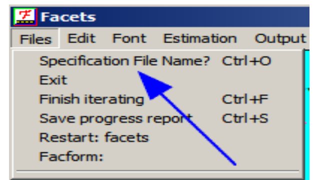
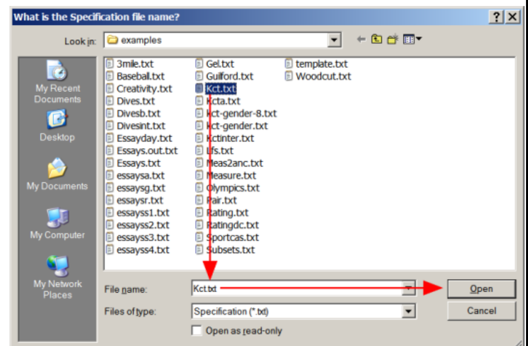
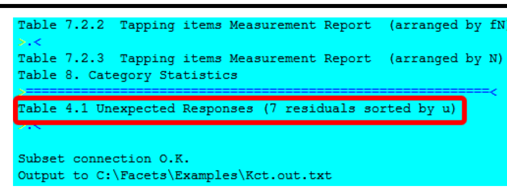
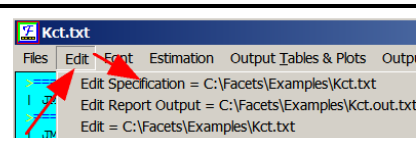
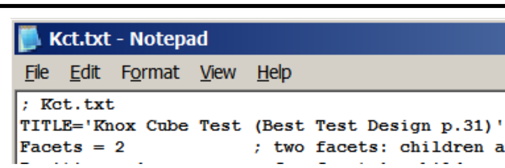
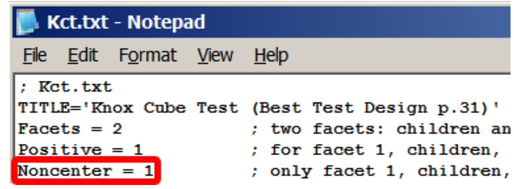
#	**Many-Facet Rasch Measurement : Facets Tutorial** **Mike Linacre - 1/2012**	
1.	**Tutorial 2. Fit analysis and Measurement Models** ***Welcome back!*** • Observations, expectations and residuals • Quality-control fit statistics elements and observations • Reliability indexes and inter-rater reliability This tutorial builds on Tutorial 1, so please go back and review when you need to.	
2.	**A. Facets Specifications and Data: The Knox Cube Test**	
3.	Let's launch *Facets* again	
4.	To start with we'll look at an analysis that's about as simple as it gets: 2-facets, dichotomous. Click on "Files" Click on "Specification File Name?"	
5.	Click on "Kct.txt" and "Open" or Double-Click on "Kct.txt" "Extra Specifications" - click on "OK" "What is the Report Output file name" - click on "Open" This is the "Knox Cube Test" data in "[Best Test Design]" (Wright & Stone, 1979, MESA Press). The Knox Cube Test was devised by Dr. Howard Knox on Ellis Island in New York harbor (next to the Statue of Liberty). It was used to screen immigrants arriving by ship from Europe. It assesses attention-span and short-term memory.	
6.	The Estimation and initial reporting completes. We will be particularly interested in Table 4.1 "Unexpected Responses", but first let's look at what this analysis is all about	
7.	Click on "Edit" menu Click on "Edit Specification = ... Kct.txt"	
8.	We know what most of this means: *;* starts a comment. I wanted to mention "*Kct.txt*", the name of the specification file. *TITLE=* is title line at the top of each output table Facets = *2* - there are two facets: children and items	

9. *Positive = 1* - the first facet (children) have positive ability: more score = more measure. The second facet, items has the default setting, negative difficulty, more score = less measure. You have probably realized that the order of the specifications doesn't matter, except that we need to specify *Facets*= early in the specification file.



Kct.txt - Notepad

File Edit Format View Help

```
; Kct.txt
TITLE='Knox Cube Test (Best Test Design p.31)'
Facets = 2            ; two facets: children an
Positive = 1          ; for facet 1, children,
Noncenter = 1         ; only facet 1, children,
```

10. *Noncenter=1*

Something that must be decided in all measurement is **where to measure from**. For short distances, we measure length from the end of the tape measure. For mountains, from sea level. For temperature, from freezing point of water for Celsius, but from freezing point of salt water for Fahrenheit. It is the same in Rasch measurement. The measuring convention is that **we measure from the center (mean) of the measures for each facet**. So item difficulties are measured from the center, *the local origin,* of the item facet. The average item has a difficulty of 0 logits. Judge severities are measured from the center, *the local origin*, of the judge facet. The average judge has a severity of 0 logits. We do this for all facets **except one**, usually the person ability facet. The person abilities are measured from the local origins of all the other facets. If the average ability is high, then the average person has a positive logit measure. If the average ability is low, then the average person has a negative logit measure.

So all facets have their local origin at their centers, except one facet.

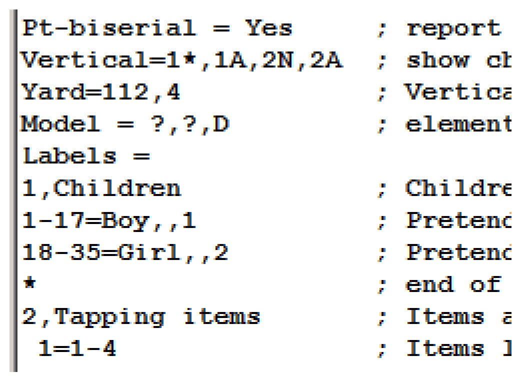
Noncenter=1 ; the first facet (children) does not have its local origin at its center.

11. *Pt-biserial = Yes* - report the point-biserial correlation in the measure tables, Table 7. These may not make much sense if the data are incomplete (so there are missing observations). This dataset is a complete rectangular dataset.

Vertical = this controls the facets to display in Table 6, the vertical rulers.

Yard = this controls the size of the display in Table 6.

Recommendation: Use the *Output Tables* pull-down menu to play with different settings of *Vertical*= and *Yard*= for Table 6, until you find settings that you like.



```
Pt-biserial = Yes      ; report
Vertical=1*,1A,2N,2A   ; show ch
Yard=112,4             ; Vertica
Model = ?,?,D          ; element
Labels =
1,Children             ; Childre
1-17=Boy,,1            ; Pretend
18-35=Girl,,2          ; Pretend
*                      ; end of
2,Tapping items        ; Items a
 1=1-4                 ; Items 1
```

12. *Model = ?, ?, D*

There is only one model specification, so it can be on the same line as *Model*=. "?" means "any element of the facet". The first "?" is for facet 1. The next "?" is for facet 2. So this model specification says: "Any element of facet one can combine with any element of facet two to produce an observation on a D-type scale". "D" means "dichotomous 0-1 scale". So *Facets* expects to see 0's and 1's in the data file. Anything else is treated as a missing value and ignored.

13. *Labels* = defines the facet names and the elements in the facets.

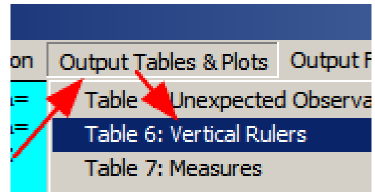
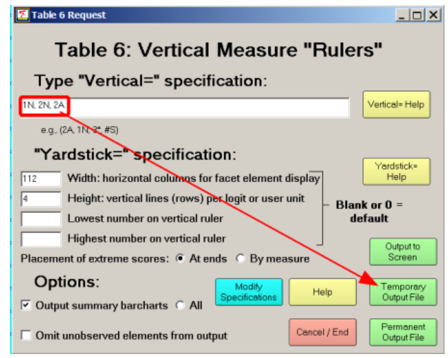
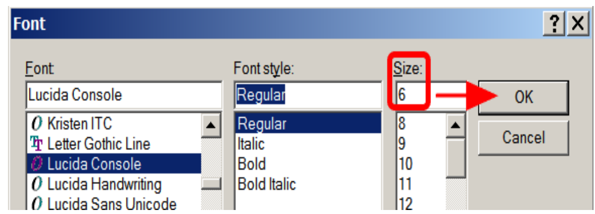
1, Children - the label or name of the first facet is "children"

1-17=Boy,,1 - after the facet name comes the list of elements. In this facet, element numbers 1 to 17 are all labeled "Boy". They could be given individual labels if desired. ",,1" means "the boys are part of element group 1". So a measure Table with totals will be produced for the boy group.

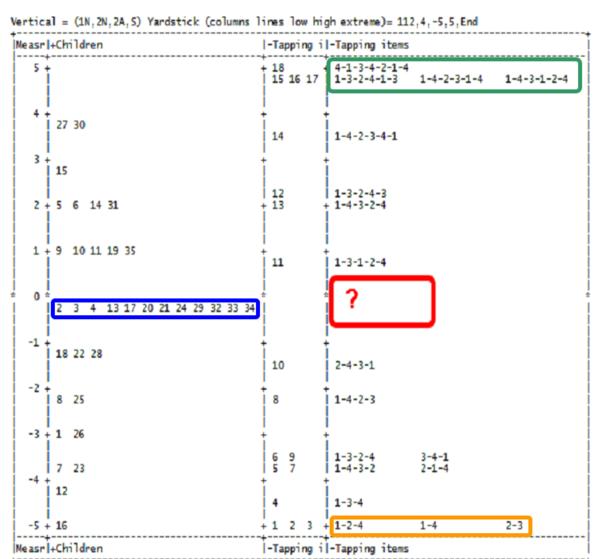
18-35=Girl,,2 - element numbers 18 to 35 are all labeled "Girl". They could be given individual labels if desired. ",,2" means "the girls are part of element group 2". So a measure Table with totals will also be produced for the girl group.

* - element lists end with "*"

14.	*2, Tapping items* - the label of the second facet is "Tapping items". The Knox Cube Test requires the participants to tap on items, (see Optional Reading at #178). *1 = 1-4* - the first item is labeled "1-4". That item requires the children to tap cube 1 and then cube 4. *Recommendation:* Choose item-labels that are meaningful to you, so that the *Facets* reports and maps have a useful message.	
15.	18=4-1-3-4-2-1-4 is the last item label. The pattern has 7 taps. * ends the element list and the facet list Data= starts the data *An example of entering the data one observation at a time:* 1,1,1 - facet 1 element 1 combines with facet 2 element 1 to produce an observation of 1. *An example of entering the data using indexing:* 1,2-18,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0 → facet 1 element 1 combines with facet 2 elements 2 to 18 to produce observations of 1 (for facet 2 element 2), of 1 (for facet 2 element 3), ... , of 0 (for facet 2 element 8), ... , of 0 (for facet 2 element 18)	18=4-1-3-4-2-1-4 * ; end of item lab Data = ; no data file na 1 ,1 ,1 ; child 1 on item 1 ,2-18,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0 2 ,1-18,1,1,1,1,1,1,0,0,1,1,1,0,0,1,0,0,0,0 3 ,1-18,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,0 4 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0 5 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,1,1,0,0,0,0 6 ,1-18,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0
16.	*Let's use indexing from here on ...* 35, 1-18, 1,1,1,1, ... - the observations for facet 1 element 35 (a girl according to the Labels=) for facet 2 elements 1 to 18, are 1, 1, 1, 1, ... The Data= instruction ends at the end of the file.	32,1-18,1,1,1,1,1,1,1,1 33,1-18,1,1,1,1,1,1,1,1 34,1-18,1,1,1,1,1,1,1,1 35,1-18,1,1,1,1,1,1,1,1
17.	Take a look at the data. Which observations accord with the Rasch model and which observations contradict it? It is usually difficult to judge by eye. red box: I've marked an observation that might be a "lucky" success .... but I'm not sure. It is for child 2, item 14. blue box: Can you pick out an "unlucky failure"? We will see how good we are doing misfit detection by eye .... **green box:** another lucky guess for child 9, item 13?	Data = ; no data file nam 1 ,1 ,1 ; child 1 on item 1 ,2-18,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0 2 ,1-18,1,1,1,1,1,1,0,0,1,1,1,0,0,1,0,0,0,0 3 ,1-18,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,0 4 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0 5 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,1,1,0,0,0,0 6 ,1-18,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0 7 ,1-18,1,1,1,1,0,0,1,0,1,0,0,0,0,0,0,0,0,0 8 ,1-18,1,1,1,1,1,0,1,0,1,1,0,0,0,0,0,0,0,0 9 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,1,0,0,0,0,0
18.	Now close the Kct.txt Edit window.	
19.		

<table>
<tr><td>20.</td><td colspan="2">B. Table 6: The Knox Cube Test Measures</td></tr>
<tr><td>21.</td><td>Let’s produce a version of Table 6, the vertical rulers which tells us what we will need to know.
In the main Facets window,
click on “Output Tables & Plots”
click on “Table 6: Vertical Rulers”</td><td>
</td></tr>
<tr><td>22.</td><td>Kct.txt has: “Vertical=1*,1A,2N,2A”
But we want to display the first facet, the children, by element number: 1N
the second facet, the items by element number, 2N
and by element label (the tapping pattern) 2A
so the specification is:
Vertical = 1N, 2N, 2A
We can specify this by typing 1N, 2N, 2A in the Vertical= box.
Then click on “Temporary Output File”</td><td>
</td></tr>
<tr><td>23.</td><td>Table 6 displays.
At first it was too big for my screen and somewhat faint, so (just like in Tutorial 1) I went to the NotePad menu bar, and used the Format pull-down menu to change the Font and the Size. Mine is “Courier New” 8-point.
For sizes smaller than 6 points, type the size into the NotePad font size box.</td><td>In NotePad:

</td></tr>
<tr><td>24.</td><td>In Table 6, at the extreme left is the measurement scale, “Measr”, in logits. It is -5 to +5, a typical range. This is not pre-set. It is estimated from the pattern of the data.

blue box: The children are shown with many of them just below 0 logits. This is the mode of their distribution. The column heading is “+Children”. “+” means “more score ↔ more measure”. So the most able (highest scoring) children are at the top. They are 27 and 30.

In the third column, the items are shown by element number. The fourth column shows them by label. The column is headed “-Tapping items”, so
“more score ↔ less measure.”
The lowest scoring items (least success by the children) are at the top. These are the most difficult items. The easiest items are at the bottom.</td><td>
<pre>
Vertical = (1N,2N,2A,S) Yardstick (columns lines low high extreme)= 112,4,-5,5,End
+--+
|Measr|+Children |-Tapping i|-Tapping items |
|-----+---------------------------+----------+----------------------------|
| 5 +   + 18   + 4-1-3-4-2-1-4 |
| | | 15 16 17 | 1-3-2-4-1-3   1-4-2-3-1-4   1-4-3-1-2-4 |
| | | | |
| 4 +   +   + |
| | 27 30 | | |
| | | 14 | 1-4-2-3-4-1 |
| 3 +   +   + |
| | 15 | | |
| | | 12 | 1-3-2-4-3 |
| 2 + 5 6 14 31   + 13   + 1-4-3-2-4 |
| | | | |
| 1 + 9 10 11 19 35   +   + |
| | | 11 | 1-3-1-2-4 |
*   0 *   *   * ?   *
| | 2 3 4 13 17 20 21 24 29 32 33 34 | | |
| -1 +   +   + |
| | 18 22 28 | | |
| | | 10 | 2-4-3-1 |
| -2 +   +   + |
| | 8 25 | 8 | 1-4-2-3 |
| -3 + 1 26   +   + |
| | | 6 9 | 1-3-2-4   3-4-1 |
| | 7 23 | 5 7 | 1-4-3-2   2-1-4 |
| -4 +   +   + |
| | 12 | | |
| | | 4 | 1-3-4 |
| -5 + 16   + 1 2 3 + 1-2-4   1-4   2-3 |
|-----+---------------------------+----------+----------------------------|
|Measr|+Children |-Tapping i|-Tapping items |
+--+
</pre></td></tr>
</table>

25. Do you notice any flaws in this version of Knox's test? Here is one ...
Red box in #24: There are 12 children in the middle of the range (children 2, 3, 4, 13, ...), but no items at their level. If the test is intended for children like those in this sample, it needs more middle-difficulty items. Can you imagine some extra items that might go in the red box in #24? They will be more difficult than the items below the red box, but easier than the items above the red box.
green box: Now look at the top of the item columns. There are 4 items that are much too difficult for this sample.
orange box: And at the bottom there are three items that are somewhat too easy. These extreme items waste everyone's time, and they may make the children frustrated or over-confident.
In their book, *"Best Test Design",* Wright & Stone improve this test.

26. **C. Table 6: Measures and Expectations**

27. Now look at person 5 in #24. His measure is +2 logits. What do we expect to happen when he encounters item 13, also at +2 logits ? The child has the same ability as the item has difficulty. We don't know what will happen. The child's probability of success is 0.5

```
+   2 +5  6  14 31                 +13        + 1-4-3-2-4
```

28. What about when child 9 of ability +1 logits attempts item 13 of difficulty +2 logits. Child 9 is less able than the item is difficult so the child will probably fail. But what is the child's exact probability of success? .4, .3, ..?

```
+   2 + 5  6  14 31                +13        + 1-4-3-2-4
|     |                            |          |
|     |                            |          |
+   1 +9  10 11 19 35              +          +
```

29. We can compute the probability of success from the Rasch model for dichotomous observations (Tutorial 1). Let's fill in the values: $B_n = +1$, $D_i = +2$

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

$$\log_e(P_{ni}/(1-P_{ni})) = +1 - +2 = -1$$

30. Rearrange the algebra. (If you are not sure about "e", please review Tutorial 1, Appendix 3).

31. The probability of success when child 9 of ability +1 logits attempts item 13 of difficulty +2 logits is p = .27.

$$P_{ni} = e^{-1} / (1 + e^{-1})$$
$$= 1/2.718 / (1 + 1/2.718)$$
$$= 0.37 / 1.37 = .27$$
= 1 success in every 4 attempts

32. **Logit-to-Probability Conversion Table**

Here is a Table to guide you when you convert dichotomous logit differences into percents (or probabilities) of success.

green text: Our difference is -1 logits. Look half-way down the right-hand pair of columns. -1.0 logits is 27% chance of success, which is the same as p=.27.

Notice these useful values:
1.1 logits difference = 75% chance of success
2.2 logits difference = 90% chance of success
3.0 logits difference = 95% chance of success

Logit diff.	% Success		
5.0	99%	-5.0	1%
4.6	**99%**	-4.6	1%
4.0	98%	-4.0	2%
3.0	**95%**	-3.0	5%
2.2	**90%**	-2.2	10%
2.0	88%	-2.0	12%
1.4	80%	-1.4	20%
1.1	**75%**	-1.1	25%
1.0	73%	-1.0	27%
0.8	70%	-0.8	30%
0.5	62%	-0.5	38%
0.4	60%	-0.4	40%
0.2	55%	-0.2	45%
0.1	52%	-0.1	48%
0	50%	-0.0	50%

33.	We have the logit measure for every child and every item. They are displayed in Table 6 (pictorially) and Table 7 (numerically). So we can use the Rasch dichotomous model to compute probability of success for every child on every item. These probabilities are the "**expected**" observations.	For dichotomous, 0 or 1, data, probability of success → the expected value of the observation
34.	Think of this in terms of frequency. What would we expect if 100 people of the ability of child 9 attempted item 13?	100 attempts at item 13 by ability of child 9. Logit difference = -1, Percent success = 27% *Expect:* 27 successes out of 100 attempts Expected value of 1 attempt = 27/100 = .27 = Rasch-model probability of success
35.		

36. **D. Rasch Theory: Observations, Expectations and Residuals: Response-level fit of the data to the Rasch model**

37. Here is the Knox Cube Test data again:

The principles of fit are easier to explain with dichotomous data than with polytomous data, so that is why we are starting here.

```
Data =                          ; no data file nam
1 ,1    ,1                      ; child 1 on item
1 ,2-18,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0
2 ,1-18,1,1,1,1,1,1,0,0,1,1,1,0,0,1,0,0,0,0
3 ,1-18,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,0
4 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0
5 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,1,1,0,0,0,0
6 ,1-18,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0
7 ,1-18,1,1,1,1,0,0,1,0,1,0,0,0,0,0,0,0,0,0
8 ,1-18,1,1,1,1,1,0,1,0,1,1,0,0,0,0,0,0,0,0
9 ,1-18,1,1,1,1,1,1,1,1,1,1,0,0,1,0,0,0,0,0
```

38. Child 9, item 13, is marked in green. The child scored "1", a success!

the observation:
$X_{ni} = X_{9,13} = 1$

39. We've already discovered in #28 that Child n=9 (ability b = 1 logit) is less able than item i=13 (difficulty d = 2 logits): b - d = -1, $P_{ni} = 0.27$

the expectation =
$E_{ni} = P_{ni} = e^{b-d}/(1+e^{b-d})$
$= e^{-1} / (1 + e^{-1}) = 0.27$

40. The difference between the **observation** and its **expectation** is the "**residual**" (what is left over). This is the part of the observation we did not expect to see ...

the residual:
$R_{ni} = X_{ni} - E_{ni} = 1 - 0.27 = 0.73$

41. We know that what we will see in the KCT data are 0's and 1's, but they are *not* the expected values. The expected values are numbers like .68 and .34. So there are almost always residuals.
There is a further question to ask, "Are we *surprised* about the size of the residual, or is it about the size of the discrepancy we were expecting to see?"
There are two aspects to what we expect:
1. The expected (average) value.
2. The expected variation of the observed around its expected value. This is called the "model variance".

Think of 100 people like child 9 attempting item 13. We expect 27 successes. The expected (average) value is 27/100 = .27. But we also expect to see 27 1's and 73 0's. So there will be residuals!
Here is a technical computation:
the **sum-of-squared-residuals** = sum of (observation - expectation)^2
= (count of successes)*(success - expected value)^2 + (count of failure)*(failure - expected value)^2
= (success count)*(1 - expected value)^2 + (failure count)*(0 - expected value)^2
$= 27*(1 - .27)^2 + 73*(0 - .27)^2 = 27 * .73^2 + 73 * .27^2 = 100 * .27 * .73 = 19.71$

the **model residual variance** = V_{ni} = sum-of-squares / count of residuals
$V_{ni} = 19.71 / 100 = 0.1971$
the **model residual standard-deviation** = square-root (variance) = $\surd(V_{ni}) = \surd(0.1971) = 0.44$
= the size of the splatter of the observations around their expected values.

42. With these numbers, we can calculate how unexpected is our residual, R_{ni} . The *standardized* residual, Z_{ni} , is as unexpected as the unit *normal deviate* [see Appendix 1. Unit Normal Deviates of this tutorial].

standardized residual =
$Z_{ni} = R_{ni} / \surd(V_{ni})$

43. In our example, $X_{13,9} = 1$, the residual $R_{ni} = 0.73$, the residual S.D. = 0.44, so that the standardized residual, Z_{ni}, is 1.66. This is as unusual as a unit *normal deviate* of 1.66, $p \approx .10$ (see Table in Appendix 1), but not unusual enough ($p<.05$) to be considered significantly misfitting the Rasch model.

$R_{ni} = 0.73$, $V_{ni} = 0.1971$, $\sqrt{(V_{ni})} = 0.44$,
$Z_{ni} = 0.73/0.44 = 1.66$; $p \approx .10$

44. Now let's look at the observation I ringed in red in #37: Child 2 on item 14. According to Table 6 (see #24), Child 2 has an ability of about -0.25 logits. Item 14 has a difficulty of about 3.37 logits.

Logit difference (child - item) =
-0.25 - 3.37 = -3.6 logits
Probability of success (Table in #32) = 3%

45. We observed a success, so $X_{ni} = 1$.
Expectation = 3% success = .03 (we are rounding the computations 2 decimal places for clarity)
Now we can compute the residual and the standardized residual.
The residual, R_{ni}, is .97 (very large)
and the standardized residual, Z_{ni}, is 5.60 (very unexpected), $p<.01$.

Observation: $X_{ni} = 1$
Expectation: $E_{ni} = P_{ni} = .03$
Residual: $R_{ni} = X_{ni} - E_{ni} = 1 - .03 = 0.97$
Model variance of the observation around its expectation:
$V_{ni} = P_{ni}*(1-P_{ni}) = .03*.97 = .03$
Standardized residual:
$Z_{ni} = R_{ni} / \sqrt{(V_{ni})} = 0.97 / \sqrt{(.03)} = 5.60$

46. **E. Table 4: Unexpected Responses**

47. We could go through this computation by hand for every observation, but it is easier to have *Facets* do it for us.
Click on the *Facets* Report output file on your Windows Taskbar
(or click on *Facets* "Edit" menu, click on "Report output file")
Scroll down to Table 4. It is the last Table. It shows the unexpected responses (or unexpected observations)

```
Knox Cube Test (Best Test Design p.31) 4/23/2009 3:12:00 AM
Table 4.1 Unexpected Responses (7 residuals sorted by u).

+---------------------------------------------------------+
| Cat  Score  Exp.  Resd StRes| Nu Chil Nu Tapping items |
|-----------------------------+---------------------------|
| 1     1       .0   1.0  6.2 |  2 Boy  14 1-4-2-3-4-1    |
| 0     0      1.0  -1.0 -6.1 |  2 Boy   7 1-4-3-2        |
| 0     0      1.0  -1.0 -6.1 | 24 Girl  7 1-4-3-2        |
| 0     0      1.0  -1.0 -4.8 | 24 Girl  6 3-4-1          |
| 1     1       .1    .9  3.5 | 24 Girl 12 1-3-2-4-3      |
| 1     1       .1    .9  3.5 | 33 Girl 12 1-3-2-4-3      |
| 0     0       .9   -.9 -3.5 | 28 Girl  5 2-1-4          |
|-----------------------------+---------------------------|
| Cat  Score  Exp.  Resd StRes| Nu Chil Nu Tapping items |
+---------------------------------------------------------+
```

48. ***Green box:*** *Facets* has done the computation for Child 2 on Item 14 more precisely than I did. It reports that the standardized residual (StRes, Z_{ni}) is 6.2. This is the most unexpected observation in these data. The observation is unexpectedly high (1) compared with its expected value (.0)

49. *Red box:* Look at the next two observations listed in Table 4. Both are on item 7 and they are equally unexpected StRes = -6.1. The minus - sign means "they did worse than we expected." Both children, 2 and 24, failed on the item when we expected them so succeed. We don't know why, but if we were serious about the children or the instrument we might inquire. The tapping pattern includes the sequence 4-3-2. Perhaps the examiner sped-up unintentionally, or perhaps he didn't clearly tap each cube so the children saw 4-2 instead of 4-3-2.

The list of unexpected responses nearly always contains useful messages about the instrument, the sample, the judges, the dataset, or whatever .....

50. **F. Table 7: Quality-control fit statistics elements and observations**

51. Looking down the list of unexpected responses is somewhat like looking at the pot-holes in a road. You want to pay some attention to them (not too much, usually), but they don't tell you much about the surface of the road as a whole. For that we need to take a wider look.

52. Scroll back up the Kct.out.txt Report Output file to Table 7.2.2. - the measure Table for Items in *fit order, descending,* or output a new copy of Table 7 from the "Output Tables" menu.
red box: You will see 4 columns: Infit and Outfit, MnSq and Zstd. These are quality-control fit statistics. They are central to the evaluation of the quality of the data for the construction of measures.

53. Table 7.2.2 Tapping items Measurement Report (arranged by fN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Nu	Tapping items
31	35	.9	.98	-3.87	.71	1.35	.8	2.30	1.2	.52	.30	7	1-4-3-2
3	35	.1	.03	3.40	.70	1.59	1.2	1.53	1.0	.60	.09	14	1-4-2-3-4-1
30	35	.9	.96	-3.41	.65	1.18	.5	.98	.6	.86	.43	6	3-4-1
6	35	.2	.08	2.25	.55	1.17	.6	1.07	.5	.82	.26	12	1-3-2-4-3
24	35	.7	.80	-1.58	.49	1.07	.3	.84	.0	.96	.46	10	2-4-3-1
12	35	.4	.27	.80	.45	1.08	.4	.79	-.1	.96	.37	11	1-3-1-2-4
31	35	.9	.98	-3.87	.71	1.05	.2	.53	.6	1.00	.47	5	2-1-4
27	35	.8	.90	-2.37	.54	.59	-1.3	.43	-.2	1.35	.65	8	1-4-2-3
7	35	.2	.10	1.96	.53	.70	-1.0	.38	-.2	1.34	.44	13	1-4-3-2-4
32	35	.9	.99	-4.45	.82	.91	.0	.35	.8	1.08	.48	4	1-3-4
30	35	.9	.96	-3.41	.65	.62	-1.0	.21	.0	1.34	.63	9	1-3-2-4
1	35	.0	.01	4.85	1.08	.76	.0	.11	1.4	1.18	.24	15	1-3-2-4-1-3
1	35	.0	.01	4.85	1.08	.76	.0	.11	1.4	1.18	.24	16	1-4-2-3-1-4
1	35	.0	.01	4.85	1.08	.76	.0	.11	1.4	1.18	.24	17	1-4-3-1-2-4
35	35	1.0	1.00	(-6.64	1.86)	Minimum					.00	1	1-4
35	35	1.0	1.00	(-6.64	1.86)	Minimum					.00	2	2-3
35	35	1.0	1.00	(-6.64	1.86)	Minimum					.00	3	1-2-4
0	35	.0	.00	(6.18	1.86)	Maximum					.00	18	4-1-3-4-2-1-4

54. How well do the observations of each element fit with the estimate of the its measure?
If the fit is good, then we can have confidence that the measure means what it says.
If the fit is bad, then the measure could mislead us.
If the fit is too good, then perhaps something is constraining the data to be too coherent.
If the fit is too bad, then those data could also be damaging the measures of other elements.

55. So we have the process of fit evaluation. In *Facets*, most fit statistics are based on summarizing the residuals that we've already thought about one at a time.

Data → Measure Estimates →
Expected Data → Residuals →
Fit statistics →
Validity of Measure Estimates

56. Imagine that we administer a dichotomous test in which the items are ordered from easy to difficult. What would we expect would happen when a typical person takes the test?

Success on the easy items ! Failure on the hard items .

And a transition zone  where the items are about as difficult as the person is able, so we expect to see some successes and some failure.
This is what has happened with the top left response-string in the Table in **#Error! Reference source not found.**:
"1110110110100000". And this is also the pattern that the Rasch model predicts: *"There is nothing so practical as a good theory" (Kurt Lewin, 1951, p. 169)*

Person Responses:
Easy -- Items -- Hard
1110110110100000

57.	So, how can we verify that this response string does match Rasch expectations? We do this using mean-square fit statistics. A mean-square is a chi-square divided by its degrees of freedom *[see **Appendix 2.** if chi-square sounds like Greek to you ...].* Let's start with chi-square fit statistics	
58.	**Chi-square fit statistics** are very useful for diagnosing the standardized residuals. The standardized residuals are modeled to be unit normal deviates. So when we square them and sum them, we expect their sum will approximate a chi-square distribution with mean equal to the count of the standardized residuals. If the chi-square value is much above the count, then the standardized residuals are further away from 0, on average, than the Rasch model predicts. The observations are farther from their expectations than the Rasch model predicts. The data are too unpredictable, "noisy". They "underfit" the Rasch model. If the chi-square value is much below the count and so much closer to zero, then the standardized residuals are closer to 0, on average, than the Rasch model predicts. The observations are closer to their expectations than the Rasch model predicts. The data are too predictable. The unexpectedness in the data is "muted". The data "overfit" the Rasch model. The value of the chi-square, along with its degrees of freedom, enable us to compute how unlikely these data are to be observed by chance when the data fit the Rasch model. When we deem the data too unlikely to have occurred by chance, then we declare that "the data misfit the model".	
59.	Chi-square statistics are useful for quantifying the fit of the data to the Rasch model, but we can make them even more convenient. The expected mean of a chi-square distribution is its "degrees of freedom", the number of independent squared unit-normal distributions it represents. If we **divide a chi-square value by its degrees of freedom, then we have a mean-square value.**	chi-square χ^2 / degrees of freedom (d.f.) = mean-square (MnSq) mean (expectation) of MnSq = 1.0 model variance of MnSq = 2 / d.f. standard deviation of MnSq = √(2/d.f.) **The expected value of a mean-square is 1.0**
60.	**G. Rasch Theory: "The data misfit the model!"**	
61.	*Are you surprised by that statement?* Many statisticians would be. Descriptive statistics are based on summarizing the data efficiently and parsimoniously. The data are considered to be the given (*Latin* "datum") truth. The statistical model (regression, ANOVA, etc.) is intended to describe the dataset. So a good **descriptive** statistical model is one which fits the data. If the model misfits the data, then try a different descriptive model. Rasch is a **prescriptive** statistical method. The Rasch model gives us what we want (additive measures in a unidimensional framework), so it is our "truth". The data may, or may not, contain the information that we need. So good data fit our Rasch model. If the data don't fit the model usefully, then the dataset as a whole doesn't support unidimensional measurement. Some part of the dataset may. In fact, usually most of a dataset does, if it is intended to capture one latent variable. *Thought:* Raw scores are the "sufficient statistics" for a Rasch analysis. If the dataset doesn't conform to Rasch analysis, then it doesn't conform to raw-score analysis either ! (But CTT analysts usually do not know this). Raw-score fit analysis tends to be superficial, so the misfit in the dataset to a raw-score Classical Test Theory model is often overlooked.	
62.	We have now had 40 years experience with mean-squares since Wright & Panchapakesan (1969) proposed them for Rasch usage (see Optional Reading at #178). The following Tables summarizes them from a Rasch measurement perspective.	

63.

H. Table 7: Interpretation of Element-level Mean-Square Fit Statistics:

Mean-square	Interpretation
>2.0	Distorts or degrades the measurement system. (The background noise is starting to drown out the music.)
1.5 - 2.0	Unproductive for construction of measurement, but not degrading. (The background noise is audible, but not intrusive to the music.)
0.5 - 1.5	Productive for measurement. (Beautiful music)
<0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations. (Music too quiet)

64.

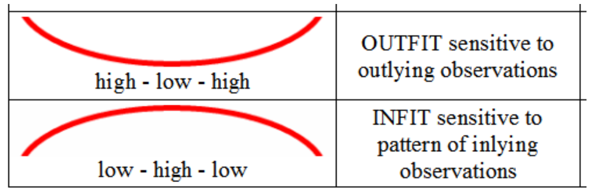
Person Responses: Easy -- Items -- Hard	Diagnosis Pattern	OUTFIT Mean-square	INFIT Mean-square	Point-Measure Correlation
111¦0110110100¦000	Modeled/Ideal	1.0	1.1	.62
111¦1111100000¦000	Guttman/Deterministic	**0.3**	**0.5**	**0.87**
000¦0000011111¦111	Miscode	**12.6**	4.3	**-0.87**
011¦1111110000¦000	Carelessness/Sleeping	**3.8**	1.0	0.65
111¦1111000000¦001	Lucky Guessing	**3.8**	1.0	0.65
101¦0101010101¦010	Response set/Miskey	**4.0**	**2.3**	**0.11**
111¦1000011110¦000	Special knowledge	0.9	**1.3**	0.43
111¦1010110010¦000	Imputed outliers *	**0.6**	1.0	0.62
111¦0101010101¦000	Low discrimination	**1.5**	**1.6**	0.46
111¦1110101000¦000	High discrimination	**0.5**	0.7	0.79
111¦1111010000¦000	Very high discrimination	**0.3**	**0.5**	**0.84**
Right ¦ Transition ¦ Wrong				
high - low - high	OUTFIT sensitive to outlying observations	>>1.0 unexpected outliers	>>1.0 disturbed pattern	
low - high - low	INFIT sensitive to pattern of inlying observations	<<1.0 overly predictable outliers	<<1.0 Guttman pattern	

65. Look back at the Table above.
The mean-squares for our imagined typical respondent are near to 1.0 - good!

Person Responses: Easy -- Items -- Hard	Diagnosis Pattern	OUTFIT Mean-square	INFIT Mean-square
111¦0110110100¦000	Modelled/Ideal	$\boxed{1.0}$	$\boxed{1.1}$

66. What about the "lucky guesser" who succeeded on the most difficult item. The OUTFIT mean-square is 3.8, much bigger than 1.0. That lucky guess has degraded the guesser's measure. It is less secure as a basis for inference. Do we really want "guessing for success"?

Person Responses: Easy -- Items -- Hard	Diagnosis Pattern	OUTFIT Mean-square	INFIT Mean-square
111¦1111000000¦001	Lucky Guessing	$\boxed{3.8}$	$\boxed{1.0}$

<table>
<tr><td>67.</td><td colspan="2">I. Table 7: Outfit vs. Infit</td></tr>
<tr><td>68.</td><td>Did you notice that the INFIT mean-square for the lucky-guesser is 1.0, its expected value?
What is going on?
The Outfit statistic is outlier-sensitive.
The Infit statistic is sensitive to patterns in the targeted responses. It is inlier-pattern sensitive.</td><td>
</td></tr>
<tr><td>69.</td><td>Take a look at “special knowledge”.
Imagine the items are in 4 cluster of difficulty: addition, subtraction, multiplication, division.
Then most children will follow the typical Rasch pattern. But those who are taught: addition, multiplication, subtraction, division will have a different pattern: fail on subtraction, succeed on multiplication.</td><td>Special knowledge or Alternative curriculum
Person Responses:
Easy -- Items -- Hard
111¦1000011110¦000
Add-Subtract-Multiply-Divide</td></tr>
<tr><td>70.</td><td>The OUTFIT statistic is 0.9 (less than 1.0). The Outfit statistic reports that responses far from the person ability are predictable.
The INFIT statistic is 1.3, reporting the patterns in the data are somewhat unpredictable. The Infit statistic detects the unexpected pattern of responses near the person ability.</td><td><table><tr><th>Person Responses:
Easy -- Items -- Hard</th><th>Diagnosis
Pattern</th><th>OUTFIT
Mean-square</th><th>INFIT
Mean-square</th></tr><tr><td>111¦1000011110¦000</td><td>Special knowledge</td><td>0.9</td><td>1.3</td></tr></table></td></tr>
<tr><td>71.</td><td>Mathematically, the OUTFIT Mean-square is the conventional statistical chi-square divided by its degrees of freedom.
The Infit statistic is an information-weighted mean-square statistic.</td><td>For the N observations that we are summarizing in the mean-square statistics:
Outfit Mean-square = $\Sigma (R_{ni}^2 / V_{ni}) / N$
Infit Mean-square = $\Sigma (R_{ni}^2) / \Sigma V_{ni}$</td></tr>
<tr><td>72.</td><td>Glance back at Table 7.2.2. It should be starting to make more sense to you.
Red arrow: Item 7 has the biggest Outfit mean-square, MnSq, statistic: 2.25. This is much bigger than the expected mean-square of 1.0. There is more “unmodeled noise” than useful “statistical information” in this item.
Green arrow: Item 14 has the second biggest Outfit MnSq: 1.48.</td><td><table><tr><th>Obsvd Score</th><th>Obsvd Count</th><th>Obsvd Average</th><th>Fair-M Avrage</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><th>Estim. Discrm</th><th>Nu Tapping items</th></tr>
<tr><td>31</td><td>34</td><td>.9</td><td>.97</td><td>-3.84</td><td>.71</td><td>1.34</td><td>.8</td><td>2.25</td><td>1.1</td><td>.54</td><td>7 1-4-3-2</td></tr>
<tr><td>3</td><td>34</td><td>.1</td><td>.03</td><td>3.37</td><td>.70</td><td>1.57</td><td>[illegible]</td><td>1.48</td><td>.9</td><td>.62</td><td>14 1-4-2-3-4-1</td></tr>
<tr><td>30</td><td>34</td><td>.9</td><td>.96</td><td>-3.38</td><td>.64</td><td>1.18</td><td>[illegible]</td><td>.97</td><td>[illegible]</td><td>.87</td><td>6 3-4-1</td></tr></table></td></tr>
</table>

73. Red box: Item 7 (biggest OUTFIT mean-square) has two unexpected responses in Table 4.1
Green box: Item 14 (second biggest Outfit MnSq) has only one response in Table 4.1, but it is the most unexpected response. This is the most outlying response.

We see that the unexpected responses in Table 4.1 can cause the Outfit MnSq statistics in Table 7 to be large. Start by looking at the Outfit statistics in Table 7 to localize the problem areas in the data. Table 4 tends to be too detailed.

```
Table 4.1 Unexpected Responses (7 residuals sorted by u).

+-----------------------------------------------------+
| Cat  Score  Exp.  Resd StRes| Nu Chil Nu Tapping items |
|----------------------------+------------------------|
| 1     1      .0   1.0   6.2 |  2 Boy  14 1-4-2-3-4-1  |
| 0     0     1.0  -1.0  -6.1 |  2 Boy   7 1-4-3-2      |
| 0     0     1.0  -1.0  -6.1 | 24 Girl  7 1-4-3-2      |
| 0     0     1.0  -1.0  -4.8 | 24 Girl  6 3-4-1        |
```

In big datasets, Table 4 can be come unmanageably long (which is why it is after Table 8 in the Report output file)

74. Diagnosing misfit:
Large OUTFIT mean-square > 1.5 - Unexpected off-target observations - Look at Table 4
Small OUTFIT mean-square < 0.5 - Off-target observations too predictable - Are there imputed data or other constraints?
Large INFIT mean-square > 1.5 - Unexpected patterns in on-target observations - Very difficult to investigate.
Suggestion: Write out the residual file to Excel. Sort on person (or item, etc.) element number and "logit". Look at patterns in responses near logit 0.
Small INFIT mean-square < 0.5 - On-target observations too predictable - Are there redundant items or response sets in rated items?

75. **J. Table 7: Misfit: Size vs. Significance: MnSq vs. Zstd**

76. We know that a large mean-square statistic flags unexpectedness in the data. But is this an unusual amount of unexpectedness, or merely a reflection of the randomness in the data which the Rasch model requires? The **Zstd** statistics (mean-squares standardized as z-statistics) answer this.
The Outfit and Infit Mean-squares are derived from chi-square statistics with their d.f.. So we know how unlikely we are to observe any particular mean-square value (or worse). This is what the Zstd statistics report.

77. We could report the probability of the mean-square. Computing the actual d.f. is complicated, so let's assume the mean-square value is a chi-square with 1 d.f.

Item 7: Outfit MnSq = 2.25
chi-square = 2.25 with d.f. = 1
probability ≈ 0.13

78. Our experience is that small probabilities become long numbers that are often difficult to think with. So instead of reporting the probability, we report the equivalent unit-normal deviate [see Appendix 1], called Zstd, "the mean-square statistic standardized like a Z-score". This is also a Student's *t*-statistic with infinite d.f.

Item 7: Outfit Zstd = 1.1
probability = 0.14

79. Reporting Zstd simplifies interpretation.
See Appendix 1 for more Zstd values.

|Zstd| ≥ 2.0 are statistically significant
|Zstd| ≥ 2.6 are highly significant

80. So, the rule-of-thumb with Outfit and Infit statistics is:
"MnSq size: large enough to be distorting; MnSq > 1.5
Zstd significance: improbable enough to be surprising. Zstd > 2.0"

81. *What about unexpectedly low mean-squares?*

These are paradoxical! Here is a Guttman response string, name after psychometrician Louis Guttman. He proposed that the ideal response string is one where someone succeeds on all the easy (for that person) items and fails on all the hard items. The result would be the response string you see here. This is also the best possible response string for Classical Test Theory. It has the highest point-biserial correlation and the highest discrimination indexes, and results in the highest test reliabilities.

Guttman's ideal response-pattern is perfect for **ordering,** but **not for measurement**. For measurement we need **uncertainty** in the responses. **The closer to the item difficulty, the more uncertain each person's responses.** This fundamental to Rasch theory.

Person Responses: Easy -- Items -- Hard	Diagnosis Pattern	OUTFIT Mean-square	INFIT Mean-square
111¦1111100000¦000	Guttman/Deterministic	0.3	0.5

Guttman/deterministic. Louis Guttman, a leading psychometrician around 1950, proposed that the ideal item would be one on which all low performers failed and all high performers succeeded. It would act like a switch. It would have infinitely high discrimination. If you knew the location on the latent variable of the switch for each person, then all person responses would be exactly determined.

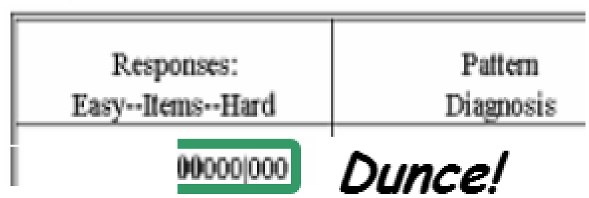
82. *So what is wrong with a Guttman pattern from a Rasch perspective?*

Rasch proposes that any reasonable subset of items should give statistically the same estimate as the full set. So let's split the test high-low:

1. According to the easy items, our respondent is a genius.
2. According to the difficult items, our respondent is a dunce.

A contradiction!

The problem is that the Rasch **transition zone** of uncertain responses to the targeted items is missing. Of course, the Rasch measure has correctly located the respondent between the easy and hard items on the latent variable, but the response string is squeezed together from the Rasch perspective.



83. ***Guttman Patterns and Low Mean-squares < 0.5:***
Guttman patterns produce low Mean-Squares.
Low mean-squares correspond to persons and items which are too predictable. They are lacking in the uncertainty Rasch needs for constructing measures.
This makes the reported standard errors (measurement precisions) too small and the reported reliabilities (measure reproducibility) too high. In general, however, **low mean-squares are not a serious problem.**

Small standard errors (high precision) and high reliability (high measure reproducibility, a consequence of high precision) are good, but only if that level of precision is really supported by the data. Here the reported standard errors (though computed correctly) are too small from a substantive perspective.
A parallel situation arises in physical measurement. Suppose you weigh yourself 100 times. Then your weight will be the average of those weights with precision (standard error of the mean) S.E.M. = S.D. of your 100 weights / 10. But do you believe this high precision about your own weight? No. It is statistically correct, but substantively misleading. You weight varies by more than that S.E.M. during each day. The calculated standard error of your weight is too small, and so may mislead you about how precisely you know your own weight.

84. In general, **low mean-squares are not a serious problem, but high mean-squares are.**
Low mean-squares rarely lead to incorrect inferences about the meaning of measures, unless they are caused by constraints which invalidate the measures.
So always investigate and remedy high mean-squares, and then re-analyze your data, before investigating low mean-squares. The overall average mean-squares are usually close to 1.0, so high mean-squares force there to be low mean-squares.

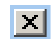
85. So what values of the mean-square statistics cause us real concern? Here is my summary table from Winsteps Help "Special Topic" "Misfit Diagnosis ..."

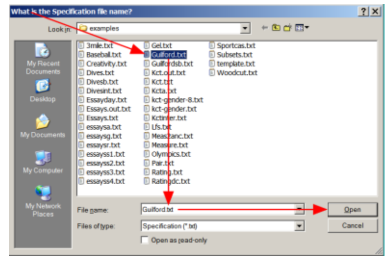
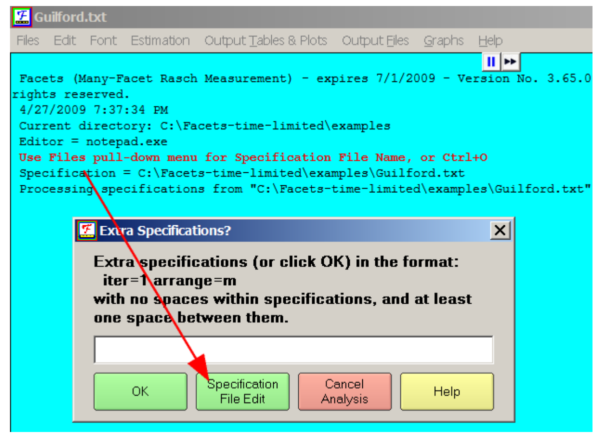
Here's a story:
When the mean-square value is around 1.0, we are hearing music! The measurement is **accurate**
When the mean-square value is less than 1.0, the music is becoming quieter, becoming muted. When the mean-square is less than 0.5, the item is providing only have the music volume (technically "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 additive) 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.**

Interpretation of mean-square fit statistics:	
>2.0	Distorts or degrades the measurement system. *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!*
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.

<table>
<tr><td>86.</td><td>Every Rasch analyst has favorite rules for identifying misfit. The Reasonable Mean-Square Fit Value is from http://www.rasch.org/rmt/rmt83b.htm

No rules are decisive, but many are helpful.</td><td><table>
<tr><th colspan="2">Reasonable Item Mean-square Ranges for INFIT and OUTFIT</th></tr>
<tr><th>Type of Test</th><th>Range</th></tr>
<tr><td>MCQ (High stakes)
MCQ (Run of the mill)
Rating scale (survey)
Clinical observation
Judged (agreement encouraged)</td><td>0.8 - 1.2
0.7 - 1.3
0.6 - 1.4
0.5 - 1.7
0.4 - 1.2</td></tr>
</table></td></tr>
<tr><td>87.</td><td>Close all open Facets windows</td><td></td></tr>
<tr><td>88.</td><td></td><td></td></tr>
</table>

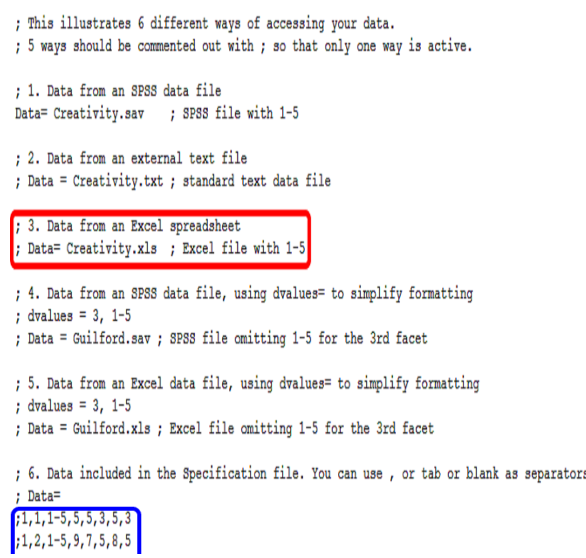
<table>
<tr><td>89.</td><td colspan="2">K. Facets Specification and Data: The Guilford Data</td></tr>
<tr><td>90.</td><td>Let’s apply what we’ve learned to some 3-facet rating data.
Launch Facets.
Click on “Files”
Click on “Specification File Name?”
Double-click on “Guilford.txt”</td><td></td></tr>
<tr><td>91.</td><td>“Extra specifications?”
Click on “Specification File Edit”</td><td>
</td></tr>
<tr><td>92.</td><td>Guilford.txt displays in a NotePad window.

Scroll down to Data=

Notice that there are alternative data files. Most are commented out with “;”

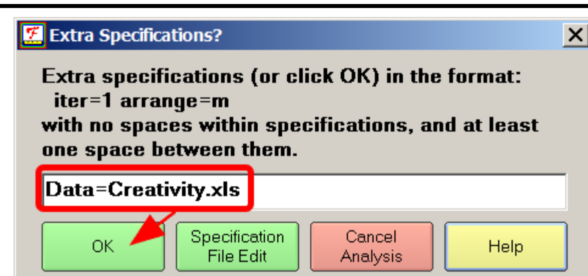
blue box: All these data files contain the same observations. You can see these in option 6.

red box: We will use the Excel file, “Creativity.xls”

Do not edit Guilford.txt - we will make the change using Extra Specifications?</td><td>

```
; This illustrates 6 different ways of accessing your data.
; 5 ways should be commented out with ; so that only one way is active.

; 1. Data from an SPSS data file
Data= Creativity.sav    ; SPSS file with 1-5

; 2. Data from an external text file
; Data = Creativity.txt ; standard text data file

; 3. Data from an Excel spreadsheet
; Data= Creativity.xls  ; Excel file with 1-5

; 4. Data from an SPSS data file, using dvalues= to simplify formatting
; dvalues = 3, 1-5
; Data = Guilford.sav ; SPSS file omitting 1-5 for the 3rd facet

; 5. Data from an Excel data file, using dvalues= to simplify formatting
; dvalues = 3, 1-5
; Data = Guilford.xls ; Excel file omitting 1-5 for the 3rd facet

; 6. Data included in the Specification file. You can use , or tab or blank as separators
; Data=
;1,1,1-5,5,5,3,5,3
;1,2,1-5,9,7,5,8,5
```
</td></tr>
<tr><td>93.</td><td>Click on your Facets analysis on the Windows task bar</td><td>
</td></tr>
<tr><td>94.</td><td>Type into the Extra Specifications? box:
Data=Creativity.xls
with no spaces

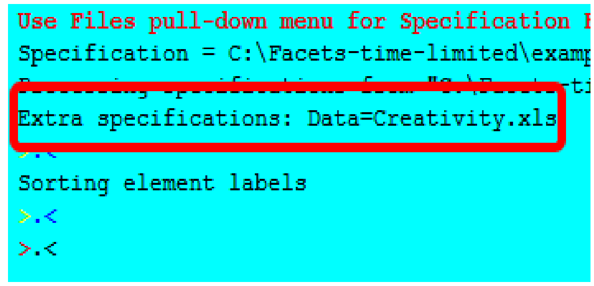
(or copy-and-paste: Ctrl+C Ctrl+V)

Click on OK</td><td>
</td></tr>
</table>

95. The Extra specification, Data=Creativity.xls is shown in the Facets Analysis window.

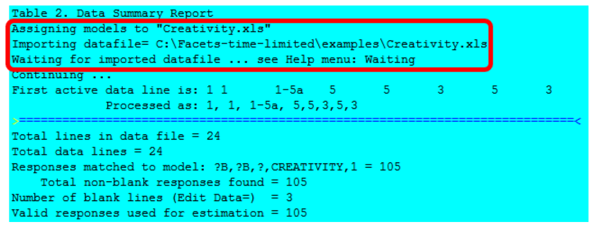
"What is the Report Output file name?"
- click on "**Open**" to accept the default value: *guilford.out.txt*

Analysis begins ....

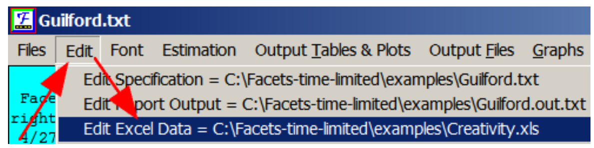


96. Notice on your *Facets* analysis window screen that the "Creativity.xls" is imported.

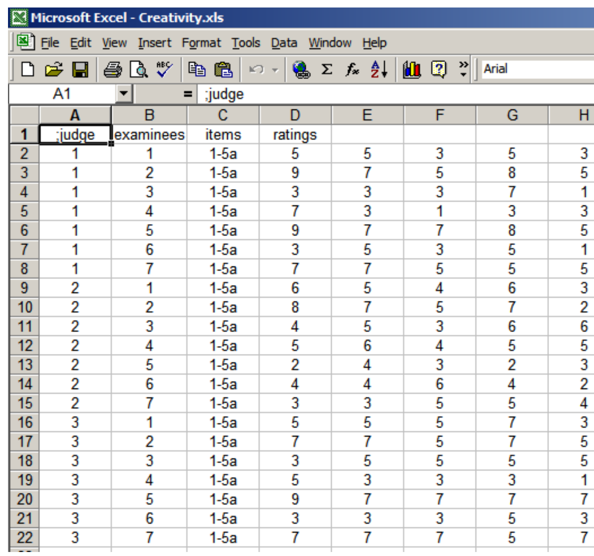
Facets launches Excel to obtain the responses. Excel is sometimes slow, so you may see the "Waiting ...." message.



97. Take a look at the Guilford data in Creativity.xls
Facets menu bar: Click on "Edit"
Click on "Edit Excel Data"
Excel launches and displays the data ...



98. The data, from "Psychometric Methods" by J.P. Guilford (1954) are of 3 Senior Scientists (the judges) rating 7 Junior Scientists (the examinees) or 5 items of Creativity. The observed range of the rating scale is 1-9. Guilford omits to tell us what the possible range was.
So row 1 of the spreadsheet is:
; judges examinees items ratings
the ";" is to tell *Facets* this row is a comment, not data.
Row 2 is the first data row, it says:
Judge 1 rated examinee 1 on 5 items, 1 to 5, and the ratings were 5, 5, 3, 5, 3
1-5 means "items 1, 2, 3, 4, 5"
"1-5a" is to prevent Excel converting 1-5 into -4. Facets ignores the "a".
There are 21 rows of data, and 105 ratings.


Microsoft Excel - Creativity.xls

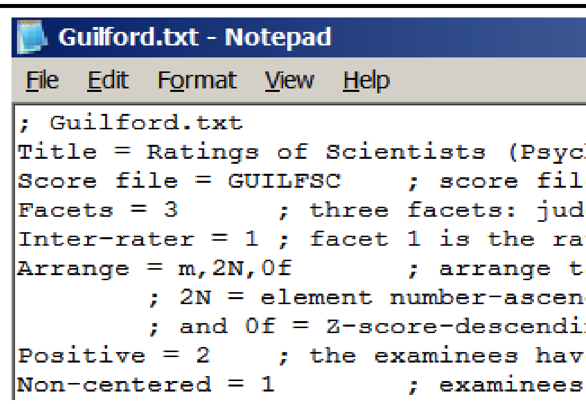
	A	B	C	D	E	F	G	H
1	;judge	examinees	items	ratings				
2	1	1	1-5a	5	5	3	5	3
3	1	2	1-5a	9	7	5	8	5
4	1	3	1-5a	3	3	3	7	1
5	1	4	1-5a	7	3	1	3	3
6	1	5	1-5a	9	7	7	8	5
7	1	6	1-5a	3	5	3	5	1
8	1	7	1-5a	7	7	5	5	5
9	2	1	1-5a	6	5	4	6	3
10	2	2	1-5a	8	7	5	7	2
11	2	3	1-5a	4	5	3	6	6
12	2	4	1-5a	5	6	4	5	5
13	2	5	1-5a	2	4	3	2	3
14	2	6	1-5a	4	4	6	4	2
15	2	7	1-5a	3	3	5	5	4
16	3	1	1-5a	5	5	5	7	3
17	3	2	1-5a	7	7	5	7	5
18	3	3	1-5a	3	5	5	5	5
19	3	4	1-5a	5	3	3	3	1
20	3	5	1-5a	9	7	7	7	7
21	3	6	1-5a	3	3	3	5	3
22	3	7	1-5a	7	7	7	5	7

99. Now let's examine the Guilford specification file.
You may have it on your Windows task bar, if not ...

Facets menu bar: Click on "Edit"
Click on "Edit Specification"

You are probably racing ahead of me, but just in case ...

; starts a comment
Title= specifies the title to print at the top of each Table.



100 **Score file=** specifies the file names to use for writing out score files for each facet.
The score files contain summary statistics for each element in each facet.
See *Facets* Help for exact details

The Score file for Facet 1:

```
GUILFSC.1.txt - Notepad
File  Edit  Format  View  Help
1 Senior scientists
    T.Score      T.Count      Obs.Avge  Fai
     171.00        35.00          4.89
     156.00        35.00          4.46
     181.00        35.00          5.17
```

101 **Facets =** specifies the number of facets in the analysis. We have 3 facets: judges, examinees and items.

Inter-rater = specifies the facet number of the rater or judge facet. This instruct *Facets* to compute rater-relevant statistics for this facet. For us, facet 1 is the judge facet, the “Senior Scientists”.

Inter-rater = will produce this in Table 7:

```
----------------------------------------
| Exact Agree. |                       |
| Obs %  Exp % | N Senior scientists   |
----------------------------------------
|  21.4   25.2 | 2 Brahe               |
|  35.7   25.8 | 1 Avogadro            |
|  37.1   25.3 | 3 Cavendish           |
----------------------------------------
```

102 **Arrange =** tells *Facets* in what order to arrange the elements when they are displayed in Table 7.
“Arrange = m” means “Arrange in measure order descending” so that the highest measure appears first in the Table. This is done for all the facets.
“Arrange = m, 2N” means: “after doing Arrange = m, then
red boxes: output a copy of Table 7 for facet 2 with the elements in numerical order ascending” so that element 1 is displayed first.

Arrange= will produce this:

```
 (arranged by 2N).

---------------------------------------------------------
it      Outfit    |Estim.|                               |
q ZStd  MnSq ZStd|Discrm|  N Junior Scientists          |
---------------------------------------------------------
4 -3.2   .23 -3.2|  1.48 | 1 Anne                        |
1 -1.1   .60 -1.2|  1.30 | 2 Betty                       |
3   .4  1.22   .7|   .84 | 3 Chris                       |
0   .9  1.37  1.0|   .87 | 4 David                       |
4  2.2  1.94  2.2|   .34 | 5 Edward                      |
9  -.8   .77  -.5|   .93 | 6 Fred                        |
5  -.3   .84  -.4|  1.37 | 7 George                      |
---------------------------------------------------------
```

103 **Positive=** defines which facets are oriented so that more score implies more measure.
In this analysis, we have chosen to do that for facet 2, the “Junior Scientists” who are the examinees.

Positive and negative facets in Table 6:

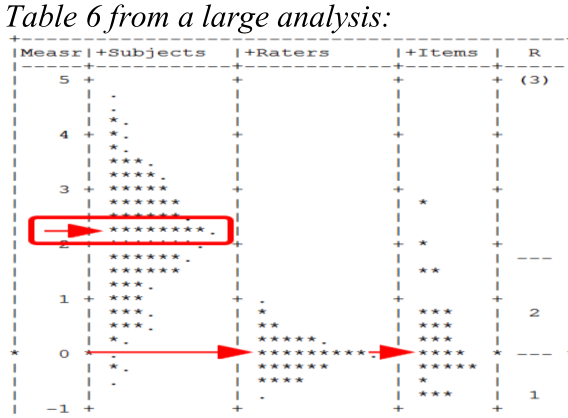
```
Table 6.0  All Facet Vertical "Rulers".

Vertical = (2N,3A,2*,1A,1A,S) Yardstick
+-----------------------------------------
|Measr|+Junior Scientists|-Traits
|-----+-------------------+---------------
```

104 **Non-centered=** specifies which facet does not have a local origin, but is measured relative to the origins of the other facets.

Example (not the Guilford analysis):
Red box: shows the subjects non-centered.
Other red arrows: the Raters and Items are centered

Guilford.txt is a study of rater behavior, so we have chosen facet 1, the “Senior Scientist” judges, to be non-centered, so that they “float” relative to the other facets.

Table 6 from a large analysis:



105 More specifications in Guilford.txt

The next two specifications are for Table 4 of “Unexpected Responses” (or observations)

```
Unexpected = 2   ; report ratings if s
Usort = (1,2,3),(3,1,2),(Z,3)    ; sor
Vertical = 2N,3A,2*,1L,1A         ;defi
Zscore = 1,2     ;report biases greate
Pt-biserial = measure ; point-measure
```

106	**Unexpected=** says how unexpected? “2” means “report responses with a standardized residual, StRes, of 2 or bigger in Table 4”.	*Table 4: Unexpected = 2:*
		`| Cat   Score   Exp.   Resd  StRes`
		`|  6     6      2.9    3.1   2.4`
		`|  2     2      6.0   -4.0  -2.7`

107	**Usort=** specifies how Table 4 is sorted. See [*Facets* Help]().	`Table 4.2 Unexpected Responses (4 residuals sorted by 3,1,2).` `| Cat  Score  Exp.  Resd StRes| N Senior sc N Junior N Traits     |` `| 2 2 6.0 -4.0 -2.7	2 Brahe 5 Edward 1 Attack |` `| 6 6 2.9 3.1 2.4	2 Brahe 6 Fred 3 Clarity |` `| 2 2 6.1 -4.1 -2.7	2 Brahe 5 Edward 4 Daring |` `| 6 6 2.9 3.1 2.4	2 Brahe 3 Chris 5 Enthusiasm |`
108	**Vertical=** defines the layout for the Table 6 “vertical rulers” **Zscore=** is for the interaction/bias analysis in Tutorial 3	`Vertical = 2N,3A,2*,1L,1A` `Zscore = 1,2     ;report bias`

109	**Pt-biserial=** is the point-biserial correlation or the point-measure correlation.
	
	PtMea is the observed point-measure correlation
	PtExp is the expected value of the point-measure correlation when the data fit the Rasch model.
	When possible, both the observed and the expected values of the correlations are reported.
		Table 7:
		`| Correlation |`
		`| PtMea PtExp |`
		`|   .49   .56 |`
		`|   .56   .58 |`
		`|   .67   .58 |`
		`|   .62   .58 |`
		`|   .51   .58 |`

110	**Model=** specifies the model. The B’s are for the interaction/bias analysis in Tutorial 3. The ?’s mean “any element”. “Creativity” is the name of a user-defined rating scale.
	Rating scale= defines the rating scale. It is called “Creativity”, and it is “R9”, a rating scale with highest category 9.
	1=lowest is the number and name of a category,
	* ends the category list
		`Model = ?B,?B,?,Creativity`
		`        ; A bias/interaction`
		`        ; senior scientists (`
		`        ; log(Pnijk/Pnijk-1)`
		`        ; Bn = ability n, Di`
		`        ; Pnijk = probability`
		`Rating scale = Creativity,R9`
		`1 = lowest      ; name of low`
		`5 = middle      ; no need to`
		`9 = highest     ; name of hig`
		`*`

111	In a model specification, "**?**" or “**$**” means "any element of this facet". "**#**" means any element of this facet, and each element has its own rating scale. So, F_k will look like: Models = ?, ?, ?, R F_{jk} will look like: Models = ?, #, ?, R ; assuming that "j" is the second facet.	Allowing each judge to have a unique (partial credit) rating scale: Model = #, ?, ?, Creativity ? does not work correctly in some versions of Windows. You can use $ instead.

112 **Labels=** specifies the facet and element names and numbers.

If you look at the Junior Scientists, you can see that the element numbers can be jumbled.

```
Labels=                 ;to
1,Senior scientists     ;na
1=Avogadro              ;na
2=Brahe                 ;th
3=Cavendish
*
2,Junior Scientists
2=Betty
5=Edward
```

113 **Data=** specifies the data.

The data always have the same layout, but they can be files of different types. Here are the current options:

Suffix	Data= Format
.txt	text file (MS-DOS or Windows)
.xls .xlsx	Excel workbook: first or only worksheet
.rda	R statistics data file
.sdata	SAS data file
.sav	SPSS data file
.dta	STATA data file
(other)	text file (MS-DOS or Windows)

```
; 1. Data from an SPSS data file
Data= Creativity.sav    ; SPSS file with 1-5

; 2. Data from an external text file
; Data = Creativity.txt ; standard text data file

; 3. Data from an Excel spreadsheet
; Data= Creativity.xls  ; Excel file with 1-5

; 4. Data from an SPSS data file, using dvalues= to simplify
; dvalues = 3, 1-5
; Data = Guilford.sav ; SPSS file omitting 1-5 for the 3rd f

; 5. Data from an Excel data file, using dvalues= to simplif
; dvalues = 3, 1-5
; Data = Guilford.xls ; Excel file omitting 1-5 for the 3rd

; 6. Data included in the Specification file. You can use ,
; Data=
;1,1,1-5,5,5,3,5,3
;1,2,1-5,9,7,5,8,5
;1,3,1-5,3,3,3,7,1
```

114 **Dvalues=**
some facet information in the data file can be specified once, instead of in every data line. This is time-saving.
More to come about this ...

```
; 5. Data from an Excel data file,
; dvalues = 3, 1-5
; Data = Guilford.xls ; Excel file
```

115. L. Facets Output Tables: The Guilford Report Output File

116. On the Windows task bar,
click on "Guilford.txt.out"
or
on the *Facets* menu bar, click on "Edit"
click on "Edit Report Output"

Table 1 reports the specifications for this analysis. The crucial details to check are the numbers of elements in each facet. If incorrect, modify your specification file.

```
Table 1. Specifications from file "C:\Facets-time-limited\examples\Guil

Title = Ratings of Scientists (Psychometric Methods p.282 Guilford 1954)
Data file = Creativity.xls
Output file = C:\Facets-time-limited\examples\Guilford.out.txt

; Data specification
Facets = 3
Non-centered =  1
Positive =  2
Labels =
1,Senior scientists ; (elements = 3)
2,Junior Scientists ; (elements = 7)
3,Traits ; (elements = 5)
Model = ?B,?B,?,CREATIVITY,1
Rating (or other) scale = CREATIVITY,R9,General,Ordinal
```

117. **Table 2** reports what happened to the data. We have 3 judges x 7 examinees x 5 traits = 105 observations. We expect them all to match our measurement model.
They do. Great!

```
Table 2. Data Summary Report.

Assigning models to "creativity.xls"
Total lines in data file = 22
Total data lines = 22
Responses matched to model: ?B,?B,?,CREATIVITY,1 = 105
```

118. **Table 3** reports the estimation process. We expect the last iteration to have very small numbers.

red box: for the largest raw-score difference, less than .5

blue box: for the biggest logit change, less than .01.
We have these. Great!

For big data sets, the maximum raw-score residual can be considerably larger without affecting the accuracy of the estimates.

```
Table 3. Iteration Report.

+-----------------------------------------------------------+
| Iteration       Max. Score Residual      Max. Logit Change |
|              Elements    %  Categories  Elements      Steps |
|-----------------------------------------------------------|
| PROX    1                                 .7405             |
| JMLE    2    21.0444  17.5    24.7419     .2372     2.2989 |
| JMLE    3    -4.1066  -3.4     -.7658    -.0648     -.0930 |
| JMLE    4    -1.4643  -1.2      .8815    -.0244     -.0517 |
| JMLE    5    -1.2319  -1.0      .3762    -.0198     -.0214 |
| JMLE    6     -.8471   -.7      .3664    -.0134     -.0218 |
| JMLE    7     -.6849   -.5      .2855    -.0102     -.0171 |
| JMLE    8     -.5835   -.4      .2359     .0081     -.0142 |
| JMLE    9     -.4905   -.3      .1932     .0067     -.0117 |
| JMLE   10     -.4082   -.3      .1592     .0055     -.0096 |
+-----------------------------------------------------------+
```

If you want to stop the iterative process early, press you Ctrl+F keys together.

119. "Subset connection O.K." so that the measures of all the elements belong to one cohesive structure. *We will discuss this in Tutorial 4.*

```
Subset connection O.K.
```

120. **Table 4** appears after Table 8

121. **Table 5** shows some global summary statistics.
For each observation:
Cat (category) is the observation
Score is the category after it has been recounted
Exp. is the expected value of the Score
Resd. is the residual = Score - Expectation
StRes is the standardized residual
Mean (average) is the average for the observations.
Count is the number of observations in the analysis.
S.D. (Population) is the standard deviation if the elements are all possible elements for the facet
S.D. (Sample) is the (larger) standard deviation if the elements are a sample of all possible elements (the population) for the facet.

```
Table 5. Measurable Data Summary.

+-----------------------------------------------+
| Cat  Score  Exp.  Resd StRes|                  |
|-----------------------------+------------------|
| 4.84  4.84  4.84   .00   .01| Mean (Count: 105) |
| 1.88  1.88  1.18  1.44  1.00| S.D. (Population) |
| 1.89  1.89  1.19  1.45  1.00| S.D. (Sample)     |
+-----------------------------------------------+
```

red box: When estimation has been successful we expect the mean residual (Resd) and the mean standardized residual (StRes) to be 0.0,
green box: and the S.D. of the standardized residual (StRes) to be 1.0.

122. An approximate global fit statistic, a log-likelihood chi-square is shown. Its degrees of freedom, d.f., are roughly (number of responses - number of elements).
The Rasch model is a model of perfection, so **we always expect to see significant misfit** to the model in empirical data, as we do here: p=.0000

```
Data log-likelihood chi-square = 331.4227
Approximate degrees of freedom = 85
Chi-square significance prob.  = .0000
```

123. *Red box:* Part of the variance in the data is explained by the Rasch measures, and, as the Rasch model predicts, part is unexplained. In these data, 41% is explained by the Rasch measures, a usual amount - even though it 41% looks low!

```
                                         Count   Mean   S.D.   Params
Responses used for estimation     =       105   4.84   1.88       20
Count of measurable responses     = 105.00
Raw-score variance of observations =  3.53 100.00%
Variance explained by Rasch measures = 1.45  41.02%
Variance of residuals             =   2.08  58.98%
```

124. **Table 6** shows the measures graphically. We can see that there is a noticeable spread among the Junior Scientists (examinees) and the Traits (items) which we want. There is also a smaller spread among the Senior Scientists (judges) which we don't usually want, but the Rasch measures have adjusted for.
The rating scale, "CREAT" is shown to the right.
Which is the most lenient judge? The column heading "-Senior Scientist" tells us. The most lenient judge will give the highest ratings. "-" means "high score implies low measure", so Cavendish is the most lenient judge.

Vertical = 2N, 3A, 2*, 1L (same as 1A), 1A, S

```
---------------------------------------------------------------------------------------------------------
|Measr|+Junior Scientists|-Traits                |+Junior Scientists|-Senior scientists|-Senior scientists|CREAT|
---------------------------------------------------------------------------------------------------------
+   1 +                  +                        +                  +                  +                  + (9) +
|     |                  |                        |                  |                  |                  |  7  |
|     |                  |                        |                  |                  |                  |     |
|     |                  |                        |                  |                  |                  | --- |
|     | 2                |                        | *                |                  |                  |     |
|     |                  | Enthusiasm             |                  |                  |                  |     |
|     | 5                |                        | *                |                  |                  |  6  |
|     | 7                |                        | *                |                  |                  |     |
|     |                  | Clarity                |                  | Brahe            | Brahe            | --- |
|     |                  |                        |                  |                  |                  |     |
*   0 *                  *                        *                  * Avogadro         * Avogadro         *  5  *
|     | 1                | Basis                  | *                | Cavendish        | Cavendish        |     |
|     | 3                |                        | *                |                  |                  | --- |
|     |                  | Attack      Daring     |                  |                  |                  |     |
|     |                  |                        |                  |                  |                  |  4  |
|     | 4                |                        | *                |                  |                  |     |
|     | 6                |                        | *                |                  |                  |     |
|     |                  |                        |                  |                  |                  | --- |
|     |                  |                        |                  |                  |                  |     |
|     |                  |                        |                  |                  |                  |  3  |
+  -1 +                  +                        +                  +                  +                  + (1) +
---------------------------------------------------------------------------------------------------------
|Measr|+Junior Scientists|-Traits                | * = 1            |-Senior scientists|-Senior scientists|CREAT|
---------------------------------------------------------------------------------------------------------
```

125. Table 6.1 is a graphical representation of the measures we see in Table 7. It is useful when we need to picture the statistics for large samples. M represents the mean, S=1 standard deviation, and Q=2 standard deviations.
The numbers represent elements. The numbers match Table 7.
Red box: In the bottom distribution for a much larger dataset, there are 28 elements at "M", the mean.
Read the numbers vertically.

```
Table 6.1  Senior scientists Facet Summary.

Logit:
                                      1    1     1
 +---------------------------Q---S--+-M----S---Q-------------------------+
-1                                    0                                   1

Infit MnSq:
                           1     1                    1
 +----------Q----------S---------M+---------S----------Q--------------+
 0                                1                                     2

Outfit MnSq:
                           1       1                     1
 +---------Q-----------S----------M-----------S----------Q------------+
 0                                1                                     2
Logit:
                               12222 2 211  1
                  22 13 79888472320 8 866981336543414 25    2
 +---------+-----Q---+--S------M------S--+---Q-----+------
-3        -2        -1         0             1         2
```

126. M. Table 7. Measure Tables

127. **Table 7** shows the scores and measures.
Measures are often reported in logits or other units unfamiliar to our audience. They often ask, "but what do they mean in terms of the scores I'm familiar with?"
Green box: This is what the "Fair Average" does. It takes the measures and shows what they imply as ratings for a standard person rated by standard judge on a standard item. "Standard" means an imaginary element with the average measure of the elements of the facet.
In this example, the data are complete, so the Observed Average rating is close to the Fair Average rating. But when there are missing data, the Fair Average adjusts for the missing data but the Observed Average does not.

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.
156	35	4.5	4.39	.24	.12
171	35	4.9	4.86	.04	.11
181	35	5.2	5.17	-.09	.11
169.3	35.0	4.8	4.81	.06	.12
10.3	.0	.3	.32	.13	.00
12.6	.0	.4	.39	.16	.00

128. In a practical assessment situation, different people may be administered different tasks and rated on different items by different judges. You encounter the difficult tasks and the severe judges. I encounter the easy tasks and the lenient judges. *Fine!* Your ability measure and my ability measure adjust for this.

But then the examination authorities say “Rasch measures are great, but when we publish the results, we want them expressed as ratings on the original rating scale!”

So we have to go from Rasch measures back to the rating scale in a way that is fair - as though you and I encountered the same judges and performed the same tasks. *Facets* does this for us by computing the ratings we would have received (according to the Rasch measures) if you and I had both performed a task of average difficulty and we were both rated by judges of average severity. This gives a “Fair Average” rating.

129. Red box: In Table 7, the “Model S.E.” is the **precision** of the measure. This indicates how fuzzy is the location of the element measure on the latent variable.

In everyday speech, the words “**precision**” and “**accuracy**” often mean the same thing, but for us they are different.
Imagine arrows being shot at a target. If the arrows form a close group, then the archery is precise. If the arrows are in the neighborhood of the center of the target, the archery is accurate. When the arrows all hit the bull’s eye, the archery is accurate and precise.

Precision | Accuracy

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
131	28	4.7	4.60	.07	.14	1.35	1.3	1.34	1.2
148	28	5.3	5.28	-.23	.13	.95	-.1	.98	.0
150	28	5.4	5.36	-.27	.13	.62	-1.6	.60	-1.7

Measurement Precision: how exact is the location on the latent variable?
Measurement Accuracy: is it the correct location?

Estimation Precision (decimal places): how closely does our estimate match the estimation criteria?
Statistics are often reported with 6 decimal places (high estimation precision) even though they are reporting only a few data points (low substantive precision).

130. **Precision** means “how reproducible is the location of the measure on the latent variable with data like these”. It is like the gradations on measurement scale. It is internal to the measuring system, and is quantified in the standard error of measurement, S.E.
The more observations of an element, the more precise will be the estimate. As carpenters say, *“Measure twice, cut once!’*

Accuracy means “how well does the measure correspond to an external standard”. In our case, the external standard is the Rasch-model ideal of invariant measure additivity.
If the data fit the Rasch model, then the parameter estimates accurately reflect the ideal additive measurement framework. For us, accuracy is quantified in the quality-control fit statistics, Infit and Outfit.

131. We can obtain **higher precision** for an element’s measure by:
1. **More observations** of the element, e.g., a person takes a longer test or is rated by more judges.
2. **Better targeting** of the element, e.g., a person takes a test that is not too easy or too high.
3. **More categories** in the rating scale, e.g., a 5-category rating scale instead of a 3-category scale, but beware of over-categorization .... which we will soon meet!

132 N. Table 7: Fit Statistics

133 **Green box**: Guilford.out.txt Table 7.1.1 shows the measures for the judges.
Blue box: Brahe (lowest Total Score) is the slightly most severe judge.
Red box: But, more importantly, look at the fit statistics. Brahe is the most misfitting (Mean-squares > 1.0). The other two judges have about the same fit.
Orange box: The average of the mean-squares is usually near 1.0, so a misfitting judge, like Brahe, forces the other judges, Avogadro and Cavendish to be reported as overfitting.

```
+---------------- --------------------------------- ---------------------+
| Total   Total  |         Model | Infit      Outfit    | |                     |
| Score   Count  |Measure  S.E.  | MnSq ZStd  MnSq ZStd| | N Senior scientists |
|---------------- +--------------+---------------------+ +---------------------|
|  156     35    |    .24   .12  | 1.42  1.7  1.47  1.8| | 2 Brahe             |
|  171     35    |    .04   .11  |  .84  -.6   .87  -.5| | 1 Avogadro          |
|  181     35    |   -.09   .11  |  .66 -1.6   .65 -1.6| | 3 Cavendish         |
|---------------- +--------------+---------------------+ +---------------------|
|  169.3   35.0  |    .06   .12  |  .97  -.2  1.00  -.1| | Mean (Count: 3)     |
|   10.3     .0  |    .13   .00  |  .33  1.4   .35  1.5| | S.D. (Population)   |
|   12.6     .0  |    .16   .00  |  .40  1.7   .42  1.8| | S.D. (Sample)       |
+---------------- --------------------------------- ---------------------+
```

Always investigate underfit (high mean-squares) before overfit (low mean-squares). Often the overfit disappears when the underfit is eliminated from the data.

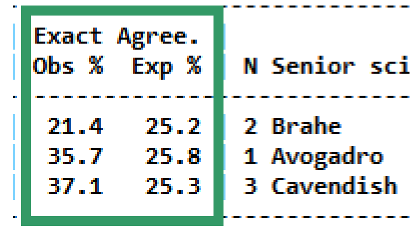
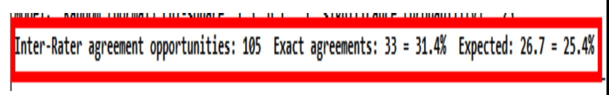
134 Notice also that the Infit and Outfit columns are similar. This is usual with long rating scales (9 categories here) so that the operational range of each item is very wide.

Under these circumstances, my choice is only to report **Outfit**, because it is the conventional statistical chi-square (divided by its d.f.) which is familiar to most statisticians, but please do report both if your audience expects to see them.

Polytomous mean-square statistics have the same characteristics as dichotomous ones, **#Error! Reference source not found.**, but are much harder to diagnose by eye.

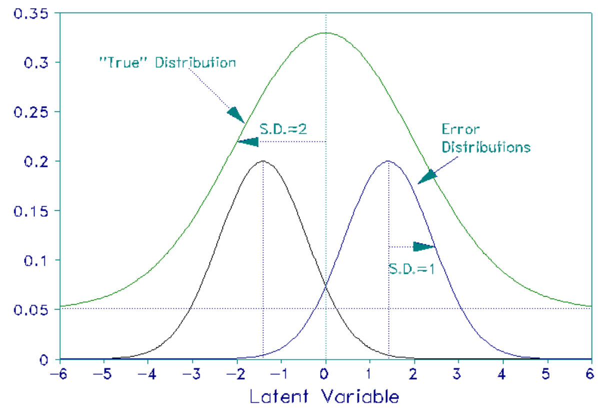
Polytomous Mean-square Fit Statistics

Response String Easy..........Hard	INFIT MnSq	OUTFIT MnSq	RPM Corr.	Diagnosis
I. modelled:				
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
IV. informative-noisy:				
32222222201111111130	.94	1.22	.55	noisy outliers
33233332212333000000	1.25	1.09	.77	erratic transitions
33133330232300101000	1.49	1.40	.72	noisy progression
33333333330000000000	1.37	1.20	.87	extreme categories
V. non-informative:				
22222222222222222222	.85	1.21	.00	one category
12121212121212121212	1.50	1.96	-.09	central flip-flop
01230123012301230123	3.62	4.61	-.19	rotate categories
03030303030303030303	5.14	6.07	-.09	extreme flip-flop
03202002101113311002	2.99	3.59	-.01	random responses
VI. contradictory:				
11111122233222111111	1.75	2.02	.00	folded pattern
11111111112222222222	2.56	3.20	-.87	central reversal
22222222223333333333	2.11	4.13	-.87	high reversal
00111111112222222233	4.00	5.58	-.92	Guttman reversal
00000000003333333333	8.30	9.79	-.87	extreme reversal

<table>
<tr><td>135</td><td colspan="2">O. Table 7: Inter-rater Statistics</td></tr>
<tr><td>136</td><td>Inter-rater= has instructed Facets to compute some rater agreement statistics.
Green box: The “Exact Agreement Observed %” report what percent of the ratings by this rater agree exactly with the ratings made by another rater. the “Exact Agreement Expected %” reports the agreement that would be seen if the data fit the Rasch model perfectly.</td><td>
Exact Agree.
Obs % Exp % N Senior sci
21.4 25.2 2 Brahe
35.7 25.8 1 Avogadro
37.1 25.3 3 Cavendish</td></tr>
<tr><td>137</td><td colspan="2">For Brahe the observed agreement is 21.4%. Is this good or bad? We would tend to expect a much higher agreement. But Facets provides a reference point. It reports that for these raters, examinees and items, the “Exact Agreement Expected %” for Brahe is 25.2%. Usually the observed agreement is slightly higher than the expected agreement, because most raters try to be “agreeable” with each other. Look at Avogadro and Cavendish, their observed agreement %’s (35.7%, 37.1%) are much higher than expected (25.8%, 25.3%). They are agreeing together against Brahe.</td></tr>
<tr><td>138</td><td>Under Table 7.1.1, the agreement statistics are summarized. In these data, the observed “exact agreement” is 31.4%, but the expected agreement is 25.4%. The judges are agreeing too well!!
Something is wrong!</td><td>
Inter-Rater agreement opportunities: 105 Exact agreements: 33 = 31.4% Expected: 26.7 = 25.4%</td></tr>
<tr><td>139</td><td colspan="2">Facets models the raters to be “independent experts”. These would produce an “exact agreement” percent, which is the same or slightly higher than the “expected agreement” percent.
But many raters are trained to behave like “rating machines”. Agreement is encouraged among the raters, and disagreements are penalized. For these raters we expect the “exact agreement” percent to be much higher than the “expected agreement” percent. When the “exact agreement” approaches 100%, the raters are behaving the same way as optical scanners do for “bubble sheets”. The raters have become part of the data-collection mechanism, they are no longer a facet of the measurement situation.</td></tr>
</table>

140	P. Table 7: Reliabilities and Separations
141	`Model, Populn: RMSE .12 Adj (True) S.D. .07 Separation .60 Strata 1.13 Reliability (not inter-rater) .26` `Model, Sample: RMSE .12 Adj (True) S.D. .12 Separation 1.02 Strata 1.69 Reliability (not inter-rater .51`
142	Under each Table 7 is a set of reliability statistics. These show the reliability of the differences between the measures in the facet. They indicate the **reproducibility** of the measures, **not the accuracy** of the measures. These reliabilities are not *inter-rater reliability* statistics (which show the rater similarity). "**Reproducible**" - we can expect the same number if we repeated the same data collection. A stopped clock is highly reproducible, so it is highly **reliable**. *Of course, it is reliably wrong!* "**Accuracy**" - the current number is near the "true" number.
143	"**Model**" means "assuming all misfit in the data is due to the randomness predicted by the Rasch model" "**Real**" (when shown) means "assuming all misfit in the data contradicts the Rasch model" "**Population**" means "assuming this set of elements is the entire population." "**Sample**" means "assuming this set of elements is a random sample from the population of interest" "**RMSE**" means "root mean-square error", a statistical average of the standard errors of the measures. "**Adj (True) SD**" means "the standard deviation of the measures, (Adj=) adjusted for measurement error", also called the *"True" standard deviation.* "**Separation**" is the True SD / RMSE. It indicates how many measurement strata could be statistically distinguishable among the measures, if the tails of the measure distribution are conceptualized to be caused *by outlying random noise.* "**Reliability**" is the ratio of the "True" variance of the measures to the observed variance. "**Strata**" is (4*Separation + 1)/3. It indicates how many measurement strata could be statistically distinguishable among the measures, if the tails of the measure distribution are conceptualized to be caused *by outlying "true" measures.*
144	This table shows the relationship between measurement variance, measurement error, and reliability.

Error RMSE	True SD	True Variance = True SD²	Observed Variance = RMSE² + True Variance	Signal-to-Noise Ratio	**Separation** = True SD / RMSE	**Reliability** = True Variance / Observed Variance	**Strata** =(4*Separation+1)/3
1	0	0	1	0	**0**	**0**	**0.3**
1	1	1	2	1	**1**	**0.5**	**1.7**
1	2	4	5	2	**2**	**0.8**	**3**
1	3	9	10	3	**3**	**0.9**	**4.3**
1	4	16	17	4	**4**	**0.94**	**5.7**

145	What does this table of separations and reliabilities means? **Here is a picture of Separation = 2.** Green curve: The larger curve is the conceptual "true" distribution. Black and blue curves: The smaller curves are the error distributions for individual measures. The x-axis on the graph locates the person measures on the latent variable. The y-axis is on the graph is the local density, i.e., what proportion of the sample we expect at the x-axis location (larger curve) or what proportion of observations of a measure with error we expect at an x-axis value (smaller curves).	 **Separation = 2 implies 2 error strata**

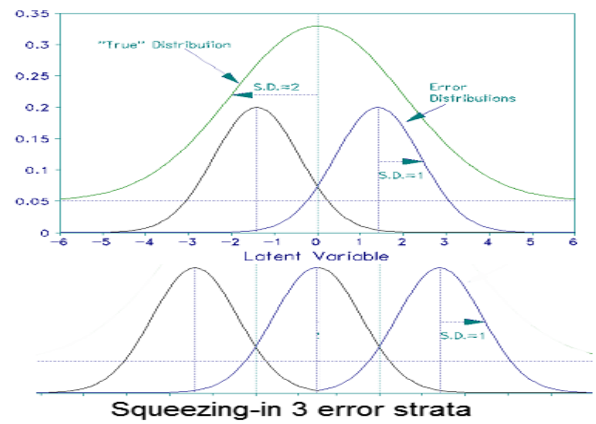
146	So the question becomes *"How many numerically different measures can we reasonably discriminate within the true distribution?"* For Separation = 2, we can see that two measures at about -1.5 and +1.5, together with their error distributions (the fuzziness of the measurement), matches the "true" distribution of the measures. The splatter around only two measures (shown by the two smaller curves) covers the whole reasonable range of the upper curve.
147	**A physical analogy:** Imagine I have a classroom of children and must report their heights. The report is overdue, so I measure their heights quickly by eye. This will yield imprecise measurements with high uncertainty. The measures of height would have big standard errors. *Question 1. How far apart must two heights be for me to be reasonably sure that the children's heights are different?* Answer: Roughly three standard errors. we are comparing two measures and both have standard errors. So, assuming the standard errors are approximately the same, the statistical value is (two error distributions)*($p<.05$)*(RMSE) = $\sqrt{2}$ * 1.96 * SE ≈ 3 *S.E. *Question 2. In the observed distribution of heights, how many "reasonably sure" height-difference strata are there, assuming there are no unusually short or tall children?* Answer: This is the "Separation", which is (True S.D. / Standard error). Practical example: My children have only two "true" heights: half are 1.5 meters high, and half are 1.7 meters high, so the true S.D. of their height measures is .1 meters. But when I record their heights by eye, the 1.5-meter-children have a range from 1.4 meters to 1.6 meters, and the 1.7 meter children have a range from 1.6 meters to 1.8 meters. So the standard error of my height measurements is about .05 meters The "separation" of my children is (true S.D./S.E.) = (.1 / .05) = 2. Two strata - exactly right. If my standard error had been larger, the two height ranges would have overlapped and the separation would have dropped. I would not have been able to distinguish tall children from short children. But if my standard error had been smaller, there would have been a gap between the two error distributions. The separation would have been bigger, perhaps 3 strata, alerting me that I could have distinguished a third strata of children of height 1.5 meters, if there had been any. It is the same with Rasch measures and their standard errors.

148 For Separation = 2, notice that the peaks of the S.E. curves are 3 units apart.

Strata: If we needed to discriminate 3 "strata", we could squeeze them in. The very top (with curve peak at 3), against the middle (with curve peak at 0), against the very bottom (with curve peak at -3).

So, a very high performer can be discriminated from a middle performer, can be discriminated from a very low performer. *But we would need to look at the empirical distribution of measures(in the "rulers" in Table 6) to see if the distribution does have long tails where those very high and very low performers would be located.*

Separation = 2 = 2 error distributions

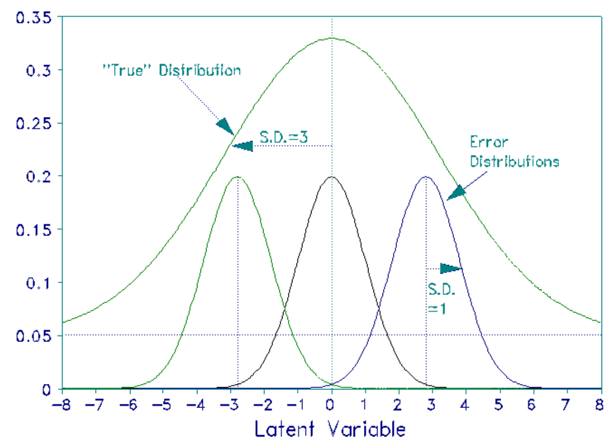


Strata = 3

149 Here's the same thing for **reliability = 0.9**, separation = 3. We can see how the narrower error distributions allow for more different measures to be squeezed into the "true" distribution.

Separation (True SD / Error SD) is more useful than reliability when reliabilities get much above 0.9. The maximum reliability is 1.0 so changes in reliability are not noticeable. Changes in the equivalent separation are always identifiable.

Strata: we could squeeze in another S.E. distribution curve, by placing the peaks at -4.5, -1.5, +1.5, -4.5



Separation =3 implies 3 error strata

150 Decision-makers say, *"We are going to use this instrument to discriminate x levels of performance."* We might respond, *"Fine! based on the sample measure distribution, and the separation pictures, this instrument can do that."* Or we might say, *"This test really does not have the discriminating power for x many levels of performance, so there will be a lot of mis-classification."*
In language testing it used to be common (maybe it still is) to try to discriminate 10 or so performance-levels based on a test that only has the statistical power to discriminate 3 or so levels. No one computed the standard errors for individual measures, so no one knew how arbitrary the classification of examinees was.

151 **How much reliability or separation do we need?** It depends what our purpose is, but we nearly always want to separate high performers from low performers, so a person separation of 2, reliability of 0.8, is often the benchmark for practical use. But reliabilities are always computed for the current sample, despite the convention of calling them the "test reliabilities". Next time the sample will be differently distributed, so the separation may be different .....

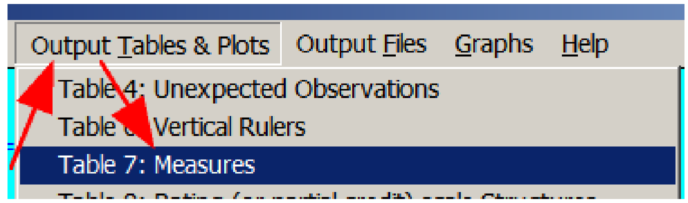
152.	**Measure summary chi-square statistics**

153.	There are two questions we may ask ourselves about the elements of a facet:
	
	1. Are the measures of the elements in a facet all statistically the same, except for measurement error? This particularly applies to raters. We want them to have the same leniency. This hypothesis is tested with the **"fixed (all same) chi-square."**
	
	2. Are the measures a random sample from a normal distribution? This particularly applies to large samples of persons. If they are, we can conveniently summarize them with a mean measure and a standard deviation. This hypothesis is tested with the **"Random (normal) chi-square"**.
	
	`Model, Fixed (all same) chi-square: 39.1  d.f.: 6  significance (probability): .00`
	`Model,  Random (normal) chi-square: 5.2  d.f.: 5  significance (probability): .39`
	
	In this example, the hypothesis that the elements have the same measure, apart from measurement error, has significance p=.00, so this hypothesis is *rejected.*
	The hypothesis that the measures are a random sample from a normal distribution has significance p=.39, so this hypothesis is *not rejected.*

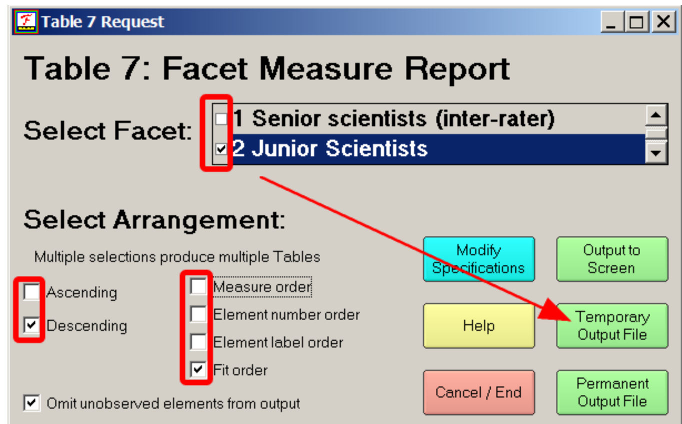
154 Rasch Fit Statistics: Are the Measures Accurate and Effective?

155 Let's produce Table 7 for "Junior Scientists", facet 2 in fit order, ascending:

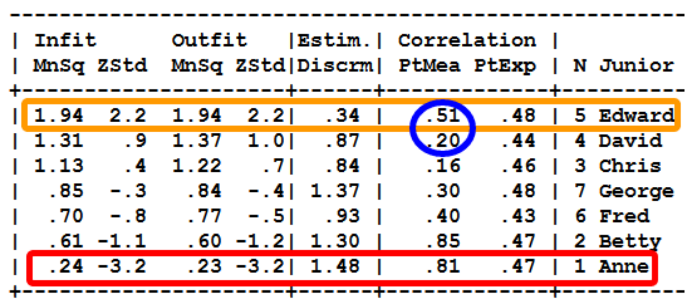
Facets analysis window for the Guilford.txt
Click on "Output Tables & Plots" menu
Click on "Table 7: Measures"



156 "Table 7 Request" dialog box:
Click on "All" to uncheck-mark it.
Scroll the list ...
Click on "2 Junior Scientists" to check-mark it
Click on "Measure" to uncheck-mark it
Click on "Fit order" check-mark it
Click on "Temporary Output File"



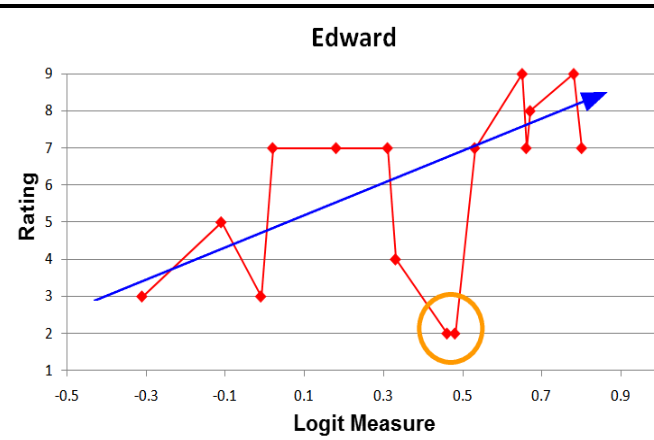
157 Here is the Table in a NotePad file.
Let us think about what this means ...
Orange box: Edward has mean-squares of 1.94, much larger than the expected 1.0. The ratings of Edward underfit the Rasch model. They are **too unpredictable from the Rasch measures or "noisy"**.
Blue circle: Edward has a high correlation - this usually means "predictable" - why?
Red box: Anne has mean-squares of .24 and .23, much lower than the expected 1.0. The ratings of Anne **overfit** the Rasch model. They are **too predictable from the Rasch measures or "muted"**.



```
---------------------------------------------------------------
| Infit      Outfit    |Estim.| Correlation |
| MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp  | N Junior
+-----------------------+------+-------------+-----------------
| 1.94  2.2  1.94  2.2|  .34 |  .51   .48  | 5 Edward
| 1.31   .9  1.37  1.0|  .87 |  .20   .44  | 4 David
| 1.13   .4  1.22   .7|  .84 |  .16   .46  | 3 Chris
|  .85  -.3   .84  -.4| 1.37 |  .30   .48  | 7 George
|  .70  -.8   .77  -.5|  .93 |  .40   .43  | 6 Fred
|  .61 -1.1   .60 -1.2| 1.30 |  .85   .47  | 2 Betty
|  .24 -3.2   .23 -3.2| 1.48 |  .81   .47  | 1 Anne
+-----------------------+------+-------------+-----------------
```

Always investigate underfit (high mean-squares) before overfit (low mean-squares). Often the overfit disappears when the underfit is eliminated from the data.

158 Blue circle: our investigation!
Hre is a plot of Edward's ratings and the logit measures that are modeled to produce them. If you have some skill with Excel, Appendix 3. Excel plots from the Residual file explains how to make this plots for yourself.
Blue line: a strong trend = high correlation.
Orange circle: two observations of "2" are surprising.
Overall, Edward's the ratings are much less predictable (from the Rasch measures) than the Rasch model expects.



159 **Edward:**
1. “Outfit MnSq = 1.94”, “Infit MnSq=1.94”.
The “Outfit mean-square” reports primarily about observations where the combined (summed) measures are far from zero.
The “Infit mean-square” reports primarily about patterns of observations where the combined (summed) measures are near to zero.
In Guilford.txt the rating scale is so long (9 categories) that the operational range of the rating scale for each item is much wider than the spread of the measures. Accordingly Outfit and Infit report essentially the same results. I prefer to report only Outfit, but some reviewers prefer Infit or both Infit and Outfit.

2. “MnSq = 1.94”
The mean-square is much greater than 1.0, so these ratings are too unpredictable. They underfit the Rasch model. They twice as much randomness as the model predicts.
Ben Wright explained fit like an old phonograph record.

> When the mean-square is close to 1.0, the music can be heard clearly.
> When the mean-square is much less than 1.0, the music is muted, muffled. It loses its rich tones.
> When the mean-square is much greater than 1.0, the music is there, but so are the pops, rumbles due to scratches and surface noise. When the mean-square is above 2.0, the noise is starting to overwhelm the music.

From the plot, we can that the data do not concur about Edward’s performance. The ratings in the orange circle say that Edward is a low performer, but other ratings say that he is a high performer. Whichever is correct, the estimated measure is a compromise, so it is an inaccurate estimate of Edward’s “true” measure.

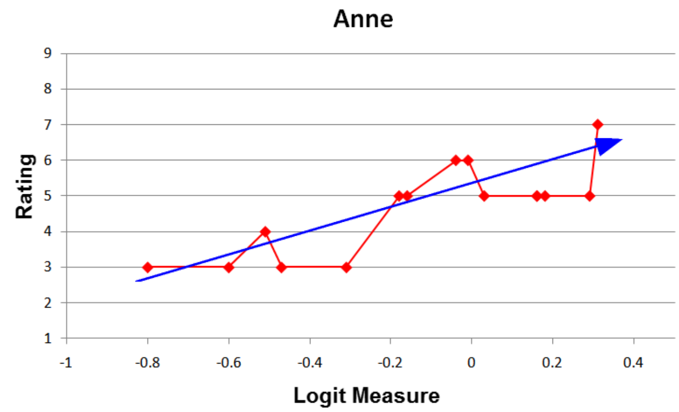
3. “Zstd = 2.2” in #157 - this is reporting the result of a statistical hypothesis test: *“These ratings conform to the Rasch model.”*
4. “Zstd = +...” - indicates that the ratings underfit (too much noise) the Rasch model
5. “= +2.2” - this value is a unit-normal deviate indicating the probability that these ratings conform to the Rasch model. It is unlikely ($p<.05$ in Appendix 1) that these ratings are the chance outcomes of a Rasch process based on the estimated measures.

6. “Do these data fit the Rasch model or not?” - the hypothesis test of fit to the Rasch model reports “They do not!”. They underfit the model: the mean-squares says the misfit is big, and the Zstd says that the misfit is unlikely to have happened by chance.

7. “What action do we take?” This depends on the circumstances.
A. The data aren’t perfect - but we expected that.
B. These data underfit the model. They are too unpredictable. Is that a cause of concern for us? Yes, the measure of Edward’s performance (based on these data) is inaccurate for practical purposes.
C. If this is our first look at the data, always examine high mean-squares (underfit) before low mean-squares (overfit). This is because the average mean-square is usually forced to be close to 1.0. So investigate Edward (MnSq = 1.94) before Anne (MnSq=0.24).
D. If we consider that the ratings in the orange circle are not representative of Edward’s general performance, we might omit Edward from the analysis, or omit those ratings. Then Edward’s idiosyncrasies won’t impact other aspects of the analysis, such as Anne’s mean-square. A later tutorial will show us how we can anchor (fix) the other measures at their good values, and measure Edward with all his ratings.
E. If this is a diagnostic test, then the orange-circled ratings may be the most important ones. They tell us where to focus our remedial action for Edward to improve his performance.

160 And here is a plot of Anne's ratings and the logit measures that are modeled to produce them. Anne has a high correlation and the highest overfit.

Notice how closely Anne's rating track along the trend line. They are more predictable (from the measures) than the Rasch model expects.



161 **Anne:**

2. "MnSq = 0.24"
The mean-square is much less than 1.0, so these ratings are too predictable. They overfit the Rasch model. They only contain one-quarter of the randomness that the model predicts. In Classical Test Theory (CTT) this would be considered good. In Rasch this indicates that these ratings contain only 24% of the measurement information that they should. These ratings are inefficient and will cause the reported standard errors to be too small and the reported reliabilities to be too high. But the measure of Anne's performance (based on these data) is accurate.

3. "Zstd = -3.2" - this is reporting the result of a statistical hypothesis test: "These ratings conform to the Rasch model."
4. "Zstd = -" - indicates that the ratings overfit.
5. "= -3.2" - this value is a unit-normal deviate indicating the probability that these ratings conform to the Rasch model. It is extremely unlikely ($p<.01$ in Appendix 1) that these ratings are the chance outcomes of a Rasch process based on the estimated measures.

6. "Do these data fit the Rasch model or not?" - the hypothesis test of fit to the Rasch model reports "They do not!". They overfit the model, highly statistically significantly.

7. "What action do we take?" This depends on the circumstances.
A. The data aren't perfect - but we expected that.
B. These data overfit the model. They are too predictable. Is that a cause of concern for us? Yes, if it is a roulette wheel. But usually No if it is the performance of a child on an educational test. The measure of Anne's performance (based on these data) is **accurate for practical purposes**.
C. If this is a standard testing situation, then overfit slightly stretches the measures (increases their range), inflates their reliability and reduces their standard errors. These are technical issues usually not of concern to anyone other than psychometricians. So it would require very strong external motivation to omit or alter this set of ratings.
D. If this a rater training situation, low mean-squares are typical of raters "playing it safe" by exhibiting central tendency or trying to agree with the ratings they think the other raters will give. This is often the result of training which emphasizes "if you disagree with the other raters too much, you will be fired!" So, before being concerned about the individual, review the training material and the instructions given to the raters. Are they explicitly or implicitly being told to agree with each other?
At the Olympic Ice-Skating, the organizers think that "rater agreement = more credibility", but to psychometricians, "excessive rater agreement = psychological pressure to agree = loss of objectivity and fairness".

162 **Q. Table 8: Rating Scale Structure**

163 Table 8 tells us about the 9-category rating scale of Creativity. It is packed with useful information about the success of our data collection - information which J.P. Guilford completely overlooked when he wrote the chapter on rating scales in his book, *"Psychometric Methods"*.

164

```
Model = ?B,?B,?,CREATIVITY
Rating (or partial credit) scale = CREATIVITY,R9,G,O
---------------------------------------------------------------------------------------------------------------
|     DATA                     | QUALITY CONTROL     |   STEP             | EXPECTATION      |  MOST  |.5 Cumul.| Cat|Response|
| Category Counts      Cum.| Avge  Exp.  OUTFIT|CALIBRATIONS | Measure at       |PROBABLE|Probabil.|PEAK|Category|
|Score   Used    %       % | Meas  Meas  MnSq |Measure  S.E.|Category  -0.5 |  from   |   at     |Prob|  Name  |
---------------------------------------------------------------------------------------------------------------
|  1        4   4%    4%| -.86  -.74   .8 |               |( -2.70)          |  low    |  low     |100%| lowest |
|  2        4   4%    8%| -.11  -.58  2.7 |  -.64    .53| -1.65   -2.21|         |  -1.75  | 17%|        |
|  3       25  24%   31%| -.36* -.40   .9 | -2.32    .39|  -.93   -1.26|  -1.48  |  -1.39  | 48%|        |
|  4        8   8%   39%| -.43* -.22   .5 |   .83    .25|  -.41    -.66|         |   -.46  | 11%|        |
|  5       31  30%   69%| -.04  -.03   .8 | -1.48    .24|   .02    -.19|   -.32  |   -.29  | 39%| middle |
|  6        6   6%   74%| -.46*  .17  4.1 |  1.71    .25|   .44     .23|         |    .34  |  9%|        |
|  7       21  20%   94%|  .45   .34   .6 | -1.00    .26|   .94     .68|    .35  |    .47  | 47%|        |
|  8        3   3%   97%|  .75   .50   .5 |  2.36    .44|  1.62    1.24|         |   1.37  | 16%|        |
|  9        3   3%  100%|  .77   .62   .8 |   .54    .60|(  2.69)  2.17|   1.45  |   1.70  |100%| highest|
-----------------------------------------------------------(Mean)---------(Modal)--(Median)-------------
```

165 Blue box: There are 9 categories, 1 to 9. Look at how they have been "Used", the category frequency counts. Do you notice anything conspicuous? *Yes you do!* Only categories 3, 5 and 7 have large counts. Perhaps the judges, the Senior Scientists, could only discriminate 3 levels of Creativity, but were told to use a 9-category scale. **Over-categorization** leads to artificially reduced standard errors, inflated reliabilities and poor fit to the Rasch model.

166 Red box: We can see evidence of poor fit in the "Average Measure" column. The rating scale is intended to represent a series of qualitative advances along the latent variable. Each category is assumed to be a quantitative advance (of a size yet to be determined) beyond the previous category. So, **higher categories should imply higher measures, and higher measures should be observed as higher categories.** *But did this happen?*

167 The "Average Measures" are the averages of the measures that combined to produce the observations in the category. We expect them to advance with category number.

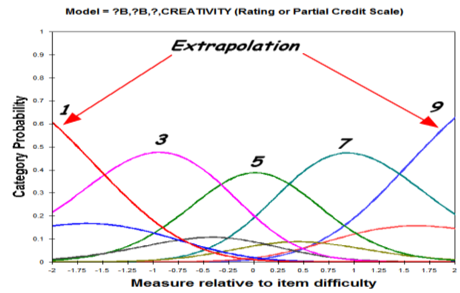
Our estimation process in Tutorial 1:
$B_n - D_i - R_r - S_s - \{F_k\} \rightarrow X_{nirs}$
Average Measure for category "j" =
Average $(B_n - D_i - R_r - S_s)$ for all $X_{nirs} = j$

168 In Table 8, we can see that the Average Measure for category 1 is **-.86.** Then for category 2 it is **-.11.** *Good so far,* categories and average measures are advancing together. But the Average Measure for Category 3 is **-.36. The Average Measure has gone backwards, and so is flagged with "*".** *This contradicts our theory about the rating scale.*
Green box: the "Expected Measure" column shows what the Average Measures would be if the data fit the Rasch model. We can see big differences, particularly for Category 6 (-.46 vs. .17). Something is seriously wrong. *What is it?*

169 Orange box: Look at the category-level "Outfit MnSq" column. We expect these to be 1.0 or less due to dependency among the categories of the rating scale. Most mean-squares are in this range. But Category 2, mean-square 2.7, and Category 6, mean-square 4.1, are showing considerably unpredictability. *Another symptom that something has gone seriously wrong with the functioning of the rating scale!*

170 Let's look at the rating scale from a graphical perspective:
On the *Facets* menu bar,
Click "Graphs"

171 The Rasch model has made as much sense of the rating scale as it can.
The x-axis is the latent variable, drawn relative to the difficulty of the item.
The y-axis is the probability of observing each category of the 9-category rating scale.
Categories 3, 5, 7 are observed more often, so they have higher probability curves. Categories 1 and 9 are the extreme categories, so the Rasch model extrapolates that they are the categories most probable to be observed outside of the range of the data. The rest of the categories have low probability of being observed.

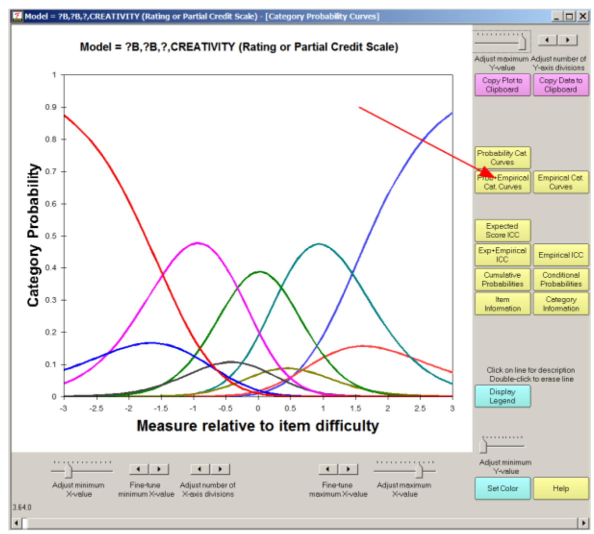


If you want to know which category a curve represents, click on the curve.
(Click on the plot background if the plot is not redrawn correctly.)

172 On the "Graphs" window,
Click on "Prob+Empirical Cat. Curves"

Prob = Probability (as predicted by the Rasch model)

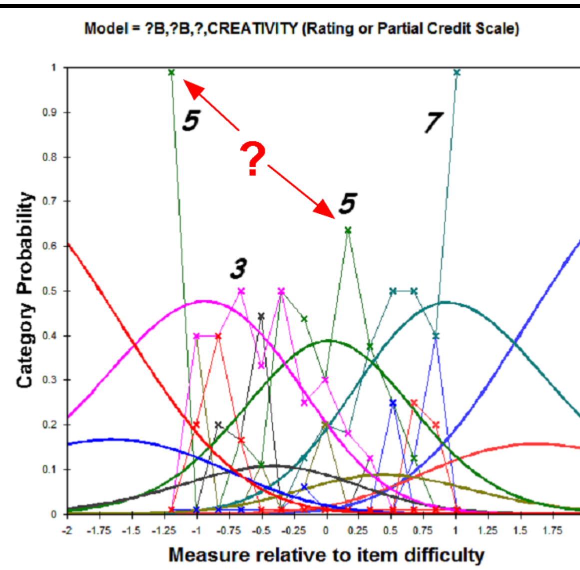
Empirical = Observed (as summarized from the dataset)

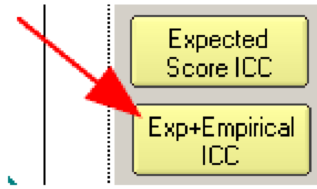
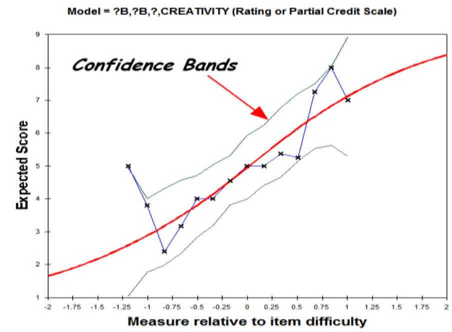


173 Prob+Empirical Cat. Curves:

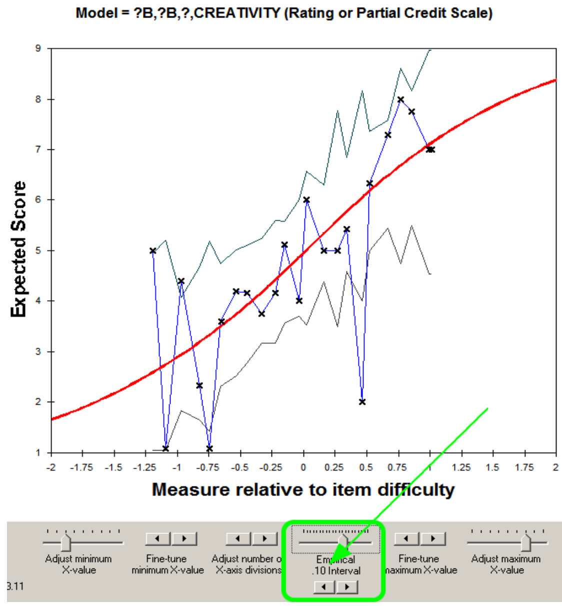
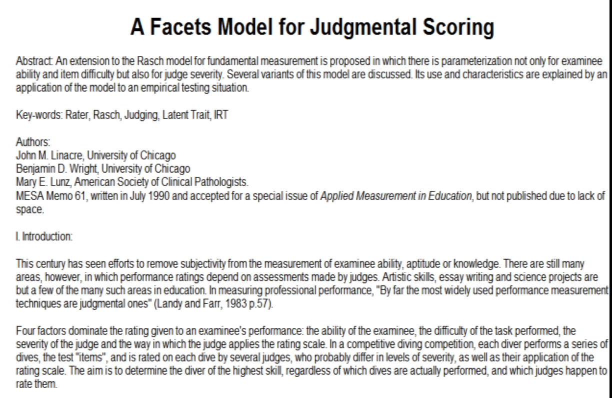
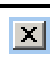
The thinner lines with x's are the empirical category frequency lines, summarizing how the rating scale categories were used. Their colors match the Rasch-model smooth curves.

Do you see that categories 3, 5, 7 are the only high-frequency categories? **Category 5** peaks in the center, where it should, but also down at the bottom - *weird!*



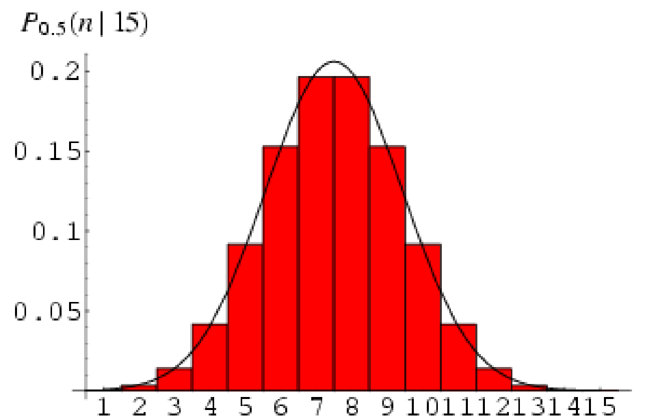
<table>
<tr><td>174</td><td>Click on “Exp+ Empirical ICC”. This will display the expected (Rasch-model) and empirical (what the data say) Item Characteristic Curves, ICCs. These show the average functioning of the rating scale along the latent variable.</td><td>
</td></tr>
<tr><td>175</td><td>The solid red line is the Rasch-Model ideal ICC for these data. The blue jagged line is what the data say. The green 95% Confidence Bands are the statistical limits of the divergence of the empirical from the ideal, as predicted by the Rasch model. We can see that the data only just remain within the lines. There is a problem at the bottom end of the empirical ICC, matching the problem with category 5 which we saw in its empirical category curve.</td><td>
</td></tr>
<tr><td>176</td><td>Green box and arrow: Move the slider below the plot to make the empirical summarizing-interval 0.10 logits. You will see that now the empirical blue line crosses over the confidence bands, which are two-sided 95% confidence intervals.

Even at this level of summarization, the misfit in the rating data are apparent, suggesting that the misfit should be investigated in greater detail in other Tables, such as Table 4. Something is seriously wrong with this Guilford dataset. J.P. Guilford did not notice it himself, but we will discover exactly what it is in the next Tutorial.

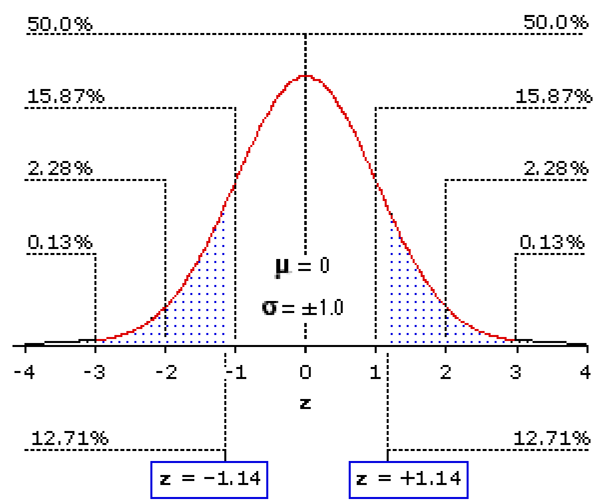
Play with the “Graphs” screen, clicking different buttons and different slider settings. Do you see anything intriguing or diagnostically useful for you?</td><td>
</td></tr>
<tr><td>177</td><td></td><td></td></tr>
<tr><td>178</td><td>Optional Reading:
#14 - Knox’s “Cube Imitation” Test - http://www.rasch.org/rmt/rmt133j.htm
#62 - Wright & Panchapakesan (1969) “A Procedure for Sample-Free Item Analysis” - http://www.rasch.org/memo46.htm
For a conceptual summary of what we have done so far, and also a glance ahead, please read “A Facets Model for Judgmental Scoring” - http://www.rasch.org/memo61.htm</td><td>
A Facets Model for Judgmental Scoring
Abstract: An extension to the Rasch model for fundamental measurement is proposed in which there is parameterization not only for examinee ability and item difficulty but also for judge severity. Several variants of this model are discussed. Its use and characteristics are explained by an application of the model to an empirical testing situation.
Key-words: Rater, Rasch, Judging, Latent Trait, IRT
Authors:
John M. Linacre, University of Chicago
Benjamin D. Wright, University of Chicago
Mary E. Lunz, American Society of Clinical Pathologists.
MESA Memo 61, written in July 1990 and accepted for a special issue of Applied Measurement in Education, but not published due to lack of space.
I. Introduction:
This century has seen efforts to remove subjectivity from the measurement of examinee ability, aptitude or knowledge. There are still many areas, however, in which performance ratings depend on assessments made by judges. Artistic skills, essay writing and science projects are but a few of the many such areas in education. In measuring professional performance, "By far the most widely used performance measurement techniques are judgmental ones" (Landy and Farr, 1983 p.57).
Four factors dominate the rating given to an examinee's performance: the ability of the examinee, the difficulty of the task performed, the severity of the judge and the way in which the judge applies the rating scale. In a competitive diving competition, each diver performs a series of dives, the test "items", and is rated on each dive by several judges, who probably differ in levels of severity, as well as their application of the rating scale. The aim is to determine the diver of the highest skill, regardless of which dives are actually performed, and which judges happen to rate them.</td></tr>
<tr><td>179</td><td>Close all windows.</td><td></td></tr>
</table>

Appendix 1. Unit Normal Deviates

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”.



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(μ, σ). Look at the plot, the values along the x-axis at labeled “z”, these are unit normal deviates. The area under the red curve indicates the probability of observing those values. http://faculty.vassar.edu/lowry/ch6pt1.html
68% of the area under the is within 1 S.D. of the mean, so we expect about 2/3 of the values we observe to be statistically close to the mean.



We are usually concerned about values far away from the mean on either side (2-sided). The Figure it says that 2.28% of the area under the curve is to the right of +2, and 2.28% is less than -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) that are conventionally regarded as indicating statistical significance.

The precise value of probability < .05 is	z > |±1.96| for p < .05
and for probability < .01 is	z > |±2.58| for p < .01

Handy table of unit normal deviates (z) and probabilities (p) for a “two sided z-test”, also called a “two-sided *t*-test with infinite degrees of freedom”

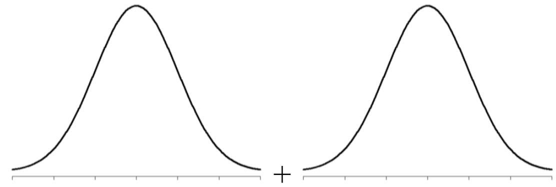
Zstd values also use this probability table:

z >	p <
±2.58	0.01
±2.33	0.02
±2.17	0.03
±1.96	0.05
±1.64	0.10
±1.28	0.20
±1.04	0.30
±0.84	0.40
±0.67	0.50

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?"***	
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) $\rightarrow N(0,1)$
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)	$\{R_{ni} / V_{ni}^{0.5}\} = \{Z_{ni}\} \rightarrow N(0,1)$

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

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?

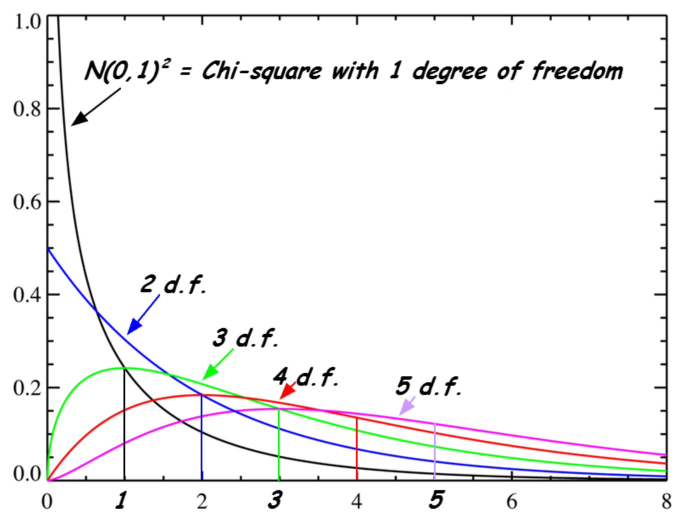


Add two unit-normal distributions:
N(0,1) + N(0,1) = N(0, 2)
The average stays the same, but they spread out more. The combined distribution has twice the variance of .a unit-normal distribution:

But what if we square the unit- normal distribution? $N(0,1)^2$ 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 add two of these $N(0,1)^2 + N(0,1)^2 = \chi^2_2$. This has two degrees of freedom, d.f., and is shown by 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).



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

Facets reports the significance (probability) of a mean-square as a unit-normal deviate (Zstd).

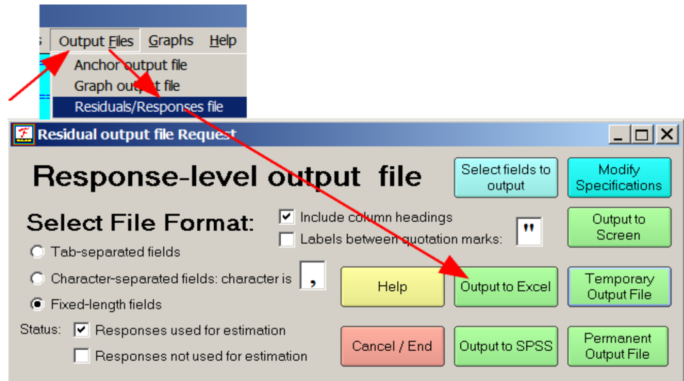
ZStd = Wilson-Hilferty (mean-square, d.f.)
see http://www.rasch.org/rmt/rmt162g.htm

Appendix 3. Excel plots from the Residual file

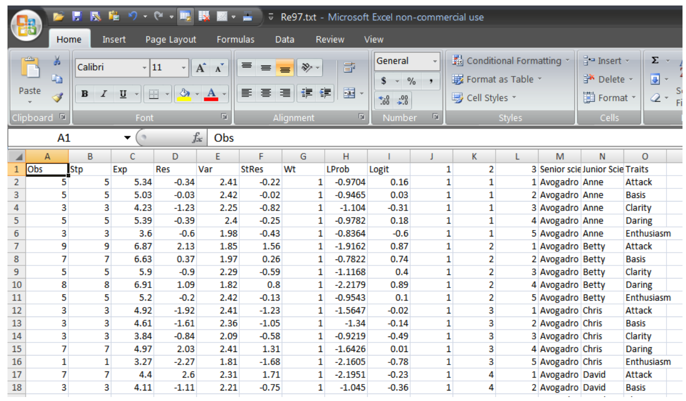
Here is how you can produce Excel plots from the Facets "Residuals file". This needs some skill with Excel.

Facets Analysis window
Click on "Output Files"
Click on "Residuals/Responses file"

"Residual output file Request" dialog box
Click on "Output to Excel"



Excel worksheet:
Click on the worksheet
"Select All" Ctrl+A
"Sort and Filter"
"Custom Sort"
Top row is headings
Sort fields:
The facet number you want, ascending, e.g., "2"
The x-axis value you want, ascending, e.g., "Logit"
OK
The worksheet is sorted

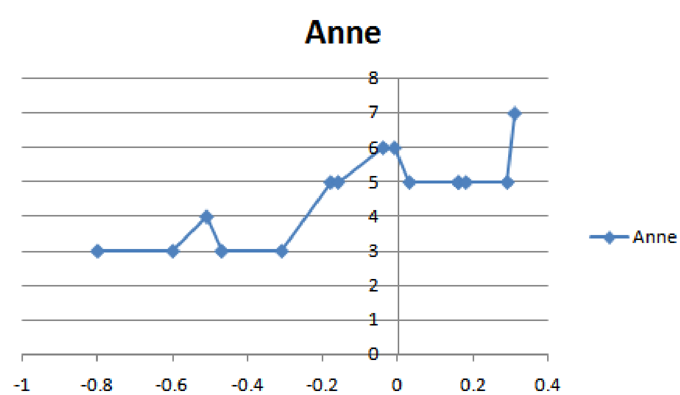


Obs	Stp	Exp	Res	Var	StRes	Wt	LProb	Logit	1	2	3	Senior scie	Junior Scie	Traits
5	5	5.34	-0.34	2.41	-0.22	1	-0.9704	0.16	1	1	1	Avogadro	Anne	Attack
5	5	5.03	-0.03	2.42	-0.02	1	-0.9465	0.03	1	1	2	Avogadro	Anne	Basis
3	3	4.23	-1.23	2.25	-0.82	1	-1.104	-0.31	1	1	3	Avogadro	Anne	Clarity
5	5	5.39	-0.39	2.4	-0.25	1	-0.9782	0.18	1	1	4	Avogadro	Anne	Daring
3	3	3.6	-0.6	1.98	-0.43	1	-0.8364	-0.6	1	1	5	Avogadro	Anne	Enthusiasm
9	9	6.87	2.13	1.85	1.56	1	-1.9162	0.87	1	2	1	Avogadro	Betty	Attack
7	7	6.63	0.37	1.97	0.26	1	-0.7822	0.74	1	2	2	Avogadro	Betty	Basis
5	5	5.9	-0.9	2.29	-0.59	1	-1.1168	0.4	1	2	3	Avogadro	Betty	Clarity
8	8	6.91	1.09	1.82	0.8	1	-2.2179	0.89	1	2	4	Avogadro	Betty	Daring
5	5	5.2	-0.2	2.42	-0.13	1	-0.9543	0.1	1	2	5	Avogadro	Betty	Enthusiasm
3	3	4.92	-1.92	2.41	-1.23	1	-1.5647	-0.02	1	3	1	Avogadro	Chris	Attack
3	3	4.61	-1.61	2.36	-1.05	1	-1.34	-0.14	1	3	2	Avogadro	Chris	Basis
3	3	3.84	-0.84	2.09	-0.58	1	-0.9219	-0.49	1	3	3	Avogadro	Chris	Clarity
7	7	4.97	2.03	2.41	1.31	1	-1.6426	0.01	1	3	4	Avogadro	Chris	Daring
1	1	3.27	-2.27	1.81	-1.68	1	-2.1605	-0.78	1	3	5	Avogadro	Chris	Enthusiasm
7	7	4.4	2.6	2.31	1.71	1	-2.1951	-0.23	1	4	1	Avogadro	David	Attack
3	3	4.11	-1.11	2.21	-0.75	1	-1.045	-0.36	1	4	2	Avogadro	David	Basis

Scatterplot:

"Series name" is the element label you want in facet 2.
x-axis values are the "Logits" (or whatever) for the element you want in facet 2
y-axis values are the "Obs(ervations)" (or whatever) for the element you want in facet 2

Excel produces a plot



Use Excel tools to customize your plot.

Be exuberant! With a little time and talent, this becomes fun.