A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling

Chong Ho Yu, Ph.Ds

Email: chonghoyu@gmail.com

Website: http://www.creative-wisdom.com

Updated: October 27, 2017

This document, which is a practical introduction to Item Response Theory (IRT) and Rasch modeling, is composed of five parts:

I.   Item calibration and ability estimation

II. Item Characteristic Curve in one to three parameter models

III. Item Information Function and Test Information Function

IV. Item-Person Map

V.   Misfit

This document is written for novices, and thus, the orientation of this guide is conceptual and practical. Technical terms and mathematical formulas are omitted as much as possible. Since some concepts are interrelated, readers are encouraged to go through the document in a sequential manner.

It is important to point out that although IRT and Rasch are similar to each other in terms of computation, their philosophical foundations are vastly different from each other. In research modeling there is an ongoing tension between fitness and parsimony (simplicity). If the researcher is intended to create a model that reflects or fits "reality," the model might be very complicated because the real world is "messy" in essence. On the other hand, some researchers seek to build an elegant and simple model that have more practical implications. Simply put, IRT leans towards fitness whereas Rasch inclines to simplicity. To be more specific, IRT modelers might use up to three parameters, but Rasch stays with one parameter only. Put it differently, IRT is said to be descriptive in nature because it aims to fit the model to the data. In contrast, Rasch is prescriptive for it emphasizes fitting the data into the model. The purpose of this article is not to discuss these philosophical issues. In the following sections the term "IRT" will be used to generalize the assessment methods that take both person and item attributes into account, as opposed to the classical test theory. This usage is for the sake of convenience only and by no means the author equates IRT with Rasch. Nevertheless, despite their diverse views on model-data fitness, both IRT and Rasch have advantages over the classical test theory.

Part I: Item Calibration and Ability Estimation

Unlike the classical test theory, in which the test scores of the same examinees may vary from test to test, depending upon the test difficulty, in IRT item parameter calibration is sample-free while examinee proficiency estimation is item-independent. In a typical process of item parameter calibration and examinee

proficiency estimation, the data are conceptualized as a two-dimensional matrix, as shown in Table 1:

Table 1. 5X5 person by item matrix.

	Item 1	Item 2	Item 3	Item 4	Item 5	Average
Person 1	1	1	1	1	1	1
Person 2	0	1	1	1	1	0.8
Person 3	0	0	1	1	1	0.6
Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2
Average	0.8	0.6	0.4	0.2	0	

In this example, Person 1, who answered all five items correctly, is **tentatively** considered as possessing 100% proficiency. Person 2 has 80% proficiency, Person 3 has 60%, etc. These scores in terms of percentage are considered tentative because first, in IRT there is another set of terminology and scaling scheme for proficiency, and second, we cannot judge a person's ability just based on the number of correct items he obtained. Rather, the item attribute should also be taken into account. In this highly simplified example, no examinees have the same raw scores. But what would happen if there is an examinee, say Person 6, whose raw score is the same as that of Person 4 (see Table 2)?

Table 2. Two persons share the same raw scores.

Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2
Person 6	1	1	0	0	0	0.4

We cannot draw a firm conclusion that they have the same level of proficiency because Person 4 answered two easy items correctly, whereas Person 6 scored two hard questions instead. Nonetheless, for the simplicity of illustration, we will stay with the five-person example. This nice and clean five-person example shows an ideal case, in which proficient examinees score all items, less competent ones score the easier items and fail the hard ones, and poor students fail all. This ideal case is known as the **Guttman pattern** and rarely happens in reality. If this happens, the result would be considered an **overfit**. In non-technical words, the result is just "too good to be true."

Table 1 5X5 person by item matrix (with highlighted average)

	Item 1	Item 2	Item 3	Item 4	Item 5	Average
Person 1	1	1	1	1	1	1
Person 2	0	1	1	1	1	0.8
Person 3	0	0	1	1	1	0.6
Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2
Average	0.8	0.6	0.4	0.2	0	

We can also make a **tentative** assessment of the item attribute based on this ideal-case matrix. Let's look at Table 1 again. Item 1 seems to be the most difficult because only one person out of five could answer it correctly. It is tentatively asserted that the difficulty level in terms of the failure rate for Item 1 is 0.8, meaning 80% of students were unable to answer the item correctly. In other words, the item is so difficult that it can "beat" 80% of students. The difficulty level for Item 2 is 60%, Item 3 is 40% … etc. Please note that for person proficiency we count the number of successful answers, but for item difficulty we count the number of failures. This matrix is nice and clean; however, as you might expect, the issue will be very complicated when some items have the same pass rate but are passed by examinees of different levels of proficiency.

Table 3. Two items share the same pass rate.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Average
Person 1	1	1	1	1	1	0	0.83
Person 2	0	1	1	1	1	0	0.67
Person 3	0	0	1	1	1	0	0.50
Person 4	0	0	0	1	1	0	0.33
Person 5	0	0	0	0	1	1	0.33
Average	0.8	0.6	0.4	0.2	0	0.8	

In the preceding example (Table 3), Item 1 and Item 6 have the same difficulty level. However, Item 1 was answered correctly by a person who has high proficiency (83%) whereas Item 6 was not (the person who answered it has 33% proficiency). It is possible that the text in Item 6 tends to confuse good students. Therefore, the item attribute of Item 6 is not clear-cut. For convenience of illustration, we call the portion of correct answers for each person "tentative student proficiency" (TSP) and the pass rate for each item "tentative item difficulty" (TID). Please do not confuse these "tentative" numbers with the item difficulty

parameter and the person theta in IRT. We will discuss them later.

In short, both the item attribute and the examinee proficiency should be taken into consideration in order to conduct item calibration and proficiency estimation. This is an iterative process in the sense that tentative proficiency and difficulty derived from the data are used to fit the model, and then the model is employed to predict the data. Needless to say, there will be some discrepancy between the model and the data in the initial steps. It takes many cycles to reach **convergence**.

Given the preceding tentative information, we can predict the probability of answering a particular item correctly given the proficiency level of an examinee by the following equation:

Probability = 1/(1+exp(-(proficiency – difficulty)))

Exp is the Exponential Function. In Excel the function is written as exp(). For example:

e0 = 1
e1 = 2.7182= exp(1)
e2 = 7.3890= exp(2)
e3 = 20.0855= exp(3)

Now let's go back to the example depicted in Table 1. By applying the above equation, we can give a probabilistic estimation about how likely a particular person is to answer a specific item correctly:

Table 4a. Person 1 is "better" than Item 1

	Item 1	Item 2	Item 3	Item 4	Item 5	TSP
Person 1	0.55	0.60	0.65	0.69	0.73	1
Person 2	0.50	0.55	0.60	0.65	0.69	0.8
Person 3	0.45	0.50	0.55	0.60	0.65	0.6
Person 4	0.40	0.45	0.50	0.55	0.60	0.4
Person 5	0.35	0.40	0.45	0.50	0.55	0.2
TID	0.80	0.60	0.40	0.20	0.00	

For example, the probability that Person 1 can answer Item 5 correctly is 0.73. There is no surprise. Person 1 has a tentative proficiency of 1 while the tentative difficulty of Item 5 is 0. In other words, Person 1 is definitely "smarter" or "better" than Item 5.

Table 4b. The person "matches" the item.

	Item 1	Item 2	Item 3	Item 4	Item 5	TSP
Person 1	0.55	0.60	0.65	0.69	0.73	1
Person 2	0.50	0.55	0.60	0.65	0.69	0.8
Person 3	0.45	0.50	0.55	0.60	0.65	0.6
Person 4	0.40	0.45	0.50	0.55	0.60	0.4
Person 5	0.35	0.40	0.45	0.50	0.55	0.2
TID	0.80	0.60	0.40	0.20	0.00	

The probability that Person 2 can answer Item 1 correctly is 0.5. The tentative item difficulty is .8, and the tentative proficiency is also .8. In other words, the person's ability "matches" the item difficulty. When the student has a 50% chance to answer the item correctly, the student has no advantage over the item, and vice versa. When you move your eyes across the diagonal from upper left to lower right, you will see a "match" (.5) between a person and an item several times. However, when we put Table 1 and Table 4b together, we will find something strange.

Table 4b (upper) and Table 1 (lower)

	Item 1	Item 2	Item 3	Item 4	Item 5	TSP
Person 1	0.55	0.60	0.65	0.69	0.73	1
Person 2	0.50	0.55	0.60	0.65	0.69	0.8
Person 3	0.45	0.50	0.55	0.60	0.65	0.6
Person 4	0.40	0.45	0.50	0.55	0.60	0.4
Person 5	0.35	0.40	0.45	0.50	0.55	0.2
TID	0.80	0.60	0.40	0.20	0.00	

	Item 1	Item 2	Item 3	Item 4	Item 5	Average
Person 1	1	1	1	1	1	1
Person 2	0	1	1	1	1	0.8
Person 3	0	0	1	1	1	0.6
Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2
Average	0.80	0.60	0.40	0.20	0.00	

According to Table 4b, the probability of Person 5 answering Item 1 to 4 correctly ranges from .35 to .50. But actually, he failed all four items! As mentioned before, the data and the model do not necessarily fit

together. This residual information can help a computer program, such as Bilog, to further calibrate the estimation until the data and the model converge. Figure 1 is an example of Bilog's calibration output, which shows that it takes ten cycles to reach convergence.

Figure 1. Bilog's Phase 2 partial output

CALIBRATION PARAMETERS
 MAXIMUM NUMBER OF EM CYCLES: 10
 MAXIMUM NUMBER OF NEWTON CYCLES: 2
 CONVERGENCE CRITERION: 0.0100
 ACCELERATION CONSTANT: 1.0000

Part II: Item Characteristic Curve (ICC)

After the item parameters are estimated, this information can be utilized to model the response pattern of a particular item by using the following equation:

P = 1/(1+exp(-(theta – difficulty)))

From this point on, we give proficiency a special name: **Theta**, which is usually denoted by the Greek symbol θ. After the probabilities of giving the correct answer across different levels of θ are obtained, the relationship between the probabilities and θ can be presented as an Item Characteristic Curve (ICC), as shown in Figure 2.

Figure 2. Item Characteristic Curve
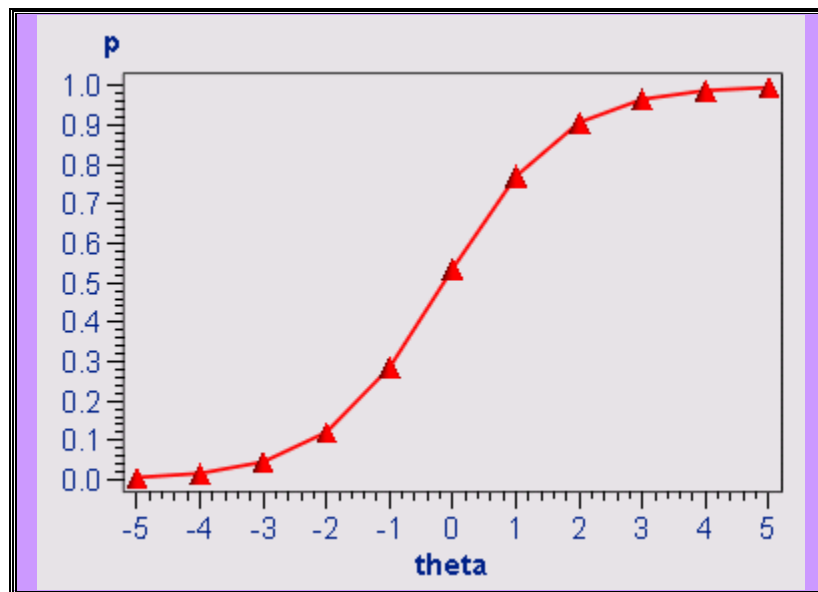
In Figure 2, the x-axis is the theoretical theta (proficiency) level, ranging from -5 to +5. Please keep in mind that this graph represents theoretical modeling rather than empirical data. To be specific, there may not be examinees who can reach a proficiency level of +5 or fail so miserably as to be in the -5 group. Nonetheless, to study the "performance" of an item, we are interested in knowing, given a person whose θ is +5, what the probability of giving the right answer is. Figure 2 shows a near-ideal case. The ICC indicates that when θ is zero, which is average, the probability of answering the item correctly is almost .5. When θ is -5, the probability is almost zero. When θ is +5, the probability increases to .99.

IRT modeling can be as simple as using one parameter or as complicated as using three parameters, namely,

A, B, and G parameters. Needless to say, the preceding example is a near-ideal case using only the B (item difficulty) parameter, keeping the A parameter constant and ignoring the G parameter. These three parameters are briefly explained as follows:

1. **B parameter**: It is also known as the **difficulty** parameter or the **threshold** parameter. This value tells us how easy or how difficult an item is. It is used in the one-parameter (1P) IRT model. Figure 3 shows a typical example of a 1P model, in which the ICCs of many items are shown in one plot. One obvious characteristic of this plot is that no two ICCs cross over each other.

Figure 3. 1P ICCs.
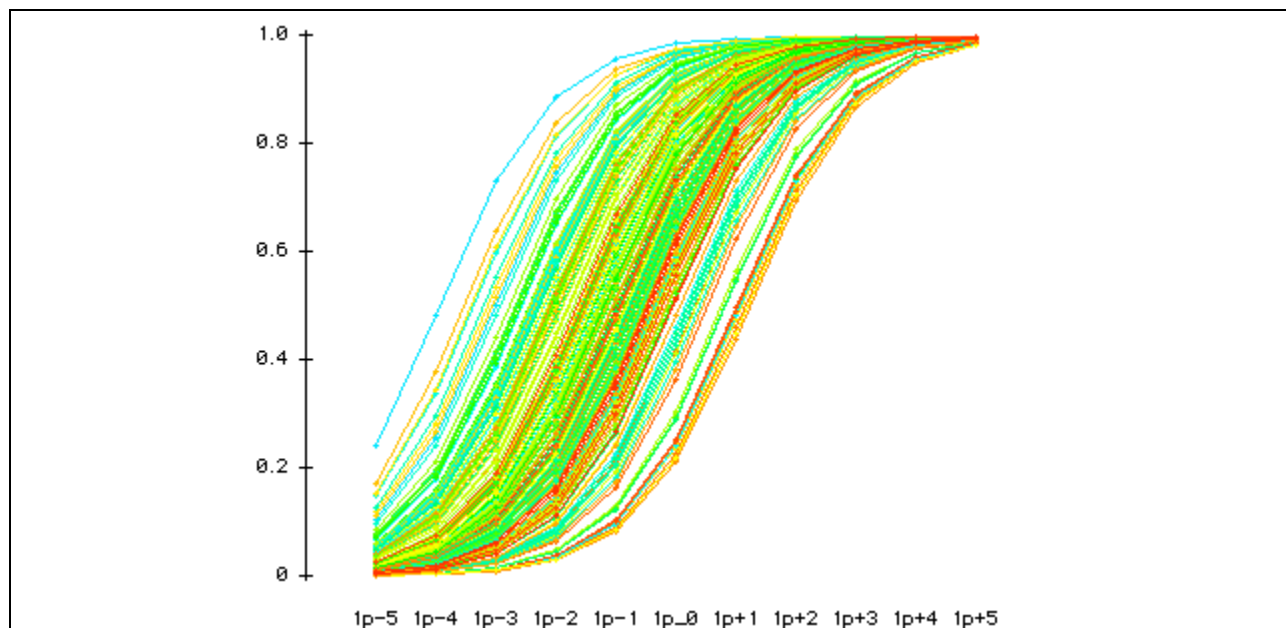
2. **A parameter**: It is also called the **discrimination** parameter. This value tells us how effectively this item can discriminate between highly proficient students and less-proficient students. The two-parameter (2P) IRT model uses both A and B parameters. Figure 4 shows a typical example of a 2P model. As you can notice, this plot is not as nice and clean as the 1P ICC plot, which is manifested by the characteristic that some ICCs cross over each other.

Figure 4a. 2P ICC
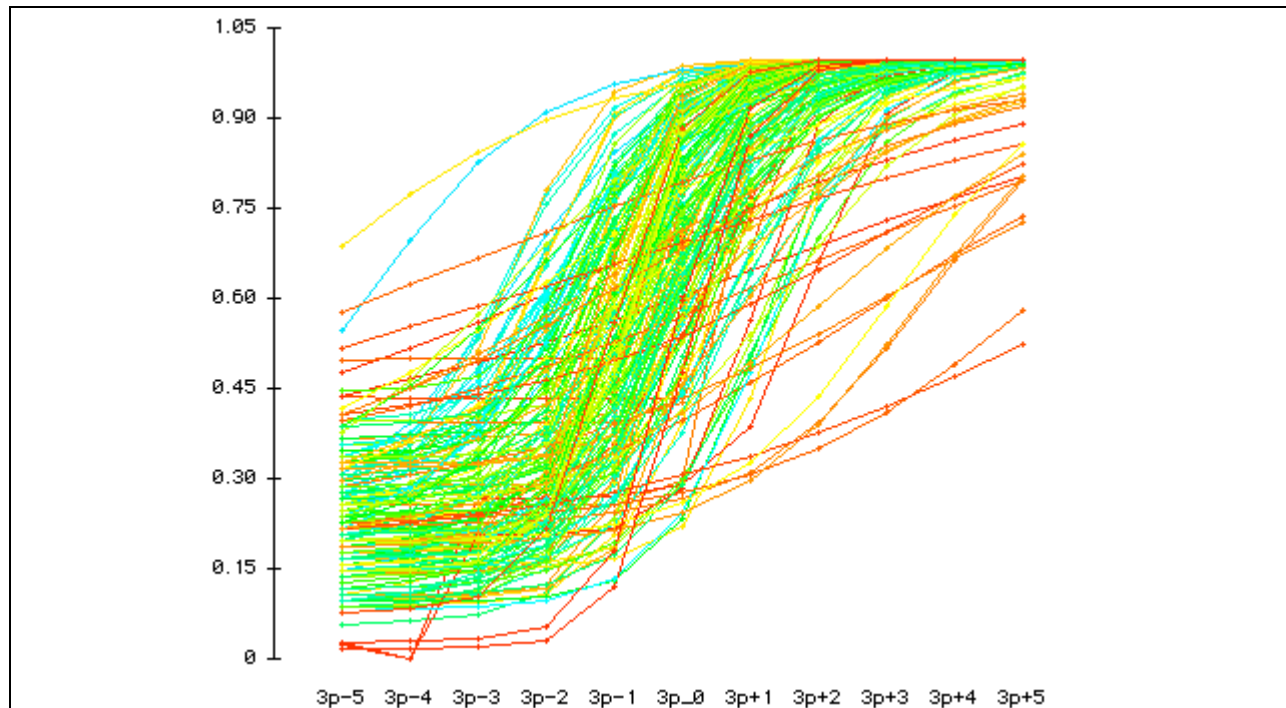
Take Figure 4b (next page) as an example. The red ICC does not have a high discrimination. The probability that examinees whose θ is +5 can score the item is 0.82, whereas the probability that examinees whose θ is -5 can score it is 0.48. The difference is just 0.82 - 0.48 = 0.26. On the other hand, the green ICC demonstrates a much better discrimination. In this case, the probability of obtaining the right answer given the θ of +5 is 1 whereas the probability of getting the correct answer given the θ of -5 is 0, and thus the difference is 1-0=1. Obviously, the discrimination parameter affects the appearance of the slope of ICCs, and that's why ICCs in the 2P model would cross over each other.

Figure 4b. ICCs of high and low discriminations.
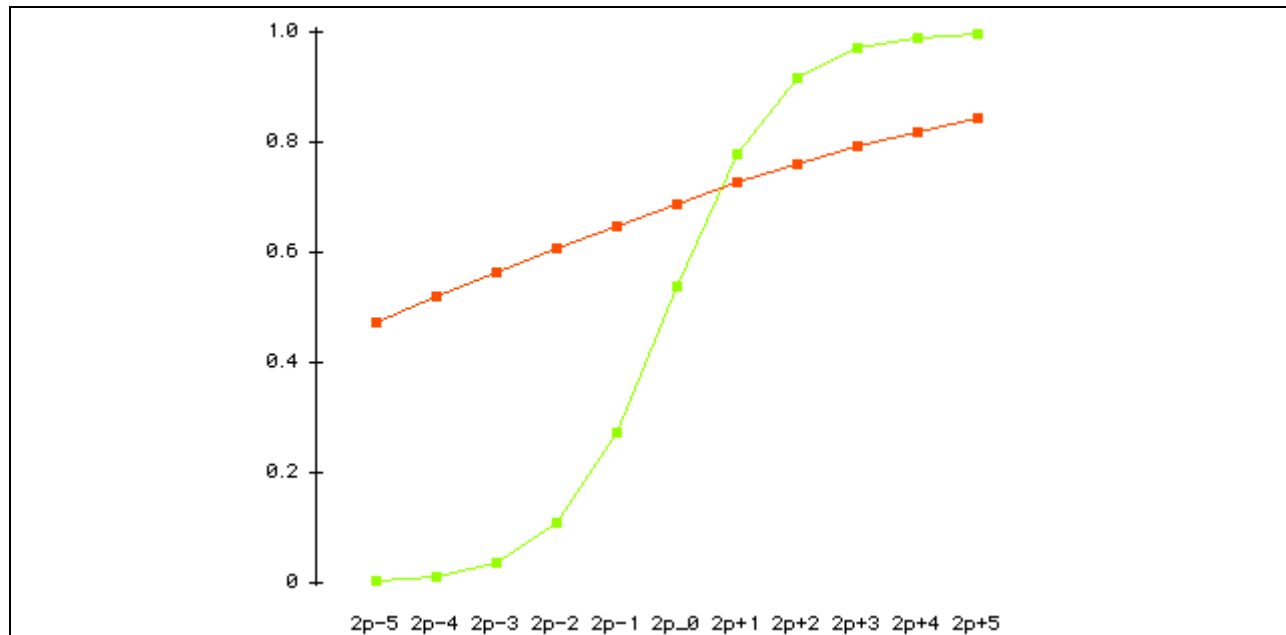
However, there is a major drawback in introducing the A parameter into the 2P IRT modeling. In this situation, there is no universal answer to the question "Which item is more difficult?" Take Figure 4b as an example again. For examinees whose θ is +2, the probability of scoring the red item is 0.7 while the probability of scoring the green item is 0.9. Needless to say, for them the red item is more difficult. For examinees whose θ is -2, the probability of answering the red item correctly is .6 whereas the probability of giving the correct answer to the green item is .1. For them the green item is more difficult. This phenomenon is called the **Lord's paradox**.

Figure 5. 3P ICCs
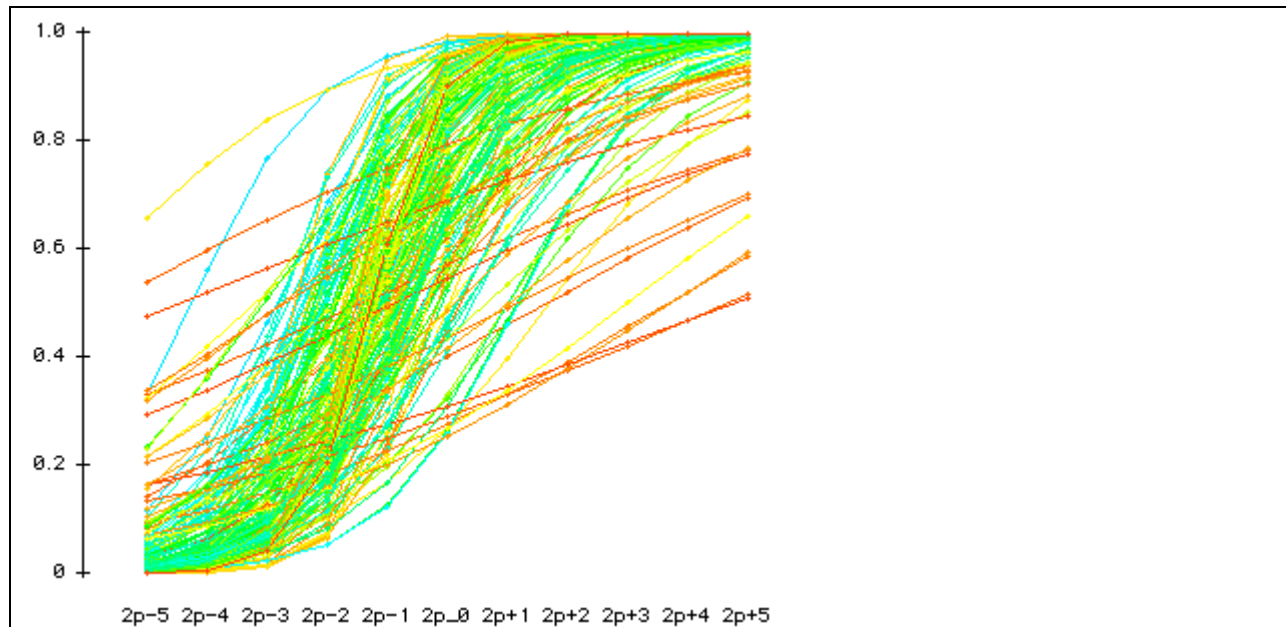
3. **C parameter**: It is also known as the **G** parameter or the **guessing** parameter. This value tells us how likely the examinees are to obtain the correct answer by guessing. A three-parameter (3P) IRT model uses A, B, and G parameters. Figures 5 and 4, which portray a 2P and 3P ICC plots using the same dataset, look very much alike. However, there is a subtle difference. In Figure 5 most items have a higher origin (the statistical term is "intercept") on the y-axis. When the guessing parameter is taken into account, it shows that in many items, even if the examinee does not know anything about the subject matter (θ=-5), he or she can still have some chances (p>0) to get the right answer.

As mentioned in the beginning, IRT modelers assert that on some occasions it is necessary to take discrimination and guessing parameters into account (2P or 3P models). However, in the perspective of Rasch modeling, crossing ICCs should not be considered a proper model because construct validity requires that the item difficulty hierarchy is invariant across person abilities (Fisher, 2010). If ICCS are crossing, the test developers should fix the items.

The rule of thumb is: the more parameters we want to estimate, the more subjects we need in computing. If there are sample size constraints, it is advisable to use a 1P IRT model or a Rasch model to conduct test construction and use a 3P as a diagnostic tool only. Test construction based upon the Item Information Function and the Test Information Function will be discussed next.

**Part III: Item Information Function and Test Information Function**

Figure 2. ICC
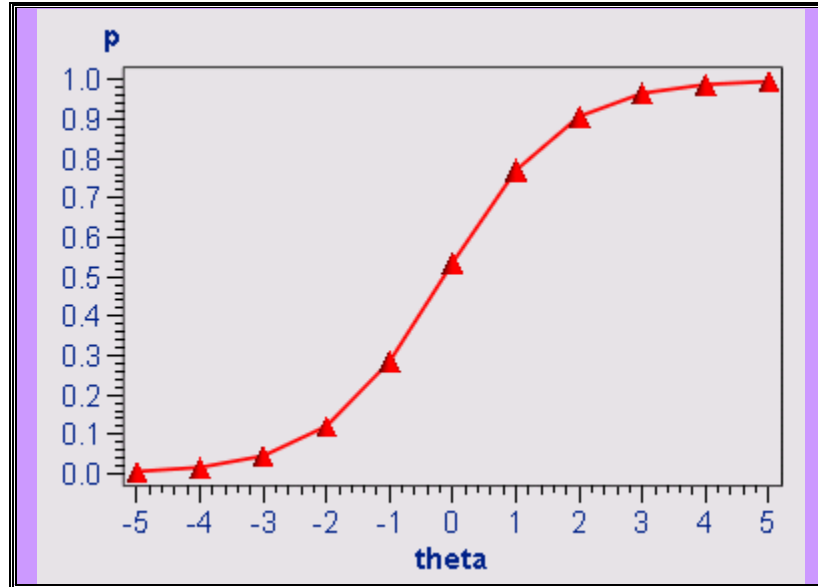
Let's revisit the ICC. When the θ is 0 (average), the probability of obtaining the right answer is 0.5. When the θ is 5, the probability is 1; when the θ is -5, the probability is 0. However, in the last two cases we have the problem of missing information. What does it mean? Imagine that ten competent examinees always answer this item correctly. In this case, we could not tell which candidate is more competent than the others with respect to this domain knowledge. On the contrary, if ten incompetent examinees always fail this item, we also could not tell which students are worse with regard to the subject matter. In other words, we have virtually no information about the θ in relations to the item parameter at two extreme poles, and less and less information when the θ moves away from the center toward the two ends. Not surprisingly, if a student answers all items in a test correctly, his θ could not be estimated. Conversely, if an item is scored by all candidates, its difficulty parameter could not be estimated either. The same problem occurs when all students fail or pass the same item. In this case, no item parameter can be computed.

There is a mathematical way to compute how much information each ICC can tell us. This method is called the **Item Information Function (IIF)**. The meaning of information can be traced back to R. A. Fisher's notion that information is defined as the reciprocal of the precision with which a parameter is estimated. If one could estimate a parameter with precision, one could know more about the value of the parameter than if one had estimated it with less precision. The precision is a function of the variability of the estimates around the parameter value. In other words, it is the **reciprocal of the variance**. The formula is as follows:

Information = 1/(variance)

In a dichotomous situation, the variance is p(1-p) whereas p = parameter value. Based on the item parameter values, one could compute and plot the IIFs for the items as shown in Figure 6.

Figure 6. Item Information Functions
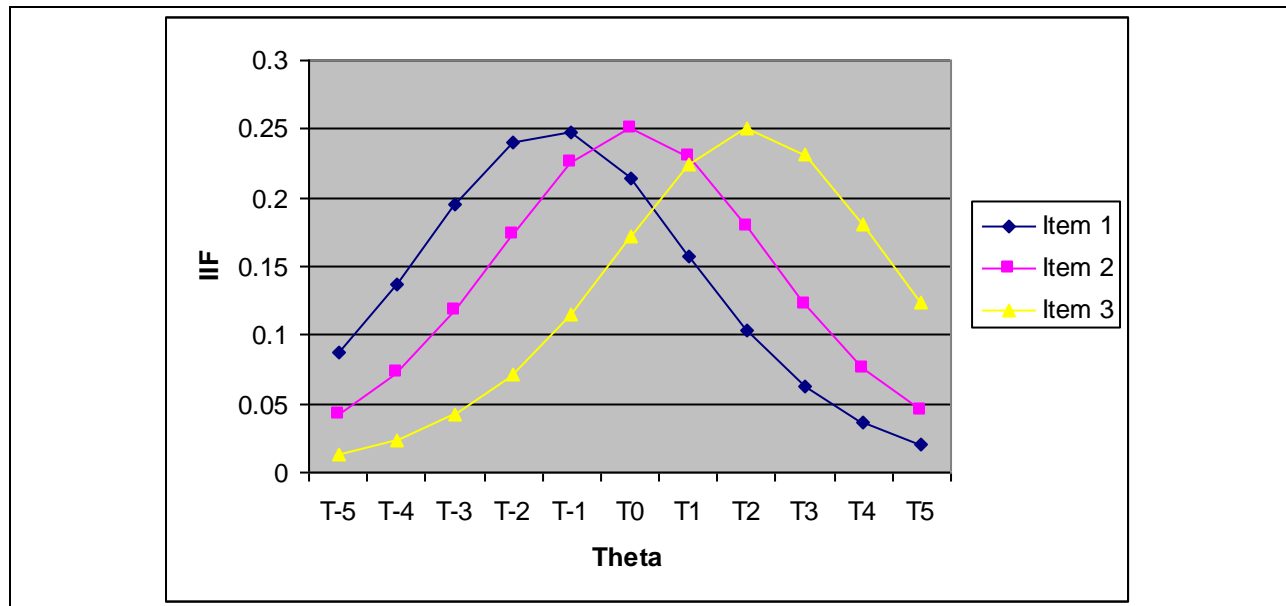
For clarity, only the IIFs of three items of a particular test are shown in Figure 6. Obviously, these IIFs differ from each other. In Item 1 (the blue line), the "peak" information can be found when the θ level is -1. When the θ is -5, there is still some amount of information (0.08). But there is virtually no information when the θ is 5. In item 2 (the pink line), most information is centered at θ zero while the amount of information in the lowest θ is the same as that in the highest θ. Item 3 (the yellow line) is the opposite of Item 1. One could have much information near the higher θ, but information drops substantively as the θ approaches the lower end.

The **Test Information Function (TIF)** is simply the sum of all IIFs in the test. While IIF can tell us the information and precision of a particular item parameter, the TIF can tell us the same thing at the exam level. When there is more than one alternate form for the same exam, TIF can be used to balance alternate forms. The goal is to make all alternate forms carry the same values of TIF across all levels of theta, as shown in Figure 7.

Figure 7. Form balancing using the Test Information Functions.
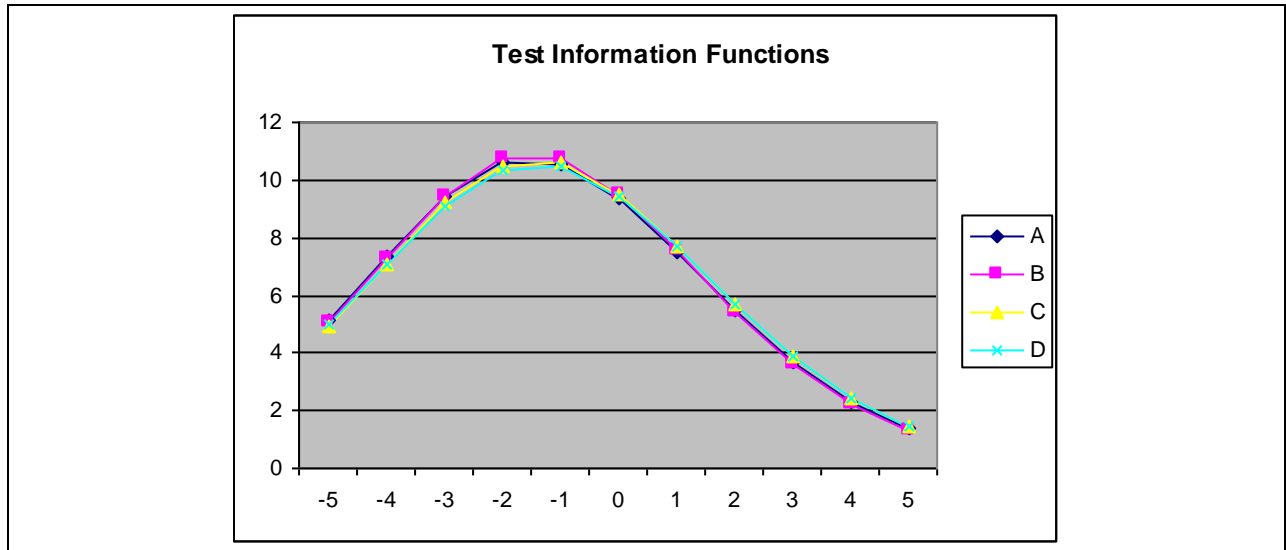
Part IV Logit and Item-Person Map

One of the beautiful features of the IRT is that item and examinee attributes can be presented on the same scale, which is known as the **logit**. Before explaining the logit, it is essential to explain the **odd ratio**. The odd ratio for the item dimension is the ratio of the number of the non-desired events (Q) to the number of the desired events (P). The formula can be expressed as: Q/P. For example, if the pass rate of an item is four of out five candidates, the desired outcome is passing the item (4 counts) and the non-desired outcome is failing the question (1 count). In this case, the odd ratio is 1:4 = .25.

The odd ratio can also be conceptualized as the probability of the non-desired outcomes to the probability of the desired outcome. In the above example, the probability of answering the items correctly is 4/5, which is 0.8. The probability of failing is 1-0.8 = 0.2. Thus, the odd ratio is 0.2/0.8 = 0.25. In other words, the odd ratio can be expressed as (1-P)/P.

The relationships between probabilities (p) and odds are expressed in the following equations:

Odds = P/(1-P) = 0.20/(1-0.20) = 0.25

P=Odds/(1+Odds) = 0.25/(1+0.25) = 0.20

The logit is the natural logarithmic scale of the odd ratio, which is expressed as:

Logit = Log(Odds)

As mentioned before, in IRT modeling we can put the item and examinee attributes on the same scale. But how can we compare apples and oranges? The trick is to convert the values from two measures into a common scale: logit. One of the problems of scaling is that spacing in one portion of the scale is not necessarily comparable to spacing in another portion of the same scale. To be specific, the difference between two items in terms of difficulty near the midpoint of the test (e.g. 50% and 55%) does not equal the gap between two items at the top (e.g. 95% and 100%) or at the bottom (5% and 10%). Take weight reduction as a metaphor: It is easier for me to reduce my weight from 150 lbs to 125 lbs, but it is much more difficult to trim my weight from 125 lbs to 100 lbs. However, people routinely misperceive that distances in raw scores are comparable. By the same token, spacing in one scale is not comparable to spacing in another scale. Rescaling by logit solves both problems. However, it is important to point out that while the concept of logit is applied to both person and item attributes, the computational method for person and item metrics are slightly different. For **persons**, the odd ratio is **P/(1-P)** whereas for **items** that is **(1-P)/P**. In the logit

scale, the original spacing is compressed, and as a result, equal intervals can be found on the logit scale, as shown in Table 5:

Table 5. Spacing in the original and the Log scale

Original	Unequal spacing		Log	Equal Spacing	
1	n/a		0	n/a	
2	2-1	1	0.30103	0.30103-0	0.30103
5	5-2	3	0.69897	0.69897-0.30103	0.39794
10	10-5	5	1	1-0.69897	0.30103
20	20-10	10	1.30103	1.30103-1	0.30103
50	50-20	30	1.69897	1.69897-1.30103	0.39794

In an IRT software application named **Winsteps**, the item difficulty parameter and the examinee theta are expressed in the logit scale and their relationships are presented in the **Item-Person Map (IPM)**, in which both types of information can be evaluated simultaneously. Figure 8 is a typical example of IPM.

In Figure 8, observations on the left hand side are examinee proficiency values whereas those on the right hand side are item parameter values. This IPM can tell us the "big picture" of both items and students. The examinees on the upper left are said to be "better" or "smarter" than the items on the lower right, which mean that those easier items are not difficult enough to challenge those highly proficient students. On the other hand, the items on the upper right outsmart examinees on the lower left, which implies that these tough items are beyond their ability level. In this example, the examinees overall are "better" than the exam items. If we draw a red line at zero, we can see that examinees who are below average would miss a small chunk of items (the grey area) but pass a much larger chunk (the pink area).

Figure 8. Item-Person Map

```
                              .  |
                            .##  |
                            .#  T|
                            .#   |   X
                          ####   |
                          .####   |
1               ########   +                               1
              .###########   |
              .############   |   XX
            .###############  S|   XXXXXX
        .####################  |T   XX
        .#####################   |   XXXX
    .########################   |
        .#######################   |   XXX
      .########################   |   XXXX
      .########################   |   XXXX
0       .#######################  M+   XXXXXX                0
      #######################  |S   XXXXXXXXXXXXX
        .####################   |   XXXXXXXXXXXXXXX
          .################   |   XXXX
        .#################   |   XXXXXXXXXX
          .###############   |   XXXXXXXXXXXX
          .############   |   XXXXXXXXXXXXXX
            .#########  S|M   XXXXXXXXXXXX
          ##########   |   XXXXXXXXXXXXXXXX
            .########   |   XXXXXXXXXXXXXXXXXXX
-1          .#########   +   XXXXXXXXXXXXXXX              -1
            .########   |   XXXXXX
            .#####   |   XXXXXXX
            .####   |   XXXXXXXXXXXXX
            .#### T|S   XXXXXXX
              .###   |   XXXXXXXXXXXXXX
              .##   |   XXXXXX
              .#   |   XXXXXXXXX
              .#   |   XX
              .#   |   X
-2               .   +   XXXX                             -2
              .  |T   X
              .   |   X
              .   |   X
              .   |
                  |   X
```

Part V: Misfit

In Figure 8 (previous page), it is obvious that some subjects are situated at the far ends of the distribution. In many statistical analyses we label them as "outliers." In psychometrics there is a specific term for this type of outliers: **Misfit**. It is important to point out that the fitness between data and model during the calibration process is different from the misfit indices for item diagnosis. Many studies show that there is no relationship between item difficulty and item fitness (Dodeen, 2004; Reise, 1990). As the name implies, a misfit is an observation that cannot fit into the overall structure of the exam. Misfits can be caused by many reasons. For example, if a test developer attempts to create an exam pertaining to American history, but accidentally an item about European history is included in the exam, then it is expected that the response pattern for the item on European history will substantially differ from that of other items. In the context of classical test theory, this type of items is typically detected by either point-biserial correlation or factor analysis. In IRT it is identified by examining the misfit indices. Table 6 is a typical output of Winsteps that indicates misfit:

Table 6. Misfit indices in Winsteps

IN.MSQ	IN.ZSTD	OUT.MS	OUT.ZSTD
1.12	3.38	1.23	3.94
1.12	3.42	1.24	3.97
0.85	-3.21	0.74	-2.65
0.85	-3.16	0.74	-2.6
0.9	-2	0.74	-2.59
0.91	-1.9	0.74	-2.56
1.23	6.48	1.38	6.51
1.23	6.44	1.37	6.38
0.93	-1.85	0.84	-2.45
0.93	-1.82	0.84	-2.37
0.82	-2.97	0.66	-2.43
0.84	-2.65	0.69	-2.15
0.98	-0.32	1.18	1.16
0.97	-0.47	1.13	0.86
1.02	0.5	1.01	0.21
1.01	0.32	1.01	0.12
0.99	-0.25	1.06	0.63
0.98	-0.36	1.06	0.6
0.87	-3.1	0.75	-2.99

Model fit

It seems confusing because there are four misfit indices. Let's unpack them one by one. "IN.ZSTD" and "OUT.ZSTD" stand for "infit standardized residuals" and "outfit standardized residuals." For now let's put

aside "infit" and "outfit." Instead, we will only concentrate on "standardized residuals". Regression analysis provides a good metaphor. In regression a good model is expected to have random residuals. A residual is the discrepancy between the predicted position and the actual data point position. If the residuals form a normal distribution with the mean as zero, with approximately the same number of residuals above and below zero, we can tell that there is no systematic discrepancy. However if the distribution of residuals is skewed, it is likely that there is a **systematic bias,** and the regression model is invalid. While item parameter estimation, like regression, will not yield an exact match between the model and the data, the distribution of standardized residuals informs us about the goodness or badness of the model fit. The easiest way to examine the model fit is to plot the distributions, as shown Figure 9.

Figure 9. Distributions of infit standardized residuals (left) and outfit standardized residuals (right)
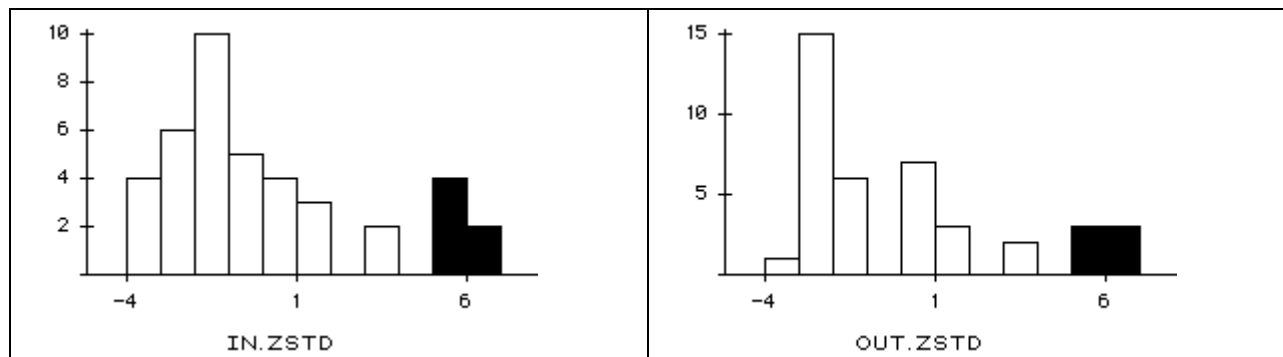
In this example, the fitness of the model is in question because both infit and outfit distributions are skewed. The shaded observations are identified as "misfits." Conventionally, while Chi-square is affected by sample size, standardized residuals are considered more immune to the issue of sample size. The rule of thumb for using standardized residuals is that a value > 2 is considered bad. However, Lai et al. (2003) asserted that standardized residuals are still sample size dependent. When the sample size is large, even small and trivial differences between the expected and the observed may be statistically significant. And thus they suggested putting aside standardized residuals altogether.

Item fit
Model fit takes the overall structure into consideration. If you remove some "misfit" items and re-run the IRT analysis the distribution will look more normal, but there will still be items with high residuals. Because of this, the "model fit" approach is not a good way to examine item fit. If so, then what is the proper tool for checking item fit? The item fit approach, of course. "IN.MSQ" and "OUT.MSQ" stand for "infit mean squared" and "outfit mean squared." In order to understand this approach, we will unpack these concepts. "Mean squared" is simply the Chi-squared divided by the degrees of freedom (df). Next, we will

look at what "Chi-squared" means using a simple example.

Table 7. 2X3 table of answer and skill level

	More skilled (theta > 0.5)	Average (theta between -0.5 and +0.5)	Less skilled (theta < -0.5)	Row total
Answer correctly (1)	10	5	15	30
Answer incorrectly (0)	5	10	5	20
Column total	15	15	20	Grand total: 50

Table 7 is a crosstab 2X3 table showing the number of correct and incorrect answers to an item categorized by the skill level of test takers. At first glance this item seems to be problematic because while only 10 skilled test-takers were able to answer this item correctly, 15 less skilled test-takers answered the question correctly. Does this mean the item is a misfit? To answer this question, we will break down how chi-squared is calculated.

Like many other statistical tests, we address this issue by starting from a null hypothesis: There is no relationship between the skill level and test performance. If the null hypothesis is true, then what percentage of less skilled students would you expect to answer the item correctly? Regardless of the skill level, 30 out of 50 students could answer the item correctly, and thus the percentage should be 30/50 = 60%.

Table 8. 3X3 table showing one expected frequency and one actual frequency

	More skilled (theta > 0.5)	Average (theta between -0.5 & +0.5)	Less skilled (theta < -0.5)	Row total
Answer correctly (1)	10	5	15 (12)	30
Answer incorrectly (0)	5	10	5	20
Column total	15	15	20	Grand total: 50

Because 20 students are classified as low skilled and if 60% of them can answer the item correctly, then the expected count (E) for students who gave the right answer belong to the low skilled group is 12 (20 X 60%) In Table 9, the number inside the bracket is the expected count assuming the null hypothesis is correct.

Table 9. 3X3 table showing two expected counts and two actual counts

	More skilled (theta > 0.5)	Average (theta between -0.5 & +0.5)	Less skilled (theta < -0.5)	Row total
Answer correctly (1)	10	5	15 (12)	30
Answer incorrectly (0)	5	10	5 (8)	20
Column total	15	15	20	**Grand total**: 50

You may populate the entire table using the preceding approach, but you can also use a second approach, which is a short cut found by using the following formula:

Expected count= [(Column total) X (Row total)]/Grand total

For example, the expected count cell of (less skilled, answer correctly) is: 20 X 20/50 = 8.

Table 10. 3X3 table showing all expected counts and all actual counts

	More skilled (theta > 0.5)	Average (theta between -0.5 & +0.5)	Less skilled (theta < -0.5)	Row total
Answer correctly (1)	10 (9)	5 (9)	15 (12)	30
Answer incorrectly (0)	5 (6)	10 (6)	5 (8)	20
Column total	15	15	20	**Grand total**: 50

Table 10 shows the expected count in all cells. You can see that there is a discrepancy between what is expected (E) and what is the observed (O) in each cell. To measure the fit between the E and O, we use the formula:

$(O-E)^2 / E$

For example, for the cell (more skilled, answer correctly): $(10-9)^2 / 9 = 0.111$

Table 11. 3X3 table showing all expected counts and all actual counts

	More skilled (theta > 0.5)	Average (theta between -0.5 & +0.5)	Less skilled (theta < -0.5)
Answer correctly (1)	0.111	1.778	0.750
Answer incorrectly (0)	0.167	2.667	1.125

Table 11 shows the computed Chi-squared in all cells. The number in each cell indicates the value of the discrepancy (residual). The bigger the number is, the worse the discrepancy is. The sum of all **(O-E) ^{2} / E** is called the Chi-square, the sum of all residuals that shows the overall discrepancy. If the Chi-square is big, it indicates that the item is a misfit. As mentioned before, the mean-squared index is the Chi-squared divided by the degrees of freedom. Explaining the degrees of freedom requires another tutorial on its own right. Please consult http://creative-wisdom.com/computer/sas/df.html (the text version) and http://creative-wisdom.com/pub/df/default.htm (the multimedia version).

Figure 10. Chi-square and degrees of freedom from a 2X3 table.

It is important to keep in mind that the above illustration is over-simplified. In the actual computation of misfit, examinees are not typically divided into only three groups and more levels may be used. There is no common consent about the optimal numbers of intervals. Yen (1981) suggested using 10 grouping intervals. Some item analysis software modules, such as (RUMM, Rasch Uni-dimensional Measurement Model) adopts 10-level grouping as the default. It is important to point out that the number of levels is tied to the

degrees of freedom, which affects the significance of a Chi-square test. The degrees of freedom for a Chi-square test is obtained by (the number of row) X (the number of column). In the preceding example df = (2-1)*(3-1)=2. Figure 10 shows the test computed at http://vassarstats.net/newcs.html. When the data are configured as a 2X3 table (competency is divided into three levels), the df is 2 and the p value is 0.0369, which is considered significant. But what would happen if the psychometrician decides to use two levels only (competent, not competent) by collapsing "more skilled" and "average" into one group ("competent")? Figure 10 shows the result from a 2X2 table, in which the df is 1 and the p value based on the Yates Chi-square is 0.1407. Additionally, the Pearson chi-square is 3.13 whereas $p = 0.0769$. But neither one is significant. In short, whether the Chi-square is significant or not highly depends on the degrees of freedom and the number of rows/columns (the number of levels chosen by the software package). Hence, the Chi-square based statistics, which will be discussed in the next section, should be adjusted by the degrees of freedom.

**Figure 11. Chi-square and degrees of freedom from a 2X2 table.**
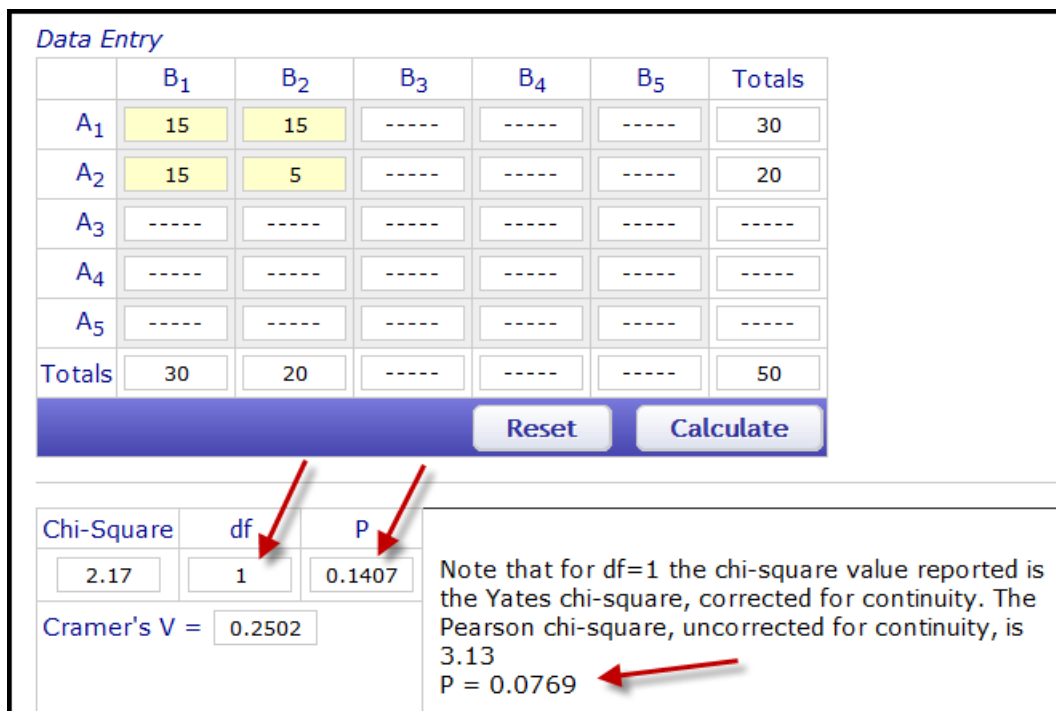
Infit and outfit

The infit mean-squared is the Chi-squared/degrees of freedom with weighting, in which a constant is put into the algorithms to indicate how much certain observations are taken into account. As mentioned before, in the actual computation of misfit there may be many groups of examinees partitioned by their skill level, but usually there are just a few observations near the two ends of the distribution. Do we care much about the test takers at the two extreme ends? If not, then we should assign more weight to examinees near the

middle during the Chi-squared computation (see Figure 12). The outfit mean squared is the converse of its infit counterpart: unweighted Chi-squared/df. The meanings of "infit" and "outfit" are the same in the context of standardized residuals. Another way of conceptualizing "infit mean square" is to view it as the ratio between observed and predicted variance. For example, when infit mean square is 1, the observed variance is exactly the same as the predicted variance. When it is 1.3, it means that the item has 30% more unexpected variance than the model predicted (Lai et al., 2003).

Figure 12. Distribution of examinees' skill level
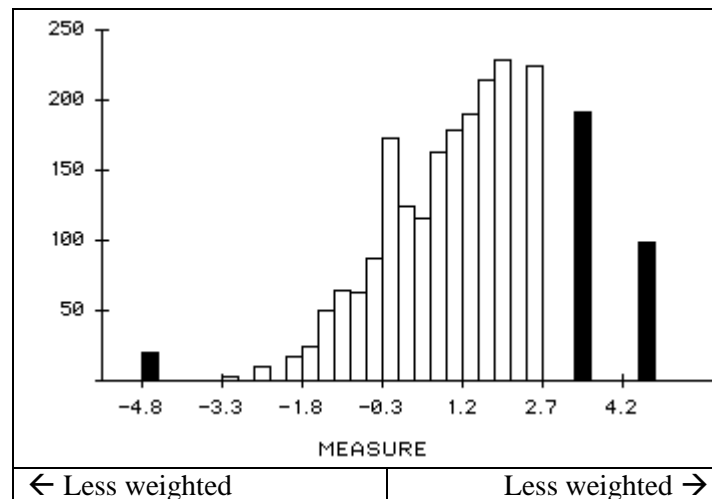
The objective of computing item fit indices is to spot misfits. Is there a particular cutoff to demarcate misfits and non-misifts? In different parts of the online manual, Winsteps recommends that any mean squared above 1.0 (Winsteps & Rasch measurement Software, 2010a) or 1.5 (Winsteps & Rasch measurement Software, 2010b) is considered too big and "noise" is noticeable, Lai et al. (2003) suggests using 1.3 as the demarcation point. The following is a summary of how different levels of mean-square value should be interpreted (Linacre, 2017):

Table 11. Interpretation of different levels of mean-square values.

Mean-square value	Implications for measurement
> 2.0	Distorts or degrades the measurement system. Can be caused by only one or a few observations. By removing them it might bring low mean-squares into the productive range.
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 high reliability and separation coefficients.

Nonetheless, many other psychometricians do not recommend setting a fixed cut-off (Wang, & Chen, 2005); instead, a good practice is to check all mean squared visually. Consider the example shown in Figure 13. None of the mean squared is above 1.5 and by looking at the numbers alone we may conclude that there is no misfit in the exam. However, by definition, a misfit is an item whose behavior does not conform to the overall. It is obvious that one particular item (depicted in an orange squared on the top) departs from the majority and thus further scrutiny for this potential misfit is strongly recommended.

Figure 13. Plot of outfit mean squared
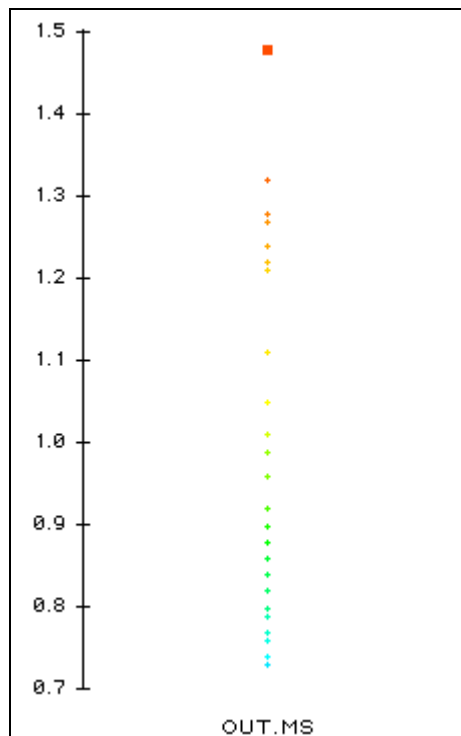
A common question to ask may be whether these misfit indices agree with each other all the time, and which one we should trust when they differ from one another. As mentioned before, the standardized residual is a measure of the model fit whereas the mean squared is for item fit. Thus, it is not necessary to compare between apples and ranges. On the contrary, comparing the infit mean squared and the outfit mean squared addresses a meaningful question. It is similarly useful to compare the infit standardized residual and the outfit standardized residual. Infit is a weighted method while outfit is unweighted. Because some difference will naturally occur, the question to consider is not about whether they are different; rather, the key questions are 1) to what degree they differ from one another, and 2) does this difference lead to contradictory conclusions regarding the fitness of certain items. One of the easiest ways to check the correspondence between infit and outfit is the **parallel coordinate**. In the parallel coordinate the

observations in two dot plots are joined to indicate whether observations from one measure substantively shifts their position in another measure. Figure 14 shows that in this example there is a fairly good degree of agreement between infit and outfit.

Figure 14. Parallel coordinate of the infit and outfit mean squared.
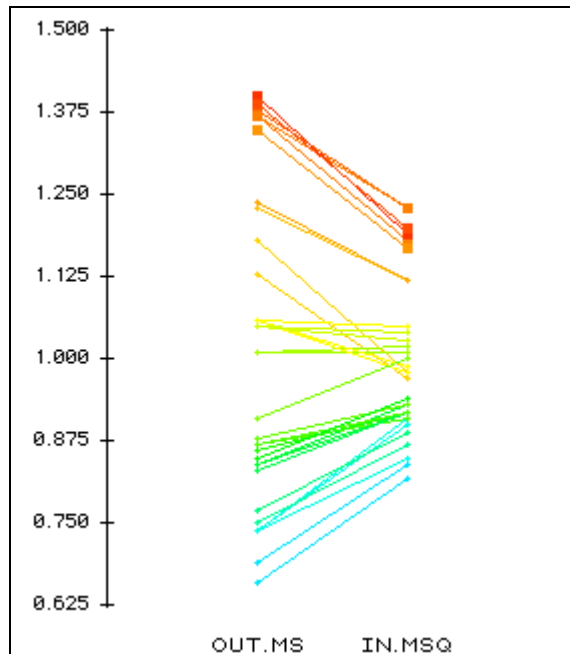
According to Winsteps and Rasch Measurement Software (2010a), if the mean squares values are less than 1.0, the observations might be too predictable due to redundancy or model overfit. Nevertheless, high mean squares are a much greater threat to validity than low mean squares. And thus it is advisable to focus on items with high mean squares while conducting misfit diagnosis. However, if you are not sure which preceding criterion should be used to identify misfits, you can simply hand over your judgment to an automated system. Winsteps is capable of performing stepwise screening unit no items showed misfit Bond & Fox, 2015).

Person fit

To add complexity to your already confused mind, please note that IRT output consists of two clusters of information: person's theta and item parameters. In the former the skill level of the examinees is estimated whereas in the latter the item attributes are estimated. The preceding illustration uses the item parameter output only, but the person's theta output may also be analyzed using the same four types of misfit indices. This measure of fit should be done before item fit analysis. It is crucial to point out that misfits among the person thetas are not simply outliers, which represent over-achievers who obtained extremely high scores or

under-achievers who obtained extremely low scores. Instead, misfits among the person thetas represent people who have an estimated ability level that does not fit into the overall pattern. In the example of item misfit, we doubt whether an item is well-written when more low skilled students (15) than high skilled students (10) have given the right answer. By the same token, if a supposedly low skilled student answers many difficult items correctly in a block of questions; one explanation for this could be an instance of cheating.

Strategy

The strategy for examining the fitness of a test for diagnosis purposes is summarized as follows:

- Evaluate the person fit to remove suspicious examinees. Use outfit mean squared because when you encounter an unknown situation, it is better not to perform any weighting on any observation. If you have a large sample size (e.g 1,000) removing a few subjects will likely not make a difference. But if a large chunk of person misfits must be deleted, it is advisable to re-compute the IRT model.

- Next, move from person output to the item output. Evaluate the overall model fit by first checking the outfit standardized residuals and second checking the infit standardized residuals. Outfit is more inclusive in the sense that every observation counts. Draw a parallel coordinate to see whether the infit and outfit model fit indices agree with each other. If there is a discrepancy, determining whether or not to trust the infit or outfit will depend on what your goal is. If the target audience of the test is examinees with average skill-level, an infit model index may be more informative.

- If the model fit is satisfactory, examine the item fit in the same order with outfit first and infit second. Rather than using a fixed cut-off for mean squared, visualize the mean squared indices in a plot to detect whether any items significantly depart from the majority, and also use a parallel coordinate to check the correspondence between infit and outfit.

- When misfits are found, one should check the key, the distracters, and the question content first. Farish (1984) found that if misfits are mechanically deleted just based on chi-square values or standardized residuals, this improves the fit of the test as a whole, but worsens the fit of the remaining items.

Conclusion

To conclude this tutorial, I would like to highlight one of the advantages of the Item Response Theory. As you may have noted in the item-person map (Figure 8), the distribution of the person logit is not highly non-normal even though the items are very easy. This is the advantage of IRT. In IRT, item difficulty parameters are independent of who is answering the item, and person thetas are independent of what items the examinees answer. Consider Figure 15, which shows the distribution of test scores in the classical sense. The distribution is skewed toward the high end. In terms of classical analysis, we may assert that most examinees have high ability. It is then obvious that estimation of ability depends on what items the examinee answers. If the test is easy, it will make the examinees look smart.

Figure 15. Raw score distribution in the classical sense
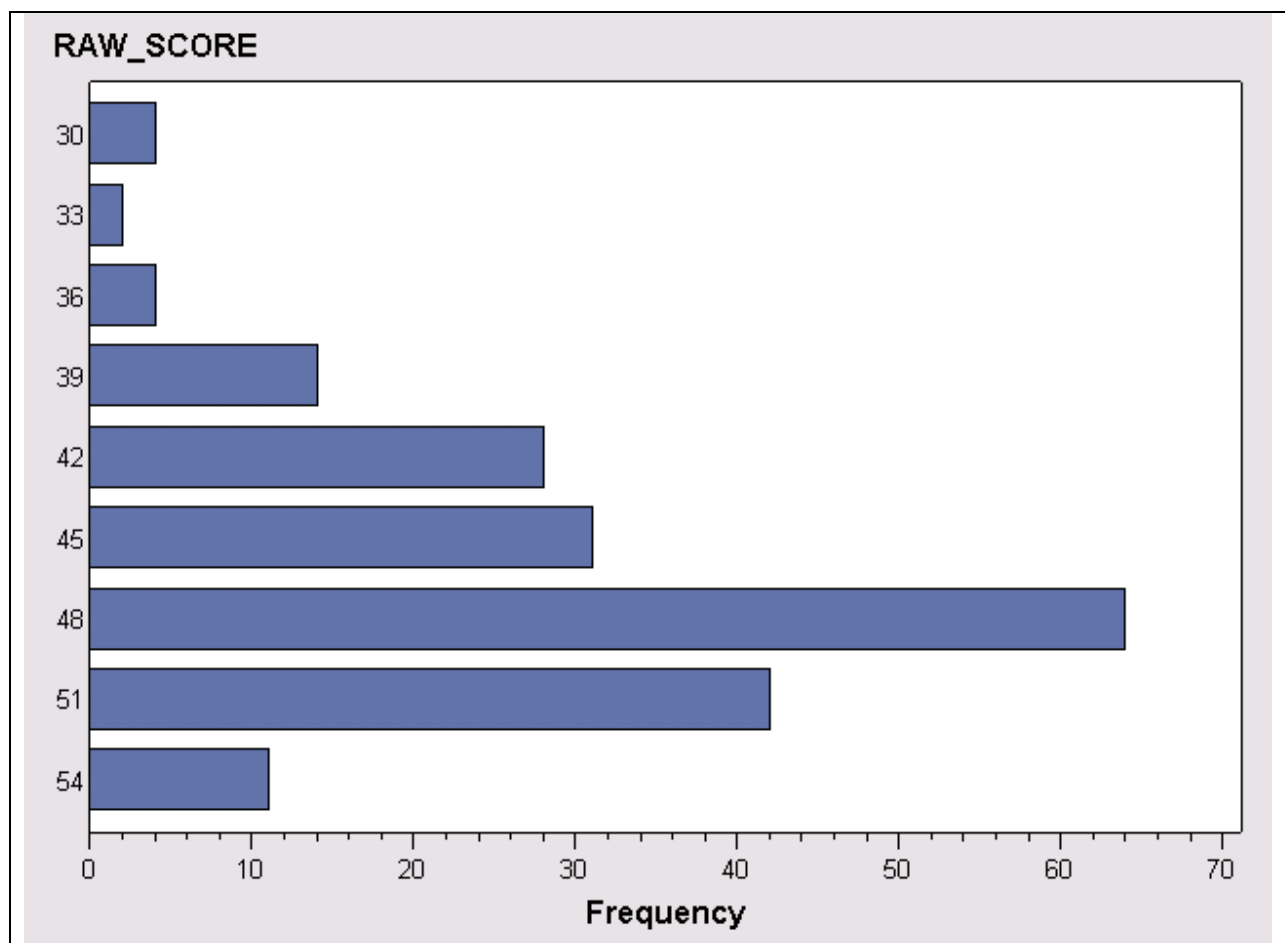
Now let's look at Figure 16, which is a histogram depicting the theta (estimation of examinee ability based upon IRT) with reference to the same exam. In contrast to Figure 15, the distribution is almost bell-shaped because ability estimation in IRT takes item difficulty into account. Even if the test is easy, your estimated

ability level will not be any higher than what it should be. In other words, ability estimation is **item-independent**. The same principle is applied to item difficulty estimation. A difficult item in which many competent people passed would not appear to be easy. Thus, item difficulty estimation is said to be **sample-independent**.

Figure 16. Theta distribution based on IRT
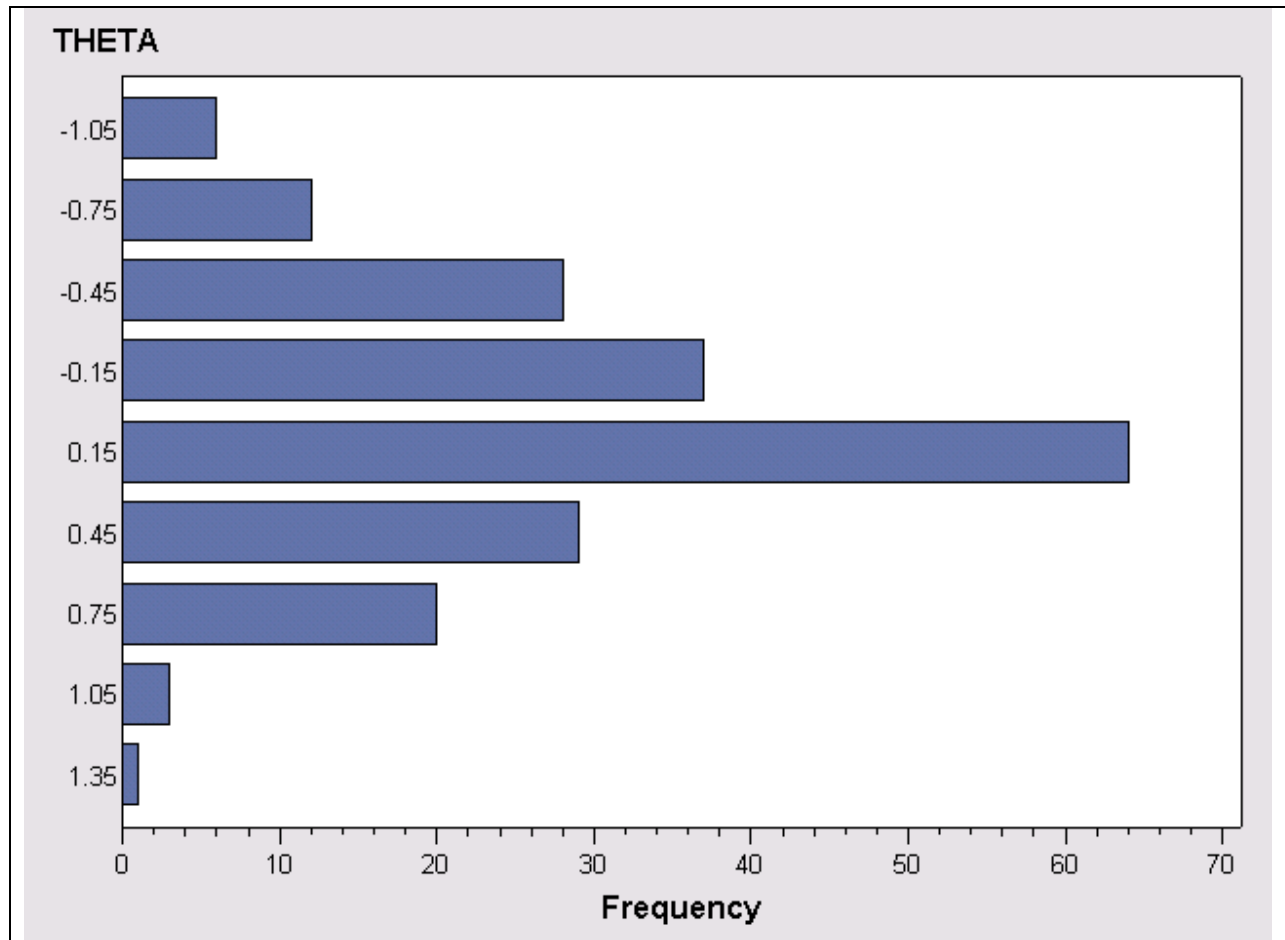
For a multimedia tutorial of IRT, please visit: http://www.creative-wisdom.com/multimedia/IRTTHA.htm

References

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement, 41*, 261-270.

Farish, S. (1984). *Investigating item stability*. (ERIC document Reproduction Service No. ED262046).

Fisher, W. (2010). IRT and confusion about Rasch measurement. *Rasch Measurement Transactions, 24*, 1288. Retrieved from http://www.rasch.org/rmt.

Lai, J., Cella, D., Chang, C. H., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten, and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue scale. *Quality of Life Research, 12*, 485–501.

Linacre, M. (2017). Teaching Rasch measurement. *Transactions of the Rasch Measurement, 31,* 1630-1631.

Reise, S. (1990). A comparison of item and person fit methods of assessing model fit in IRT. *Applied Psychological Measurement, 42*, 127-137.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps Program for the family of Rasch models. *Educational and Psychological Measurement, 65*, 376 – 404.

Winsteps & Rasch measurement Software. (2010a). Misfit diagnosis: Infit outfit mean-square standardized. Retrieved from http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm.

Winsteps & Rasch measurement Software. (2010b). Item statistics in misfit order. Retrieved from http://www.winsteps.com/winman/index.htm?table10_1.htm.