liking for Science" data:
the "person ability" = the child's "likingness for science"
the "item difficulty" = the item's "dislikeability as a science activity"

33.	**Predictive validity:** do the children's measures match our expectations? The **blue box** shows the distribution of the children's measures. The "1" on the extreme left in the red circle is the one child who liked these activities the least. The "1' on the right in the **green** circle is the one child who liked these activities the most. The average of the children's liking measures is indicted by "M" (= Mean). The "3" indicates that there are 3 children at this location. "S" is one standard deviation from the mean. "T' is two standard deviations.	
34.	When there are more than 9 persons in one location then the digits are printed vertically, so that 19 becomes 1 over 9.	
35.	Do you notice that the "M" (for Mean) under the person distribution is at about +1? This is one logit above the zero-point on the measurement scale, its "local origin". The local origin at 0 is set at the average difficulty of the items. This would be the location at which the average response to the survey questions is "1", "Neutral". So, in this survey, the average child is responding 1 logit above neutral, towards "Like". To understand more about the contents of this Table we need to think more about the Rasch model.	

36.	Summary: What do we learn from item-person maps? **1. Distributions:** Persons: we usually expect a normal distribution, or a distribution skewed in a certain way. Do we see this? Items: we usually expect the items to be uniformly distributed (like marks on a ruler) or clustered at pass-fail points. Do we see this? **2. Targeting:** Persons-Items: Is the test too easy or too hard for the sample? For educational tests, we expect about 80% success (= 1.5 logit difference between the person and items for dichotomous data). On surveys we may expect 70% "agreement" due to the normal psychological process of "compliance". **3. Predictive validity:** Persons: Are the people ordered as we would expect based on other information about them. Do the experienced people have higher measures? Do the healthier people have higher measures? Do the more educated people have higher measures? **4. Construct validity:** Items: Are the items ordered as we would expect based on what we intend to measure? Is "division" generally more difficult than "addition"? Is "climbing stairs" more difficult than "eating"? Is "hitting a home run" more difficult than "hitting a single"? **5. Inference:** Persons: The Rasch measure is like a person's height or weight. It is an independent number which we can then use to predict or construct further information about the person. If the person measures as "healthy" or "happy", then we expect a longer life than someone who measures "unhealthy" or "unhappy". Items: From the item hierarchy we often learn more about the underlying variable. It was this which brought Trevor Bond (of Bond & Fox) in contact with Rasch measurement. He needed to strengthen some aspects of the Piagetian theory of child development. CTT couldn't help him. Rasch did. From the stucture of the Rasch hierarchy, Trevor adjusted some aspects of Piagetian theory and was able (apparently for the first time) to compare the sizes of the gaps between the Piagetian stages. The JPS Rasch Analysis Homepage: www.piaget.org/Rasch/jps.rasch.home.html
37.	

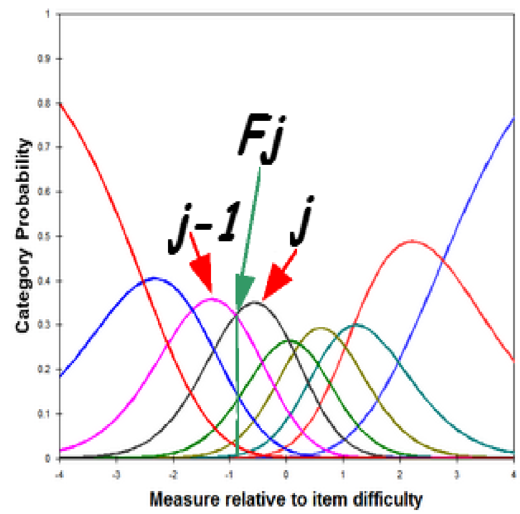
38.	**C. Rasch Polytomous Models**	
39.	We have already encountered dichotomous data, such as “Right or Wrong”. “Dicho-tomous” means “two cuts” in Greek. In performance assessment and attitude surveys, we encounter rating scales, such as “none, some, plenty, all” and “strongly disagree, disagree, neutral, agree, strongly agree”. This is a “Likert” (Lick-urt) scale, popularized by Rensis Likert, 1932. There are several Rasch measurement models for rating scales so we will call them “polytomous models”. “Poly-tomous” means “many cuts” in Greek. In the literature you will also see them also called “polychotomous” models - an example of what etymologists call “mistaken back-formation”!	
40.	**The Rasch-Andrich Rating Scale Model**	
41.	David Andrich (now the Chapple Professor at the University of Western Australia) published a conceptual break-through in 1978. He perceived that a rating scale could be thought of as a series of Rasch dichotomies. – See Lesson 1 for the Rasch dichotomous model. *Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 357-74.*	
42.	The **Rasch-Andrich Rating Scale Model** specifies the probability, P_{nij}, that person *n* of ability B_n is observed in category *j* of a rating scale applied to item *i* of difficulty D_i as opposed to the probability $P_{ni(j-1)}$ of being observed in category *(j-1)*. So, in a Likert scale, *j* could be “agree” then *j-1* would be “neutral”.	$\log_e(P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_j$ F_j is the “Rasch-Andrich threshold” also called the “step calibration” or “step difficulty”
43.	What is the model all about? Let’s start with our friend, the Rasch dichotomous model from Lesson 1. We can write it this way:	$\log_e(P_{ni1} / P_{ni0}) = B_n - D_i$
44.	As we go along the latent variable, we can plot the probability of scoring a correct answer. For someone of very low ability at the left-end of the latent variable, the probability of a wrong answer, of scoring 0, is very high (red line), and the probability of a correct answer, or scoring 1, is very low (blue line). For someone of very high ability at the right-hand end, the probability of scoring 0 (red line) is very low and the probability of scoring 1 (blue line is very high). For someone whose ability exactly matches the item’s difficulty (green arrow), the probability of scoring 1 and scoring 0 are the same. So the Measure points to the ability Bn which matches the difficulty Di of the item, In this case, the item difficulty, Di is about 0.8 logits. The point where the 0 and 1 probability curves cross is called the Rasch-Andrich Threshold.	

<table>
<tr><td>45.</td><td>Now let’s look at the same thing for a rating scale. Here it is for the “Liking for Science”.
Click on the Winsteps menu bar, “Graphs menu”,
click on “Probability Category Curves”.</td><td>
</td></tr>
<tr><td>46.</td><td>Let’s see these probability curves relative to the latent variable.
Click on “Click for Absolute x-axis”.
Here “Absolute” means “relative to zero-point of the latent variable”.
The label on the button changes to “Click for Relative x-axis”, which means “relative to the difficulty of the item being displayed”.</td><td> 
</td></tr>
<tr><td>47.</td><td>Let’s rewrite the Rasch-Andrich model for the relationship between categories 0 and 1. It is the same as the dichotomous model, but with one more parameter, F_1</td><td>$$\log_e(P_{ni1} / P_{ni0}) = B_n - D_i - F_1$$</td></tr>
<tr><td>48.</td><td>Here is the 3-category “Liking for Science” data. “Watch Birds”.
Look closely at the relationship between the red “0” line and the blue “1” line.
At the left-hand side it definitely looks the same as the dichotomous model: high probability of 0 and low probability of 1.
Then we reach the Rasch-Andrich threshold (green arrow) where the probability of 0 and 1 is the same. This location is at “the item difficulty + the first threshold” = measure $D_i + F_1$.
After that, the probability of 1 increases but then falls. Meanwhile the probability of 0 falls even faster.
At the right hand side, when we compare the probability of scoring 1 to the probability of scoring 0, the probability of scoring 1 always increases relative to the probability of 0, exactly as in the dichotomous case.</td><td>
</td></tr>
<tr><td>49.</td><td>We can tell the same story about categories 1 and 2. Here is the Rasch-Andrich model for these categories</td><td>$$\log_e(P_{ni2} / P_{ni1}) = B_n - D_i - F_2$$</td></tr>
</table>

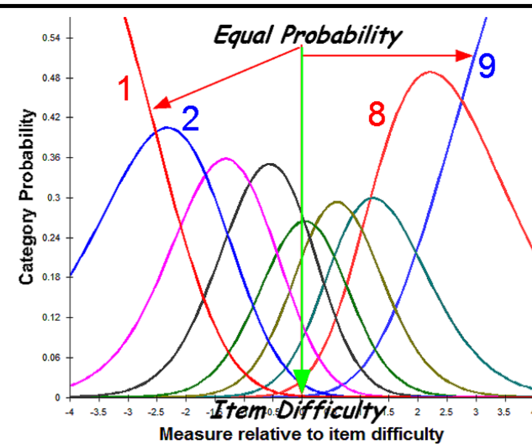
50.	Look closely at the relationship between the blue “1” line and the mauve “2” line. At the left-hand side, the probability of 1 is higher than the probability of 2. In fact, 1 and 2 have relatively the same probabilities as 01 and 1 in the dichotomous model. Then we reach the Rasch-Andrich threshold (green arrow) where the probability of 1 and 2 is the same. This location is at “the item difficulty + the 2nd threshold” = measure $D_i + F_2$. At the right-hand side it definitely looks the same as the dichotomous model: low probability of 1 and high probability of 2.	
51.	So we have two dichotomous relationships: 0-1 and 1-2. When we put them together using the fact that probabilities always sum to 1, we see the 3-category picture.	$\log_e(P_{ni1} / P_{ni0}) = B_n - D_i - F_1$ $\log_e(P_{ni2} / P_{ni1}) = B_n - D_i - F_2$ $P_{ni0} + P_{ni1} + P_{ni2} = 1$
52.	And where is the item difficulty, D_i ? It is located where the top and bottom categories are equally probable, the black arrow. This item is a less difficult to like (less challenging) item, so the rating scale structure is located around the item difficulty of –0.4 logits, below the average item difficulty of 0 logits. In the Rasch-Andrich model, the rating scale structure, parameterized by {Fj}, is defined to be the same for all items. This is ideal for many applications, such as Likert scales (Strongly Agree, Agree, Neutral, Disagree Strongly Disagree), where the response structure for each item is intended to be the same. The rating scale structure slides up and down the latent variable for each item to match that item’s difficulty.	
53.	In the Rasch-Andrich model, the rating scale structure, is defined to be the same for all items, for all different values of D_i. So the picture looks the same for every item, relative to its item difficulty. The Rasch-Andrich thresholds (green arrows) are in the same place relative to the item difficulty (black arrow) for all items. This is ideal for many applications, such as Likert scales (Strongly Agree, Agree, Neutral, Disagree Strongly Disagree), where the response structure for each item is intended to be the same. The rating scale structure slides up and down the latent variable for each item to match that item’s difficulty.	

54. As with the Rasch's original dichotomous model, B_n is the person ability, attitude, capability, etc. D_i is the item difficulty, challenge, impediment.
But we now have a new location on the latent variable, F_j. This is called the "Rasch-Andrich threshold", also called the "step calibration" or "step difficulty" or "tau", Greek letter: τ. It is the point on the latent variable (relative to the item difficulty) where the probability of being observed in category *j* equals the probability of being observed in category *j-1*.

This plot shows the probability curves for each category of a 9-category rating scale according to the Rasch-Andrich model. The F_j at about –0.9 logits relative to the item difficulty at zero is the location of equal-probability between the 3^{rd} and 4^{th} categories, so is F_4.



55. You may be asking yourself: "How does this rating scale probability structure align with item difficulty?" The answer is simple: **the item difficulty is located at the point where the highest and lowest categories are equally probable.**
In this Figure, the probability curves of a 9-category (1, 2, 3, 4, 5, 6, 7, 8, 9) rating scale are drawn relative to the item difficulty, so the item difficulty is at "0" on the x-axis. This is the point (green arrow) where the highest, "9", and lowest, "1" categories are equally probable.

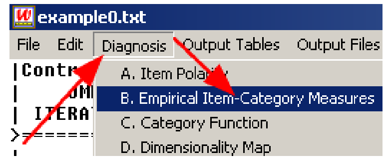
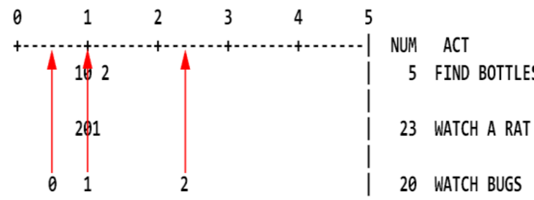
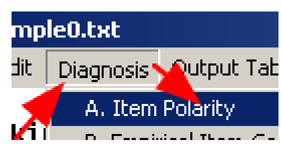


56. **Parameter estimation and fit statistics:**
Rasch assumes that the incoming data represent qualitively-ordered observations on the intended latent variable. Based on this assumption, the Rasch measures are computed.
After the parameters are estimated:
Pnij is the model probability of observing category j based on those estimates.
ΣjPnij is the expected value of each observation.
The expected values are compared with the observed values, and fit statistics are produced. It is here that a large category misfit could indicate that the original qualitative category ordering was incorrect.

Parameter Estimation:

$$\log_e(P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_j$$

j are the specified qualitative-levels
Bn is estimated from the person raw score.
Di is estimated from the item raw score.
Fij is estimated from the category frequency.

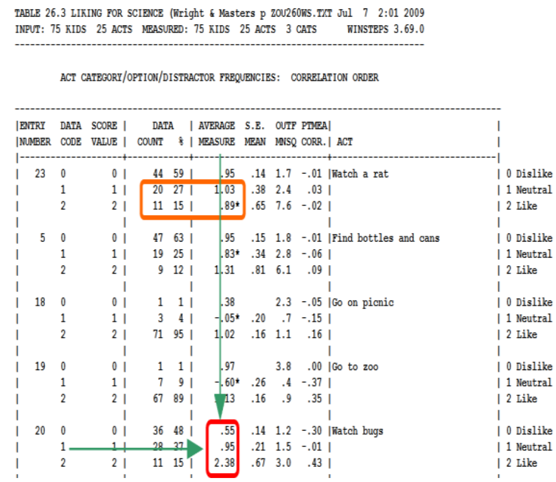
57.	D. **Diagnosis B. Empirical Item-Category Average Measures**	
58.	Click on “02-…” on your Windows Task bar, if it is still there, or Click on the Winsteps Analysis Window Click on Diagnosis menu Click on B. Empirical Item-Category Measures	
59.	Let’s see how the values in this Table 2.6 are computed. For each response in the data file there is a person ability and item difficulty. In this Table, each item is positioned vertically roughly according to its difficulty, D_i. Each response category for each item is positioned horizontally according to the average of the abilities (“liking for science”) of the children who responded that category for the item. So, in this example, look at the third line “Watch Bugs”. The average “liking” ability of the children who chose “Dislike” (0) in response to “Watch Bugs” is 0.5 logits. The average ability of those who chose “Neutral” (1) is 1 logit. And the average ability of those who chose “Like” (2) is 2.3 logits. **These values are descriptions of the sample. They are not Rasch-model estimates of F_j.**	 Look at the first item, “Find bottle” and at the second item, “Watch a rat”. Is there something wrong? Something seriously wrong? We expect to see: 0 1 2
60.	It’s useful to be able to look at the pictures and then refer to the corresponding numbers. We can find these in Table 26 and many other item Tables. Click on “26-…” on your Windows Task bar, if it is still there, or Click on Diagnosis menu Click on A. Item Polarity	

61.

Table 26 displays.
Scroll down to Sub-Table 26.3
Red box: The Average Measures we saw plotted in Table 2.6 are listed here. We can see that their exact values are .55, .95 and 2.38.

This is a general approach in Winsteps:
1. Look at a picture to identify something interesting.
2. Look at a Table of numbers to see exact details.

Notice that this Table gives considerable statistical information about each response code for each item. **Average measure** reports the average ability of the children who selected the response. **We expect higher categories to be chosen by children with higher "liking" abilities.** When that doesn't happen, there is a * (orange rectangle) to warn us that the average abilities are out of order.
S.E. Mean gives the standard error of the mean of the distribution of the abilities of the people who responded in this category. It is useful if you want to use a t-test in investigate whether there is a statistically significant difference between the levels of ability for (children who choose 0, children who choose 1, children who choose 2) for each item. See adjacent panel →
OUTF MNSQ is the Outfit Mean-square statistic for observations in the category, useful for detecting unexpected responses.
PTMEA CORR. is the point-measure correlation between scored responses and ability measures: each correlation is computed with a response in this category scored as "1", and responses in other categories scored "0". We expect the highest category will have a strong positive correlation with ability, and the lowest category to have a strong negative correlation with ability.



```
TABLE 26.3 LIKING FOR SCIENCE (Wright & Masters p ZOU260WS.TXT Jul  7  2:01 2009
INPUT: 75 KIDS  25 ACTS  MEASURED: 75 KIDS  25 ACTS  3 CATS     WINSTEPS 3.69.0
--------------------------------------------------------------------------------

      ACT CATEGORY/OPTION/DISTRACTOR FREQUENCIES:  CORRELATION ORDER

-----------------------------------------------------------------------------------
|ENTRY   DATA  SCORE |     DATA    | AVERAGE  S.E.  OUTF PTMEA|                  |
|NUMBER  CODE  VALUE |  COUNT    % | MEASURE  MEAN  MNSQ CORR.| ACT              |
|--------------------+-------------+-------------------------+-------------------|
|   23   0         0 |     44   59 |      .95   .14  1.7  -.01 |Watch a rat       | 0 Dislike
|        1         1 |     20   27 |     1.03   .38  2.4   .03 |                  | 1 Neutral
|        2         2 |     11   15 |      .89*  .65  7.6  -.02 |                  | 2 Like
|                    |             |                          |                  |
|    5   0         0 |     47   63 |      .95   .15  1.8  -.01 |Find bottles and cans | 0 Dislike
|        1         1 |     19   25 |      .83*  .34  2.8  -.06 |                  | 1 Neutral
|        2         2 |      9   12 |     1.31   .81  6.1   .09 |                  | 2 Like
|                    |             |                          |                  |
|   18   0         0 |      1    1 |      .38        2.3  -.05 |Go on picnic      | 0 Dislike
|        1         1 |      3    4 |     -.05*  .20   .7  -.15 |                  | 1 Neutral
|        2         2 |     71   95 |     1.02   .16  1.1   .16 |                  | 2 Like
|                    |             |                          |                  |
|   19   0         0 |      1    1 |      .97        3.8   .00 |Go to zoo         | 0 Dislike
|        1         1 |      7    9 |     -.60*  .26   .4  -.37 |                  | 1 Neutral
|        2         2 |     67   89 |     1.13   .16   .9   .35 |                  | 2 Like
|                    |             |                          |                  |
|   20   0         0 |     36   48 |      .55   .14  1.2  -.30 |Watch bugs        | 0 Dislike
|        1         1 |     28   37 |      .95   .21  1.5  -.01 |                  | 1 Neutral
|        2         2 |     11   15 |     2.38   .67  3.0   .43 |                  | 2 Like
```

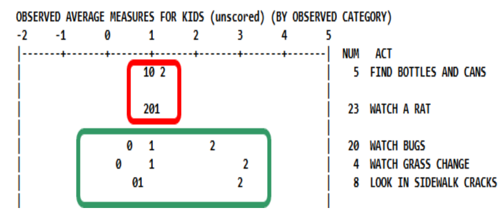
Orange box: Let's compare Average Measures for item 23, categories 1 and 2. For each category, there is an ability distribution with a count n, an average ability µ and a S.E. of the mean. So this is an independent-samples, unequal variances, *t*-test:

$$t = (\mu 2 - \mu 1) / \sqrt{(SE2^2 + SE1^2)}$$
with (n2 + n1 - 2) d.f.
$$= (.89 - 1.03) / \sqrt{(.65^2 + .38^2)}$$
with approximately (11+20-2) d.f.
$$= -0.19 \text{ with } 29 \text{ d.f.}$$

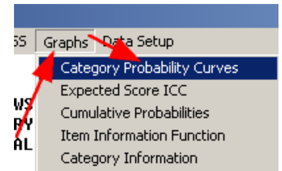
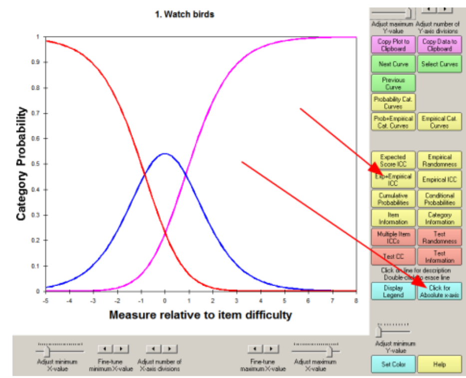
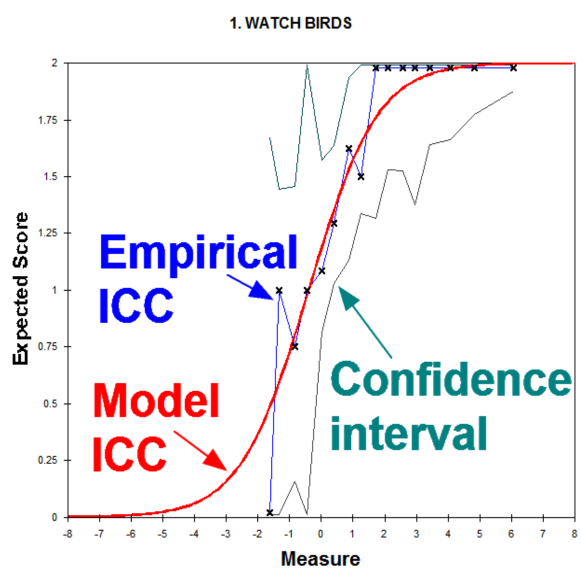
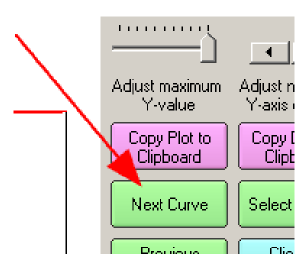
We conclude that category 1 and category 2 have samples whose average ability does not differ significantly.
Since this computation involves approximations, interpret significance tests conservatively (so that we need bigger-than-usual *t*-statistics to convince ourselves that the result is statistically significant).

62.

On the Windows Task Bar, click back to Table 2.6
We can now see that the sample of children has behaved on most items how we would expect, in the **green** box, "higher category → higher average measure". But some items are contradicting our theory. In the **red** box, we don't yet know what is the message contained in this empirical category order, but it definitely doesn't concur with the Rasch definition of the latent variable.

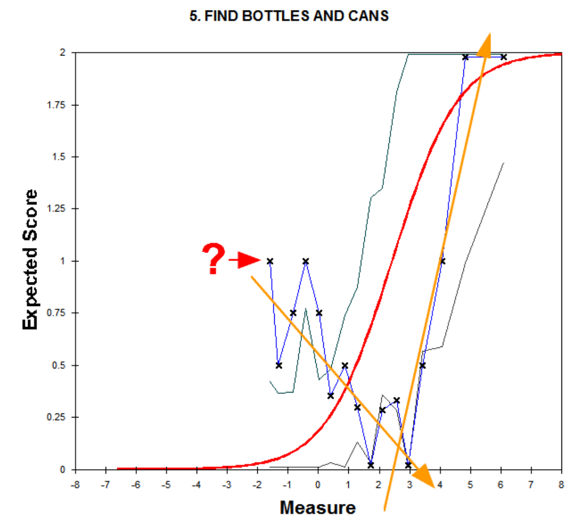


```
OBSERVED AVERAGE MEASURES FOR KIDS (unscored) (BY OBSERVED CATEGORY)
-2     -1      0      1      2      3      4      5
|------+------+------+------+------+------+------|  NUM   ACT
|                   10 2                          |    5  FIND BOTTLES AND CANS
|                                                 |
|                   201                           |   23  WATCH A RAT
|                                                 |
|              0   1         2                    |   20  WATCH BUGS
|             0    1            2                 |    4  WATCH GRASS CHANGE
|               01             2                  |    8  LOOK IN SIDEWALK CRACKS
|                                                 |
```

63.	**E. Model and Empirical Item Characteristic Curves (ICCs)**	
64.	Let’s investigate the items in the **red** box. Click on the Graph window, “1. Watch”, on the Windows Task Bar. If it is not there, Click on the *Winsteps* menu bar, “Graphs menu”, click on “Probability Category Curves”.	 
65.	The Graph window displays. If the **blue** button says “Click for Absolute x-axis”, then please click on it. We want, “Measure” as the title of the x-axis. Then click on “Exp+Empirical ICC”	
66.	The Graph now shows 4 lines. The red line is the item characteristic curve as expected by the Rasch model. It is the Rasch-model prediction of how children at different measures along the latent variable (x-axis) would score on the item (y-axis) on average. The blue line is the empirical ICC. Each “x” summarizes the responses of children with measures near the measure of “x” on the x-axis. We can see that the blue line approximates the red line. The green-gray lines are two-sided 95% confidence bands. These are 1.96 standard errors vertically away from the red line. The more observations in an interval, the closer the green lines are to the red line. The Graph for Item 1. Watch Birds looks about as good as it gets.	
67.	Click on “Next Curve” several times until you get to our first suspect item - **Item 5. Find Bottles and Cans** As you click, notice that the blue lines for items 2, 3, 4 are within the confidence intervals. The red lines for all the items have the same shape. This is a characteristic of the Andrich Rating Scale Model. But the red lines move left and right on the latent variable, depending on the difficulty of the item displayed.	

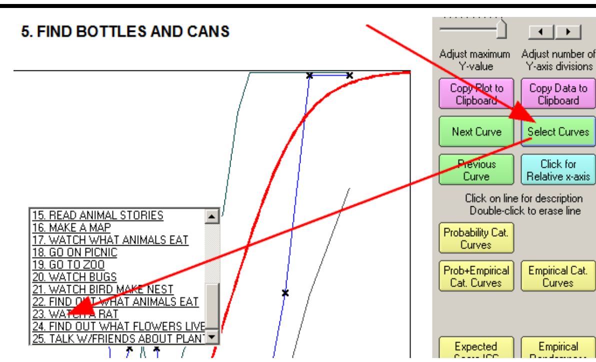
68. For Item 5, the red line is the same again. And the confidence intervals for the model-predicted dispersion of the observations around their expectations are also shown. We can see that observations are now outside the confidence intervals ... somewhat surprising if the data fit the model.

But look at the blue empirical ICC! It is composed of two parts. I've marked them in **orange.** The right-hand upward arrow is what we would expect for an item about 2 logits more difficult than this one. The left-hand downward arrow tells an opposite story. This item is two items: one item for children who like science activities, and an opposite item for children who don't.



69. Let's look at another suspect item, Item 23.

Click on "Select Curves"
Scroll down inside the item list box
Click on 23. "Watch a Rat"

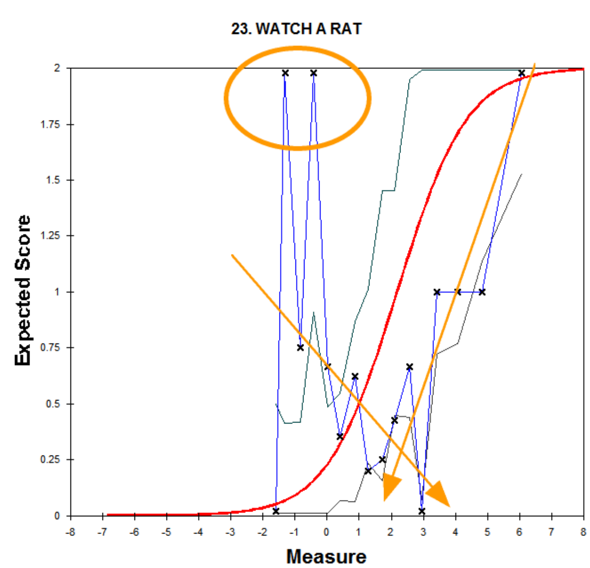


70. We see the same pattern we saw with Item 5, but more exaggerated. Some children whom we predict to dislike this item (solid red curve) instead like it (orange circle).

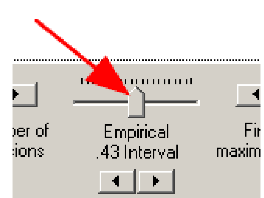
So now we have two items telling us a different story. We have the basis for concluding there is a second dimension in this test.

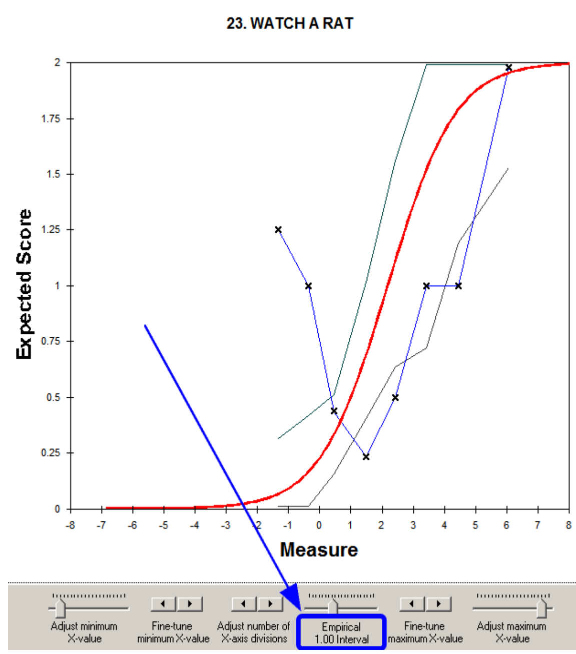
The general rule is: "**All items must be about the same thing,** our intended latent variable, **but then be as different as possible**, so that they tell us different things about the latent variable."

But when two or more items tell us the **same "different thing"**, then we have indications of a **secondary dimension.**



71. The display of the Empirical ICCs is sensitive to the width of the interval corresponding to each "x" on the latent variable (x-axis).
Move the width slider left-and-right to see what happens to the blue empirical ICC. Do this by left-clicking your mouse and dragging the slide's pointer.



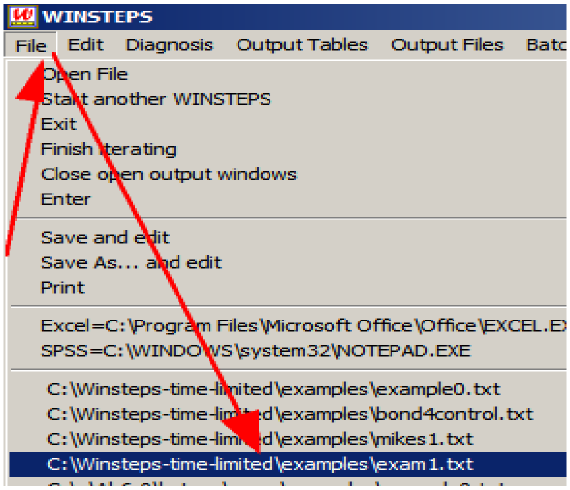
72.	For a final report, you would choose a reasonable slider setting which conveys the message forcefully. Here I've chosen 1 logit. We can see clearly the U-shape of the empirical ICC. ***What do you think shows the pattern of responses best?*** Play with the other sliders and settings. Click the other buttons. *Wow! This is almost a video game!* So we now have strong evidence for deleting these two items from the analysis. But we will keep them for the present .... Pictures are great, but investigating them can be time-consuming, and we may not know what to look for. So Tables of number can be helpful.	
73.	We'll back-track to a dichotomous analysis, so close all Winsteps windows.	
74.		

75. **F. Dichotomous Rasch Fit Statistics**

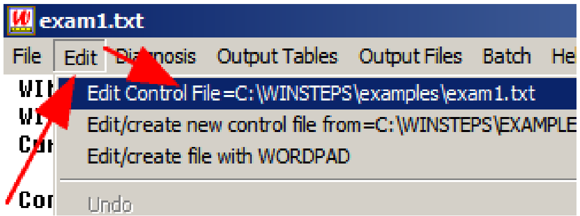
76. Let's take another look at the Knox Cube Test.
Launch Winsteps

77. When Winsteps displays,
Click on "File"
Click on ".... exam1.txt" which you should see on the "most recently used" list.

If it is not there, then
Click on "Open File" and
Open "exam1.txt"



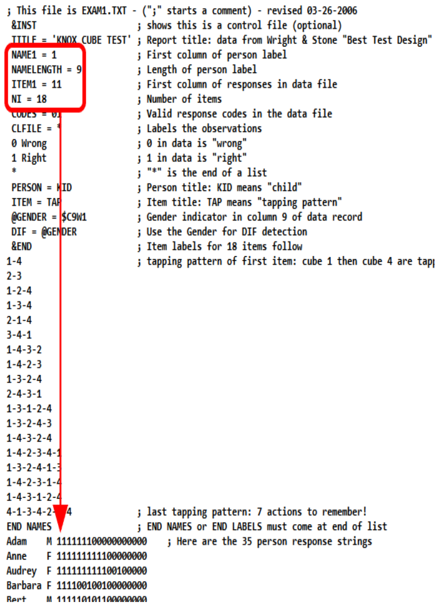
78. We want to examine the control file before we perform the analysis, so
Click on "Edit"
Click on "Edit Control File"



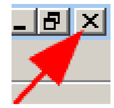
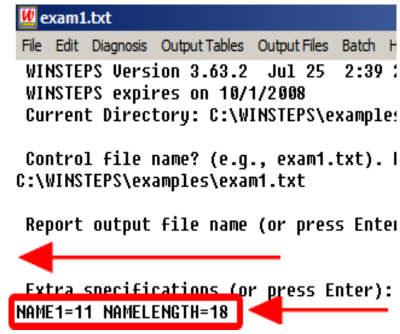
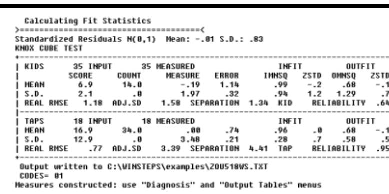
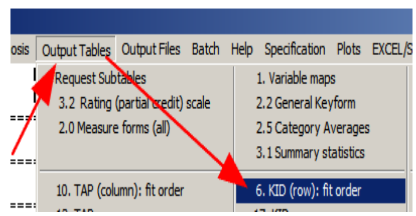
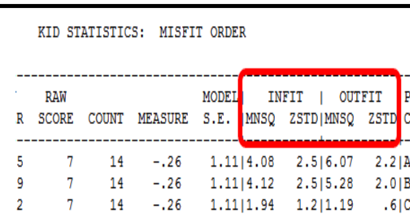
79. The Winsteps Control file for the KCT data displays. It should be somewhat familiar to you.

What we are going to do is to make the responses into the person labels! This is so that we can see them on the person Tables.
ITEM1=11 The responses start in column 11
NI=18 There are 18 responses

So, if we want these to be the person label, we need:
NAME1=11 The person label starts at column 11
NAMELENGTH=18 The person label is 18 columns wide.
But don't change anything now!
You've noticed that NAMELENGTH=, NAMELEN=, NAMEL= are all the same control variable. Winsteps only looks at the first few letters. Enough to make the variable name unique.

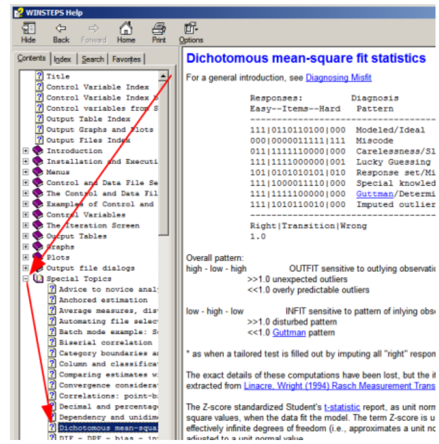


```
; This file is EXAM1.TXT - (";" starts a comment) - revised 03-26-2006
 &INST                       ; shows this is a control file (optional)
 TITLE = 'KNOX CUBE TEST' ; Report title: data from Wright & Stone "Best Test Design"
 NAME1 = 1                   ; First column of person label
 NAMELENGTH = 9              ; Length of person label
 ITEM1 = 11                  ; First column of responses in data file
 NI = 18                     ; Number of items
 CODES = 01                  ; Valid response codes in the data file
 CLFILE = *                  ; Labels the observations
 0 Wrong                     ; 0 in data is "wrong"
 1 Right                     ; 1 in data is "right"
 *                           ; "*" is the end of a list
 PERSON = KID                ; Person title: KID means "child"
 ITEM = TAP                  ; Item title: TAP means "tapping pattern"
 @GENDER = $C9W1             ; Gender indicator in column 9 of data record
 DIF = @GENDER               ; Use the Gender for DIF detection
 &END                        ; Item labels for 18 items follow
1-4                          ; tapping pattern of first item: cube 1 then cube 4 are tap
2-3
1-2-4
1-3-4
2-1-4
3-4-1
1-4-3-2
1-4-2-3
1-3-2-4
2-4-3-1
1-3-1-2-4
1-3-2-4-3
1-4-3-2-4
1-4-2-3-4-1
1-3-2-4-1-3
1-4-2-3-1-4
1-4-3-1-2-4
4-1-3-4-2-1-4                ; last tapping pattern: 7 actions to remember!
END NAMES                    ; END NAMES or END LABELS must come at end of list
Adam    M 111111100000000000    ; Here are the 35 person response strings
Anne    F 111111111100000000
Audrey  F 111111111100100000
Barbara F 111100100100000000
Bert    M 111110101100000000
```

80.	Close the Edit window	
81.	In the Winsteps Analysis window, Report Output? Press Enter Extra Specifications? NAME1=11 NAMELENGTH=18 *No spaces within instructions, but a space between.* This enables us to enter more control variables without changing the control file. Useful for once-only changes. Press Enter.	
82.	The Analysis is performed	
83.	Now let's look at the Fit statistics - how well the data match the Rasch model's expectations. Click on Output Tables Click on 6. KID (row): fit order	
84.	Table 6 is displayed in a NotePad window. We are interested in the INFIT and OUTFIT statistics.	

85. G. Exploring INFIT and OUTFIT Statistics

86. Let’s look at some patterns of misfit we would want to identify and diagnose.
To see them:
On the Winsteps Menu Bar
Click on Help
Click on Contents
In the Contents panel,
Click on Special Topics
Click on Dichotomous Mean-Square Fit Statistics



87. Here they are:
In this Table, we imagine that the items have been arranged from easy to hard (as they are on the Knox Cube Test) and have been administered in ascending order of difficulty as a multiple-choice (MCQ) test with a time limit. A type of test familiar to all school children in the USA. The items are scored “1” and “0”

```
Responses:          Diagnosis               INFIT   OUTFIT
Easy--Items--Hard    Pattern                 MnSq    MnSq
-----------------------------------------------------------
111|0110110100|000  Modeled/Ideal            1.1     1.0
000|0000011111|111  Miscode                  4.3    12.6
011|1111110000|000  Carelessness/Sleeping    1.0     3.8
111|1111000000|001  Lucky Guessing           1.0     3.8
101|0101010101|010  Response set/Miskey      2.3     4.0
111|1000011110|000  Special knowledge        1.3     0.9
111|1111100000|000  Guttman/Deterministic    0.5     0.3
111|1010110010|000  Imputed outliers *       1.0     0.6
-----------------------------------------------------------
Right|Transition|Wrong                   Expectation: 1.0
```

88. How do we expect a child of medium ability to respond? We expect the child to get the easy items almost always correct (green box) and the hard items almost always incorrect (red box). In between, is a “transition” zone where the item difficulties are targeted on the child’s ability. Here we expect the child to succeed on some items and fail on others (blue box). If an item’s difficulty exactly corresponds to the child’s ability, then the child’s probability of success is 0.5, and we expect success or failure (1 or 0) equally.
This is the response pattern predicted by the Rasch model. We can see that this response pattern produces INFIT and OUTFIT mean-square (MnSq) statistics near 1.0.

```
Responses:          Diagnosis        INFIT  OUTFIT
Easy--Items--Hard    Pattern          MnSq   MnSq
---------------------------------------------------
111 0110110100 000  Modeled/Ideal     1.1    1.0
```

Transition points for dichotomies are at approximately p=0.25, 0.75, which is ±1.1 logits difference between the item difficulty and the person ability

89. What about guessing - a common problem on MCQ items? The only guessing that is of great concern is when the guess is lucky - a correct answer to a hard item (red circle). This is an unexpectedly correct response - an outlier. The OUTFIT statistic is sensitive to outliers. Its value is now 3.8, much bigger than its baseline value of 1.0. INFIT statistics are relatively insensitive to outliers. Its value is the baseline 1.0.

```
111|1111000000|001  Lucky Guessing     1.0    3.8
```

90.	And careless mistakes? These are incorrect answers to easy items (red circle). Again this is an unexpected response - an outlier. So the OUTFIT statistic is again high, 3.8, but the INFIT statistic is relatively unchanged at its baseline value of 1.0. Values of fit statistics greater than 1.0 are termed "underfit" - the responses are too unpredictable from the Rasch model's perspective.	011|1111110000|000 Carelessness/Sleeping 1.0 3.8
91.	Let's think about a different behavior: **the plodder.** He works slowly and carefully through each item, double-checking his answers. He succeeds on every item (green box). But then time runs out. He is automatically scored incorrect (red box) on all the remaining harder items. If we know the cut-point (blue arrow) we can predict all the child's responses exactly. Psychometrician Louis Guttman proclaimed that this is the ideal response pattern. The child's responses seem to tell us that his ability is exactly at the blue arrow. *But, where is his "transition zone" predicted by the Rasch model?* What we do see is a response pattern that is too predictable. There is no area of uncertainty in it. Accordingly both the INFIT mean-square of 0.5 and the OUTFIT of 0.3 are less than 1.0. This is termed "overfit". The responses are too predictable from the Rasch-model perspective.	111|1111100000|000 Guttman/Deterministic 0.5 0.3 "Plodder" ? Is the distance between the red box and the green box near or far? If all the data are "Guttman-predictable", so that they look like this, then data don't tell us. If the entire data set has a Guttman pattern, then we can exactly order all the persons and items on the latent variable, but we have no information to estimate how close together they are.
92.	Let's imagine this situation: most schools teach addition ➔ subtraction ➔ multiplication ➔ division, but my school teaches addition ➔ multiplication ➔ subtraction ➔ division. So when I take a standard arithmetic test, I succeed on the addition items. Fail on the subtraction items (red box). Succeed on the multiplication items (green box) and fail on the division items. Compare this response string to the others. We are not surprised by a failure or two on the subtraction items, or by a success or two on the multiplication items. It is the overall pattern that is surprising. This is what INFIT identifies. So the INFIT mean-square is 1.3, greater than 1.0, indicating underfit, "too much unpredictability". But the OUTFIT mean-square is 0.9, less than 1.0, indicating overfit, my performance on the easy "addition" items and hard "division" items is slightly too predictable.	111|0000|111|000 Special knowledge 1.3 0.9 "Alternative curriculum"

93. So what values of the mean-square statistics cause us real concern? Here is my summary table from Winsteps Help “Special Topic” “Misfit Diagnosis ...”

Here’s a story:
When the mean-square value is around 1.0, we are hearing music! The measurement is **accurate**
When the mean-square value is less than 1.0, the music is becoming quieter, becoming muted. When the mean-square is less than 0.5, the item is providing only have the music volume (technically “statistical information”) that it should. But mutedness does not cause any real problems. Muted items aren’t efficient. The measurement is less accurate.

When the mean-squares go above 1.0, the music level stays constant, but now there is other noise: rumbles, clunks, pings, etc. When the mean-square gets above 2.0, then the noise is louder than the music and starting to drown it out. The measures (though still forced to be linear) are becoming distorted relative to the response strings. So **it is mean-square values greater than 2.0 that are of greatest concern. The measurement is inaccurate.**

But be alert, the ***explosion*** *caused by only one very lucky guess can send a mean-square statistic above 2.0. Eliminate the lucky guess from the data set, and harmony will reign!*

Interpretation of parameter-level mean-square fit statistics:	
>2.0	Distorts or degrades the measurement system.
1.5 - 2.0	Unproductive for construction of measurement, but not degrading.
0.5 - 1.5	Productive for measurement.
<0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations.

There are other rules at
Reasonable Mean-Square Fit Values
http://www.rasch.org/rmt/rmt83b.htm

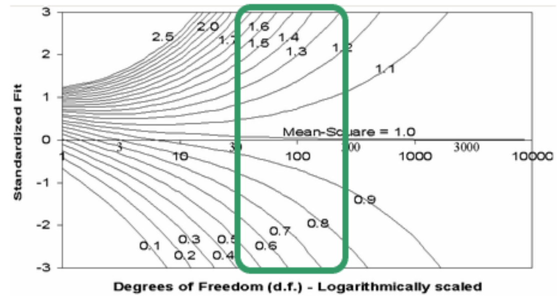
Reasonable Item Mean-square Ranges for INFIT and OUTFIT	
Type of Test	Range
MCQ (High stakes)	0.8 - 1.2
MCQ (Run of the mill)	0.7 - 1.3
Rating scale (survey)	0.6 - 1.4
Clinical observation	0.5 - 1.7
Judged (agreement encouraged)	0.4 - 1.2

94. **Please answer Assignment 2, Question 5**

95.	**H. Computing INFIT and OUTFIT “MnSq” Mean-Square Statistics** Fortunately, computers do the tedious computations for us, but we do need some understanding of what the computers are doing	
96.	Let’s start with the Rasch dichotomous model we met in Lesson 1. The **Rasch dichotomous model** specifies the probability, P, that person *n* of ability B_n succeeds on item *i* of difficulty D_i	$\log_e(P_{ni}/(1-P_{ni})) = B_n - D_i$
97.	Then the average expected response is E_{ni}	For dichotomous data, $E_{ni} = P_{ni}$
98.	And the variance of the observed responses around their expectation is W_{ni}. Its computation is shown in the adjacent box. For each **observation, X_{ni},** the Rasch model provides an **expectation, E_{ni}**, and the **model variance**, W_{ni}, of the observation around its expectation. INFIT and OUTFIT are combinations of X_{ni}, E_{ni} and W_{ni}.	
99.	We imagine a dichotomous situation in which the probability of scoring 1 is Pni, so that the probability of scoring 0 is (1-Pni). The expected value of the response, its expectation, is **Eni** = 1 * Pni + 0 * (1-Pni) = Pni. We partition the variance of the observations around their expectation, Eni. The part due to scoring 1, V(1) = (probability of scoring 1) * (distance of 1 from the expected value)^2 = Pni * (1-Eni)^2 Similarly, the part due to scoring 0, V(0) = (probability of scoring 0) * (distance of 0 from the expected value)^2 = (1-Pni) * (0-Eni)^2 So the total "model" variance of a dichotomous observation around its expectation = **Wni** = V(1) + V(0) = Pni * (1-Eni)^2 + (1-Pni) * Eni^2 = Pni * (1-Pni)	
100.	The first combination is the **Residual, R_{ni}** = Observation - its Expectation. There is a Residual for each observation.	$R_{ni} = X_{ni} - E_{ni}$
101.	The model variance of the observation around its expectation is W_{ni}, so its square-root is the standard deviation of the model “observation distribution”. This leads to the **Standardized Residual, Z_{ni}**, which roughly quantifies the unexpectedness of the observation as a “unit normal deviate”. If this term is new to you, please read Appendix 1, *“Unit Normal Deviates”*.	$Z_{ni} = R_{ni} / \sqrt{(W_{ni})}$

102	The OUTFIT mean-square statistic is the average of the squared standardized-residual for the responses by a person, U_n, or on an item U_i. It is called "mean-square" because it is the average value of the squared values. It is also the equivalent chi-square statistic divided by its degrees of freedom. If this is new to you, please study Appendix 2, *"Chi-square, mean-square and degrees of freedom"*. OUTFIT is a conventional Pearson chi-square fit statistic divided by its degrees of freedom. So, when choosing whether to report OUTFIT or INFIT, report OUTFIT. It will be more familiar to most statisticians. *OUTFIT means "Outlier-sensitive fit statistic".*	$U_n = \sum_{i=1}^{L} Z_{ni}^2 \Big/ L \quad U_i = \sum_{n=1}^{N} Z_{ni}^2 \Big/ N$
103	The INFIT mean-square is the information-weighted average of the squared residuals. INFIT means "Inlier-pattern-sensitive fit statistic", or more technically, "Information-weighted fit statistic".	$U_n = \frac{\sum_{i=1}^{L} Z_{ni}^2 W_{ni}^2}{\sum_{i=1}^{L} W_{ni}^2} \quad U_i = \frac{\sum_{n=1}^{N} Z_{ni}^2 W_{ni}^2}{\sum_{N=1}^{N} W_{ni}^2}$
104	**Computing INFIT and OUTFIT "ZSTD" Fit Statistics**	
105	Mean-square statistics indicate the size of the misfit, but statisticians are usually more concerned with the improbability of the misfit, its "significance". So corresponding to each mean-square there is a ZSTD statistic showing the probability of the mean-square as a unit-normal deviate (again, see Appendix 1 if you don't know about these). The ZSTD is the probability associated with the null hypothesis: "These data fit the Rasch model". In conventional statistics, when p<.05, i.e., ZSTD is more extreme than ±1.96, then there is "statistical significance", and the null hypothesis is rejected.	Wilson-Hilferty transformation: $q^2 = 2/\text{d.f.}$, where d.f. ≈ MnSq divisor $\text{ZSTD} = (\text{MnSq}^{1/3} - 1)(3/q) + (q/3)$ *Computers do this computation for us!*

106	ZSTD means "Standardized like a Z-score", i.e., as a unit-normal deviate. So we are looking for values of 2 or more to indicate statistically significant model misfit.	`------------------------`
		`|   INFIT   |  OUTFIT  |`
		`|MNSQ  ZSTD |MNSQ  ZSTD|`
		`-+----------+----------+`
		`.|4.08   2.5|6.07   2.2|.`

<table>
<tr><td>107</td><td>The relationship between significance (ZSTD) and size (MnSq) is controlled by the degrees of freedom (d.f.). See the plot in Winsteps Help “Misfit Diagnosis ..” or http://www.winsteps.com/winman/diagnosingmisfit.htm
We can see that if the d.f. (x-axis) are too small (less than 30) even huge misfit is statistically insignificant, but if the d.f. are too large (greater than 300), then substantively trivial misfit is statistically significant. Notice that mean-squares greater than 1, noisy underfit, are reported with positive ZSTD, but mean-squares less than 1, muted overfit, are reported with negative ZSTD.</td><td>


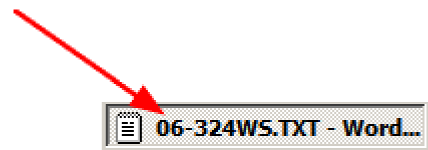
</td></tr>
<tr><td>108</td><td>When sample sizes become huge, then all misfit becomes statistically significant (red boxes). Here the sample sizes are in the thousands. Even the substantively trivial mean-square of 1.12 is reported as statistically significant.</td><td>
<table>
<tr><th>ENTRY NUMBER</th><th>RAW SCORE</th><th>COUNT</th><th>MEASURE</th><th>MODEL S.E.</th><th>INFIT MNSQ</th><th>INFIT ZSTD</th><th>OUTFIT MNSQ</th><th>OUTFIT ZSTD</th></tr>
<tr><td>1</td><td>4000</td><td>14000</td><td>-3.06</td><td>.03</td><td>.61</td><td>-9.9</td><td>.29</td><td>-9.9</td></tr>
<tr><td>2</td><td>7000</td><td>14000</td><td>-.27</td><td>.04</td><td>.18</td><td>-9.9</td><td>.08</td><td>-9.9</td></tr>
<tr><td>3</td><td>8000</td><td>14000</td><td>.98</td><td>.03</td><td>1.03</td><td>1.3</td><td>.32</td><td>-9.9</td></tr>
<tr><td>4</td><td>3000</td><td>14000</td><td>-3.73</td><td>.03</td><td>1.23</td><td>9.9</td><td>.90</td><td>-1.6</td></tr>
<tr><td>5</td><td>5000</td><td>14000</td><td>-2.34</td><td>.03</td><td>1.12</td><td>7.9</td><td>.51</td><td>-9.9</td></tr>
</table>
</td></tr>
<tr><td>109</td><td></td><td></td></tr>
</table>

110	**I. Investigating Fit Statistics**

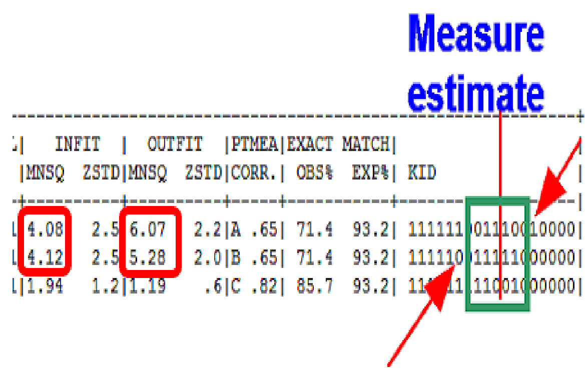
111 *The general rules are:*

1. Unexpected outlying observations (OUTFIT) before unexpected inlying patterns of observations (INFIT).
2. Size before significance: Mean-squares before ZSTD
3. Underfit (noise) before Overfit (muted): high mean-squares before low mean-squares, positive ZSTD before negative ZSTD.
4. Mean-squares are forced to average near one, so that high mean-squares force low mean-squares.
5. Start from the worst item or person and them work in towards the model-fitting ones. Stop when you lose interest because there is nothing remarkable about the item or person.
6. After eliminating the "worst" item or person, there is always another "worst" item or person who may look yet worse in the new, more model-fitting context. So don't eliminate mechanically or there will be no items or persons left!
7. If in doubt, compare the person (or item) measures with and without the doubtful items (or persons). If there is no noticeable difference, then the misfit doesn't matter. We'll see how to do this later.

112 Let's look at some examples.
Click on Table 6 for the analysis of exam1.txt on your Windows Task Bar, or
Click on Output Tables
Click on 6. KID (row): fit order



113 Table 6 displays. Here are the first 3 lines, the worst-fitting children. Notice the large mean-squares, and that only obviously problematically large mean-squares are indicated as significant.
Remember that we entered at Extra Specifications?
NAME1=11 NAMELENGTH=18
so that the response strings shows as the person label.
The green box indicates the approximate dividing line between INFIT (inside) and OUTFIT (outside).
The red arrows indicate the unexpected observations causing the large OUTFIT statistics.
The problems in the data that cause large OUTFIT are usually easy to identify, diagnose and remedy (if desired).



114 The large INFIT statistics are due to the unexpected response patterns in the green box. Do these look unusually strange to you? *Probably not.* This is typical. **The problems causing large INFIT statistics are usually very difficult (or impossible) to identify and diagnose, and almost always impossible to remedy.**
But INFIT is a greater threat to the substantive validity of the measures than OUTFIT. This is because INFIT reports misfit in the region where the item is supposed to be most useful for measurement, or the region in which the person's ability estimate lies. We saw this with the "plodder" in our earlier example at #

115 Scroll down to the bottom of Table 6.1.

There are conspicuously small mean-squares. The smallest possible values are 0.00 and these are close. But don't panic! Look at the mean (average) mean-squares, they are not far from 1.00:

```
.99   -.2| .68
```

The huge mean-squares at the top of this Table have forced the small mean-squares at the bottom of it.

```
 -------------------------
 |   INFIT   |  OUTFIT   |
 |MNSQ  ZSTD|MNSQ  ZSTD|
 +----------+----------+

 | .18  -1.3| .08   -.7|
 | .18  -1.3| .08   -.7|
 | .18  -1.3| .08   -.7|
 | .18  -1.3| .08   -.7|
 | .18  -1.3| .08   -.7|
 | .18  -1.3| .08   -.7|
 | .18  -1.3| .08   -.7|
 +----------+----------+
```

116 We don't usually have the responses conveniently in the person (or item label), so we need to look elsewhere to identify the unexpected responses.
Scroll down in Table 6 to Table 6.5, the "Most Unexpected Responses".
This Sub-Table shows the responses that could trigger large OUTFIT mean-squares. In the red box, the rows are children (persons), ordered by ability measure from top down, and the columns are items, ordered by item difficulty with the easiest item at the left.
The "."s mean that the observation was not surprising. 0 or 1 means that this response was surprising, and the scored response is shown.
The most misfitting child (first in Table 6.1) is child 25. We can see that the large OUTFIT is caused by two unexpectedly incorrect answers (to "easier to remember" items 7 and 8) and one unexpected correct answer (to "harder to remember" item 14 - its number is printed vertically).

```
TABLE 6.5 KNOX CUBE TEST
INPUT: 35 KIDS  18 TAPS  MEASURED: 35 KIDS  18
--------------------------------------------

MOST UNEXPECTED RESPONSES
KID                          MEASURE  |TAP
                                      |     111
                                      |47568024
                                  high
   32 111111111110100110    3.73 D|......0.
   33 111111111101010000    1.94 E|.......1
   35 111111111100110000    1.94 F|.......1
   12 111111111001000000    -.26 C|......1.
   19 111110011111000000    -.26 B|.0.0..1.
   25 111111001110010000    -.26 A|.0..0..1
   14 111101111110000000   -1.37 J|..0.....
    9 111011111000000000   -2.23 M|0.......
    4 111100100100000000   -3.61 I|.....1..
                                      |----low-
                                      |47568111
                                      |     024
```

Misfitting observations are shown as 0 and 1.
Other observations are shown as "."

117 Scroll down a little further. There is more information about the unexpected responses in Table 6.6.
We can see here that the most surprising response was the correct answer by child 25 to item 14. Its "standardized residual" (Z_{ni}) is 6.14. This is so extreme, as a unit-normal deviate, that it is not even in shown in my copy of *"CRC Standard Mathematical Tables"!*
Do we believe this observation? Was it a clerical error? Did the child use a trick to do it correctly (like remembering a tune with the same tapping rhythm)?

We could change this observation to missing data.
Unethical? **No. Our purpose is to measure the child meaningfully, not to get the child a high score.** We are in charge of the data; the data are not in charge of us.

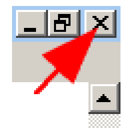
```
TABLE 6.6 KNOX CUBE TEST                              ZOU324WS.TXT Jul 25
INPUT: 35 KIDS  18 TAPS  MEASURED: 35 KIDS  18 TAPS  2 CATS          WIN
--------------------------------------------------------------------------

MOST UNEXPECTED RESPONSES

+-------------------------------------------------------------------------
| DATA |OBSERVED|EXPECTED|RESIDUAL|ST. RES.|MEASDIFF| TAP    | KID     |
|------+--------+--------+--------+--------+--------+--------+---------+
|    1 |      1 |    .03 |    .97 |   6.14 |  -3.63 |     14 |      25
|    0 |      0 |    .97 |   -.97 |  -5.97 |   3.57 |      7 |      25 |
|    0 |      0 |    .97 |   -.97 |  -5.97 |   3.57 |      7 |      19 |
```

118 Take a look at the equivalent Table for the items, Table 10. In what ways is the story in Table 10 different from the story in Table 6? Does it lead us to the same or different conclusions?

```
---------------------------------------------------------------------
|   INFIT    |  OUTFIT    |PT-MEASURE |EXACT MATCH|
|MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| TAP
+----------+----------+-----------+-----------+-------------
|1.33    .9|2.21   1.1|A .40   .54| 94.1  91.7| 1-4-3-2
|1.56   1.2|1.49    .8|B .22   .38| 85.3  92.0| 1-4-2-3-4-1
|1.17    .6| .96    .4|C .53   .57| 91.2  90.0| 3-4-1
|1.16    .6|1.06    .4|D .42   .48| 85.3  86.7| 1-3-2-4-3
|1.07    .4| .79   -.1|E .55   .56| 76.5  79.1| 1-3-1-2-4
|1.06    .3| .83    .0|F .61   .62| 79.4  83.0| 2-4-3-1
|1.04    .2| .52    .1|G .55   .54| 88.2  91.7| 2-1-4
| .90    .0| .35   -.2|g .55   .51| 94.1  94.0| 1-3-4
| .74   -.1| .11   -.6|f .32   .24| 97.1  97.0| 1-3-2-4-1-3
| .74   -.1| .11   -.6|e .32   .24| 97.1  97.0| 1-4-2-3-1-4
| .74   -.1| .11   -.6|d .32   .24| 97.1  97.0| 1-4-3-1-2-4
| .70  -1.0| .38   -.4|c .60   .50| 88.2  84.6| 1-4-3-2-4
| .62  -1.0| .21   -.6|b .68   .57| 91.2  90.0| 1-3-2-4
| .59  -1.3| .43   -.4|a .72   .61| 94.1  86.5| 1-4-2-3
```

119 Have you discovered that **a misfitting observation causes both the item and the person to misfit?** So when we discover misfit in our data, we must investigate: Is the problem due to the person (guessing, carelessness, ...) or due to the item (poor wording, miskey, off-dimension, ...)

120 Let’s go a little deeper ... Close all windows.



121 **J. Polytomous Fit Statistics and Scalograms**

122 Our polytomous fit statistics, quantifying how well the data fit the model, are OUTFIT and INFIT again, the same as for dichotomies, but now they are more challenging to diagnose.
To start with, there are many more possibilities ....

Here is the first part of the diagnostic table in "Polytomous Mean-Square Fit Statistics", a "Special Topic" in Winsteps Help, also at http://www.winsteps.com/winman/polytomous.htm

Take a look at the full table. "RPM" is our friend, the "point-measure correlation" we saw in *Diagnosis A. Item Polarity.*

Polytomous mean-square fit statistics

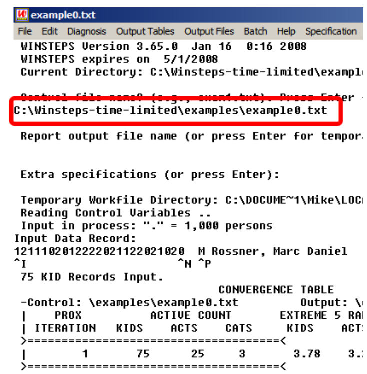
Response String Easy....................Hard	INFIT MnSq	OUTFIT MnSq	RPM (PTMEA) Corr.	Diagnosis
I. modeled:				
33333132210000001011	.98	.99	.78	*Stochastically*
31332332321220000000	.98	1.04	.81	*monotonic in form,*
33333331122300000000	1.06	.97	.87	*strictly monotonic*
33333331110010200001	1.03	1.00	.81	*in meaning*
II. overfitting (muted):				
33222222221111111100	.18	.22	.92	Guttman pattern
33333222221111100000	.31	.35	.97	high discrimination
32222222221111111110	.21	.26	.89	low discrimination
32323232121212101010	.52	.54	.82	tight progression
III. limited categories:				
33333333332222222222	.24	.24	.87	high (low) categories
22222222221111111111	.24	.34	.87	central categories
33333322222222211111	.16	.20	.93	only 3 categories

123 There is too much to remember here, so let's look at this for the "Liking for Science" data. So

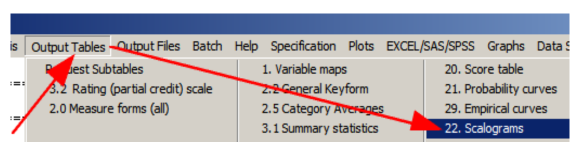
Launch Winsteps

Run the analysis for "example0.txt" - you know how to do this!

This uses a 3-category: 0, 1, 2 rating scale, analyzed with the Andrich Rating Scale model.



124 When the dataset is small, it is often useful to look at the data in "Scalogram" format, a layout popularized by psychometrician Louis Guttman around 1950.
Winsteps Menu Bar
Click on Output Tables
Click on 22. Scalogram



125 Table 22: A scalogram orders the persons from high measure to low measure as rows, and the items from low measure (easy) to high measure (hard) as columns. Here it is:

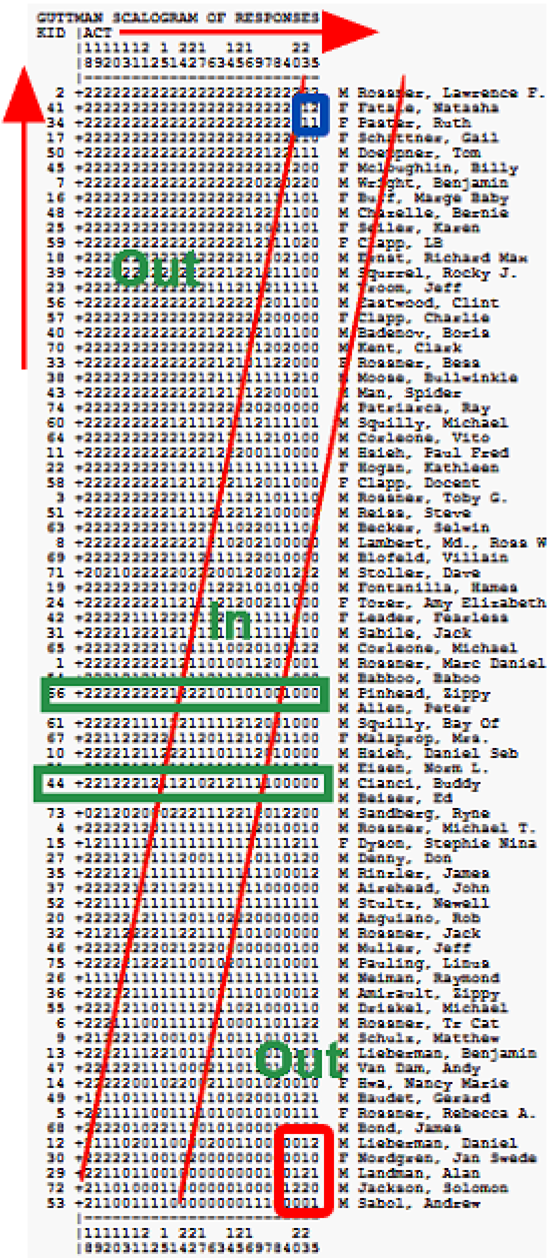
Top left corner:
where the "more able" (more liking) children respond to the "easier" (to like) items. So we expect to see responses of "Like" (2). *We do!*

Top right corner:
(blue box) where the "most liking" children and the "hardest to like items" meet - you can see some ratings of 1.

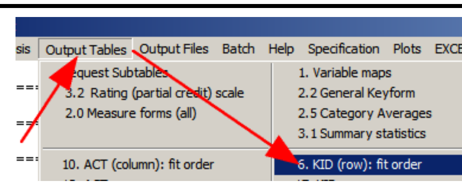
Bottom right corner:
where "less able" (less liking) children respond to the "harder" (to like) items. So we expect to see responses of "Dislike" (0). ***But do we??***
Something has gone wrong! There are 1's and 2's where we expected all 0's. The fit statistics should tell us about this!

Transition zone
To the left of the left red diagonal we expect 2's. In this zone *Outfit* is more sensitive to unexpected responses.
To the right of the right red diagonal we expect 0's. In this zone. *Outfit* is more sensitive to unexpected responses.
Between the red diagonals lies the *transition zone* where we expect 1's. In this zone *Infit* is more sensitive to unexpected patterns of responses.

More categories in the rating scale means a wider transition zone. Then the transition zone can be wider than the observed responses. Infit and Outfit will report the same fit values, so you only need to report Outfit.



126 Let's look at the children in fit order.
Winsteps Menu Bar
Click on Output Tables
Click on 6. KID (row): fit order



127 Table 6 displays. Let's diagnose some of the problems. Compare between Table 6 and Table 22.

First off, person 72.

Large Outfit mean-square (5.16), smaller Infit mean-square (2.02). With polytomous response the distinction between Outfit and Infit is much less than for dichotomies. With long rating scales or narrow samples the distinction can disappear. So, when Infit and Outfit are almost the same, only report Outfit, because that is a conventional chi-square statistic divided by its d.f.

So our diagnosis here is "problem with outlying observations" - and that is what we have here! Unexpectedly high ratings to difficult items by less "liking" children.

```
TABLE 6.1 LIKING FOR SCIENCE (Wright & Masters p. ZOU108WS.T
INPUT: 75 KIDS  25 ACTS  MEASURED: 75 KIDS  25 ACTS  3 CATS
--------------------------------------------------------------
KID: REAL SEP.: 2.67  REL.: .88 ... ACT: REAL SEP.: 5.32  RE

       KID STATISTICS:  MISFIT ORDER

+-------------------------------------------------------------
|ENTRY   RAW                       MODEL|   INFIT  |  OUTFIT  |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|
|------------------------------------+----------+----------+
|    72     14     25   -1.32     .37|2.02   2.9|5.16   5.7|

 29 +2211011001000000000100121  M LANDMAN, ALAN
 72 +2110100011000000100011220  M JACKSON, SOLOMON
 53 +2110011110000000011100001  M SABOL, ANDREW
    |-------------------------
    |1111112 1 221  121    22
    |8920311251427634569784035
```

128 What about child 73?

Same again, but this time the problem focuses more on an outlier at the "easy" end of the response string.

Child 73 responds "0" (dislike), but we expect "2" (like). This item is difficult for this child.

```
+-------------------------------------------------------------
|ENTRY   RAW                        MODEL|   INFIT  |  OUT
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ
|------------------------------------+----------+-----
|    72     14     25   -1.32     .37|2.02   2.9|5.16
|    71     33     25     .97     .35|3.15   5.4|4.95
|    73     28     25     .38     .34|2.88   5.0|4.84

 62 +2222222222002022002000000
 73 +0120200022211122212012200
  4 +2222120111111111112010010
```

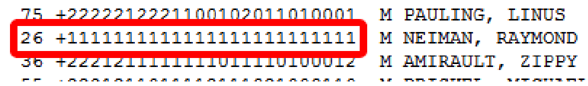
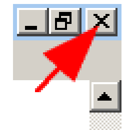
129 For child 7, the Infit is bigger than the Outfit.

We can see that the pattern is extreme 0's and 2's where we expect to see 1's, i.e., where the "liking difficulty" of the items is targeted on the "likeability" of the items.

```
|     6     24     25    -.08     .34|1.62   2.1|2.39   3.5|
|     7     44     25    2.71     .48|1.84   1.9|1.10    .4|
|     9     24     25    -.08     .34|1.41   1.5|1.83   2.4|

 45 +2222222222222222222221200  F MCLOUGHLIN, BILLY
  7 +2222222222222222220220220  M WRIGHT, BENJAMIN
 16 +2222222222222222222111101  F BUFF, MARGE BABY
```

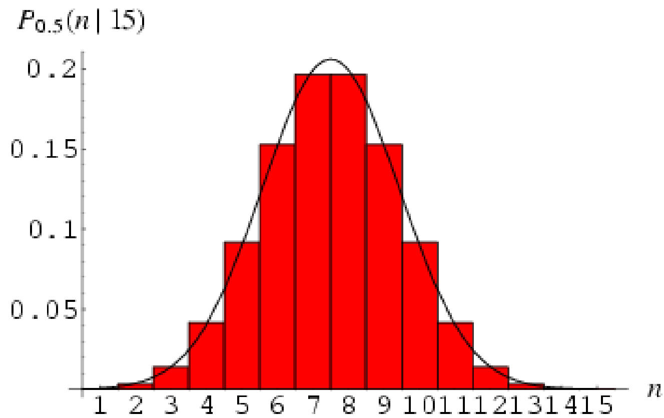
130 Let's look at the other extreme of fit: Overfit - at the bottom of Table 6.1.
Here is child 21 with Outfit and Infit well below 1.0.
Notice that child 21 has a very predictable pattern going from 2 ➔ 1 ➔ 0, almost exactly in accord with what the Rasch-Andrich model predicts. Obviously this isn't bad. In fact, if we needed a child whose responses best summarize those of the other children, child 21 would be the one! But this also means that child 21 gives us the least new information about the relative likeability of the items. We can see that because all child 21's responses of "1" blur those items together as "neutral".

```
|   21    28    25    .38    .34|.29  -3.9|.31

 10 +2222121122211101112010000
 21 +2222121111111111111111000
 44 +2212221211210212111100000
```

131	*Now it's your turn.* Look at the response strings in Table 22, and the fit statistics in Table 6. Do you see anything interesting? *Fit statistic interpretation is a central aspect of Rasch analysis.*	How about Child 26? What do you think about him? 
132	Just enough time left for one short topic. Close all windows.	
133		
134	**Supplemental Reading**	
135	B&F chapter 6 focuses on our work here.	
136	"Rating Scale Analysis", (Wright & Masters) chapter 2	
137	"Best Test Design", (Wright & Stone) chapter 4	

138 **Appendix 1. Unit Normal Deviates**

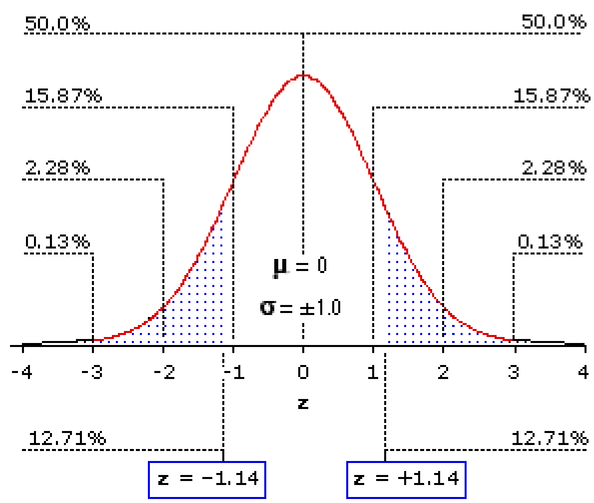
139 The “normal” distribution is fundamental to statistics. It describes what happens when events happen “normally”, purely by chance. The Figure shows the probability of different numbers of “heads” when a coin is tossed 15 items in the red bars:
http://mathworld.wolfram.com/NormalDistribution.html
We can see that the overall pattern follows a bell-shaped curve the continuous black line. This pattern gets closer to a smooth line, the more coins we toss. The black continuous line for an infinite number of tosses is the “normal distribution”.



140 We are interested in a special case of the normal distribution. We want the one when its mean is zero, and its standard deviation is 1.0. This is called the “unit normal distribution”, abbreviated N(0,1). Statisticians use the Greek letter mu, μ, for the mean or average, and the Greek letter sigma, σ, for the standard deviation or spread, so the general normal distribution is $N(\mu, \sigma^2)$.

Look at the plot, the x-axis is labeled “z”. “z” means that these values are “z-scores” also called “unit normal deviates”. They are possible values of the unit normal distribution. The y-axis indicates the probability of observing the z values. Looking at the red curve, values of z near 0 have high probability. Values of z outside ±3 have very low probability.

The area under the red curve indicates the cumulative probability of observing z values. 68% of the area under the red curve is within ±1, i.e., within 1 S.D. of the mean of the unit normal distribution. So we expect about 2/3 of the values we observe by chance to be statistically close to the mean.



http://faculty.vassar.edu/lowry/ch6pt1.html

141 We are usually concerned about values far away from the mean on either side (a 2-sided test). This Figure says that 2.28% of the area under the curve is to the right of +2, and 2.28% is to the left of -2. So, when we sample from random behavior modeled this way, we expect to encounter values outside of ±2.0 only 2.28%+2.28% = 4.56% of the time. This is less than the 5% (in other words, $p<.05$) which is conventionally regarded as indicating statistical significance, i.e., to be contradicting the idea that everything is random.

142 The precise value of $p < .05$ is — $z > |\pm 1.96|$ for $p<.05$

143	and for $p < .01$ is	$z > \lvert\pm 2.58\rvert$ for $p<.01$
144	But, remember, just because a value is statistically significant doesn't mean that it is wrong. We do expect to see those values occasionally. The question to ask ourselves is "Why now?"	
145	What if we don't have a unit-normal distribution? We can often approximate it by taking our set of numbers, our data, subtracting from them their mean (arithmetic average) and dividing them by their standard deviation)	(the data - their mean) / (their standard deviation) $\rightarrow N(0,1)$
146	Residuals from our data, $\{R_{ni}\}$, have a mean of zero, and a modeled standard deviation of $V_{ni}^{0.5}$ so the standardized residuals $\{Z_{ni}\}$ should approximate $N(0,1)$	$\{R_{ni} / V_{ni}^{0.5}\} = \{Z_{ni}\} \rightarrow N(0,1)$

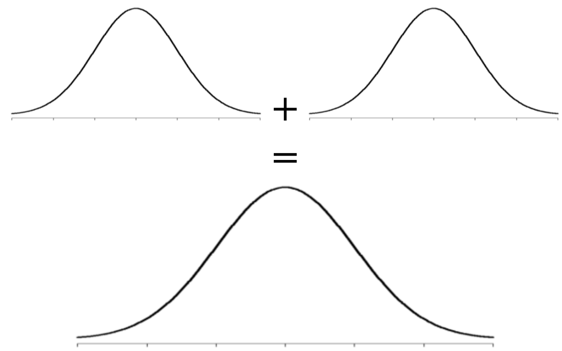
147 **Appendix 2. Chi-square, mean-square and degrees of freedom**

148 We talked about the unit-normal distribution in Appendix 1. And have discovered that the standardized residuals $\{Z_{ni}\}$ approximate N(0,1), the unit-normal distribution. So, what happens when we accumulate them?

Add two unit-normal distributions:

$$N(0,1) + N(0,1) = N(0, 2)$$

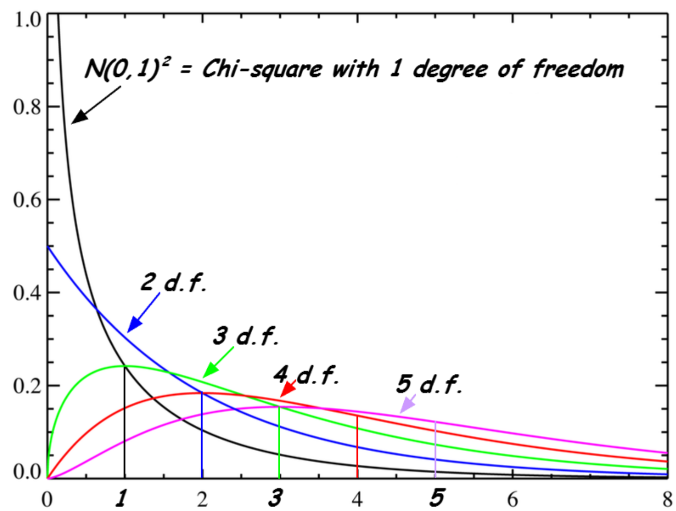
The average stays the same, but they spread out more.



149 But what if we square the values in a unit-normal distribution? The values in a unit normal distribution have a mean of 0, a range of -∞ to +∞ and a variance of 1. When we square these values, we have a distribution with a mean of 1, a range of 0 to +∞, and a variance of 2. This is called the "chi-square distribution with 1 degree of freedom", shortened to χ^2_1. It is the black curved line on the plot. Its mean is its degrees of freedom, indicated by the black vertical line going up from 1.

We can sum two of these square unit normal distributions: $N(0,1)^2 + N(0,1)^2 = \chi^2_2$. This has two degrees of freedom, d.f., and is the blue curve on the plot.

We can keep adding more. So, when we have added "k" squared (unit normal distributions) we have a chi-square distribution with k d.f., χ^2_k. It has a mean of k and a variance of 2k, so a standard deviation of √ (2k).



150 Since the mean of chi-square statistic is its d.f., it is convenient to divide the chi-square by its d.f., so that its value can be compared with 1.0. This makes scanning a Table of fit statistics much easier than when chi-square statistics with their d.f. are reported.

Mean-square = χ^2_k / k

Mean-square << 1 is over-fit, dependency, over-parameterization, over-predictability

Mean-square >>1 is under-fit, noise, misfit, lack of predictability

151 *Winsteps* reports the significance (probability) of a mean-square as a unit-normal deviate (ZSTD).

ZStd = "standardized like a z statistic"= Wilson-Hilferty (mean-square, d.f.)

see http://www.rasch.org/rmt/rmt162g.htm