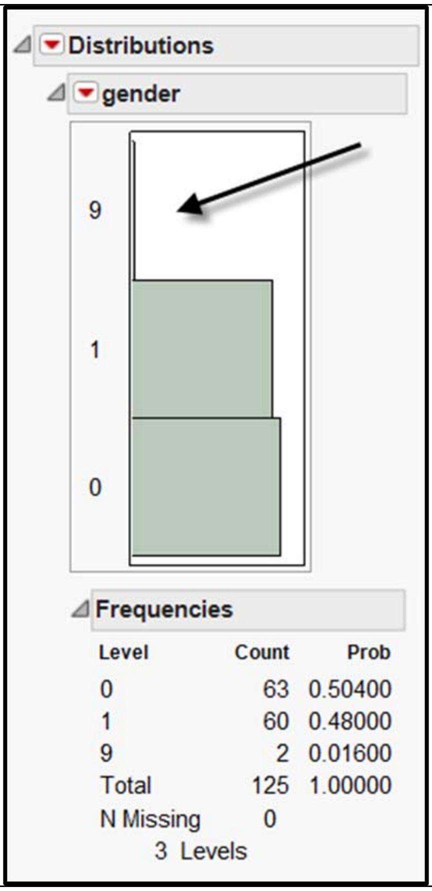
Comparing IRT scores and raw scores in JMP

Chong Ho Yu, Ph.D. (2014)
Azusa Pacific University
chonghoyu@gmail.com
http://www.creative-wisdom.com/computer/sas/sas.html

The objective of this write-up is to explain why scoring by Item Response Theory (IRT) is better than using the conventional raw scores. Although JMP has a nice graphical user face and its learning curve is not as steep as that of Winsteps, RUMM, and Bilog, please keep in mind that JMP is a general-purposed statistical package rather than a specialized assessment tool, and thus its IRT information is limited. Nevertheless, for those who want immediate results without going through scripting and programming, JMP is a good start. Also, as a general package, JMP allows you to do other things, such as computing descriptive and inferential statistics, data mining, six-sigma, experimental design, and many others.

In the following I will use a hypothetical data set consisting of 125 observations. This is the result of a 21-item test taken by students from three universities: Azusa Pacific University (APU), California State University (CSU), and the University of California at Los Angeles (UCLA).

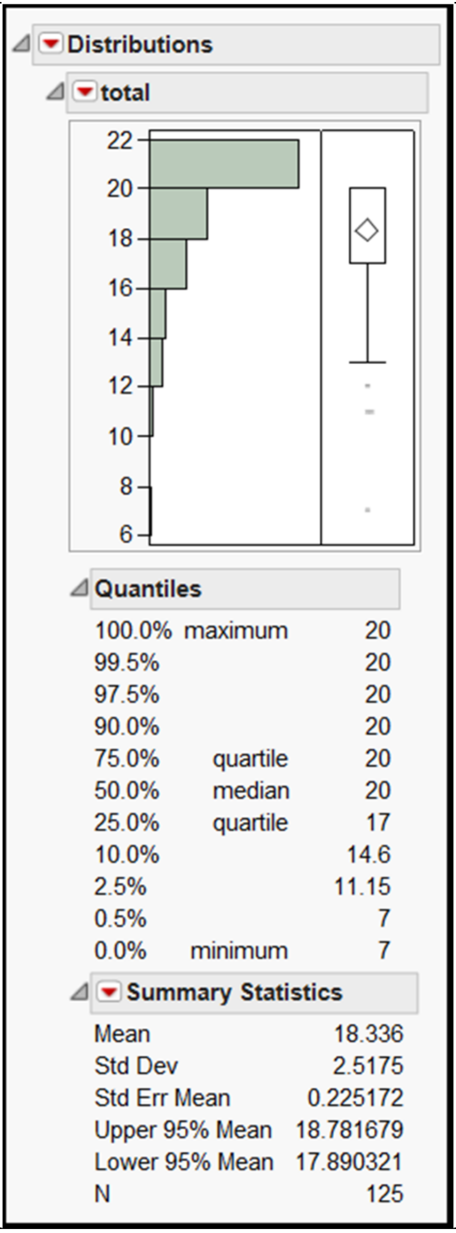
Before any analysis is performed, one must verify that the data are clean. The easiest way to do so is simply plotting the distribution of each variable, including demographic variables and item responses. If there is any anomaly, you can spot it right away. For example, the gender plot should show males and females only. If there is a "third sex," you have to go back to the original data to clean that up. By the same token, every item score is either "1" or "0." If there is a "9," chances are "9" denotes a missing value. To plot the data, go to Analyze - Distribution. Next, select the variables that you want to visualize to the Y box. You can select a block of variables by holding the shift key. Alternatively, you can select multiple variables by holding the control key.



Frequencies

Level	Count	Prob
0	63	0.50400
1	60	0.48000
9	2	0.01600
Total	125	1.00000
N Missing	0	

3 Levels



```
Quantiles
100.0%  maximum      20
99.5%                20
97.5%                20
90.0%                20
75.0%     quartile   20
50.0%     median     20
25.0%     quartile   17
10.0%              14.6
2.5%              11.15
0.5%                  7
0.0%      minimum     7

Summary Statistics
Mean               18.336
Std Dev            2.5175
Std Err Mean     0.225172
Upper 95% Mean  18.781679
Lower 95% Mean  17.890321
N                     125
```

By looking at the histogram and the descriptive statistics alone, I can tell that reporting raw scores will be problematic. The distribution is negatively skewed. The median is 20 whereas the mean is 18.336. In other words, most students did very well and needless to say the test is very easy.

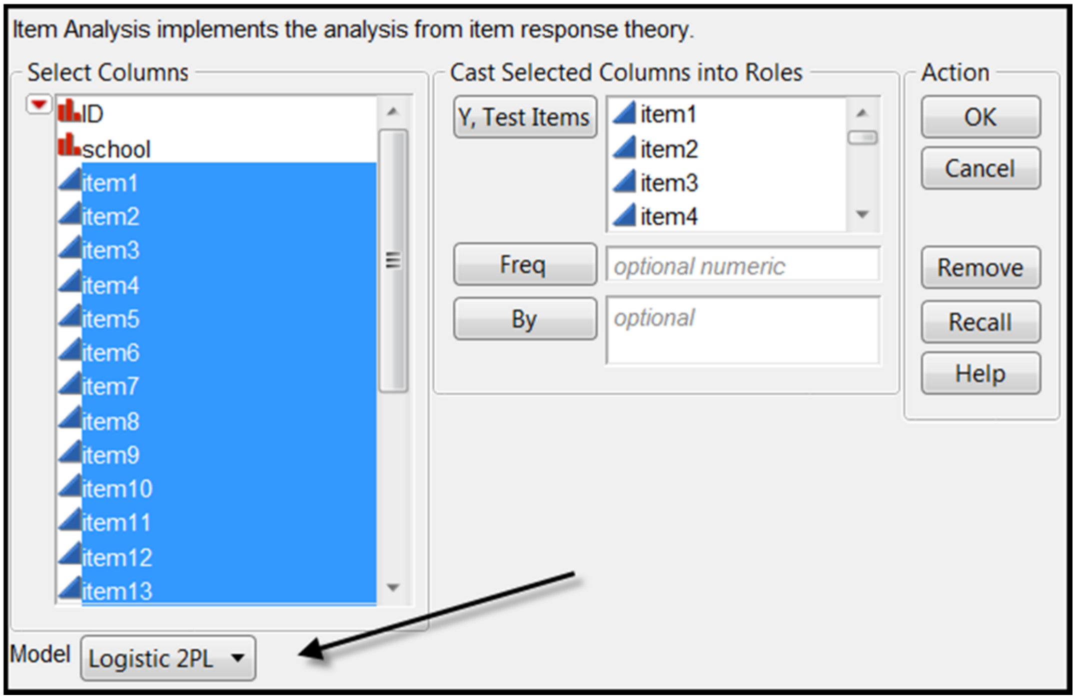
It is a common practice for instructors to adjust the curve when most examinees receive poor scores. As a result, their grades are typically improved by one letter (e.g. C→B, B→A). At first glance it sounds reasonable, but indeed it is unfair.

Students demand norming and curving when the test is difficult or their scores are not desirable. But in this example when the majority achieved high scores, should all the grades be adjusted downward by one letter? I can foresee immediate protest.

Another common approach is that the instructor looks at the scantron statistics to spot difficult items. For example, if 80% of the examinees fail a particular question, the instructor will give full credit to everyone. First, it is unfair to those well-prepared students who answered the tough item correctly. Second, when some question is so easy that 90% of the students could score it, would the grader take away the point from the student?

There is a better alternative to the preceding two approaches: IRT –scoring. Item Response Theory estimates the ability of the examinees by taking item difficulty into account. For more information, please visit http://www.creative-wisdom.com/multimedia/IRTTHA.htm. This ability estimate is also known as theta.

To run IRT, go to **Analyze – Consumer Research – Item analysis**. Select all test items, drag, and drop them into “Y, Test Items.” In the model pop-up menu, select 1-, 2-, or 3-parameter logistic (PL) model (In this example I used the 2 PL model). As mentioned before, JMP is not as versatile as Winsteps, RUMM, and Bilog. You are confined to use logit and no option for probit is available. Next, click on **OK**.

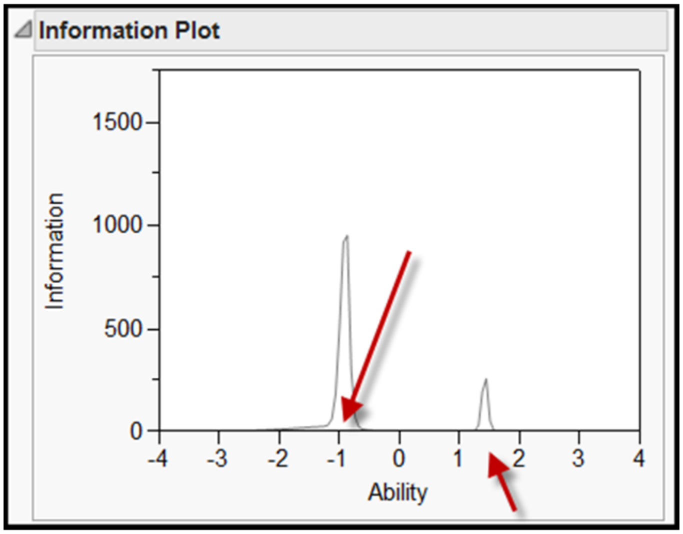


Item Analysis implements the analysis from item response theory.

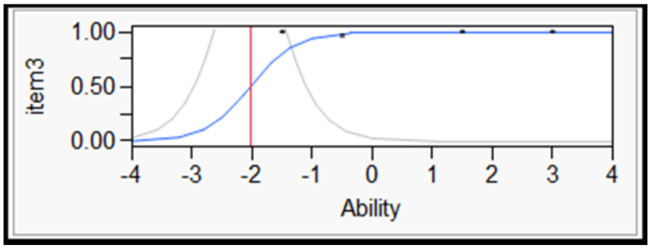
Y, Test Items: item1, item2, item3, item4

Model: Logistic 2PL

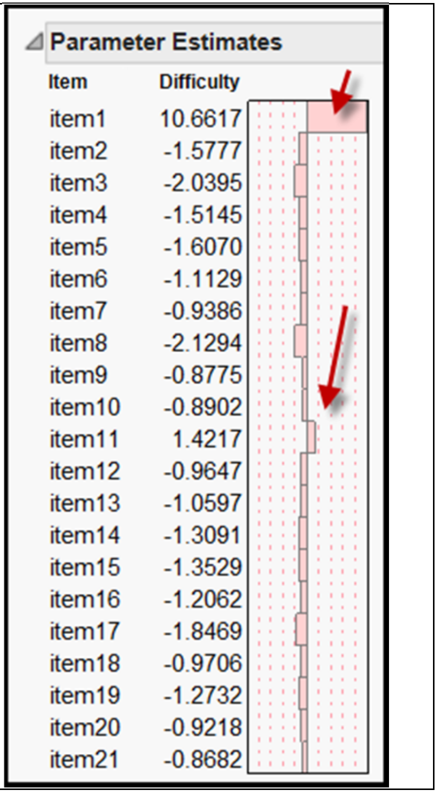
In the resulting page, you can show or hide information by clicking on the blue triangle, and request more information by clicking on the red triangle. There is an item characteristic curve (ICC) and item information function (IIF) for each item. And there is the test information function (TIF) for all items together. ICC tells you the probability of answering the item correctly at different levels of student ability, whereas IIF informs you how much reliable information about the student you can obtain at different levels of student ability. TIF is simply the sum of all IIFs in a test. For more information about ICC, IIF, and TIF, please consult http://www.creativewisdom.com/multimedia/IRTTHA.htm or http://www.creativewisdom.com/computer/sas/IRT.pdf . The following plot is the TIF. If this test is given, the most reliable information that we can obtain is the information from the students whose ability is around -1 or +1.5.



Please note that in JMP each ICC has a vertical red line. The red line shows the intersection of the probability and the ability when P = .5. In Question 3, students whose ability level is about -2 have 0.5 probability of answering the item correctly. In other words, if the item is easy, the red line leans toward left; if it is hard, it leans toward right. By looking at the location of the red line, the user can tell which item is a challenger and which one is a give-away.



IRT centers the degree of item difficulty at zero. Any item that has a difficulty value below 0 is considered easy whereas any question that has a difficulty parameter above 0 is regarded as challenging. In this test almost all items are easy, as indicated by their negative difficulty parameters, except Item 1 and Item 11. Consider this scenario: two students have the same scores (e.g. 18). If student A scored Item 1 and Item 11, but Student B missed both, who is a better student? The answer is obvious.

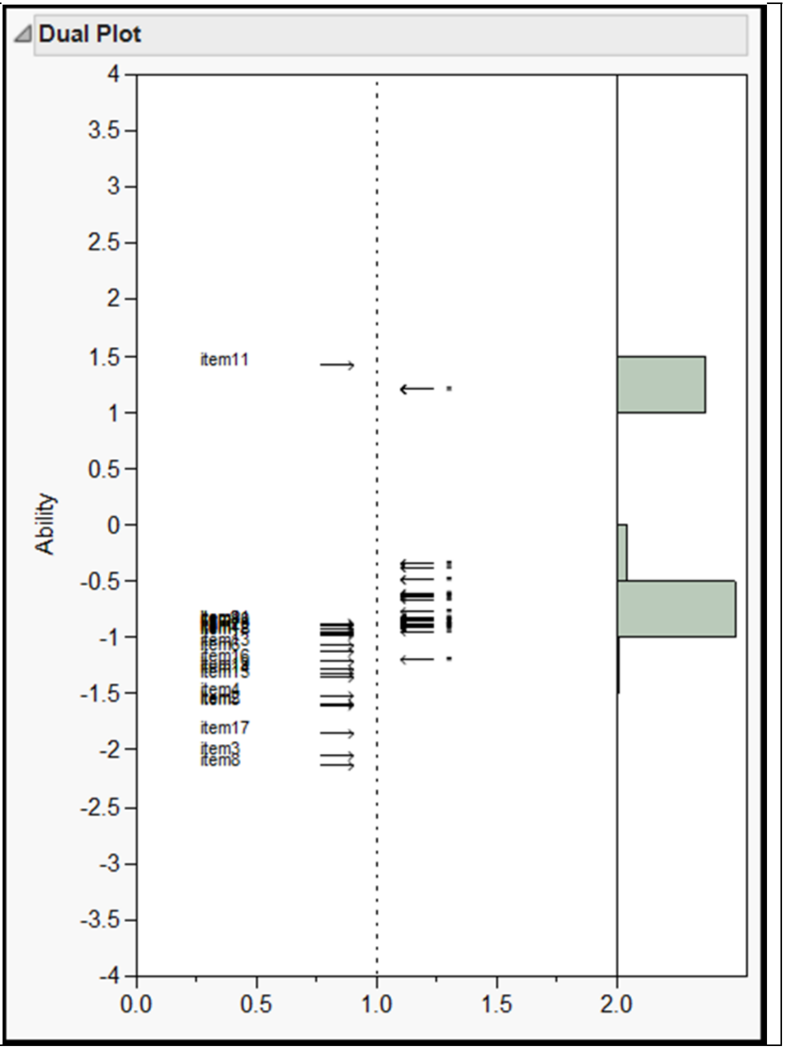


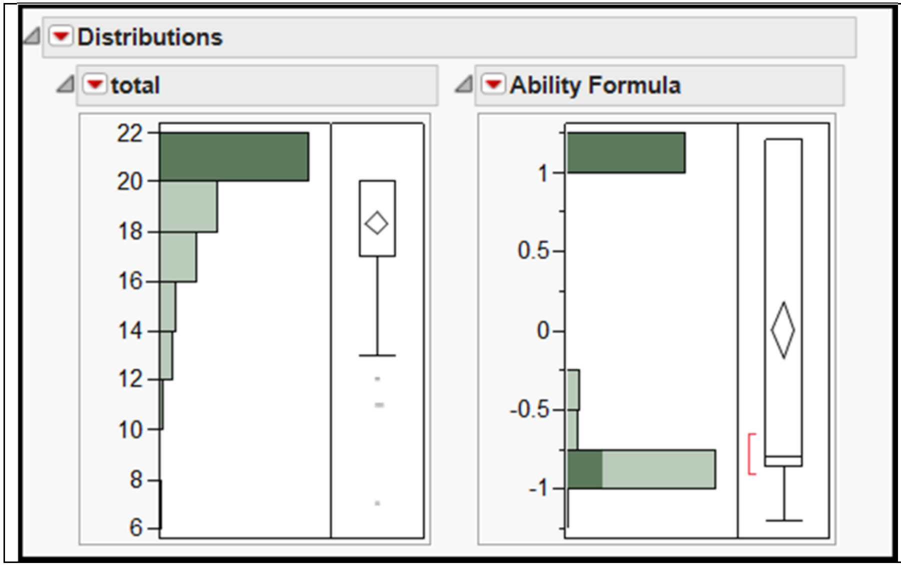
Parameter Estimates

Item	Difficulty
item1	10.6617
item2	-1.5777
item3	-2.0395
item4	-1.5145
item5	-1.6070
item6	-1.1129
item7	-0.9386
item8	-2.1294
item9	-0.8775
item10	-0.8902
item11	1.4217
item12	-0.9647
item13	-1.0597
item14	-1.3091
item15	-1.3529
item16	-1.2062
item17	-1.8469
item18	-0.9706
item19	-1.2732
item20	-0.9218
item21	-0.8682

The graph on the right hand side is a dual plot, which is equivalent to the item-person map (IMP) in Rasch Unidimensional Measurement Modeling (RUMM). The attributes of all items and students are re-scaled in terms of logit, and therefore they can be compared side by side. JMP's graphs are dynamic and interactive. Logit is the natural log if the odds ratio. If you want to identify the students who are above average (> 0), you can select the points and the corresponding rows in the spreadsheet are highlighted simultaneously.

Typically, the primary goal of IRT is to examine the quality of items. One should not make a firm judgment about student ability until items are validated. Nevertheless, one can conduct initial analysis using the ability estimates yielded from IRT modeling. To append the ability estimates to the original table, click the red triangle and choose **Save Ability Formula**.

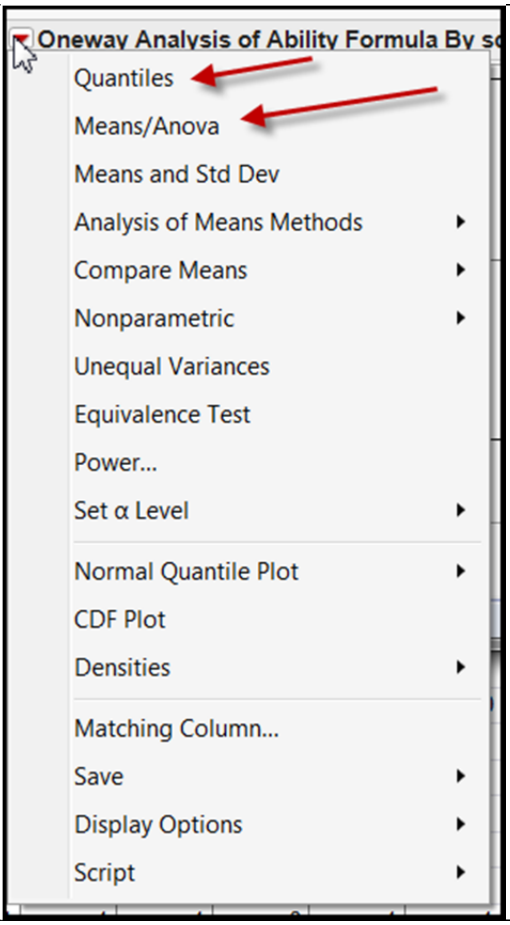




Theta (ability estimate) and raw scores are not necessarily corresponding to each other. The panel on the lefts shows the histograms of raw score (sum total of all 21 items) and ability. While most students who earned 20 points (highlighted bar) have the highest estimated ability, some of them are classified as average or low ability (between -0.5 and -1)!

After the ability estimates are saved, one can perform various exploratory analyses. For example, to detect whether there is a significant performance gap between different school, one can use Fit Y and X by putting ability formula into Y and school into X. To show the boxplots and the diamond plots, select **Quantiles** and **Means/ANOVA** from the red triangle.

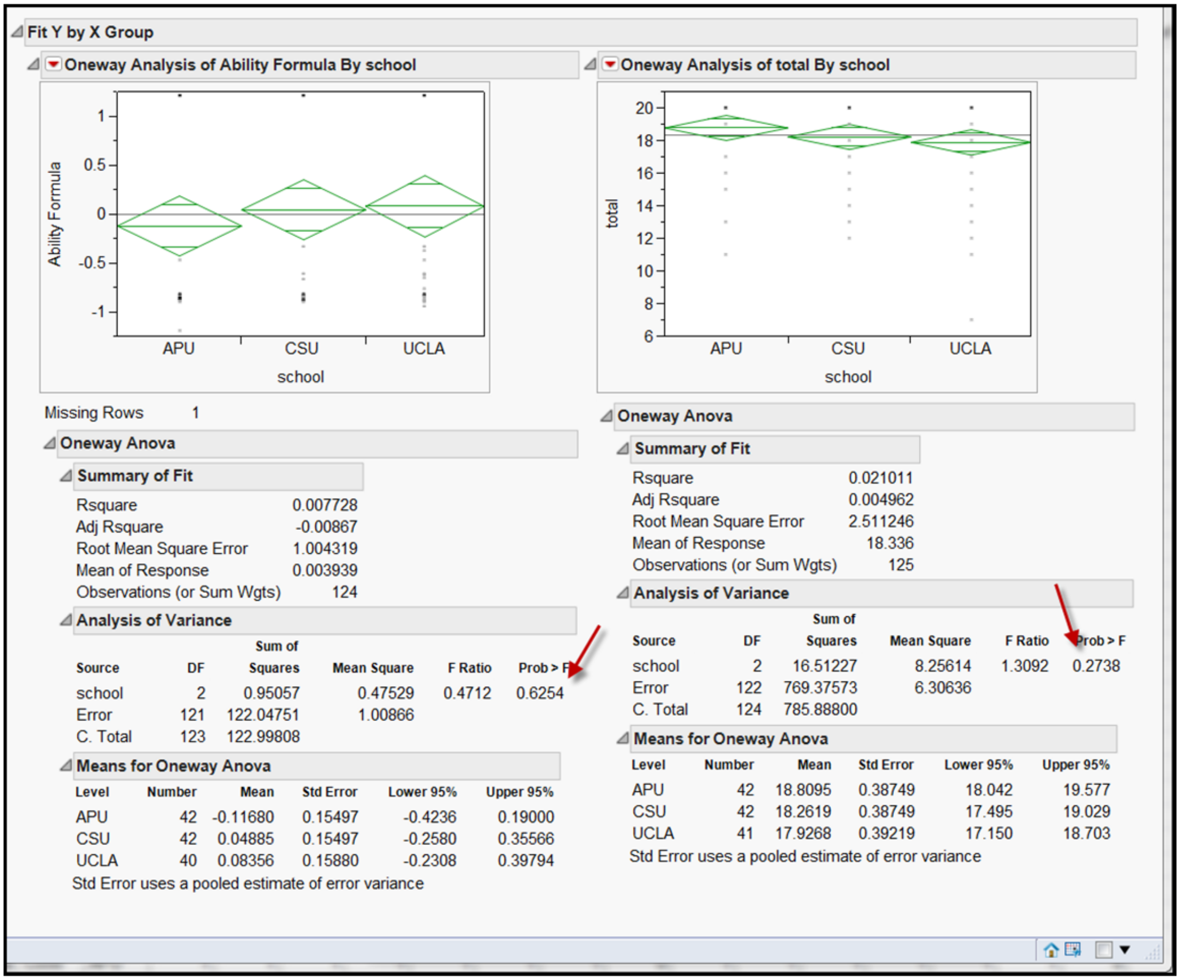
Although parametric procedures such as t-tests and F-tests are widely used for between-group comparison, these procedures based upon centrality may mislead the researcher, especially in the case of heterogeneity of variance. As a remedy, more and more researchers endorse the use of confidence intervals (CI). By using CI, the researcher not only looks at the group differences by means, but also by variability. JMP provides a powerful tool named diamond plot to visualize variability, as demonstrated in the following diamond plots.



The diamond plot condenses a lot of important information:

- **Grand mean**: represented by a horizontal line. In IRT ability estimates, the mean is always zero.
- **Group means**: the horizontal line inside each diamond is the group means.
- **Confidence intervals**: The diamond is the CI for each group.

Data analysis utilizing theta and that using raw scores would yield vastly different results. The following panel displays two ANOVA results and diamond plots based on theta and raw scores, respectively. On the right hand side, APU students outperform their counterparts in CSU and UCLA in terms of raw scores, but the conclusion is reversed when theta is used on the left hand side. But don't take it seriously. First, both p values are not significant. Second, these are only hypothetical data.



Fit Y by X Group

Oneway Analysis of Ability Formula By school

Missing Rows 1

Oneway Anova

Summary of Fit

Rsquare	0.007728
Adj Rsquare	-0.00867
Root Mean Square Error	1.004319
Mean of Response	0.003939
Observations (or Sum Wgts)	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
school	2	0.95057	0.47529	0.4712	0.6254
Error	121	122.04751	1.00866		
C. Total	123	122.99808			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
APU	42	-0.11680	0.15497	-0.4236	0.19000
CSU	42	0.04885	0.15497	-0.2580	0.35566
UCLA	40	0.08356	0.15880	-0.2308	0.39794

Std Error uses a pooled estimate of error variance

Oneway Analysis of total By school

Oneway Anova

Summary of Fit

Rsquare	0.021011
Adj Rsquare	0.004962
Root Mean Square Error	2.511246
Mean of Response	18.336
Observations (or Sum Wgts)	125

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
school	2	16.51227	8.25614	1.3092	0.2738
Error	122	769.37573	6.30636		
C. Total	124	785.88800			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
APU	42	18.8095	0.38749	18.042	19.577
CSU	42	18.2619	0.38749	17.495	19.029
UCLA	41	17.9268	0.39219	17.150	18.703

Std Error uses a pooled estimate of error variance

Summary

Running IRT is no longer a highly technical task for psychometricians only. Unlike syntax-based IRT software packages, such as Bilog and Winsteps, JMP allows users to obtain quick results by pointing and clicking, dragging and dropping. More importantly, every report in JMP is presented in a graphical fashion; you can make accurate and meaningful interpretations without invoking numeric-based statistical reasoning. Happy JMPing!