Recent Developments in Rasch Measurement

WANG Wen Chung

Chair Professor of Educational and Psychological Measurement
The Hong Kong Institute of Education

The Hong Kong Institute of Education
香 港 教 育 學 院

The Hong Kong Institute of Education
10 Lo Ping Road, Tai Po, Hong Kong, China

This paper was firstly presented in a lecture of the Chair Professors Public Lecture Series of The Hong Kong Institute of Education on 28 January 2010.

Contents

About the Author *iii*

Recent Developments in Rasch Measurement

Abstract *1*

1. Introduction *2*

2. Dichtomous Items *6*

3. Model Extensions *23*

4. Conclusion *42*

References *43*

Refereed Papers *47*

Selected Recent Research Projects *59*

About the Author

wcwang@ied.edu.hk

WANG Wen Chung（王文中）

is the Director of Assessment Research Center, Chair Professor of Educational and Psychological Measurement of the Hong Kong Institute of Education (HKIEd), and Adjunct Professor of Beijing Normal University. Prior to joining HKIEd in August 2008, he was a Distinguished Professor and Head, Department of Psychology at National Chung Cheng University in Taiwan.

Professor WANG obtained a PhD degree from University of California at Berkeley (UCB) in 1994. He was a Fulbright grantee in 2005 and visiting professor of UCB in 2005 and 2007. His research interests include Rasch measurement, item response theory, and psychometrics. He received Research Excellence Awards from National Science Council and National Chung Cheng University, and the Mu-Dou Award（木鐸獎）for his contribution to education.

He has published more than 120 refereed journal articles (including more than 30 SSCI journal articles) and several books and chapters. He is an editorial board member of several journals and a reviewer of more than 30 journals. He has completed supervision of 6 PhD students and 20 master students. When staying in Taiwan, he was a principal investigator of many external projects funded by National Science Council. In 2009/2010 he was granted a GRF project.

Recent Developments in Rasch Measurement

WANG Wen Chung

Abstract

Rasch measurement has been widely applied in the human sciences, including education, psychology, health sciences, sports, management, sociology and political sciences. The major beauty of Rasch measurement is that it diagnoses noise in test or survey data and converts ordinal item response or test raw score into a linear measure such that subsequent parametric statistical analysis (e.g., t-test, ANOVA, correlation and regression) becomes feasible, and intra-person growth and inter-person difference can be quantified. Recent decades have witnessed the blooming of Rasch measurement. In this paper, I highlight several important developments where Rasch models have been extended to deal with complicated testing situations:

(a) polytomous items (e.g., constructed-response items)
(b) multiple facets (e.g., rater effect)
(c) multilevels (e.g., gender difference in math, school effect)
(d) mixture models (i.e., latent class plus latent trait)
(e) testlet items (i.e., a set of items are connected by a common stimulus of passage or figure)
(f) multiple dimensions (e.g., tests with subtests)
(g) hierarchical latent traits (e.g., Quality of life includes physical, psychological, social and environmental domains, and each domain may include subdomains.)
(h) structural equation modeling with categorical data
(i) differential item functioning (i.e., items function differently for different groups of test-takers, an issue of test fairness)
(j) computerized adaptive testing / computerized classification testing
(k) person-item interaction in rating scale items or Likert items (e.g., my strongly agree is equal to your agree, which is equal to his neutral)

1. INTRODUCTION

Tests (including inventories, questionnaires, systematic observations and interviews) have been widely used in the human sciences to measure hypothetical constructs or attributes, like ability, personality, attitude or interest. For example, achievement tests are used to measure subject proficiency, self-reports inventories are used to measure anxiety or quality of life. These attributes are not directly observable; rather, they have to be inferred from observable events. Hence, they are referred to as latent traits. Actually, many attributes in the natural sciences are latent traits. Gravity is a good example.

There are two major purposes in measurement. One is to quantify inter-individual difference, for example, who is more proficient, more outgoing, or happier. The other is to quantify intra-individual difference (i.e., growth), for example, whether a test-taker is more capable in mathematics than last semester, or more satisfied than last month. If a test consists of only a single item, then its reliability and validity will be too low to be useful. A test often contains multiple items in order to increase reliability and validity. Although items in a test were designed to measure the same construct, empirical evidence is needed to assess whether this purpose is fulfilled, which is the major task of item analysis.

If through item analysis, items in a test are found to measure the same construct, that is, the assumption of unidimensionality is met, then the next step is often to sum up item scores to form a test raw score (or its linear transformation) and use it to describe a test-taker's level. For example, in an ability test with dichotomous items (scored as 0 or 1 in an item), test raw score is the sum of individual item scores. The higher the raw score is, the more proficient the test-taker. Likewise, in an inventory with rating scale or Likert items (strongly disagree = 1, disagree = 2, agree = 3, strongly agree = 4), test raw score is used to depict a test-taker's level on a construct (e.g., anxiety, happiness). The higher the raw score, the higher level the construct is. If items in a test do not measure the same construct (i.e., the assumption of unidimensionality is not met), then item scores should not be summed because the resulting raw score is meaningless.

For the sake of communication, in this paper "ability" test is used to represent any kind of tests, the word "ability" is used to represent any latent trait, item "difficulty" is used to replace item "threshold." The reader can easily generalize the concepts and methods that are introduced in this paper to non-ability test contexts.

Every measure consists of some measurement error. In classical test theory (CTT; Lord & Novick, 1968), it is assumed that an observed score is the sum of a true score and an error score. In addition, assumptions about the error score are made. In CCT, scores are assumed to be interval (Stevens, 1946). This assumption may hold in physical measures (e.g., height and body temperature). However, it may not hold for test scores in the social sciences. In practice, raw scores (or their linear transformations) are often used to describe inter- and intra-individual difference or correlation between latent traits. If the assumption of interval data does not hold, then raw scores cannot be operated arithmetically (e.g., to compute mean and variance) such that the aforementioned analysis is misleading.

Some may view testing hypothesis or estimating confidence interval with standard computer programs like SPSS or SAS (e.g., correlation, regression or ANOVA) as an application of CTT. Actually, it is not. The major idea of CTT is measurement error. However, measurement error is not considered in ordinary data analysis. If measures contain a very small amount of measurement error (e.g., height and weight), then ignoring measurement error by using ordinary data analysis does little harm. Unfortunately, measures in the social sciences often (if not always) consist of a great amount of measurement error. Ignoring measurement error can cause serious mistakes in hypothesis testing and confidence interval estimation (Lord & Novick, 1968). An even more serious problem is that raw scores are not interval and should not be treated as such.

1. 1 Properties of Raw Scores

Assume a mathematic test has 50 dichotomous items. A test-taker receives a score of 20. We may consider this person as not proficient in mathematics. If an easier test is administered and the same person receives a score of near 50, then we may consider the same person as highly proficient. In other words, whether this person is proficient or not depends on which test is administered. That is, the judgment of person ability level with raw score is test dependent.

In addition to person ability level, we are also interested in item difficulty. If a test is administered to 100 persons and 90 of them answer item 1 correctly, we may consider item 1 very easy. If the same test is administered to another group of 100 persons who are less proficient and only 20 of them answer item 1 correctly, then we may consider item 1 as very difficult. Therefore, the judgment of item difficulty with passing rate is sample dependent.

If the judgment of person ability depends on test difficulty, and the judgment of item (test) difficulty depends on person ability, then the goal of measuring person ability and calibrating item difficulty is not achieved. Simply put, it is not appropriate to use raw score and passing rate to describe person ability and item difficulty, respectively.

It is not appropriate to use raw score to describe ability distance between persons, either. In the previous example of mathematics test, suppose person A scores 10 points higher than person B, in a test with a maximum score of 50 points, such a difference is moderate and we may consider the ability distance between them moderate. If test developers can create many items that person A can answer correctly but person B cannot, then the difference in their raw scores can be as high as nearly 50 points. Under such a case, we may consider the ability distance between them very large. On the other hand, if test developers create items that are so easy (or difficult) such that both persons answer them correctly (or incorrectly), then the difference in their raw scores will be close to zero, indicating little ability distance between them. Therefore, the judgment of ability distance between persons with raw scores is test dependent, too.

The example of ability distance between persons applies to group difference (e.g., gender difference in mathematics) or treatment effect (e.g., the experimental group has a mean 10 points higher than the control group). The implication is that the distance between persons or groups or treatment effects can be controlled by test developers, if raw scores are used.

Growth measurement with raw scores has the same problem of test dependent. For example, person A receives a score of 30 in a pretest before an instructional treatment and a score of 31 in a posttest after the treatment. An increment of 1 single point in a test with a maximum point of 50 suggests a very small treatment effect. If a clever test developer is recruited and he/she creates items that are so difficult that almost none of the persons can answer them correctly before the treatment but in the same time so easy that almost everyone will answer them correctly after the treatment, then the score difference before and after the treatment will be nearly the perfect score of 50. This suggests that difference in raw scores between treatments or time points cannot describe person's growth appropriately.

Another interesting question is: Suppose person A receives a score 10 points higher than person B, who receives a score 10 points higher than person C, would this suggest that the ability distance between persons A and B is equal to that between persons B and C? That is, whether raw scores are interval? The answer is obviously no, because as mentioned previously, the judgment in

ability distance between persons with raw scores is test dependent. Using a different test will in general produce different distances in raw scores between persons, for example, person A receives a score 5 point higher than person B, who receives a score 20 points higher than person C in this new test. This explains that test scores are not interval.

Similar problems of interval data occur in response time. In a cognitive ability experiment where response time needed to accomplish a task is recorded, suppose person A takes 10 seconds longer than person B, who takes 10 second longer than person C, can we claim that the ability distance between persons A and B is equal to that between persons B and C? Obviously, we cannot, because if another cognitive task is used, the same distances will not generally found, for example, for a new task person A may take 5 seconds longer than person B, who may take 20 seconds longer than person C. In other words, although "seconds" are interval data in the natural sciences, they are not interval in the social sciences.

1. 2 Data Analysis of Raw Scores

Raw scores are often treated as interval and analyzed accordingly. There are two major reasons for doing so. One is that users do not realize raw scores are not interval. The other is that users, although realizing raw scores are not interval and should not be treated as such, do not have access to appropriate methods and thus have to follow conventional methods of data analysis. Since the 1960s, many researchers have realized that the unit of data analysis in the social sciences should be item response rather than test score. Original item responses are the first-hand data and should be analyzed from them. Item responses are categorical and ordinal, rather than continuous and interval. Adding up item scores to form test raw scores can at most create ordinal data, not to mention that item scores should not be summed if these items do not measure the same latent trait. With this consensus, researchers have developed a class of models for item responses, which is referred to as item response theory (IRT; Lord, 1980). This research trend declares clearly that the unit of analysis should be switched from CTT's test score to IRT's item response.

The popularity of IRT and Rasch measurement can be demonstrated by the numbers of articles included in academic databases. Searching abstracts for "item response theory" or "Rasch" up to January 2010 yields the following numbers of articles: PsycINFO (3,188), ERIC (2,597), MEDLINE (1,709), SportDiscus with Full Text (503), ABI/INFORM (502), and Sociological Abstracts (100).

2. DICHTOMOUS ITEMS

2.1 The Rasch Model

As Sir Isaac Newton discovered gravity when his head was hit with a falling down apple, Georg Rasch (1901 - 1980) developed the famous Rasch measurement model (Rasch, 1960) when he questioned what caused an incorrect or a correct answer to an item. To introduce Rasch's basic idea, let us treat item response as an effect or a dependent variable and consider what major causes or independent variables will be. Let P_{ni1} and P_{ni0} denote the probabilities of being scoring 1 and 0 on item i for person n, respectively. Define the odds as the ratio of these two probabilities. Rasch proposed that there are two major causes that affect the odds: one is person n's ability θ_n^*, and the other is item i's difficulty δ_i^*, and their relationship is:

$$odds_{ni} \equiv P_{ni1} / P_{ni0} = \theta_n^* / \delta_i^* . \quad (1)$$

Equation 1 is the Rasch model for dichotomous items. As θ_n^* and δ_i^* are dividable, they are at a ratio scale. We shall come back to this issue later.

The Rasch model has other expressions. Taking the natural logarithm of both sides of Equation 1 leads to:

$$\log\left(odds_{ni}\right) = \log\left(P_{ni1} / P_{ni0}\right) = \log\left(\theta_n^* / \delta_i^*\right) . \quad (2)$$

Because

$$\log\left(\theta_n^* / \delta_i^*\right) = \log\left(\theta_n^*\right) - \log\left(\delta_i^*\right), \quad (3)$$

and define $\log\left(\theta_n^*\right) \equiv \theta_n$, $\log\left(\delta_i^*\right) \equiv \delta_i$, and $\log\left(odds_{ni}\right) \equiv \text{logit}_{ni}$, one has:

$$\text{logit}_{ni} = \theta_n - \delta_i . \quad (4)$$

Equation 4 is the common expression of the Rasch model. In the equation, ability θ and difficulty δ are additive, meaning that they are at the same logit unit and are interval. Their values are between negative infinity and positive infinity. In practice, most of them are within ± 3.

Because of $P_{ni1} + P_{ni0} = 1$, one has:

$$P_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} , \quad (5)$$

$$P_{ni0} = \frac{1}{1 + \exp(\theta_n - \delta_i)} , \quad (6)$$

where $\exp(x)$ is the exponential function of x, and $\exp(1)$ is approximately

2.718. If a person has an ability level of -2 logits, and the item has a difficulty of 0 logit, then the probability of being correct for that person on that item is: $\frac{\exp[-2-0)]}{1+\exp[-2-0]} = 0.12$.

According to Equation 5, once θ and δ are both known, then the probability of being correct can be computed directly. In other words, the response of a person to an item becomes predictable, which is not applicable in CTT.

Figure 1 shows the relationship between the probability (of being correct) and the distance between θ and δ. When $\theta - \delta = 0$ (i.e., the person's ability level is equal to the item's difficulty), the probability is a half. When $\theta - \delta > 0$ (i.e., the person has an ability level higher than the item's difficulty), the probability is greater than 0.5, and the greater the distance is the higher the probability. When $\theta - \delta < 0$ (i.e., the person has an ability level lower than the item's difficulty), the probability is smaller than 0.5, and the greater is the distance (in absolute value) the lower the probability.

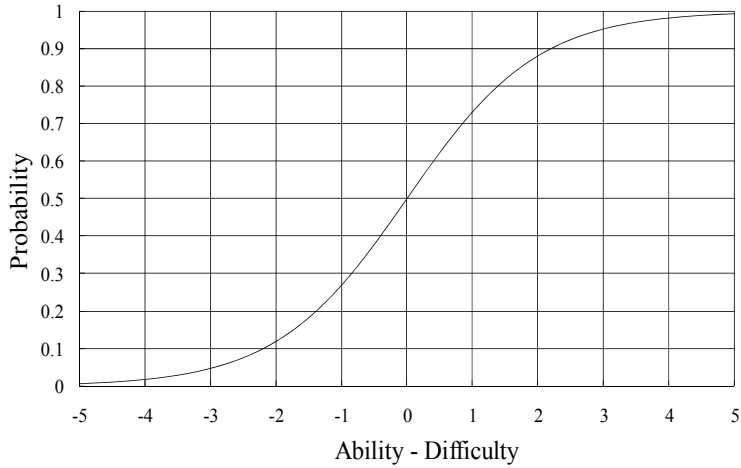

Figure 1. Probability and the difference between person ability and item difficulty

2.2 Properties of the Rasch Model

Although the Rasch model allows for prediction of item responses, are θ and δ

free from being test dependent and sample dependent? Below we use Newton's Second Law of Motion as an analogy to describe the property of θ and δ in the Rasch model. According to the Law, the net force F on an object with mass m is equal to the product of the object's mass and its acceleration a:

$$F = ma. \tag{7}$$

Suppose we want to compare the force of two persons, we may ask them to push the same object (e.g., ball) and measure their accelerations. According to the Law, we find:

$$F_1 = ma_1, \tag{8}$$

$$F_2 = ma_2. \tag{9}$$

Taking a division leads to:

$$\frac{F_1}{F_2} = \frac{ma_1}{ma_2} = \frac{a_1}{a_2}. \tag{10}$$

Mass has been cancelled out from the equation, indicating that the comparison of two forces is independent of mass. Therefore, the comparison is "objective." In addition, no matter whether the two forces are very large (e.g., these two persons are adults) or very small (e.g., they are kids), the ratio of two forces are a constant of a_1 / a_2. That is, F is at a ratio scale.

What will happen if the comparison of two forces is dependent of mass? For example, the ratio of the accelerations for two persons is 2 for a heavy ball, 3 for a light ball, and 1 for a very light ball. If this is the case, then the comparison of two forces is not possible.

Taking the natural logarithm of both sides of Equation 7 leads to:

$$\log(F) = \log(ma) = \log(m) + \log(a). \tag{11}$$

Define $F^* \equiv \log(F)$, $m^* \equiv \log(m)$ and $a^* \equiv \log(a)$. Then,

$$F^* = m^* + a^*. \tag{12}$$

Now, ask two persons to push the same ball and then measure their accelerations. According to Equation 12, one has:

$$F_1^* = m^* + a_1^*, \tag{13}$$

$$F_2^* = m^* + a_2^*. \tag{14}$$

Subtracting Equation 14 from Equation 13 leads to:

$$F_1^* - F_2^* = (m^* + a_1^*) - (m^* + a_2^*) = a_1^* - a_2^*. \tag{15}$$

Mass has been cancelled out from the equation, indicating that the comparison of two forces is independent of mass and is objective. In addition, no matter whether the two forces are very large or very small, the distance of the two forces is a constant of $a_1^* - a_2^*$. Hence, F^* is at an interval scale.

Assume two persons' ability levels θ_1^* and θ_2^* are to be compared. They are asked to respond to the same item. According to Equation 1, one has:

$$odds_1 = \theta_1^* / \delta^* , \tag{16}$$

$$odds_2 = \theta_2^* / \delta^* . \tag{17}$$

The ratio of the two equations is:

$$\frac{odds_1}{odds_2} = \frac{\theta_1^* / \delta^*}{\theta_2^* / \delta^*} = \frac{\theta_1^*}{\theta_2^*} . \tag{18}$$

Item difficulty δ^* has been cancelled from Equation 18, suggesting the comparison of two persons' ability levels is test-free and objective. In addition, no matter whether these two persons have very high ability level or not, their ratio is a constant of $odds_1 / odds_2$. That is, θ^* is at a ratio scale.

Next consider the comparison of two item difficulties. Let the same person respond to two items. According to Equation 1, one finds:

$$odds_1 = \theta^* / \delta_1^* , \tag{19}$$

$$odds_2 = \theta^* / \delta_2^* . \tag{20}$$

The ratio of the two equations is:

$$\frac{odds_1}{odds_2} = \frac{\theta^* / \delta_1^*}{\theta^* / \delta_2^*} = \frac{\delta_2^*}{\delta_1^*} . \tag{21}$$

Person ability has been cancelled out from the ratio, meaning that the comparison of two item difficulties is sample-free and objective. No matter whether these two items are very difficult or easy, their ratio is a constant of $odds_1 / odds_2$. Hence, δ^* is at a ratio scale.

The use of Equation 4 leads to the same conclusion. Let two persons with ability level of θ_1 and θ_2 respond to the same item. According to Equation 4, one finds:

$$\text{logit}_1 = \theta_1 - \delta , \tag{22}$$

$$\text{logit}_2 = \theta_2 - \delta . \tag{23}$$

Taking a subtraction for them leads to:

$$\text{logit}_1 - \text{logit}_2 = (\theta_1 - \delta) - (\theta_2 - \delta) = \theta_1 - \theta_2 . \tag{24}$$

Item difficulty has been cancelled out from the equation, suggesting that the comparison of ability levels is test-independent and objective. Besides, no matter whether the two persons have high or low ability levels, their distance is a constant of $\text{logit}_1 - \text{logit}_2$, indicating θ is at an interval scale. Likewise, for the comparison of two item difficulties, let the same person respond to two

items. According to Equation 4, one finds:

$$\text{logit}_1 = \theta - \delta_1 , \tag{25}$$

$$\text{logit}_2 = \theta - \delta_2 . \tag{26}$$

Taking a subtraction leads to:

$$\text{logit}_1 - \text{logit}_2 = (\theta - \delta_1) - (\theta - \delta_2) = \delta_2 - \delta_1 . \tag{27}$$

Person ability is cancelled out, suggesting that the comparison of two item difficulties is sample independent and objective. No matter these two items are very difficult or easy, their distance is a constant of $\text{logit}_1 - \text{logit}_2$, suggesting an interval scale.

In the Rasch model, θ and δ can be separated from each other and are thus test-independent and sample-independent. Rasch called this property of parameter separation as "specific objectivity."

2. 3 The Two- and Three-Parameter Models

In the Rasch model, each item has only a single parameter called difficulty. Hence, the model is also called the one-parameter model. About the same time when Rasch developed his measurement model, American researchers Allan Birnbaum, Frederic M. Lord and others proposed similar models. For example, Birnbaum (1968) proposed the two-parameter model as:

$$P_{ni1} = \frac{\exp[a_i(\theta_n - \delta_i)]}{1 + \exp[a_i(\theta_n - \delta_i)]} , \tag{28}$$

and the three-parameter model as:

$$P_{ni1} = c_i + (1 - c_i) \times \frac{\exp[a_i(\theta_n - \delta_i)]}{1 + \exp[a_i(\theta_n - \delta_i)]} , \tag{29}$$

where a_i is the slope parameter, c_i is the asymptotic parameter, and δ_i is the location parameter, of item i. If $c_i = 0$ for every item, then Equation 29 becomes Equation 28. If $a_i = 1$ for every item, then Equation 28 becomes Equation 5, the Rasch model. In this regard, the one-parameter model is a special case of the two-parameter model, which is a special case of the three-parameter model.

Do the parameters in the two- or three-parameter model share the same property of specific objectivity as the Rasch model? Take the two-parameter model as an example. Equation 28 can be written as:

$$\log\left(odds_{ni}\right) \equiv \log\left(P_{ni1} / P_{ni0}\right) = a_i\left(\theta_n - \delta_i\right) . \tag{30}$$

In order to compare two persons' ability levels, let them respond to the same item. According to Equation 30, one finds:

$$\text{logit}_{1i} \equiv a_i(\theta_1 - \delta_i)\,, \tag{31}$$

$$\text{logit}_{2i} \equiv a_i(\theta_2 - \delta_i)\,, \tag{32}$$

Taking a subtraction leads to:

$$\begin{aligned} &\log_{1i} - \log_{2i} \equiv a_i(\theta_1 - \delta_i) - a_i(\theta_2 - \delta_i) = a_i(\theta_1 - \theta_2) \\ &\Rightarrow \theta_1 - \theta_2 = (\text{logit}_{1i} - \text{logit}_{2i}) / a_i \end{aligned} \tag{33}$$

Obviously, the distance between the two persons, $\theta_1 - \theta_2$, changes across items. As a result, the comparison of ability is no longer objective. Only when a_i is a constant (e.g., 1), will the item parameter a_i be cancelled out from Equation 33 and the comparison of ability be objective. Under such a case, the model is actually the Rasch model. Similar conclusions can be drawn for the three-parameter model.

The parameters in the Rasch model, θ and δ, have the same logit unit. We may say a person has an ability of 2 logits or an item has a difficulty of 3 logits. However, in the two- or three-parameter models, the person and the item parameters do not have the same unit. It is thus not applicable to say a person has an ability of 2 logits or an item has a difficulty of 3 logits.

Some introductory textbooks of item response theory claim that the parameters in the two- or three-parameter models are test-independent and sample-independent. In fact, this independence in the two- or three-parameter models applies to only parameter estimation, rather than parameter separation. Independence in parameter estimation is not a unique property of IRT models. Ordinary linear models have this property. For example, in a simple regression $\hat{Y} = a + bX$, the estimation of parameters *a* and *b* does not depend on the range of *X*. That is, given the model is true, the estimation of *a* and *b* when *X* is low will be equivalent to that when *X* is high. However, CTT does not have independence even in parameter estimation, because CTT does not formulate any functional relationship between persons and items.

2.4 Item Characteristic Curve

Figure 2 shows probabilities in three items with difficulties -2, 0, and 1 logit across ability levels under the Rasch model. The curve is called item characteristic curve or item response function. Two fundamental properties can be found:

1. For any item, the higher is the ability the higher the probability. When the ability approaches positive infinity, the probability approaches 1; when the

ability approaches negative infinity, the probability approaches 0. That is, the probability is monotonically increasing across person ability levels.

2. For any ability level, the higher is the item difficulty, the lower the probability. For example, the probability of being correct in item 1 is always higher than that in item 2, which is always higher than that in item 3. That is, the probability is monotonically increasing across item "easiness" levels.

Note that these item response functions are nonlinear, because a linear function will eventually lead to a probability greater than 1 as the ability increases, or smaller than 0 as the ability decreases, which is theoretically impossible.

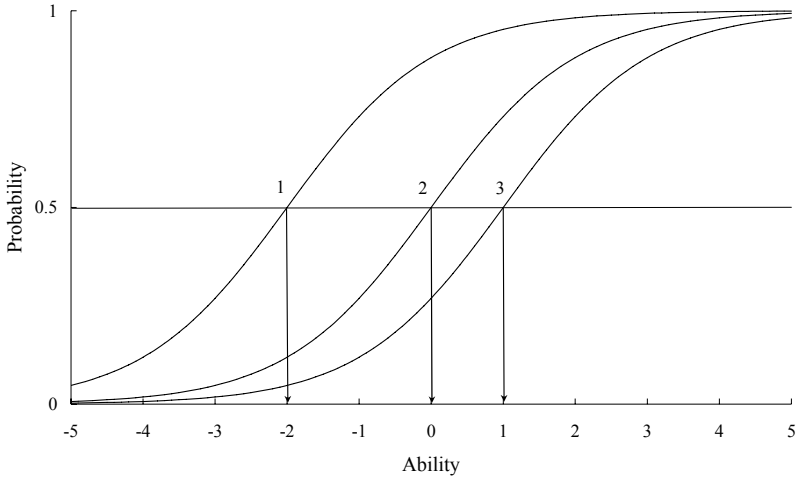

Figure 2. Item characteristic curves for three items under the Rasch model

Figure 3 shows item characteristic curves of three items under the two-parameter model. The three items have a location parameter of 0, and a slope parameter of 2, 1, and 0.5, respectively. At the point of 0, item 1 has the largest slope, followed by item 2, and item 3 has the smallest slope. For ability levels near the location parameter (here, 0), the increment in probability via a unit of increment in ability is the largest for item 1, followed by the item 2, and is the smallest for item 3. This is why the slope parameter is also called the discrimination parameter. An item that has the highest discrimination power at

the ability levels near the location parameter, can have the lowest discrimination power at other ability levels. For example, among the three items, item 1 has the highest discrimination power for those persons with an ability level around 0, and the lowest discrimination power for those persons with an ability level far away from 0 (e.g., -2 or 2).

For the Rasch model, the following two fundamental conditions hold: (1) for any item, the higher the ability is, the higher the probability; and (2) for any person, the easier the item is, the higher the probability. For the two- or three-parameter models, only condition 1 holds, but condition 2 does not. As shown in Figure 3, for those persons with ability levels below 0, the rankings of the probabilities for the three items are 3 > 2 > 1; for those persons with ability levels above 0, the rankings are 1 > 2 > 3; for those persons with ability of 0, the probabilities are the same across items. It is not possible to tell which item is more difficult. Thus, the location parameter cannot be interpreted as difficulty, or the slope parameter as discrimination.

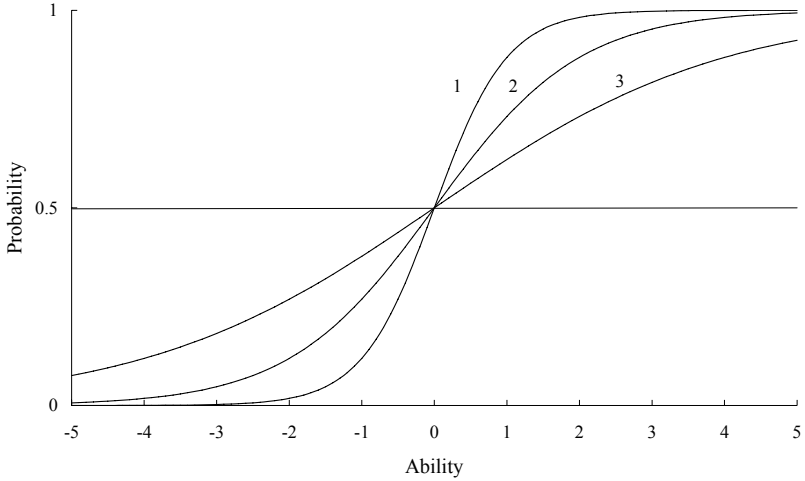

Figure 3. Item characteristic curves for three items with an identical location parameter but different slope parameters under the two-parameter model

Figure 4 shows item characteristic curves of three items under the three-parameter model. The δ, *a*, *c* parameters are 0, 2, and 0.2 for item 1; 0, 1, and 0 for item 2; and 0, 0.5, and 0.3 for item 3. The *c*-parameter denotes the probability for a person with an ability of negative infinity. Such a probability

is 0.2 for item 1, 0 for item 2, and 0.3 for item 3. The three-parameter model has been widely applied to multiple-choice items, because many practitioners intuitively assume people with very low ability will randomly select an option from all options in a multiple-choice item. Therefore, the *c* parameter is also called the pseudo-guessing parameter.

The three-parameter model inherits the properties of the two-parameter model: Condition 1 holds but condition 2 does not. In the three items in Figure 4, all the δ parameters are 0. However, the rankings of the probabilities for the three items vary across ability levels. Hence, the location parameter cannot be interpreted as difficulty, the slope parameter cannot be interpreted as discrimination, and the asymptotic parameter cannot be interpreted as pseudo-guessing.

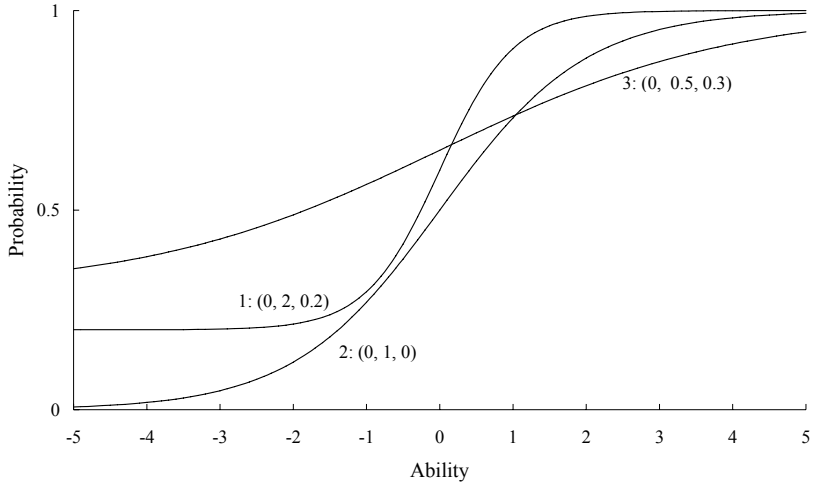

Figure 4. Item characteristic curves for three items with an identical location parameter under the three-parameter model

2.5 Statistics Perspective vs. Measurement Perspective

In the 1960s, Rasch, Birnbaum, and Lord proposed their models for test data. However, their models are different, which was because they adopted different perspectives. Rasch, adopting the measurement approach, attempted to establish a measurement model that yields objective measurement. Birnbaum

and Lord, adopting the statistics prospective, wished to describe the relationship between person and item. In the statistics prospective, data are perfect and should not be edited. The major task for a data analyst is to develop a statistical model that fits the data. If the model-data fit is not good enough, another more complicated model should be pursued. This step repeats until the model has a good fit. In this perspective, the one-parameter model is only a special case of the two- or three-parameter model, and there is little academic merit in Rasch measurement. If the one-parameter model does not have a good fit, then one should try the two-parameter model, if it still does not have a good fit, then try the three-parameter model. In fact, when one adopts the statistics approach seriously, then one will realize that no model can have a perfect fit to the data, and that the best model is the data set itself.

The value of a model is not on its truth or fault, because any model is a simplified theory and cannot have a perfect fit to data. The value of a model is on its usefulness. The Rasch model is not developed only to fit data, rather it is developed to diagnose data and to clean data in order to yield objective scales for persons and items. Anyone who has developed tests will be concerned whether every item is appropriately written. Anyone who has administered tests will recognize that there may be a large amount of noise in raw data to be cleaned. Test-takers may misunderstand items, and they may be careless or even cheating. Data recoding is another source of noise. We need an efficient measurement model to diagnose noise so as to yield meaningful and objective measurement. The Rasch model as well as its associated technique is such a tool. Although the two- and three-parameter models can detect noise in data to some extent, they fail to yield objective measurement due to their model limitations.

Many Rasch scholars do not agree to put the Rasch model under the IRT category, because the Rasch and the two- or three-parameter models were developed from different perspectives and different goals. However, more and more researchers agree to do so, because after all these models are developed to fit item responses.

Many testing companies or users prefer the two- or three-parameter model, mainly because the Rasch model is too simple to fit their data. Given that the data cannot be revised (e.g., high-stakes examinations), a poor fit between the Rasch model and the data would cause an instant crisis, because objective measurement is not possible. Even when the data can be revised to some extent, the revision is often very costly and time-consuming. Under such a case, adopting a more complicated model like the two- or three-parameter model to improve model-data fit will be much easier and more feasible.

After tests are developed and administered to test-takers, many test analysts may incline to adopt the two- or three-parameter model because of a better fit. If, on the other hand, a new test is to be developed or an old test is to be revised, then the use of Rasch technique to monitor test development, improve test quality, and yield objective measurement should be encouraged (Wilson, 2005).

2.6 Parameter Estimation

When item responses are collected, the next step is to estimate person ability and item difficulty. Many computer programs can be used. Here, one of the common estimation methods, the maximum likelihood estimation (MLE), is introduced briefly. Assume there are five items with difficulty of -2, -1, 0, 1 and 2, respectively, and a person has a response pattern of (1,1,1,0,0) on the five items. What is the ability of that person? His/her ability may be high, median, or low. The MLE principle is to select an ability level that is the most likely to generate such a response pattern. The Rasch model is taken as an example to explain the MLE method.

For a person with $\theta = -3$, the probability on item 1 ($\delta = -2$) is

$$\frac{\exp(\theta_n - \delta_i)}{1+\exp(\theta_n - \delta_i)} = \frac{\exp[-3-(-2)]}{1+\exp[-3-(-2)]} = 0.2689 .$$

Likewise, the probability on item 2 ($\delta = -1$) is 0.1192 and that on item 3 ($\delta = 0$) is 0.0474. The probability (of being incorrect) on item 4 ($\delta = 1$) is

$$\frac{1}{1+\exp(\theta_n - \delta_i)} = \frac{1}{1+\exp(-3-1)} = 0.9820$$

and that on item 5 ($\delta = 2$) is 0.9933. Thus, the likelihood of generating a response pattern of (1,1,1,0,0) is the product of the five probabilities: $0.2689 \times 0.1192 \times 0.0474 \times 0.9820 \times 0.9933 = 0.0015$.

Similar computations can be conducted for other ability levels, and they are summarized in Table 1 and Figure 5. The largest (maximum) likelihood is 0.2402 and it is located at the ability level around 0.6 (or 0.591 more precisely). We thus claim that the MLE estimate for that person's ability level is 0.6.

Table 1. Probabilities of the five items and their likelihood under the Rasch model

Item	1	2	3	4	5	
δ	-2	-1	0	1	2	
Score	1	1	1	0	0	
Ability	Prob.	Prob.	Prob.	Prob.	Prob.	Likelihood
-3.0	0.2689	0.1192	0.0474	0.9820	0.9933	0.0015
-2.8	0.3100	0.1419	0.0573	0.9781	0.9918	0.0024
-2.6	0.3543	0.1680	0.0691	0.9734	0.9900	0.0040
-2.4	0.4013	0.1978	0.0832	0.9677	0.9879	0.0063
-2.2	0.4502	0.2315	0.0998	0.9608	0.9852	0.0098
-2.0	0.5000	0.2689	0.1192	0.9526	0.9820	0.0150
-1.8	0.5498	0.3100	0.1419	0.9427	0.9781	0.0223
-1.6	0.5987	0.3543	0.1680	0.9309	0.9734	0.0323
-1.4	0.6457	0.4013	0.1978	0.9168	0.9677	0.0455
-1.2	0.6900	0.4502	0.2315	0.9002	0.9608	0.0622
-1.0	0.7311	0.5000	0.2689	0.8808	0.9526	0.0825
-0.8	0.7685	0.5498	0.3100	0.8581	0.9427	0.1060
-0.6	0.8022	0.5987	0.3543	0.8320	0.9309	0.1318
-0.4	0.8320	0.6457	0.4013	0.8022	0.9168	0.1586
-0.2	0.8581	0.6900	0.4502	0.7685	0.9002	0.1844
0.0	0.8808	0.7311	0.5000	0.7311	0.8808	0.2073
0.2	0.9002	0.7685	0.5498	0.6900	0.8581	0.2252
0.4	0.9168	0.8022	0.5987	0.6457	0.8320	0.2365
0.6	0.9309	0.8320	0.6457	0.5987	0.8022	0.2402
0.8	0.9427	0.8581	0.6900	0.5498	0.7685	0.2359
1.0	0.9526	0.8808	0.7311	0.5000	0.7311	0.2242
1.2	0.9608	0.9002	0.7685	0.4502	0.6900	0.2065
1.4	0.9677	0.9168	0.8022	0.4013	0.6457	0.1844
1.6	0.9734	0.9309	0.8320	0.3543	0.5987	0.1599
1.8	0.9781	0.9427	0.8581	0.3100	0.5498	0.1349
2.0	0.9820	0.9526	0.8808	0.2689	0.5000	0.1108
2.2	0.9852	0.9608	0.9002	0.2315	0.4502	0.0888
2.4	0.9879	0.9677	0.9168	0.1978	0.4013	0.0696
2.6	0.9900	0.9734	0.9309	0.1680	0.3543	0.0534
2.8	0.9918	0.9781	0.9427	0.1419	0.3100	0.0402
3.0	0.9933	0.9820	0.9526	0.1192	0.2689	0.0298

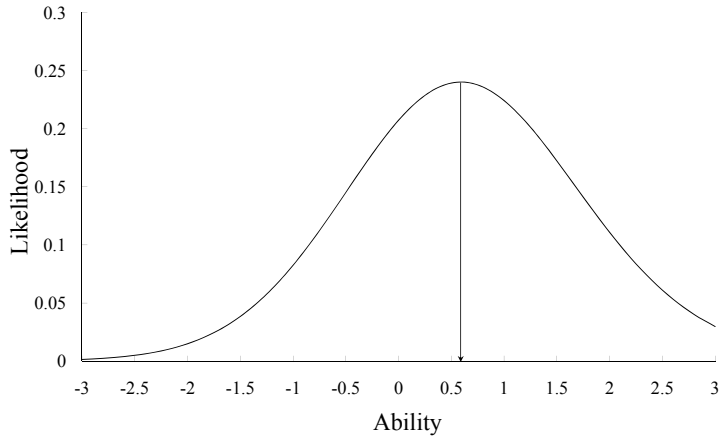

Figure 5. Likelihood of being a response pattern of (1,1,1,0,0) on five items under the Rasch model

In the above calculation, item difficulties are assumed to be known in advance. In reality, both person ability and item difficulty are often unknown and should be estimated jointly from data. Under such a case, one may give an initial guess to the item difficulty such as $\log (P_0 / P_1)$, and then based on these initial guess, yield an MLE for every person. Then, based on the initial MLEs for persons, yield an MLE for every item, then update an MLE for every person again. This step repeats until the change in estimates between two consecutive iterations is very small. Most computer programs adopt more efficient ways of parameter estimation, but the general principle remains unchanged.

2.7 Raw Score and Person Measure

Under the Rasch model, raw score is a sufficient statistic for person ability. In other word, raw score and the Rasch person measure has a one-to-one correspondence. Two persons with the same raw scores receive the same Rasch person measures. A higher raw score corresponds to a higher Rasch person measure, and a low raw score to a lower Rasch person measure. As shown in Table 2, where the five items have difficulties of -2, -1, 0, 1, 2 logits, the Rasch person measures are all -1.93 for those response patterns that have a raw score of 1; -0.59 for a raw score of 2, 0.59 for a raw score of 3; and 1.93 for a raw score of 4.

A close visit to the likelihoods in Table 2 reveals that some response patterns are aberrant. For example, among those response patterns with a raw score of 1, the likelihood is 0.3017 when the easiest item is answered correctly; but is only 0.0055 when the most difficult item is answered correctly. In general, the smaller the likelihood is, the more aberrant the response pattern. Table 2 does not show the person measures for a zero score or a perfect score, which is because the MLE is located at negative infinity and positive infinity, respectively. This means that we are not able to yield a finite estimate for these response patterns with the current test. Under such a case, easier (or more difficult) items should be administered until the score is neither zero nor perfect.

In the two- or three-parameter model, raw score and the ability estimate do not have a one-to-one correspondence, meaning that a lower raw score may receive a higher ability measure. As shown in Table 2, under the two-parameter model, a raw score of 1 can receive an ability measure of -2.29 or 0.51, whereas a raw score of 2 can receive an ability measure of -0.71. This will cause practical problems. Imagine how serious it will be when in high-stakes tests (e.g., college entrance examinations) one of the two test-takers with the same raw score is accepted but the other is rejected; or a test-taker with a lower raw score is accepted but another test-taker with a higher raw score is rejected. Those who advocate the two- or three-parameter model might argue that two test-takers with the same raw score can receive different measures because the one who answered a more difficult item correctly should receive a higher credit than the one who answered an easier item correctly. This sounds reasonable. Actually, it is not. When two test-takers have the same raw score, one answering a more difficult item correctly must at the same time have answered easier items incorrectly. If a correct answer to a more difficult item should receive a higher credit, then an incorrect answer to an easier item should receive a higher debt, too. Actually, when a person answers more difficult items correctly but easier items incorrectly, then the response pattern is aberrant and further investigation is needed.

Table 2. Response patterns, likelihoods, and ability estimates under the Rasch and two-parameter models

Response Pattern					Rasch Model		Two-Parameter Model	
					Likelihood	Measure	Likelihood	Measure
1	0	0	0	0	0.3017	-1.93	0.2758	-2.29
0	1	0	0	0	0.1110	-1.93	0.1673	-2.29
0	0	1	0	0	0.0408	-1.93	0.0507	-0.71
0	0	0	1	0	0.0150	-1.93	0.0069	0.51
0	0	0	0	1	0.0055	-1.93	0.0009	0.51
1	1	0	0	0	0.2402	-0.59	0.2273	-0.71
1	0	1	0	0	0.0884	-0.59	0.1195	0.06
1	0	0	1	0	0.0325	-0.59	0.0265	0.85
1	0	0	0	1	0.0120	-0.59	0.0036	0.85
0	1	1	0	0	0.0325	-0.59	0.0725	0.06
0	1	0	1	0	0.0120	-0.59	0.0161	0.85
0	1	0	0	1	0.0044	-0.59	0.0022	0.85
0	0	1	1	0	0.0044	-0.59	0.0161	1.15
0	0	1	0	1	0.0016	-0.59	0.0022	1.15
0	0	0	1	1	0.0006	-0.59	0.0012	1.71
1	1	1	0	0	0.2402	0.59	0.2289	0.51
1	1	0	1	0	0.0884	0.59	0.0722	1.15
1	1	0	0	1	0.0325	0.59	0.0098	1.15
1	0	1	1	0	0.0325	0.59	0.0833	1.43
1	0	1	0	1	0.0120	0.59	0.0113	1.43
1	0	0	1	1	0.0044	0.59	0.0084	2.01
0	1	1	1	0	0.0120	0.59	0.0505	1.43
0	1	1	0	1	0.0044	0.59	0.0068	1.43
0	1	0	1	1	0.0016	0.59	0.0051	2.01
0	0	1	1	1	0.0006	0.59	0.0092	2.36
1	1	1	1	0	0.3017	1.93	0.3005	1.71
1	1	1	0	1	0.1110	1.93	0.0407	1.71
1	1	0	1	1	0.0408	1.93	0.0412	2.36
1	0	1	1	1	0.0150	1.93	0.0912	2.87
0	1	1	1	1	0.0055	1.93	0.0553	2.87

2.8 Model-Data Fit

Only when there is a good fit between the data and the Rasch model, will the resulting person and item measures be objective and interval. Hence, model-data fit is a critical issue in Rasch analysis. After parameter estimation, that is, measures for every person and every item become known, we can compute the probability of a person on an item, which is the expected item score. Subtracting the expected item score from the observed item response (either 0 or 1) forms the residual score. If the residual score is large (meaning that the observed score is very different from the expected score), then the model-data fit is poor.

The major task of residual analysis are twofold: (a) person fit, whether a person's response pattern has a good fit; and (b) item fit, whether an item's response pattern has a good fit. For example, person A responds to a test with 20 items. After person A's measure and the difficulties for the 20 items are calibrated, we can compute the person's expected scores on all items and thus their residual scores. The next step is to examine whether the null hypothesis of a good model-data fit can be statistically rejected. If so (e.g., the person answered easy items incorrectly but difficult items correctly), then person A's response pattern does not match the model's expectation, indicating a poor fit. Likewise, we can examine whether an item has a good fit. For example, assume 100 persons respond to item 1. After these persons' measures and the item difficulty are calibrated, the expected scores and the residual scores for the 100 persons on that item can be obtained. If the null hypothesis of a good fit is rejected, then the item does not have a good fit.

In practice, it is common that some persons or items do not have a good fit. There are many causes. For instance, test-takers may cheat, be careless, too nervous or tired. They may use unexpected skills to solve the problems, or some of the items were just taught in cram schools. Whatever the reason may be, we have to admit that our measurements for these persons are not successful and we are not able to quantify their ability levels. From clinical points of view, these aberrant response patterns deserve follow-up investigation. A new explanation or theory may thus be created. There are many reasons for a poor item fit. Items may not be clearly written. They may measure dimensions that are different from that measured by other items in the same test. For example, in a mathematics test with word problems, the wording in some items may be too difficult for some test-takers to understand.

A poor-fit item is often simply removed from the test. However, it should be noted that the item is removed not because it is unimportant, but because it

does not work harmoniously with the other items. If the construct measured by a poor-fit item is very important, then a stand-alone test should be developed such that the construct can be measured more precisely.

3. MODEL EXTENSIONS

After 40 years of development, Rasch measurement is very mature and is widely used in practice. Below, several extensions of the Rasch model are introduced.

3.1 Polytomous Items

Tests may contain polytomous items (e.g., essays, rating scale items, Likert items). Item scores are ordinal, not interval. The Rasch model for dichotomous items can be extended to fit polytomous items. Let P_{nij} and $P_{ni(j\text{-}1)}$ denote the probabilities of scoring j and $j-1$ on item i for person n, respectively, θ_n denote person n's ability, and δ_{ij} denote the j-th step difficulty of item i. Under the partial credit model (Masters, 1982), it is assumed:

$$\text{logit}_{nij} \equiv \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = \theta_n - \delta_{ij} = \theta_n - (\delta_i + \tau_{ij}), \tag{34}$$

where after reparameterization, δ_i is the mean of the step difficulties in item i and is called the overall difficulty, and τ_{ij} is the j-th deviation from the mean and is called the j-th threshold for item i. Suppose item i has $M+1$ categories and they are scored as 0, 1, …, M, then there will be M step difficulties of δ_{ij} for that item. The first step difficulty describes how difficult it is by moving from category 1 (scoring 0) to category 2 (scoring 1), the second step difficulty describes how difficult it is by moving from category 2 (scoring 1) to category 3 (scoring 2), and so on, the M-th step difficulty describes how difficult it is by moving from category M to category $M+1$. The step difficulties correspond to the points on the ability scale where two successive item response category characteristic curves intersect. These step difficulties can be reparameterized as a mean difficulty of δ_i and M thresholds of τ_{ij}, given that the M thresholds sum to zero.

If items in a test are scored according to the same rubric, for example, rating scale items or Likert items, it is justifiable that all the items share the same set of thresholds. This is the rating scale model (Andrich, 1978):

$$\text{logit}_{nij} = \theta_n - (\delta_i + \tau_j), \tag{35}$$

where τ_j does not have the subscript of i, suggesting all the items share the same set of thresholds. The partial credit model has been widely applied to constructed-response items, whereas the rating scale model to rating scale items.

Figure 6 shows the item characteristic curves for a 4-point item. The three step difficulties are -3, -2, and 2, respectively. Note that the first and the second category characteristic curves interest at -3, the first step difficulty; the second and the third category characteristic curves interest at -2, the second step difficulty; and the third and the fourth category characteristic curves interest at 2. The mean of the three step parameters is -1, which is the overall difficulty. The three thresholds are thus -2, -1 and 3, respectively. If items in a test follow the rating scale model, only the overall difficulty will vary but the thresholds will be identical, across items. In other words, the patterns of item characteristic curves remain unchanged across items, but the locations may shift horizontally.

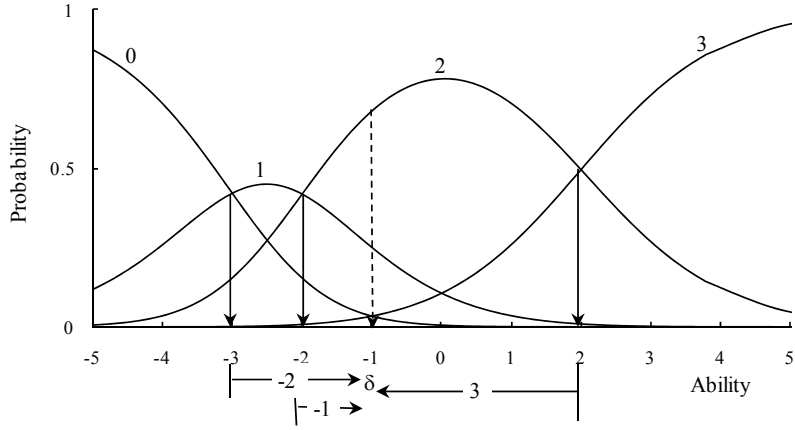

Figure 6. Item characteristic curves for a 4-point item with steps difficulties -3, -2 and 2

3.2 Many Facets and Linear Decomposition

In the aforementioned models, it is assumed that only two major factors (also called facets) govern item responses: person and item. The other facets are treated as random errors. Hence, they are referred to as two-facet models. In some testing situations, additional facets may be involved. For example, item responses to open-ended items are often scored by raters. In addition to the usual facets of person and item, a third facet of "rater" can be involved. The same item response will receive a lower score when it is judged by a severe rater than by a lenient rater. If so, rater effect should be consider as the third

facet. The facets model (Linacre, 1989) was developed for such data:

$$\log\left(\frac{P_{nijk}}{P_{ni(j-1)k}}\right) = \theta_n - (\delta_i + \tau_j + \eta_k), \tag{36}$$

where P_{nijk} and $P_{ni(j-1)k}$ are the probabilities of scoring j and $j-1$ on item i for person n when judged by rater k; η_k is the severity of rater k; and the others are defined as those in the rating scale model. The larger is η_k, the more difficult to receive a high score from rater k. There are three facets in Equation 36: person ability, item difficulty and rater severity. Thus, it is a three-facet model. The model can be easily generalized to involve more facets.

If raters do exhibit different degrees of severity, then they should be directly considered in the model. Ignoring rater effects by fitting standard two-facet models will produce biased person measures and thus ruin test fairness. Nowadays, the facets model has been widely used to examine rater effects, especially in language testing where raters are involved.

One may adopt the concept of analysis of variance and treat item response as a dependent variable, and person ability, item difficulty and rater severity as three independent variables. In such a three-way factorial design, it is assumed only the three main effects exist and the two-way or three-way interaction effects do not. The assumption of no interaction effects is to ensure objective measurement. If interaction effects do exist in the data (e.g., item difficulty depends on persons), then the meaning of difficulty is vague, so too the meaning of ability.

The basic principle in the facets model is to linearly decompose item parameters. Consider that items are generated from a combination of multiple features, for example, items of figure rotation are constructed by (a) number of lines, (b) complexity of shapes, (c) rotation angles, and (c) number of dimensions. It is thus justifiable that item difficulty is a linear composition of these features (Embretson, 1998). That is, the difficulty for a dichotomous item can be formulated as:

$$\delta_i = \boldsymbol{\beta}' \mathbf{X}_i = \beta_1 X_{i1} + \cdots + \beta_1 X_{ip}, \tag{37}$$

where δ_i is the difficulty of item i; $\boldsymbol{\beta}$ is the regression vector of $\beta_1, \ldots, \beta_p$; and $\mathbf{X}_i$ is the design vector of the p features, $X_{i1}, \ldots, X_{ip}$. Hence, the item response model becomes:

$$\text{logit}_{ni} = \theta_n - (\beta_1 X_{i1} + \cdots + \beta_1 X_{ip}), \tag{38}$$

which is the linear logistic test model (Fischer, 1973). This model has been generalized to polytomous items (Fischer & Parzer, 1991; Fischer & Pononcy, 1994)

The facets model intends to add more facets into standard two-facet models. The linear logistic test model intends to linearly decompose the item facet into several sub-facets (features). The basic logics of these two approaches are actually identical.

The major advantage of linear decomposition of item difficulty is to reveal the contribution of each feature to the difficulty. For example, if the Rasch model is fit to items of figure rotation, an item receives a difficulty estimate. What contributes to the difficulty remains unknown. In contrast, if the linear logistic test model is fit, then not only each item receives a difficulty estimate, but also each feature receives a regression weight to depict its contribution to the item difficulty. Suppose the regression weight for the feature "number of dimensions" is much larger than that for "number of lines", meaning that "number of dimensions" contributes more to the difficulty than "number of lines", then more resource should go into teaching the former feature than the latter feature. Moreover, the difficulties of new items that are generated from these features can be directly computed from Equation 37, no need of empirical administration. This is ideal in test development: Difficulty can be theoretically derived without empirical administration. One of the exciting examples is the lexile framework for English reading comprehension (visit http://www.lexile.com). The difficulty of reading a text can be directly computed from two components: length of each sentence and frequency of each word.

3.3 Multilevel Models

Most standard IRT or Rasch models do not have a multilevel structure. Assume we are interested in estimating gender difference in some latent trait measured by some test (e.g., mathematics). Following standard procedures, we would first fit an IRT model to the test data to obtain person ability estimates, and then apply an ordinary regression or an independent sample *t*-test to the person ability estimates. In doing so, the person ability estimates are treated as true values and their measurement errors are ignored. As measurement errors in the social sciences are often too large to ignore, a multilevel modeling that takes into account measurement error is needed. At the first level, an IRT model is fit to describe the relationship between items and persons. The model can be the Rasch model for dichotomous items, or the partial credit model or the rating scale model for polytomous items, or the facets model for rater data. At the second level, the person measures are treated as a criterion variable which is then regressed on a set of predictors (e.g., gender and age):

$$\theta_n = \boldsymbol{\lambda}' \mathbf{W}_n + \varepsilon_n , \tag{39}$$

$$\varepsilon_n \sim N(0, \sigma_\varepsilon^2) , \tag{40}$$

where $\mathbf{W}_n$ is a vector of observed predictors for person n, and $\boldsymbol{\lambda}$ is the vector of regression parameters. The parameters in first level model and the second level model can be estimated simultaneously, such that measurement errors are taken into account (Adams, Wilson, & Wu, 1997).

The regression coefficients $\boldsymbol{\lambda}$ can be further regressed on another set of predictors to form a three-level model, such as when persons (students) are nested within organizations (schools), and so on for more levels, as shown in Figure 7. Curves in Figure 7 are purposely used to depict nonlinear IRT functions between items and persons. Multilevel modeling can be applied to the facets models and others (Wang & Jin, 2010; Wang & Liu, 2007).

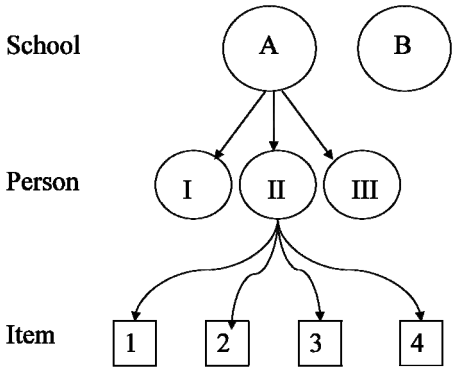

Figure 7. Graphical representation of a three-level item response model

3.4 Mixture Models

When the Rasch model is found to have a poor fit to a data set, it does not necessarily mean that the test involves more than one dimension. It may be because there are sub-populations in the test-takers and different sub-populations of test-takers treat the test differently. Cognitive psychology has shown that individuals at the same proficiency level may differ qualitatively in the mechanisms underlying their performance. The patterns of item difficulty for different groups of test-takers can reveal their qualitative differences. For example, items can be hard if one strategy is adopted to solve

the item, but much easier when another strategy is adopted. Within this context, one may argue that the Rasch model holds only within sub-populations but not across sub-populations. The membership of sub-populations for each test-taker is unknown (latent) and should be estimated from data.

Rost (1990) proposed the mixed Rasch model to jointly estimate the item and person parameters as well as the group membership. The model is mixed, because it integrates both latent trait models (i.e., IRT) and latent class models. The mixed Rasch model is given as follows:

$$P_{ni1} = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_n - \beta_{ig})}{1 + \exp(\theta_n - \beta_{ig})}, \tag{41}$$

where P_{ni1} is the probability of scoring 1 in item i for person n, β_{ig} is the difficulty of item i for group g ($g = 1, \ldots, G$), and π_g is the class size parameter or mixing proportion. The primary parameters in the model are the class proportions and the item difficulties within each class of g. Class membership for individual test-takers are estimated post hoc from their relative likelihood in the different classes. Equation 41 has been extended to fit polytomous items (Rost, 1991). The mixed Rasch model has been applied to both ability tests and non-ability tests (Maij-de Meij , Kelderman, & van der Flier, 2008; Rost, Carstensen, & Von Davier, 1997).

3.5 Testlet Response Models

Testlet-based items (Wainer, 1995), where a set of items share a common stimulus, e.g., a reading comprehension passage or a figure, have been widely used in educational and psychological tests. As different persons may have differential perspective or background knowledge on the common stimulus, the assumption of local item independence between items within a testlet may be violated. Testlet response models have been proposed to take into account local dependence within items in a testlet by adding a random-effect variable into IRT models, one for each testlet (Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005b):

$$\text{logit}_{ni} = \theta_n - \delta_i - \gamma_{nd(i)}, \tag{42}$$

$$\gamma_{nd(i)} \sim N(0, \sigma^2_{\gamma_d}), \tag{43}$$

where $\gamma_{nd(i)}$ is the interaction between person n and item i within testlet d, and it is assumed to be normally distributed; and the others are defined as those in the Rasch model. For testlet polytomous items, Equation 42 can be extended as:

$$\text{logit}_{nij} = \theta_n - (\delta_i + \tau_{ij}) - \gamma_{nd(i)}. \tag{44}$$

The magnitude of $\sigma^2_{\gamma_d}$ describes the testlet effect: the larger is $\sigma^2_{\gamma_d}$, the stronger the testlet effect. If $\sigma^2_{\gamma_d}$ is zero for every testlet, then Equations 42 and 44 become the Rasch model and the partial credit model, respectively. Testlet response models can be represented as Figure 8 where items 1 and 2 are independent items, items 3 and 4 belong to the first testlet, and items 5 and 6 belong to the second testlet.

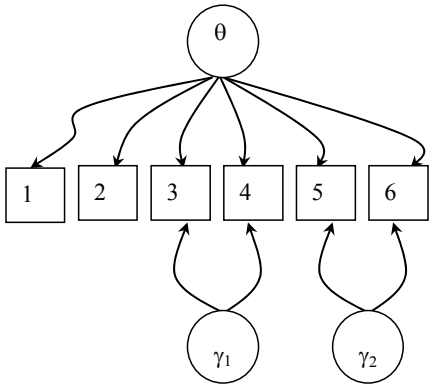

Figure 8. Graphical representation of a testlet response model

3.6 Multidimensional Models

In developing educational and psychological tests, there is an inevitable tension between the desire for precise measurement and the desire for a wide range of measures. In any given test, a choice must be made between measuring a very specific attribute with a high degree of accuracy, and sampling a vaster range of attributes with much less accuracy. Since actual testing time is typically limited, test developers often have to sacrifice accuracy and develop several short tests in order to cover as many important attributes as testing time allows. Given that scores of short tests can be terribly unreliable (low measurement precision), it would be very desirable if the reliability of scores for a short test could be increased to a more satisfactory level by the adoption of a more efficient statistical method.

Standard IRT models are unidimensional. If a test consists of several subtests and each subtest measures a distinct latent trait, it may be analyzed with unidimensional IRT models in two ways. First, the whole test is assumed to

measure a single latent trait and analyzed accordingly, or, second, each subtest is assumed to measure a distinct latent trait and analyzed separately, one subtest at a time. The first "composite" unidimensional approach violates the test's claim of subtest structure, and thus is difficult to validate. The second "consecutive" unidimensional approach, although incorporating the subtest structure, ignores the potential inter-correlations between related but not identical latent traits, and is likely to yield highly imprecise measures and thus cannot be considered reliable. In reality, there are always non-zero correlations between latent traits, meaning that, at least in theory, the multidimensional approach (Figure 9) is more accurate than the unidimensional one. Moreover, the greater the correlations, the greater the number of subtests, then the greater the measurement precision in employing the multidimensional approach (Wang, Chen, & Cheng, 2004).

There are two kinds of multidimensionality: between-item and within item. In between-item multidimensionality each item measures a single dimension and a set of items measure multiple dimensions. For example, a test with several subtests and each subtest measures a distinct dimension. In within item multidimensionality, an item may measure more than one dimension simultaneously. For example, an essay can be used to measure both "content knowledge" and "language skill". An item like "loss of interests" reflects a person's degree of depression; an item like "anxious foreboding" reflects a person's degree of anxiety; and an item like "worrying" can reflect a person's degrees of both depression and anxiety. In such a case, the item "worrying" is multidimensional. When a test contains multidimensional items, the dimensionality of that test is called "within-item" multidimensionality, which is depicted on the right-hand side of Figure 9.

Multidimensional Rasch models (Adams, Wilson, & Wang, 1997; Kelderman, 1996; Rost & Carstensen, 2002) can help minimize the validity and reliability problems encountered when unidimensional Rasch models are applied to a test containing multiple subtests. Multidimensional models preserve the subtest structure, and simultaneously calibrate all subtests and thus utilize the correlations between subtests to increase precise measurement of each subtest. In many cases, the correlations between latent traits are also of great interest, for instance, they are often used to evaluate internal validity, concurrent validity, or predictive validity. In the multidimensional approach, measurement error is taken into account, and the correlation is estimated directly and thus is free from attenuation.

The multidimensional random coefficients multinomial logit model (Adams, Wilson, & Wang, 1997) deserves attention because it contains many multidimensional or unidimensional Rasch models as special cases and can be

applied to both between-item and within item multidimensionality. Let person n's levels on the L latent traits be denoted as $\boldsymbol{\theta}_n^{\mathrm{T}} = (\theta_{n1},...,\theta_{nL})$, which is considered to be randomly sampled from a population with a multivariate normal density function $g(\boldsymbol{\theta}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the multivariate normal distribution. Under the MRCMLM, the probability of a response in category j of item i for person n is defined as

$$P_{ij}(\boldsymbol{\theta}_n) = \frac{\exp\left(\mathbf{b}_{ij}^{\mathrm{T}}\,\boldsymbol{\theta}_n + \mathbf{a}_{ij}^{\mathrm{T}}\,\boldsymbol{\xi}\right)}{\sum_{u=1}^{K_i}\exp\left(\mathbf{b}_{iu}^{\mathrm{T}}\,\boldsymbol{\theta}_n + \mathbf{a}_{iu}^{\mathrm{T}}\,\boldsymbol{\xi}\right)}, \tag{45}$$

where K_i is the number of categories in item i; $\boldsymbol{\xi}$ is a vector of location parameters that describe the items; $\mathbf{b}_{ij}$ is a score vector (known a priori) given to category j of item i across the L latent traits; and $\mathbf{a}_{ij}$ is a design vector given to category j of item i that describes the linear relationship among the elements of $\boldsymbol{\xi}$.

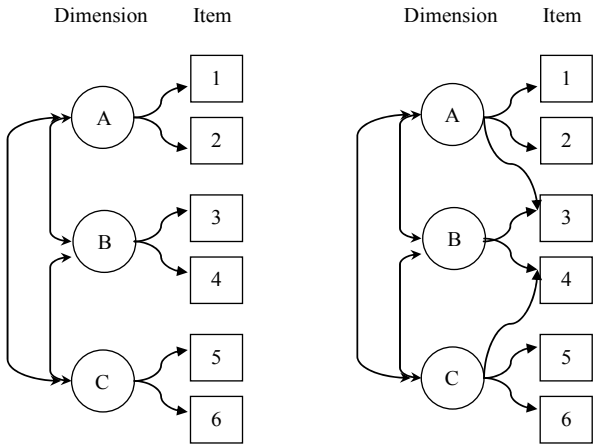

Figure 9. Between-item (left) and within-item (right) multidimensionality

3.7 Hierarchical Latent Traits

Many survey questionnaires and educational and psychological tests measure multiple latent traits that have a hierarchical structure. For example, the Basic Competency Assessment in Hong Kong covers three subjects: Chinese, English,

and Mathematics, each of which includes several domains. A three-order hierarchical structure can be formulated, with the domain abilities in the first level, the three subjects in the second one, and basic competency in the third. Two examples of hierarchical structures are the Revised NEO Personality Inventory, which contains five broad factors, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, each comprising six domains, and the WHOQOL-100, which measures health-related quality of life and consists of six domains, Physical, Psychological, Level of Independence, Social Relationships, Environment, and Spirituality.

Unfortunately, this theory of hierarchical structures cannot be empirically tested with the existing non-hierarchical models. If data with hierarchical latent traits are analyzed using traditional non-hierarchical approaches, then the person measures and their rankings or classifications are incorrect, and subsequent decisions or policies (e.g., college admissions, diagnosis of diseases) based on those findings will be flawed. Recently, researchers have developed item response models that account for hierarchical multidimensionality (de la Torre & Song, in press; Sheng & Wikle, 2008). A major advantage of this hierarchical approach is that person measures of the latent traits in every level can be estimated simultaneously and accurately, which is not possible in non-hierarchical Rasch models.

De la Torre and Douglas (2004) proposed hierarchical latent trait models in the context of cognitive diagnosis. Sheng and Wikle (2008) and De la Torre and Song (in press) developed second-order IRT models that simultaneously account for overall (second-order) and domain (first-order) latent traits. In these models, the first-order latent trait is assumed to be a weighted function of the second-order latent trait:

$$\theta_l^{(1)} = \beta_l^{(2)}\theta^{(2)} + \varepsilon_l^{(1)}, \tag{46}$$

where $\theta_l^{(1)}$ is the first-order *l*-th latent trait; $\theta^{(2)}$ is the second-order latent trait; $\beta_l^{(2)}$ is a regression weight of the second-order latent trait on the first-order *l*-th latent traits, and $\varepsilon_l^{(1)}$ is assumed to be normally distributed. The relationship between the first-order latent trait and item response can follow any IRT function, such as the Rasch model for dichotomous items or the rating scale model or the partial credit model for polytomous items.

In the second-order IRT model, shown as Figure 10, person measures of the first-order and the second-order latent traits are estimated simultaneously. However, the existing second-order IRT model is actually limited to dichotomous items and are not applicable for polytomous items. Such a model may be still too simple to fit the complexity of real testing situations. Further

work is needed to develop more complicated hierarchical models that go beyond second-order.

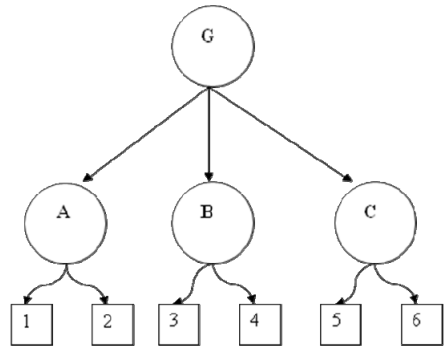

Figure 10. Graphical representation of a second-order model

3.8 Nonlinear Structural Equation Models

Structural equation modeling (SEM) comprises two components, a measurement model and a structural model. The measurement model relates observed responses (indicators) to latent variables or sometimes observed covariates. The structural model specifies relations among latent variables or regressions of latent variables on observed covariates. SEM can take into account the modeling of interactions, nonlinearities, correlated independents, measurement error, correlated error terms, multiple latent independents each measured by multiple indicators, and one or more latent dependents also each with multiple indicators. Standard SEM requires interval data. In practice, standard SEM has been widely applied to item responses, which are actually categorical and ordinal. To resolve this problem, several researchers have extended SEM to categorical data (Muthén, 1984, 2002; Skrondal & Rabe-Hesketh, 2004, 2005). When observed responses are categorical, the conventional measurement model for continuous responses should be modified. The main idea is that the relationship between categorical variables and latent traits should become nonlinear (Glöckner-Rist & Hoijtink, 2003), whereas the structural relationship among latent traits remains linear.

When SEM is applied to item responses, the measurement model is better established within the IRT context than in the CTT context, because IRT describes the relationship between item responses and latent traits more

appropriately than CCT. Figure 11 presents a nonlinear SEM in which IRT function is imposed between items and latent traits. Computer programs such as Mplus (Muthén & Muthén, 2004) and GLLAMM (Rabe-Hesketh, Pickles, & Skrondal, 2001) make nonlinear SEM applicable. It is expected applications of nonlinear SEM will become popular in the near future (Su & Wang, 2007)

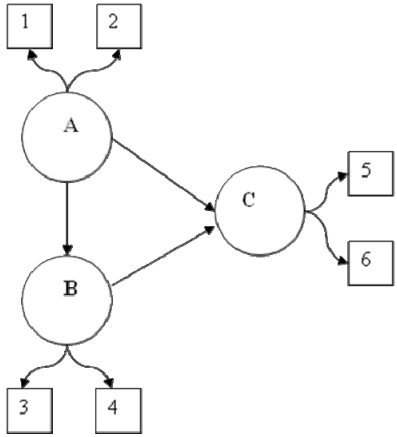

Figure 11. Graphical representation of a nonlinear structural equation model

3.9 Differential Item Functioning

Test fairness is a logical and moral imperative for the makers and users of tests. One potential threat to test fairness is differential item functioning (DIF), which occurs when test-takers with identical latent trait levels have different probabilities of endorsing (or answering correctly) an item, because of different group memberships. In DIF assessment, scores of all test-takers have to be placed on a common metric through the use of a matching variable, so that test-takers' responses to a studied item can be compared for evidence of DIF. Only when a matching variable contains exclusively DIF-free items will subsequent DIF analyses be correct. In most or all practical cases of DIF studies, a matching variable must be established by using the studied test itself. If the studied test, serving as a matching variable contains DIF items, then DIF analysis is based on a biased matching variable and the subsequent DIF assessment is incorrect. On the other hand, if the matching variable is free of DIF (meaning that the studied test does not contain DIF items), then DIF

analysis is no longer necessary. This is a circular problem. Apparently, establishment of a clean matching variable is fundamental in DIF assessment. Unfortunately, existing DIF assessment methods fail to resolve this fundamental problem successfully; this leads to unfair tests and a waste of time and money in test development.

There is a variety of methods for establishment of common metrics, which can be classified into three major categories (Wang, 2004): the equal-mean-difficulty (EMD) method, the all-other-item (AOI) method, and the constant-item (CI) method. In the EMD method, the mean item difficulties of the test for the two groups are constrained to be equal, so that the parameters for the two groups are placed on a common metric for subsequent DIF assessment. This method has been implemented in the IRT computer programs ConQuest. In the AOI method, all but the studied item serve as a matching variable so that the studied item can be assessed for DIF. The AOI method has been implemented in the IRT computer programs Winsteps. In the CI method, the user has to specify a set of items to serve as a matching variable, and the other items are tested for DIF.

By definition, the assumption of EMD between groups holds only when either: (a) the test does not contain any DIF items, or (b) the test contains multiple DIF items, among which some favor the reference group and the others favor the focal group to exactly the same extent, so that the mean difficulties for the two groups are identical. Obviously, these two conditions hardly exist in practice. A direct consequence of employing the EMD method to assess DIF in any imperfect tests (real tests are always imperfect) is that approximately half of the items are classified as favoring the reference group and the others as favoring the focal group (when the estimation error is put aside).

The AOI method assumes that all but the studied item are DIF-free. This assumption holds only when either (a) the test is perfect, or (b) the studied item is the only DIF item in the test. As the number of DIF items in the test increases, the degree of violation of assumption increases, and the AOI method performs worse.

To implement the CI method, a set of items has to be chosen in advance to serve as a matching variable. It is essential to ensure the matching variable is as clean as possible (i.e., consisting exclusively of DIF-free items). Given that the matching variable is clean, (a) the CI method yields appropriate DIF assessment even when the test contains as many as 40% DIF items; (b) anchoring one single item can yield appropriate DIF assessment, but the longer the matching variable, the higher the power of DIF detection; and (c) a matching variable of 4 or more items in length (around 10% ~ 20% of the test)

is generally enough to yield a high power (Shih & Wang, in press; Wang, 2004, Wang & Yeh, 2003). In other words, a pure short matching variable is better than a long but contaminated (by the inclusion of DIF items) matching variable.

Since a matching variable can be seriously contaminated by inclusion of DIF items, it is desirable to apply scale purification procedures to remove DIF items from a matching variable. It has been found that DIF assessment methods with scale purification often outperform those without scale purification in reducing inflated Type I error rates, and increasing deflated power when tests contain DIF items. Unfortunately, scale purification procedures cannot guarantee appropriate DIF assessment because they may not be able to remove all DIF items from a matching variable, especially when tests contain many DIF items.

In short, despite the popularity of the EMD, AOI, and scale purification methods, it was observed that: (a) the EMD and AOI methods generally yield misleading DIF assessment, (b) methods with scale purification perform better than those without scale purification, although they may yield inappropriate DIF assessment when many items in the test have DIF, and (c) the CI method produces appropriate DIF assessment when a pre-specified set of DIF-free items serves as a matching variable.

It becomes apparent that DIF assessment should involve two steps. In Step 1, select a small set of items (e.g., 10% ~ 20% of items) from a studied test that are least likely to have DIF. In Step 2, use these selected items to establish a matching variable to assess DIF in other items. This is called the DIF-free-then-DIF strategy (Wang, 2008), because a set of items are selected first and other items are then tested for DIF. There are two important contributions made by the new strategy: (a) It resolves the circular problem in DIF assessment by selecting a small set of DIF-free items to serve as a matching variable, and (b) it is more accurate in DIF assessment than traditional methods, especially when there are many DIF items in a studied test (Shih & Wang, 2009; Su & Wang, 2009).

3.10 Computerized Adaptive Testing / Computerized Classification Testing

Computerized adaptive testing (CAT) has been largely implemented. Major advantages of CAT are shorter, quicker tests, flexible testing schedules, increased test security, better control of item exposure, better balancing of test content areas for all ability levels, quicker test item updating, quicker reporting, and a better test-taking experience for the test-taker. However, CAT has some

disadvantages, including equipment and facility expenses, limitations of much current CAT administration software, unfamiliarity of some test-takers with computer equipment, apparent inequities of different test-takers taking different tests, and difficulties of administering certain types of test in CAT format (Linacre, 2000; Wainer, Dorans, Eignor, Flaugher, Green, Mislevy, Steinberg, & Thissen, 2000).

In some cases, CAT users are not very interested in point estimates of examinees' latent trait levels, rather they may be more interested in classifying examinees into a limited number of categories (e.g., fail and pass; liberal and conservative; normal, marginal, and abnormal;). Standard CAT algorithms can be modified to attain this classification goal, which is called computerized classification testing (CCT). The sequential probability ratio test has been successful for CCT (Eggen & Straetmans, 2000; Spray & Reckase, 1996). Weissman (2007) further proposed a general approach for item selection in adaptive multiple-category classification tests. It uses mutual information, which is a special case of the Kullback-Leibler distance, or relative entropy. It was found that mutual information works efficiently with the sequential probability ratio test and alleviates the difficulties encountered with using other local and global information measures in the multiple-category classification setting.

Traditional CAT or CCT is based on unidimensional IRT models. As multidimensional item response theory begins to receive recognition and computerized adaptive testing becomes popular in practice, the merger of these two, which is called multidimensional CAT or CCT is a direction to explore. Several researchers have pioneered this direction. Segall (1996) formulated a Bayesian procedure for latent trait estimation and adaptive item selection. van der Linden (1999) derived an algorithm that minimizes the asymptotic variance of the maximum likelihood estimator when a linear combination of multiple latent traits, rather than individual latent traits, is of interest. Wang and Chen (2004) conducted a series of simulations to compare the measurement efficiency of multidimensional CAT with that of unidimensional CAT. The results showed that the higher the correlation between latent traits, the more latent traits there are, and the more scoring levels there are in the items, the more efficient multidimensional CAT is than the unidimensional CAT. In addition to multidimensional models, CAT and CCT can be developed under other types of IRT models, for example, testlet response models, hieratical models, and unfolding models (Lee & Wang, 2009; Liu & Wang, 2009; Shih & Wang, 2008).

3.11 Person-Item Interaction in Rating Scale Items or Likert Items

Standard IRT models assume that persons do not interact with items. In some cases, person may interact with items. For example, person-item interaction has been commonly found in testlet-based items and testlet response models are thus proposed to take into account person-item interaction within a testlet. Person-item interaction is also likely to exist in rating scale items (e.g., seldom, sometimes, often, always). A person might consider the gap between "seldom" and "sometimes" along the latent trait continuum large but the gap between "sometimes" and "often" small; whereas another person might have a quite different perspective on the two gaps. As in testlet response models, one may add a set of random-effect parameters into standard IRT models to account for such a person-item interaction. Wang, Wilson and Shih (2006) proposed the random-effect rating scale model to directly take into account person-item interaction in rating scale items:

$$\text{logit}_{nij} = \theta_n - (\delta_i + \tau_j) - \gamma_{nj}, \tag{47}$$

$$\gamma_{nj} \sim N(0,\ \sigma_j^2), \tag{48}$$

where γ_{nj} denotes the interaction between person n and threshold j, which is assumed to be normally distributed with mean zero and variance σ_j^2; and the others are defined as those in the rating scale model. The larger is σ_j^2, the larger the person-threshold interaction. If $\sigma_j^2 = 0$ for all j, then the model becomes the rating scale model.

The random-effect approach to person-item interaction can be applied to describe rater effect. In the facets model, each rater is given a fixed-effect parameter to depict the severity. That is, each rater is assumed to hold a constant degree of severity across ratings. In reality, a rater's severity may change over ratings. It is thus more flexible to treat rater severity as a random-effect (Wang & Wilson, 2005a):

$$\text{logit}_{nijk} = \theta_n - (\delta_i + \tau_j) - (\eta_k + \gamma_{nk}), \tag{49}$$

$$\gamma_{nk} \sim N(0,\ \sigma_k^2), \tag{50}$$

where γ_{nk} denotes the interaction between rater k when judging person n, which is assumed to be normally distributed with mean zero and variance σ_k^2; and the others are defined as those in Equation 36. The magnitude of σ_j^2 depicts the

intra-rater reliability for rater k: the larger is σ_j^2, the lower the intra-rater reliability for rater k. If $\sigma_k^2 = 0$ for all k, every rater holds a constant degree of severity across his/her ratings, then the random-effect facets model becomes the standard facets model.

Person-item interaction is likely to exist in Likert items (e.g., strongly disagree, disagree, agree, and strongly agree) as well. Like the above models, one may apply the random-effect approach to modeling such a person-item interaction. In recent years, many IRT models have been applied to fit Likert items, which can be classified as two approaches: the dominance approach and the ideal-point approach. It has been argued that responses to Likert items are more consistent with the ideal-point approach than the dominance approach. This argument implies that attitude measures based on disagree-agree responses are more appropriately developed from unfolding models of the ideal-point approach than from cumulative models of the dominance approach. Several IRT-based unfolding models have been developed to fit responses of attitude items, including the (generalized) hyperbolic cosine model (Andrich, 1996), and the (generalized) graded unfolding model (Roberts, Donoghue, & Laughlin, 2000).

There are four basic premises about the response process in these unfolding IRT models. The first premise is that when persons are asked to express their agreement with an attitude statement, they tend to agree with the item to the extent that it is located close to their position on a unidimensional latent attitude continuum. The second premise is that persons select an observed response category for either of two reasons. For example, a person might disagree with an item in either a very negative or a very positive way. If the item is located far below the person's position on the trait continuum (i.e., the item's content is much more negative than the person's attitude), then the person "disagrees from above" the item. In contrast, if the item is located far above the person's position, then the person "disagrees from below" the item. Hence, there are two possible latent responses, "disagree from above" and "disagree from below," associated with the single observed response of "disagree."

The third premise is that "latent responses" follow a cumulative item response model, for example, the rating scale model or the partial credit model. For a four-point Likert scale (e.g., strongly disagree, disagree, agree, strongly agree), there will be eight latent responses, one pair per point. Standard IRT models are applied to define latent responses. However, the model must ultimately be defined in terms of the observed response categories associated with the graded

agreement scale. The two latent responses corresponding to a given observed response category are mutually exclusive. Therefore, the probability for a person to select a particular category is the sum of the probabilities associated with the two corresponding latent responses.

Figure 12 illustrates the probability functions for eight latent responses which come from a hypothetical item with four observed response categories of strongly disagree, disagree, agree, and strongly agree. Figure 13 displays the probability functions for the four observed responses of the same item in Figure 12. When person-item interaction exists in Likert items, one can add a random-effect to standard unfolding models (Wu & Wang, 2009), as being done in the testlet response model and the random-effect rating scale and facets models.

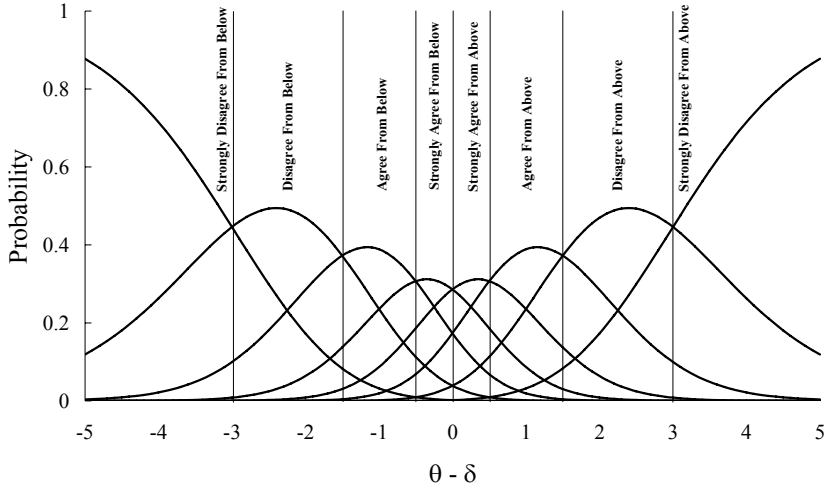

Figure 12. Latent response curves for a hypothetical 4-category item as a function of $\theta - \delta$

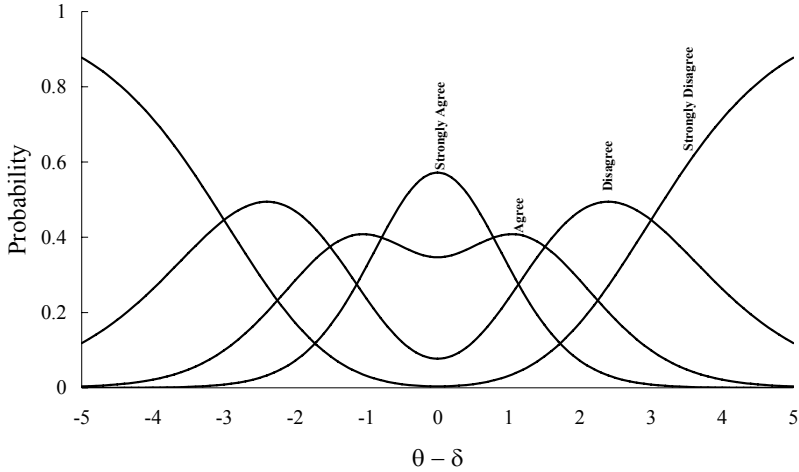

Figure 13. Observed response curves for a hypothetical 4-category item as a function of $\theta - \delta$

4. CONCLUSION

Since the 1960s, researchers have realized that the unit of data analysis should be item response rather than test score. A summation of item scores does not automatically produce an interval scale. Item responses are categorical and ordinal, not continuous or interval. Standard statistical methods, such as ANOVA, regression, factor analysis or SEM, require interval data and are not appropriate for item responses. CTT, although taking into account measurement error, is not appropriate for ordinal data like item responses. Besides, in CTT the judgments of person ability and item difficulty are mutually confounded. We need a theory to treat item responses appropriately and to produce objective measurement and interval data. Rasch measurement is such a theory that has the property of specific objectivity (i.e., parameter separation between person ability and item difficulty) and yields interval measures. Multi-parameter IRT models, although taking a further step beyond CTT, fail to produce objective measurement.

Due to space constraints, this paper highlights only some of the recent developments in Rasch measurement, including polytomous items, multiple facets, multilevels, mixture models, testlet design, multiple dimensions, hierarchical latent traits, nonlinear SEM, DIF, CAT/CCT, and person-item interaction. Rasch measurement not only is theoretically sound but also practically important to test development and data analysis. We believe that Rasch measurement will continue to grow and its applications will continue to be widespread.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D. (1996). Hyperbolic cosine latent trait models for unfolding direct-responses and pairwise preferences. *Applied Psychological Measurement, 20*, 269-290.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait for cognitive diagnosis. *Psychometrika, 69*, 333-353.

de la Torre, J., & Song, H. (in press). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinee into three categories. *Educational and Psychological Measurement, 60*, 713-734.

Embretson, S. E. (1998). A cognitive-design system approach to generating valid tests: Applications to abstract reasoning. *Psychological Methods, 3*, 380-396.

Fischer, G. H. (1973). The linear logistic test model as instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H., & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of treatment effects. *Psychometrika, 56*, 637-651.

Fischer, G. H., & Pononcy, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59*, 177-192.

Glöckner-Rist, A., & Hoitjink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation modeling, 10*, 544-565.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement, 20*, 155-168.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices*. (2nd ed.). New York: Springer-Verlag.

Lee, K.-H., & Wang, W.-C. (2009, July). *Multidimensional computerized classification test.* Paper presented at the Pacific Rim Objective Measurement Symposium. Hong Kong.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA.

Linacre, J. M. (2000). Computer-adaptive testing CAT: A methodology whose time has come [Electronic Version]. *MESA Memorandum*, 69. Retrieved January l 0, 2010 from http://www.rasch.org/memo69.pdf

Liu, C.-W., & Wang, W.-C. (2009, July). *The application of the random-threshold generalized graded unfolding model to computerized adaptive testing and computerized classification testing*. Paper presented at the International Meeting of the Psychometric Society. Cambridge.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent indicators. *Psychometrika, 49*, 115-132.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). GLLAMM: A general class of multilevel models and a Stata program. *Multilevel Modeling Newsletter, 13*, 17–23.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75-92.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*, 42-56.

Rost, J., Carstensen, C., & Von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Munster, Germany: Waxmann.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological measurement, 68*, 413-430.

Shih, C.-L., & Wang, W.-C. (2008, March). *A comparison of item selection strategies in computerized adaptive testing for testlet-based items*. Paper presented at the annual meeting of National Council on Measurement in Education. New York.

Shih, C.-L., & Wang, W.-C. (2009, July). *Selecting DIF-free items to serve as anchors for assessment of differential item functioning: The MIMIC method*. Paper presented at the International Meeting of the Psychometric Society. Cambridge.

Shih, C.-L., & Wang, W.-C. (in press). DIF detection using the MIMIC method with a pure short anchor. *Applied Psychological Measurement*.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL. Chapman & Hall/CRC.

Skrondal, A., & Rabe-Hesketh, S. (2005). Structural equation modeling: Categorical variables. In Everitt, B. and Howell, D. (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 1-8). London: Wiley.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*, 405-414

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Su, C.-M., Wang, W.-C. (2007, April). *Nonlinear structural equation models: An item response modeling approach to categorical response variables*. Paper presented at the annual meeting of National Council on Measurement in Education. Chicago.

Su, Y.-H., & Wang, W.-C. (2009). *The 'DIF-free-then-DIF' strategy applied to the logistic regression procedure for DIF assessment*. Paper presented at the International Meeting of the Psychometric Society. Cambridge.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181-195.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398-412.

Wainer, H., Dorans, D. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9*, 387-408.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295-316.

Wang, W.-C., & Jin. K.-Y. (2010). Multilevel, two-parameter, and random-weights generalizations of the model with internal restrictions on item difficulty. *Applied Psychological Measurement*, 34, 46-65.

Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement, 67*, 583-605.

Wang, W.-C., & Wilson, M. R. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*, 296-318.

Wang, W.-C., & Wilson, M. R. (2005b). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126-149.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

Wang, W.-C., Wilson, M. R., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43*, 335-353.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67*, 41-58.

Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wu, S.-L., & Wang, W.-C. (2009, July). *The random-threshold generalized graded unfolding model.* Paper presented at the Pacific Rim Objective Measurement Symposium. Hong Kong.

Refereed Papers

A. International Journals

(*Corresponding author; IF = 5-year impact factor)

1. Chien, T.-W., Wang, W.-C., Chien, C.-C., & Hwang, W.-S.* (in press). Rasch analysis of positive changes following adversity in cancer patients attending community support groups. *Psycho-Oncology*. SCI, IF = 3.945
2. Chien, T,-W.*, Wang, W.-C., Lin, S.-B., Lin, C.-Y., Guo, H.-R., & Su, S.-B. (in press). KIDMAP, a Web based system for gathering patients' feedback on their doctors. *BMC Medical Research Methodology*. SCI, IF = 2.82 (unofficial)
3. Cheng, Y. Y.*, Wang, W. C., Liu, K. S., & Chen, Y. L. (in press). Effects of association instruction on fourth graders' poetry creativity in Taiwan. *Creativity Research Journal*. SSCI, IF = 1.261
4. Wang, W.-C.*, & Shih, C.-L. (in press). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*. SSCI, IF = 1.092
5. Wang, W.-C.,* & Jin, K.-U. (in press). A generalized model with internal restrictions on item difficulty for polytomous items. *Educational and Psychological Measurement*. SSCI, IF = 1.507
6. Shih, C.-L., & Wang, W.-C.* (in press). DIF detection using the MIMIC method with a pure short anchor. *Applied Psychological Measurement*. SSCI, IF = 1.092
7. Wang, W.-C.*, & Jin. K.-Y. (2010). Multilevel, two-parameter, and random-weights generalizations of the model with internal restrictions on item difficulty. *Applied Psychological Measurement*, *34,* 46-65. SSCI, IF = 1.092
8. Wang, W.-C.* & Shih, C.-L. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*, 713-731. SSCI, IF = 1.507
9. Chien, T.-W., Wu, H.-M., Wang, W.-C., Castillo, R. V., & Chou, W.* (2009). Reduction in patient burdens with graphical computerized adaptive testing on the ADL scale: Tool development and simulation. *Health and Quality of Life Outcomes, 7*, 39-44. SSCI, IF = 3.200

10. Cheng, Y.-Y., Wang, W.-C.*, & Ho, Y.-H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth-fidelity dilemma. *Educational and Psychological Measurement, 69*, 369-388. SSCI, IF = 1.507

11. Shih, C.-L., & Wang, W.-C.* (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199. SSCI, IF = 1.092

12. Lin, T.-K., Weng, C.-Y., 2, Wang, W.-C., Chen, C.-C., Lin, I.-M., & Lin, C.-L.* (2008). Hostility trait and vascular dilatory functions in healthy Taiwanese. *Journal of Behavioral Medicine, 31*, 517-524. SSCI, IF = 2.785

13. Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9,* 387-408.

14. Wang, W.-C.* (2008). A critique of Raju and Oshima's prophecy formulas for assessing the reliability of item response theory-based ability estimates. *Applied Psychological Measurement, 32,* 261-266. SSCI, IF = 1.092

15. Chen, C.-T., & Wang, W.-C.* (2007). Effects of ignoring item interaction on item parameter estimation and detection of interaction items. *Applied Psychological Measurement, 31,* 388-411. SSCI, IF = 1.092

16. Wang, W.-C.*, & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement, 67*, 583-605. SSCI, IF = 1.507

17. Hsieh, C.-L.*, Jang, Y., Yu, T.-Y., Wang, W.-C., Sheu, C.-F., & Wang, Y.-H. (2007). A Rasch Analysis of the Frenchay Activities Index in patients with spinal cord injury. *Spine*, *32,* 437-442. SCI, IF = 3.526

18. Su, Y.-H., Sheu, C.-F., & Wang, W.-C.* (2007). Computing confidence intervals of item fit in the family of Rasch models using the bootstrap method. *Journal of Applied Measurement*, *8*, 190-203.

19. Koh, C.-L, Hsueh, I.-P., Wang, W.-C., Sheu, C.-F., Yu, T.-Y., Wang, C.-H., & Hsieh, C.-L.* (2006). Validation of the Action Research Arm Test using item response theory in stroke patients. *Journal of Rehabilitation Medicine, 38*, 375-380. SCI, IF = 3.057

20. Wang, W.-C.*, Wilson, M. R., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43,* 335-353. SSCI, IF = 1.144

21. Hsueh, I.-P., Wang, W.-C., Wang, C.-H., Lin, J.-H., Sheu, C.-F., & Hsieh, C.-L.* (2006). A simplified stroke rehabilitation assessment of movement instrument. *Physical Therapy, 86,* 936-943. SCI, IF = 2.844

22. Wang, W.-C.*, Yao, G., Tsai, Y.-J., Wang, J.-D., & Hsieh, C.-L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research, 15*, 607-620. SCI, IF = 2.985

23. Lin, J.-H., Wang, W.-C., Sheu, C.-F., Lo, S.-K., Hsueh, I.-P., & Hsieh, C.-L.* (2005). A Rasch analysis of a self-perceived change in quality of life scale in patients with mild stroke. *Quality of Life Research, 14,* 2259-2263. SCI, IF = 2.985

24. Sheu, C.-F.*, Chen, C.-T., Su, Y.-H., & Wang, W.-C. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods, 37,* 202-218. SSCI, IF = 2.611

25. Su, Y.-H., & Wang, W.-C* (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods for detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18,* 313-350. SSCI, IF = 0.44

26. Wang, W.-C.*, & Chen, C.-T. (2005). Item parameter recovery, standard error estimate and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65,* 376-404. SSCI, IF = 1.507

27. Wang, W.-C.*, & Wilson, M. R. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65,* 549-576. SSCI, IF = 1.507

28. Wang, W.-C.*, & Wilson, M. R. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*, 296-318. SSCI, IF = 1.092

29. Wang, W.-C.*, & Wilson, M. R. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149. SSCI, IF = 1.092

30. Wang, W.-C.*, Cheng, Y.-Y., & Wilson, M. R. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, *65*, 5-27. SSCI, IF = 1.507

31. Hsueh. I.-P., Wang, W.-C., Sheu. C.-F., & Hsieh. C.-L.* (2004). Rasch analysis of combining two indices to assess comprehensive ADL function in stroke patients. *Stroke*, *35(3),* 721-726. SCI,IF = 6.872

32. Wang, W.-C*., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF Detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17,* 113-144. SSCI, IF = 0.44

33. Wang, W.-C*., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9,* 116-136. SSCI, IF = 6.119

34. Wang, W.-C.* (2004). Direct estimation of correlation as a measure of association strength using multidimensional item response models. *Educational and Psychological Measurement*, *64*, 937-955. SSCI, IF = 1.507

35. Wang, W.-C.* (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*, 221-261. SSCI, IF = 1.493

36. Wang, W.-C.*, & Chen, H.-C. (2004). The standardized mean difference within the framework of item response theory. *Educational and Psychological Measurement, 64*, 201-223. SSCI, IF = 1.507

37. Wang, W.-C.*, & Su, Y.-H. (2004). Factors Influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28,* 450-480. SSCI, IF = 1.092

38. Wang, W.-C.*, & Wu, C.-I. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement*, *64,* 758-780. SSCI, IF = 1.507

39. Wang, W.-C.*, & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295-316. SSCI, IF = 1.092

40. Wang, W.-C.*, & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27,* 479-498. SSCI, IF = 1.092

41. Wang, W.-C.*, & Cheng, Y.-Y. (2001). Measurement issues in screening outstanding teachers. *Journal of Applied Measurement, 2,* 171-186.

42. Wang, W.-C*. (2000). Factorial modeling of differential distractor functioning in multiple-choice items. *Journal of Applied Measurement, 1*, 238-256.

43. Wang, W.-C*. (2000). Modeling effects of differential item functioning in polytomous items. *Journal of Applied Measurement, 1*, 63-82.

44. Wang, W.-C*. (2000). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online, 5(1)*, 51-76.

45. Wang, W.-C. * (1999). Direct estimation of correlations among latent traits within IRT framework. *Methods of Psychological Research Online, 4*, 2, 47-68.

46. Wang, W.-C.* (1998). Rasch analysis of distractors in multiple-choice items. *Journal of Outcome Measurement, 2*, 43-65.

47. Wang, W.-C.*, Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with Rasch models. *Journal of Outcome Measurement, 2*, 240-265.

48. Adams, R. J.*, Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23. SSCI, IF = 1.092

49. Wilson, M*, & Wang, W-C. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*, 51-72. SSCI, IF = 1.092

B. Taiwan Journals

(*Corresponding author; TSSCI = Taiwan Social Sciences Citation Index)

1. Chien, T,-W., Wang, W.-C., Uen, & Lin, H.-J.*(in press). Rasch analysis of monthly abnormal diagnosis volume by real-time emergency department-based detection for pediatrics. *Journal of Healthcare Management*.

2. Chien, T.-W.*, & Wang, W.-C., & Su, S.-B. (in press). Web-KIDMAP as a knowledge structure map for physicians accurately explaining health examination reports. *Hospital*.

3. Chien, T.-W.*, Chen, M.-C., Wang, W.-C., & Wen, Y.-H. (in press). DRGs assessment and detection using item response theory. *Journal of Formosa Healthcare*.

4. Chien, T.-W.*, Wang, W.-C., & Liang, O.-Y. (in press). Analyzing performance with Rasch model on medical service quality reports. *Journal of Formosa Healthcare*.

5. Chien, T.-W., Liu, O., & Wang, W.-C.* (2009). Determining admission priority for examinees with identical test scores using Rasch analysis. *Psychological Testing, 56,* 129-151. TSSCI

6. Chien, T,-W., & Wang, W.-C.* (2009). Unidimensionality and threshold analysis of balanced scorecard implemented in hospitals. *The Journal of Taiwan Association for Medical Informatics, 18*(2), 1-13.

7. Chien, T,-W., Wang, W.-C., Uen, Y.-H*., &Chen, M.-J.(2009). Survival analysis applied to comparing length of stays in diseases among hospitals. *Journal of Healthcare Management, 10,* 23-35.

8. Chen, P.-H.*, Huang, H.-Y., & Wang, W.-C. (2008). The influences of the features of testlet on computerized adaptive testing. *Psychological Testing, 55*, 129-150. TSSCI

9. Chien, T.-W.*, Wang, W.-C., & Liang, O.-Y. (2008). Performance analysis of service for quality report card used in hospital under global budgeting. *Formosa Journal of Healthcare Administration, 4(1)*, 15-24.

10. Chien, T.-W.*, Wang, W.-C., Yang, M.-H., & Liang, O.-Y. (2008). A study of assessment on revenue impacts of hospitals while changing DRG relative weights. *Hospital, 41(1)*, 30-41.

11. Chien, T.-W., Chen, M.-J., Wang, W.-C., & Uen, Y.-H.* (2008). DRGs assessment and detection using item response theory. *Formosa Journal of Healthcare Administration, 4(1)*, 25-37.

12. Chien, T.-W., Wang, W.-C.*, & Uen, Y.-H.* (2008). Rasch analysis of hospital bed vacancy. *Bulletin of Diwan College of Management, 3*, 15-32.

13. Chien, T.-W., Wang, W.-C., & Hu, C.-W.* (2008). Clinical assessments on activities of daily living (ADL) function via computerized adaptive testing. *Journal of Taiwan Association for Medical Informatics, 17(1)*, 1-14.

14. Chien, T.-W., Wang, W.-C., Chen, Y.-C., & Uen, Y.-H.* (2008). Visualization analysis of hospital vacant beds through computer science. *Journal of Taiwan Association for Medical Informatics, 17(1)*, 41-52.

15. Chien, T.-W., Wang, W.-C., Liang, O.-Y.* (2008). Item response theory provides cost analysts with Rasch-KIDMAP to monitor unusual cases. *Hospital, 41(1)*, 1-13

16. Tsai, L.-T., Yang, C.-C.*, Wang, W.-C., & Shih, C.-L. (2008). A simulation study on using MIMIC model to assess the accuracy of differential item functioning. *Psychological Testing, 55*, 287-312. TSSCI

17. Uen, Y.-H., Chen, M.-J., Chien, T.-W.*, Wang, W.-C., & Lin, M.-J. (2008). Study on impacts and initiatives of DRGs implementation for general surgery. *Hospital, 41(4)*, 23-35.

18. Wen, Y.H., Chien, T.-W.*, & Wang, W.-C. (2007). Standardized residual analysis onto hospital volume of outpatient visits. *Taiwan Health Insurance Magazine, 4(1)*, 82-90.

19. Chao, H.-Y., Wang, W.-C.*, & Yeh, B.-Y. (2007). Development and item response analysis of the Positive, Negative, and Physiological Hyperarousal Scales. *Psychological Testing, 54*, 223-258. TSSCI

20. Chien, T.-W.*, & Wang, W.-C. (2007). Influence and impact on medical centers under DRGs: CTT vs. IRT. *Bulletin of Diwan College of Management, 2*, 89-106.

21. Chien, T.-W., & Wang, W.-C.* (2007). Using Rasch analysis to assess quality indicators in different level of hospitals. *Journal of Healthcare Management, 8(3)*, 249-258.

22. Chien, T.-W., Wang, W.-C.*, & Liang, O.-Y. (2007). Improving reliability and measuring growth rate of hospital service subscales using multidimensional Rasch analysis. *Hospital, 40(6)*, 32-50.

23. Chien, T.-W., Wang, W.-C., Liang, O.-Y.* (2007). Outlier payments under DRGs being suggested: a study based on Rasch analysis. *Journal of Formosa Healthcare, 3(1),* 55-66.

24. Chien, T.-W., Wang, W.-C., & Liang, O.-Y.* (2007). Using Rasch approach analyzing abornormal data collected from hospital web-transaction through smart IC cards. T*aiwan Health Insurance Magazine, 3(2)*, 1-20.

25. Chien, T.-W.*, Wang, W.-C., Liang, O-Y., & Lin, H.-J. (2007). Study on behavior of coefficient of variation for Taiwan's newly version DRGs. *Hospital, 40(1)*, 40-52.

26. Chien, T.-W., Wang, W.-C., Chen, N.-S., & Lin, H.-J.* (2007). Prediction and management of medical fees when patients receiving cares in hospitals under DRGs. *Journal of Taiwan Association for Medical Informatics, 16(4)*, 13-24.

27. Chien, T.-W., Wang, W.-C., Chen, N.-S., & Lin, H.-J.* (2007). A module for screening out the most unexpected reimbursement under DRGs. *Journal of Taiwan Association for Medical Informatics, 16(1)*, 15-26.

28. Chien, T.-W., Wang, W.-C., Chen, N.-S., & Su, S.-B.* (2007). A module for releasing check-up messages from health examination reports. *Journal of Taiwan Association for Medical Informatics, 16 (2)*, 13-28.

29. Chien, T.-W., Wang, W.-C., Su, S.-B., & Chen, Y.-C.* (2007). Detecting abnormalities of questionnaire surveys: Online satisfaction analyses onto physicians. *Journal of Taiwan Association for Medical Informatics, 16(3)*, 71-82.

30. Chien, T.-W., Ken, S.-T.*, Wang, W.-C., & Cheng, T.-Y. (2007). Rasch analysis of accreditation for Taiwan's elderly welfare institutes in 2004. *Hospital, 40(2)*, 1-17.

31. Chien, T.-W., Chiu, C.-F., Wang, W.-C., & Su, S.-B.* (2007). Using Rasch method to analyze abnormal results of annual labors' health examination. *Chinese Journal of Occupational Medicine, 14(1)*, 29-42.

32. Chien, T.-W., Hsu, S.-C.*, Wang, W.-C., & Liang, O.-Y. (2007). The best head nurse awarded first prize-using Rasch analysis. *Hospital, 40(3)*, 37-54.

33. Chien, T.-W., Hwang, S., Wang, W.-C., & Liua, S.-Y.* (2007). Rasch analysis of fine medical record writing appraised by judge scores. *Hospital, 40(4)*, 22-36.

34. Chien, T.-W., Yang, M.-H.*, & Wu, H.-L., & Wang, W.-C. (2007). Detecting DRG re-sequencing with information module and reimbursement consequences of re-sequencing. *Taiwan Health Insurance Magazine, 4(1),* 32-48.

35. Chien, T.-W.*, Liang, O.-Y., Wang, W.-C. (2006). Rasch analysis of employee questionnaire for screening roll names as focused interviews with CEO. *Hospital, 39 (2)*, 27-39.

36. Chien, T.-W.*, Liang, O.-Y., Wang, W.-C., & Lin, H.-J. (2006). Rasch analysis decided who won the award for year of MVP physician on health care giving. *Taiwan Health Insurance Magazine, 2(2)*, 72-85.

37. Chien, T.-W.*, Wang, W.-C., Chen, N.-S., & Lin, H.-J. (2006). Improving hospital indicator management with the Web-KIDMAP module: THIS as an Example. *Journal of Taiwan Association for Medical Informatics, 15(4)*, 15-26.

38. Chien, T.-W.*, Wang, W.-C., Liang, O.-Y., & Lin, H.-J. (2006). Rasch analysis of major diagnosis category of health care for inpatients among medical centers in Taiwan. *Hospital, 39(5)*, 26-42.

39. Chien, T.-W.*, Wang, W.-C., Yang, M.-H., & Liang, O.-Y. (2006). A study of CMI impacts on hospital for Taiwan’s 3rd Version DRG Reform. *Taiwan Health Insurance Magazine, 3(1)*, 79-96.

40. Chien, T.-W.,* Su, S.-B., Wang, W.-C., & Lin, H.-J. (2006). Rasch analysis of sleep quality for inhabitant populace of apartment buildings. *Bulletin of Diwan College of Management, 1*, 331-361.

41. Shih, C.-L., & Wang, W.-C.* (2006). Rasch analysis of rating scale data. *Journal of Education and Psychology, 29*, 399-421. TSSCI

42. Wang, W.-C.,* & Chen, C.-T. (2005). A comparison of scale scores -- Based on the Senior Examination for Physicians. *National Elite, 1(1)*, 137-156.

43. Hung, L.-F., & Wang, W.-C.* (2005). Multilevel modeling for testing whether items have good discrimination. *Chinese Journal of Psychology, 47*, 197-209. TSSCI

44. Huang, T.-Y., & Wang, W.-C.* (2005). Development of the Coping Strategy of Adolescent Dating Conflict Inventory. *Journal of Education and Psychology, 28*, 469-494. TSSCI

45. Chien, T.-W.,* Wang, W.-C., Liang, O.-Y., & Lin, H.-J. (2005). Rasch analysis answers the question of how to calculate DRG relative weights. *Taiwan Health Insurance Magazine, 2 (1)*, 73-86.

46. Chien, T.-W.,* Lin, H.-J., Wang, W.-C., & Liang, O.-Y. (2005). Rasch analysis of 2003 England Inpatient Questionnaire of the Picker Institute Europe. *Hospital, 38(6)*, 27-39.

47. Chien, T.-W.,* Liang, O.-Y., & Wang, W.-C. (2005). Effects on costs of medical care and length of stay with IRT computer-generated informational messages directed to physicians. *Taiwan Health Insurance Magazine, 2(1)*, 87-102.

48. Wang, W.-C.* (2004). Rasch measurement theory and application in education and psychology. *Journal of Education and Psychology, 27*, 637-694. TSSCI

49. Liang, O.-Y., Chien, T.-W.,* Lin, H.-J., & Wang, W.-C. (2004). Study on DRG applied to self-management in reimbursing hospitals for inpatient services. *Journal of Healthcare Management, 5,* 120-133.

50. Chen, P.-H.,* & Wang, W.-C. (2004). Influences of item exposure control on reliability of ability estimation in multidimensional adaptive testing: Using the empirical data of the 2001 Basic Competency Test for Junior High School. *Journal of Education and Psychology, 27*, 181-213. TSSCI

51. Wang, W.-C.*, & Wu, C.-I. (2003). Scaling issues in longitudinal studies: The Symptom Checklist-90-Revised as an empirical example. *Chinese Journal of Mental Health, 16(3)*, 1-30. TSSCI

52. Weng C.-Y.*, Wang, W.-C., & Wu, Y.-C. (2003). Development and psychometric properties of the Self-Evaluation Profile Inventory for Chinese chronically disabled patients. *Chinese Journal of Mental Health, 16(4)*, 83-109. TSSCI

53. Wang, W.-C.*, & Hung, L.-F. (2002). Multilevel modeling for testing whether items have good discrimination. *Chinese Journal of Psychology, 44*, 253-262. TSSCI

54. Cheng, Y.-Y., & Wang, W.-C. (2002). Factors that influence creativity behavior for awarded-winning teachers in scientific competition. *Research in Applied Psychology, 15*, 163-190.

55. Hung, L.-F., & Wang, W.-C.* (2001). Information matrix tests for population distribution hypothesis in finite samples. *Chinese Journal of Psychology, 43*, 35-44. TSSCI

56. Wang, W.-C.*, & Cheng, Y.-Y. (2000). Development and item response analysis of the Creativity Development Inventory. *Psychological Testing, 47*, 153-173. TSSCI

57. Cheng, Y.-Y.,* Wang, W.-C., Yang, S.-C., & Yeh, Y.-C. (2000). The study of evaluation system of teaching effectiveness for the college of management in the National Sun Yat-sen University. *Journal of Sunology: A Social Science Quarterly, 2(1*), 125-159.

58. Wang, W.-C.,* & Chen, H.-C. (1999). Distractibility analysis of multiple-choice items: Application to the English test of the 1997 Joint College Entrance Examination. *Psycholigical Testing, 46(2),* 113-128. (in English) TSSCI

59. Wang, W.-C.,* & Chang, C.-H. (1998). Rasch likelihood ratio test of item differential functioning. *Chinese Journal of Psychology, 40*, 15-32. TSSCI

60. Wang, W.-C.*, & Chen, H.-C. (1998). Development and item response analysis of the Teaching Perspectives Inventory. *Research in Applied Psychology, 2*, 181-207.

61. Wang, W.-C.,* & Chen, H.-C. (1998). Analysis of core jobs of secondary school beginning teachers. *Journal of Education and Psychology*, *22*, 87-112. TSSCI

62. Hung, L-F., & Wang, W.-C.* (1998). Multi-level meta-analysis of central tendency and variability of effect sizes. *Psychological Testing, 46*, 61-72. TSSCI

63. Chen. P.-H.,* & Wang, W.-C. (1998). The development of the Quality of Life Inventory. *Psychological Testing, 46(1),* 57-74. TSSCI

64. Wang, W.-C.* (1997). Objectivity of ratings and objectivity of ability estimates: A comparison between traditional approaches and item response modeling. *Psychological Testing, 44*, 29-52.TSSCI

65. Wang, W.-C.* (1997). Test construct: Factor analysis or Rasch analysis? *Survey Research, 3*, 129-166.

66. Li, Y.-J., Wu, J.-J., Kuo, J.-S., & Wang, W.-C. (1997). Self-regulation of mood: Strategies for changing a bad mood and their correlates. *Journal of National Chengchi University, 75*, 173-206.

67. Wang, W.-C.* (1996). Some controversial issues about the Rasch measurement model. *Journal of Education and Psychology, 19*, 1-26.TSSCI

68. Cheng, Y.-Y., Wang, W.-C., Wu, J.-J., & Hwang, C.-K. (1996). A preliminary report of the construction of the Critical Thinking Scale. *Psychological Testing, 4*3, 213-226. TSSCI

69. Wang, W.-C.* (1995). Some item response models for measuring changes. *Psychological Testing, 42*, 395-414. TSSCI

70. Wang, W.-C.* (1995). Causal inferences in studies of school effectiveness: Use of hierarchical linear models as an example. *Journal of Education and Psychology, 18*, 51- 82. TSSCI

71. Wang, W.-C.* (1994). Goodman's association models for contingency tables. *Psychological Testing, 41*, 253-278. TSSCI

72. Wang, W.-C. * (1993). Detecting rater's severities through item response modeling. *Journal of Education and Psychology, 16*, 83-106. TSSCI

73. Wu, J.-J.*, Cheng, Y.-Y., & Wang, W.-C. (1992). Revision of the Watson-Glaser Critical Thinking Appraisals. *Journal of Education and Psychology, 15,* 39-78. TSSCI

74. Chen, L.-W.*, Wang, W.-C., & Wu, J.-J. (1990). Gender, grades, self-disclosure, and loneliness. *Journal of National Chengchi University, 61*, 213-236.

75. Chen, L.-W.*, Wu, J.-J., & Wang, W.-C. (1990). The Self-consciousness Scale: A revised version for use with Chinese high school and college students. *Psychological Testing, 37*, 211-226. TSSCI

Selected Recent Research Projects

No.	Role	Project	Source	Duration
1	PI	Person-Item Interaction in Likert Items: Model Development and Operational Issues	GRF	2010.01 ~ 201212
2	Co-I	Quantitative Psychology and Psychometrics	NSC	2008.12 ~ 2010.11
3	PI	Integration of SEM and IRT: Assessment of Differential Item Functioning with the MIMIC Procedures	NSC	2008.08 ~ 2009.07
4	PI	General Item Response Models for Person-Item Interaction	NSC	2008.08 ~ 2009.07
5	Co-I	Developing a Short-Form Functional Assessment System for Stroke Patients	NSC	2008.08 ~ 2009.07
6	Co-I	Quantitative Psychology and Psychometrics	NSC	2007.12 ~ 2008.11
7	PI	Integration of SEM and IRT: Assessment of Differential Item Functioning with the Mimic Procedures	NSC	2007.08 ~ 2008.07
8	PI	General Item Response Models for Person-Item Interaction	NSC	2007.08 ~ 2008.07
9	Co-I	Developing a Short-Form Functional Assessment System for Stroke Patients	NSC	2007.08 ~ 2008.07

No.	Role	Project	Source	Duration
10	Co-I	Development of a Refined Version of the Stroke Rehabilitation Assessment of Movement Scale for Stroke Patients (III)	NSC	2007.08 ~ 2008.07
11	Co-I	Development of a Comprehensive ADL Scale for Stroke Paints (II)	NSC	2006.08 ~ 2007.07
12	Co-I	Developing a Short-Form Functional Assessment System for Stroke Patients	NSC	2006.08 ~ 2007.07
13	Co-I	Creative Instruction Experiment and Subject Matter Development in Natural Science for Sixth Grade in Elementary School (II)	NSC	2006.08 ~ 2007.07
14	PI	Multidimensional and Multilevel Item Response Modeling of Rating Scale Data (I)	NSC	2006.08 ~ 2007.07
15	Co-I	Development of a Refined Version of the Stroke Rehabilitation Assessment of Movement Scale for Stroke Patients (II)	NSC	2006.08 ~ 2007.07
16	PI	Testlet Response Theory: Model Development and Application (II)	NSC	2006.08 ~ 2007.07
17	Co-I	Self-Evaluation and Measure System on Total Patient Centered Care for Patient Safety	DoH	2006.01 ~ 2006.12
18	PI	Item Response Models for Non-Ignorable Missing Data	NSC	2005.08 ~ 2006.07
19	Co-I	Development of a Comprehensive ADL Scale for Stroke Patients (I)	NSC	2005.08 ~ 2006.07

No.	Role	Project	Source	Duration
20	Co-I	Creative Instruction Experiment and Subject Matter Development in Natural Science for Sixth Grade in Elementary School (I)	NSC	2005.08 ~ 2006.07
21	Co-I	The Discrete Mixtures of IRT Models: An Extension of the Multidimensional Random Coefficients Multinomial Logit Model	NSC	2005.08 ~ 2006.07
22	Co-I	Development of a Refined Version of the Stroke Rehabilitation Assessment of Movement Scale for Stroke Patients (I)	NSC	2005.08 ~ 2006.07
23	PI	Testlet Response Theory: Model Development and Application (I)	NSC	2005.08 ~ 2006.07
24	Co-I	Development of a Comprehensive ADL Scale for Stroke Patients	NSC	2004.08 ~ 2005.07
25	Co-I	Using the Rasch Model to Develop a Balance Scale for Stroke Patients (II)	NSC	2004.08 ~ 2005.07
26	PI	Formulating Rasch Family Models as Generalized Linear Mixed Models: Implementations and Applications (II)	NSC	2004.08 ~ 2005.07
27	Co-I	Creative Instruction Experiment and Subject Matter Development in Science：Natural Science in Primary School (Ⅲ)	NSC	2004.08 ~ 2005.07
28	PI	The Pearson Correlation Coefficient as an Effect Size Measure Within the Framework of Item Response Theory	NSC	2004.08 ~ 2005.07

No.	Role	Project	Source	Duration
29	Co-I	Comparison of Testlet Selection Strategy in Computerized Adaptive Testing	NSC	2004.08 ~ 2005.07
30	Co-I	A Bayesian Approach to the Testlet Model Where Latent Traits Are Correlated with Random Testlet Effects	NSC	2004.08 ~ 2005.07
31	Co-I	An Integrated Auditing Model for the Evaluation on Medical Quality in Hospitals at All Levels under Global Budgeting	DoH	2004.01 ~ 2004.12
32	Co-I	Using the Rasch Model to Develop a Balance Scale for Stroke Patients (I)	NSC	2003.08 ~ 2004.07
33	PI	Formulating Rasch Family Models as Generalized Linear Mixed Models: Implementations and Applications (I)	NSC	2003.08 ~ 2004.07
34	Co-I	The Research on Relationships between Physical Fitness Quotient and Intelligence Quotient of Teenagers	NSC	2003.08 ~ 2004.07
35	Co-I	The Effect of Thinking Style of Teachers and Students on Their Interactions (II)	NSC	2003.08 ~ 2004.07
36	PI	Evaluation of Academic Journals in Psychology	NSC	2002.11 ~ 2003.03
37	Co-I	Study of The Concept of Environment Sustainable Development for Teachers-Humanistic Aspect (II)	NSC	2002.08 ~ 2003.07

No.	Role	Project	Source	Duration
38	PI	Detection and Correction of Local Item Dependency (II)	NSC	2002.08 ~ 2003.07
39	Co-I	Study of the Concept of Environment Sustainable Development for Teachers---Humanistic Aspect (I)	NSC	2001.08 ~ 2002.07
40	PI	Detection and Correction of Local Item Dependency (I)	NSC	2001.08 ~ 2002.07
41	Co-I	A Study to Characteristics of Family Violence Perpetrators and Evaluation Instrument for Treatment	MoI	2001.07 ~ 2002.07
42	Co-I	The Effect of Thinking Style of Teacher and Students on Their Interactions (I)	NSC	2000.08 ~ 2001.07
43	PI	Between-Item and Within-Item Multidimensional Computerized Adaptive Testing (II)	NSC	2000.08 ~ 2001.07
44	Co-I	A Study of Creative Thinking and Its Correlates for Awarded Teachers in Scientific Competitions	NSC	1999.08 ~ 2000.07
45	PI	Between-Item and Within-Item Multidimensional Computerized Adaptive Testing	NSC	1999.08 ~ 2000.07
46	Co-I	A Study of Creative Thinking and Its Correlates for Awarded Teachers in Scientific Competitions	NSC	1998.08 ~ 1999.07
47	PI	Analysis of Differential Distractor Functioning Parameters in Multiple-Choice Items	NSC	1998.08 ~ 1999.07

No.	Role	Project	Source	Duration
48	PI	Evaluate, Revise and Develop Verbal Tests for Science Gifted Students Using Item Response Theory (II)	NSC	1997.08 ~ 1999.07
49	Co-I	Cognitive-Psychometric Modeling in Computerized Adaptive Testing	NSC	1997.08 ~ 1998.07
50	PI	Evaluate, Revise, and Develop Verbal Tests for Science Gifted Students	NSC	1996.08 ~ 1997.07
51	PI	A Pilot Study of Developing an Evaluation System for College Students' Career Planning for Becoming Teachers	NSC	1996.08 ~ 1997.07
52	PI	Rater Severity in Non-Multiple-Choice Items	NSC	1995.08 ~ 1996.07

Note:

NSC = National Science Council

MoI = Ministry of Interior,

DoH = Department of Health, Executive Yuan, Taiwan