Research Report

# A Study of Confidence and Accuracy Using the Rasch Modeling Procedures

*Insu Paek*

*Jihyun Lee*

*Lazar Stankov*

*Mark Wilson*

*July 2008*

*ETS RR-08-42*

# A Study of Confidence and Accuracy Using the Rasch Modeling Procedures

Insu Paek, Jihyun Lee, and Lazar Stankov
ETS, Princeton, NJ

Mark Wilson
University of California at Berkeley

July 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS' constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html



**Abstract**

This study investigated the relationship between students' actual performance (accuracy) and their subjective judgments of accuracy (confidence) on selected English language proficiency tests. The unidimensional and multidimensional IRT Rasch approaches were used to model the discrepancy between confidence and accuracy at the item and test level and to assess disattenuated strength of association between accuracy and confidence. The analysis results indicate a pattern of overconfidence bias (i.e., overestimation of success rate), which was related to item difficulty. In addition, the strength of association between accuracy and confidence dimension was relatively high: The confidence dimension explained 45% and 52% of the variability in the accuracy dimension for the two tests employed in this study.

Key words: Confidence, Rasch model, accuracy, multidimensional Rasch model

**Acknowledgments**

We would like to thank Dr. Frank Rijmen for his comments on an early draft of this paper. We are also grateful to Jennifer Minsky for her involvement in many stages of data collection, Cathy Trapani and her team for data cleaning, and Dr. Amy Schmidt for her comments at a later stage of writing this paper.

## Table of Contents

Page

Introduction ..... 1

Data ..... 2

How Confident Are You That Your Answer Is Correct? ..... 2

Method ..... 2

Results ..... 7

1. Are People's Confidence Levels Comparable to Their Accuracy Levels? ..... 7

2. Do People's Confidence Levels Change as Their Accuracy Levels Change? ..... 9

3. Are People's Confidence Levels Related to the Difficulty Level of a Specific Task? ..... 13

4. What Is the Strength of Association Between People's Accuracy and Confidence Levels? ..... 14

Discussion ..... 15

Summary ..... 17

References ..... 19

Notes ..... 22

Appendixes

A. Item Difficulty Estimate and Item Fit Statistic ..... 23

B. Local Item-Level Discrepancy Index, Local Test-Level Discrepancy Index, and Local Discrepancy Percentage ..... 25

## List of Tables

Page

Table 1. Item Discrepancy Index (IDI) and Test-Level Discrepancy Index (TDI) for Listening and Reading ..... 8
Table 2. Frequency Distributions for Ability Estimate on Accuracy ..... 13
Table 3. Latent Correlation Between Confidence and Accuracy Dimensions ..... 15

## List of Figures

Page

Figure 1. Item plots for listening..... 9
Figure 2. Item plots for reading. ..... 10
Figure 3. Confidence and accuracy at the test score level. ..... 11
Figure 4. Average standard deviation plot for subjective probability. ..... 12
Figure 5. Relationship between item difficulty and item discrepancy index (IDI). ..... 14

## Introduction

Statistical modeling based on item response theory (IRT) has become one of the most widely used psychometric methods in educational testing. The Rasch (Rasch, 1980), two-parameter, and three-parameter logistic models (Birnbaum, 1968) have become increasingly popular over the past 20 years, in particular for scaling, equating, item banking, computerized adaptive testing, standard setting, and test assembly purposes (see Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Wright & Stone, 1979; Yen & Fitzpatrick, 2006).

In the field of decision making, measures of confidence and accuracy have been studied for economic and weather forecasting. Forecasters' confidence and their actual performance are typically compared. In psychological and educational assessments, a concept of bias scores[1] (i.e., differences between accuracy and confidence scores) was introduced (see Stankov & Crawford, 1996; Stankov & Lee, 2007) and the relationship between confidence and accuracy has been examined, for the most part, by employing the calibration curves approach (see Hakstian & Kansup, 1975; Keren, 1991; Lichtenstein, Fischoff, & Phillips, 1982).

The IRT approach has not been used in the studies of confidence and accuracy, but its methodological components can be very useful in helping to understand these constructs. IRT modeling utilizes a latent construct at the ability level, either at a local point or within a certain range, and can provide the means to estimate the discrepancy between confidence and accuracy at the item or test levels. The IRT item parameters (e.g., item difficulty) model stimulus characteristics, allowing us to further examine the stimulus properties and latent constructs of confidence and accuracy.

For this current investigation, we employed one of the popular IRT models, the Rasch model, which exhibits additive conjoint measurement (Perline, Wright, & Wainer, 1979). Instruments fitting the Rasch model have sufficiency properties of simple observed statistics for model parameters, separablility of person and item parameters, and specific objectivity (Fischer, 1995; Hoijtink & Boomsma, 1995; Molenaar, 1995; Rasch, 1966; Wright & Master, 1982). We explored in this study four substantive questions, as follows:

1. Are people's confidence levels comparable to their accuracy levels?
2. Do people's confidence levels change as their accuracy levels change?
3. Are people's confidence levels related to the difficulty level of a specific task?

4. What is the strength of association between people's accuracy and confidence levels?

A unidimensional Rasch model calibration was used to answer the first three questions, and a multidimensional Rasch model calibration was used for the last question. We also developed indices to express people's confidence levels relative to their accuracy levels. Those indices were based on the unidimensional Rasch calibration results. Although these four questions could be investigated in a much more substantive manner, the focus of this paper is on illustrating the application of Rasch modeling and on proposing potentially useful tools in the context of researching the nature of the relationship between confidence and accuracy.

## Data

The data for accuracy were based on the item responses from the Reading and Listening sections of an English language proficiency test ($N = 820$). We used 24 multiple-choice (MC) Reading items and 17 MC Listening items for this study. Each MC item had four options. All the MC items were dichotomously scored as either 0 (incorrect) or 1 (correct).

The data for confidence were obtained from the participants' self-rating on their level of confidence about getting each item correct. They were asked to give their subjective rating (expressed in percentage terms) immediately after responding to a test item. The question below along with the subjective rating range was given to the participants to rate their confidence:

***How Confident Are You That Your Answer Is Correct?***

This approach of collecting confidence levels has been employed by Crawford and Stankov (1996), Juslin (1994), Keren (1991), and Stankov and Crawford (1996). In Crawford and Stankov's (1996) and Stankov & Crawford's (1996) studies, the participants were allowed to use any integer number as a percentage to express their confidence levels (for instance, 53%). However, they used only numbers rounded to the nearest 10. With a convention of assigning 20% as the lowest probability value to incorporate a guessing factor in MC questions with five options, a subjective probability range of 20% to 100% was used in this study.

## Method

In this study, we have defined *accuracy* as the dimension underlying the performance on the English language proficiency test, and *confidence* as the dimension producing the subjective probability (i.e., a person's subjective judgment on the correctness of each item, expressed as a

percentage). The Rasch model was fitted to the dichotomously scored item responses. The general form of the Rasch model for binary item responses is:

$$P(X_i = x_i|\theta) = \frac{\exp[x_i(\theta - \delta_i)]}{1 + \exp(\theta - \delta_i)}, \quad (1)$$

where $X_i$ is an indicator variable for the $i$th item ($1 \le i \le I$), $x_i$ is its realization (1 for correct and 0 for incorrect), $\theta$ is a person's latent score (latent ability), and $\delta_i$ is the $i$th item difficulty. $P(X_i = x_i \mid \theta)$ is a probability of getting the $i$th item response $x_i$ for a person's latent score $\theta$. From this Rasch calibration, we obtained the objective probability for answering each item correctly for each person and his or her latent accuracy score.

For research questions 1 and 2 (Are people's confidence levels comparable to their accuracy levels? and Do people's confidence levels change as their accuracy levels change?), the objective probability was compared with the subjective probability for each item and at the total test level. The probability comparisons and their interpretations are:

$$\text{Objective probability} = P_i \equiv E(X_i = 1|\theta_{ac}) = P(X_i = 1|\theta_{ac}), \quad (2)$$

$$\text{Subjective probability} = P_i^* \equiv E(\pi_i|\theta_{ac}) = \frac{\sum_{n:\,\hat{\theta}_{ac(n)}=\theta_{ac}} \pi_{ni}}{\sum_n I[\hat{\theta}_{ac(n)} = \theta_{ac}]}, \quad (3)$$

$$\text{Overconfidence} = P_i^* - P_i > 0, \text{ and} \quad (4)$$

$$\text{Underconfidence} = P_i^* - P_i < 0, \quad (5)$$

where $P_i$ is defined through Equation 1, $\theta_{ac}$ is a latent accuracy score, $\pi_{ni}$ is the $n$th person's subjective probability on the $i$th item ($1 \le n \le N$), $I[\cdot]$ is an indicator function (1 if $\hat{\theta}_{ac(n)} = \theta_{ac}$; 0 otherwise), and $\hat{\theta}_{ac(n)}$ is $n$th person's estimate on the latent accuracy dimension. Equation 3 is a regression of the subjective rating onto accuracy and is estimated by averaging the subjective rating for all participants with the same ability on the accuracy dimension. Objective probability and person latent scores are obtained through the unidimensional Rasch model calibration for the accuracy data. Subjective probability is estimated by Equation 3. A summary statistic for the discrepancy between confidence and accuracy at the item level can be expressed as:

$$\text{Item-level discrepancy index (IDI}_i) = \sum_{\theta_{ac}} (P_i^* - P_i) w(\theta_{ac}), \quad (6)$$

where $w(\theta_{ac})$ is a relative frequency of a person's latent score distribution at $\theta_{ac}$, that is,

$$w(\theta_{ac}) = \frac{N_{\theta_{ac}}}{N}, \quad (7)$$

with $N$ = total sample size and $N_{\theta_{ac}}$ = number of persons at $\theta_{ac}$,

$$\text{Overconfidence: IDI}_i > 0 \text{ for the } i\text{th item, and} \quad (8)$$

$$\text{Underconfidence: IDI}_i < 0 \text{ for the } i\text{th item.} \quad (9)$$

In this study, we classify the size of IDI by the differential item functioning (DIF) effect size on the probability scale suggested by Dorans and Holland (1993). In their classification, sizes |0.05| and |0.10| are used as thresholds for negligible, medium, and large DIF. With this rule applied to the discrepancy between accuracy and confidence, the following criteria for overconfidence can be established:

$$\text{Large discrepancy} = \text{IDI}_i > 0.10, \quad (10)$$

$$\text{Medium discrepancy} = 0.05 < \text{IDI}_i \leq 0.10, \text{ and} \quad (11)$$

$$\text{Small or negligible discrepancy} = 0 < \text{IDI}_i \leq 0.05. \quad (12)$$

For underconfidence, the thresholds −0.05 and −0.10 are used for classification. These IDI cutoff points can be seen as arbitrary choices, but no such theoretical cutoff points had been established at the time that this paper was written. We introduce these rules as a start, and their usefulness can be verified in future studies.

An index for the discrepancy between confidence and accuracy at the test level in a given $\theta_{ac}$ is defined as follows:

$$\text{Objective expected score} = \sum_i P_i, \quad (13)$$

$$\text{Subjective expected score} = \sum_i P_i^*, \quad (14)$$

Overconfidence at the total test score level = $\sum_i P_i^* - \sum_i P_i > 0$, and (15)

Underconfidence at the total test score level = $\sum_i P_i^* - \sum_i P_i < 0$. (16)

A summary statistic for the test-level discrepancy index (TDI) for a given $\theta_{ac}$ is defined as:

$$\text{TDI} = \sum_{\theta_{ac}} [\sum_i P_i^* - \sum_i P_i] w(\theta_{ac}), \quad (17)$$

Overconfidence on average: TDI > 0, and (18)

Underconfidence on average: TDI < 0. (19)

A more flexible version of the TDI may be expressed as:

$$\text{Discrepancy percentage (DP)} = \frac{\sum_{\theta_{ac}} [\sum_i P_i^* - \sum_i P_i] w(\theta_{ac})}{\text{Test Length}} \times 100, \quad (20)$$

where the test length is the number of items used in the calculation. The DP is a signed measure of underconfidence or overconfidence standardized by the test length, which is an adjustment for comparisons between tests with different test lengths.

For the person latent score estimation, the expected a posteriori (EAP; Bock & Mislevy, 1982) estimator was used. The EAP is an expectation of a person's posterior probability on the ability distribution given the person's item response string $\mathbf{X} = (x_1, x_2, \cdots, x_I)'$ and the item parameters $\{\delta_i\}$. Its mathematical expression is:

$$E(\theta | \mathbf{X}, \{\delta_i\}) = \frac{\int \theta P(\mathbf{X} | \theta, \{\delta_i\}) g(\theta) d\theta}{\int P(\mathbf{X} | \theta, \{\delta_i\}) g(\theta) d\theta} \quad (21)$$

The prior distribution for $\theta$, $g(\theta)$, was assumed to follow a univariate normal distribution for the unidimensional Rasch calibrations and a bivariate normal distribution for the multidimensional Rasch calibrations.

To answer our third research question (Are people's confidence levels related to the difficulty level of a specific task?), the item difficulties obtained from the unidimensional Rasch calibrations were compared to the item-level discrepancy index. In order to answer the fourth

research question (What is the strength of association between people's accuracy and confidence levels?), the variance-covariance matrix for the accuracy and confidence dimensions was estimated by a multidimensional Rasch modeling. Because the participants provided their confidence ratings in a probabilistic response metric, we simulated the dichotomous item responses for the confidence dimension. Assuming that the confidence rating is the subjective probability defined by the item response function, IRF($\theta_{ac}$, $\theta_{co}$), where $\theta_{co}$ is a latent confidence dimension, ($\theta_{ac}$, $\theta_{co}$) that follows a bivariate distribution with mean vector $\boldsymbol{\gamma}$ and variance-covariance $\sum$, and assuming strict monotonicity of IRF with regard to the latent dimension for confidence $\theta_{co,}^{2}$ the subjective probability rating can be expressed as:

$$\text{Subjective probability rating} \equiv \text{IRF}_{ni}(\theta_{ac},\theta_{co}) = \text{IRF}_{ni}(\theta_{co}) = P(Y_{ni}=1|\theta_{co}), \tag{22}$$

$$Y_{ni} \text{ for } \theta_{co} = 1 \text{ if } P(Y_{ni}=1|\theta_{co}) > u \sim \textit{uniform}(0,1), \text{ or } 0 \text{ otherwise}, \tag{23}$$

where $Y_{ni}$ is $n$th person's $i$th item response from the confidence dimension $\theta_{co}$, and $u$ is a random draw from the standard uniform distribution. The equality between $\text{IRF}_{ni}(\theta_{ac},\theta_{co})$ and $\text{IRF}_{ni}(\theta_{co})$ in Equation 22 is due to the simple structure (i.e., an item loads onto a single dimension) in the model data fitting.

The multidimensional Rasch model used in this study has the following form:

$$P(X_{id}=x_{id}|\theta_d) = \frac{\exp[x_{id}(\theta_d-\delta_{id})]}{1+\exp(\theta_d-\delta_{id})}, \tag{24}$$

where $d$ is an indicator (for either accuracy or confidence) on which dimension the $i$th item loads, and $\delta_{id}$ is the $i$th item difficulty on dimension $d$. The variance-covariance matrix was modeled through ($\theta_{ac}$, $\theta_{co}$) ~ MVN($\boldsymbol{\mu}$, $\Sigma$), where MVN stands for multivariate normal distribution, $\boldsymbol{\mu}$, is 2 × 1 mean vector, and $\Sigma$ is 2 × 2 variance-covariance matrix. The variances and correlations (or covariances) are model parameters and are estimated directly from the item response matrix. The program ConQuest (Wu, Adams, & Wilson, 1998) was used for both unidimensional and multidimensional Rasch model calibrations. ConQuest is capable of estimating all possible model variants based on the multidimensional random coefficient multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997). Thus, this program can perform the fitting of a variety of Rasch family models, such as the rating scale model (Andrich,

1978), the partial-credit model (Masters, 1982), the facet model (Linacre, 1989), and the linear logistic test model (LLTM; Fisher, 1989). ConQuest uses marginal maximum likelihood (MML) estimation with the expectation-maximization (EM) algorithm (Bock & Aitkin, 1981; Dempster, Laird, & Rubin, 1977).

The investigation of item misfit for the Rasch model was performed through the weighted mean squared (WMNSQ) fit statistic, using ConQuest. The WMNSQ fit statistic is based on the difference between the observed and predicted responses. It approaches 1 when the model fits the data well (Wright & Masters, 1982; Wu, 1997). Adams and Khoo (1996) and Wilson (2005) suggested the interval of [0.75, 1.33] of the WMNSQ for the item fit criterion. All items used for this study were within this interval. The WMNSQ fit statistics are provided in Appendix A, along with the item difficulties and their standard errors.

## Results

We organized the results of this study by the four research questions presented in the introduction to this paper.

### *1. Are People's Confidence Levels Comparable to Their Accuracy Levels?*

The results of the calculations for IDI, TDI, DP, and the classifications of the IDI values are shown in Table 1. The Listening section had nine small IDIs, three medium IDIs, and five large IDIs. The IDI values for items with medium or large IDIs were all positive and ranged from 0.065 to 0.207; these positive IDIs indicated overconfidence expressed on those items. Within the Listening section, 47% [100*(8/17)] of the items showed evidence of overconfidence. At the test level, the TDI was 0.82, indicating that the participants overpredicted their Listening section score by 0.82 score points on average.

The Reading section had 10 small IDIs, 4 medium IDIs, and 10 large IDIs. The IDI values for items with medium or large IDIs were all positive and ranged from 0.061 to 0.633; these positive IDIs indicated overconfidence expressed on those items. Within the Reading section, 58% [100*(14/24)] of the items showed evidence of overconfidence. At the test level, the TDI was 2.89, indicating that the participants overpredicted their Reading section score by 2.89 score points on average.

**Table 1**

***Item Discrepancy Index (IDI) and Test-Level Discrepancy Index (TDI) for Listening and Reading***

| | Listening | | | Reading | |
|---|---|---|---|---|---|
| | IDI | IDI class | | IDI | IDI class |
| Item 1 | 0.119 | L | Item 1 | 0.048 | |
| Item 2 | −0.002 | | Item 2 | 0.001 | |
| Item 3 | 0.004 | | Item 3 | 0.149 | L |
| Item 4 | 0.064 | M | Item 4 | 0.004 | |
| Item 5 | 0.023 | | Item 5 | 0.042 | |
| Item 6 | −0.012 | | Item 6 | 0.179 | L |
| Item 7 | −0.003 | | Item 7 | 0.016 | |
| Item 8 | 0.065 | M | Item 8 | 0.299 | L |
| Item 9 | 0.141 | L | Item 9 | 0.046 | |
| Item 10 | 0.111 | L | Item 10 | 0.078 | M |
| Item 11 | −0.045 | | Item 11 | 0.099 | M |
| Item 12 | −0.027 | | Item 12 | 0.010 | |
| Item 13 | 0.066 | M | Item 13 | 0.109 | L |
| Item 14 | 0.207 | L | Item 14 | 0.162 | L |
| Item 15 | 0.112 | L | Item 15 | 0.141 | L |
| Item 16 | −0.004 | | Item 16 | 0.010 | |
| Item 17 | 0.001 | | Item 17 | 0.089 | M |
| | | | Item 18 | 0.290 | L |
| | | | Item 19 | 0.061 | M |
| | | | Item 20 | 0.257 | L |
| | | | Item 21 | 0.004 | |
| | | | Item 22 | 0.047 | |
| | | | Item 23 | 0.118 | L |
| | | | Item 24 | 0.633 | L |
| TDI | 0.820 | | TDI | 2.894 | |
| DP | 4.8 | | DP | 12.1 | |

*Note*. DP = discrepancy percentage, M = medium size IDI, L = large size IDI.

DPs were 4.8 for the Listening section and 12.1 for the Reading section, which reveals that the participants showed more pronounced overconfidence in the Reading section than in the Listening section.

### *2. Do People's Confidence Levels Change as Their Accuracy Levels Change?*

Figures 1 and 2 show a few example items from the Listening and Reading sections, respectively. In these plots, the values of $P_i$ (objective probabilities for accuracy shown in Equation 2) and $P_i^*$ (subjective probabilities for confidence shown in Equation 3) were plotted to see how people's confidence may change according to their accuracy at the item level. Vertical dotted lines in Figures 1 and 2 represent the IRT Rasch item difficulties. A higher difficulty value represents a more difficult item. Figure 1 (Listening) shows three example items, each of which represents an item with a different size of IDI: Item 7 with a small IDI (IDI = −0.003), Item 13 with a medium IDI (IDI = 0.066), and Item 14 with a large IDI (IDI = 0.207). The rest of the items with medium or large IDIs showed the same pattern as Items 13 and 14. In general, the tendency to show overconfidence decreased as ability on the accuracy dimension increased; that is, the participants with higher accuracy scores tended to predict their actual performance more accurately than those with lower accuracy scores did.

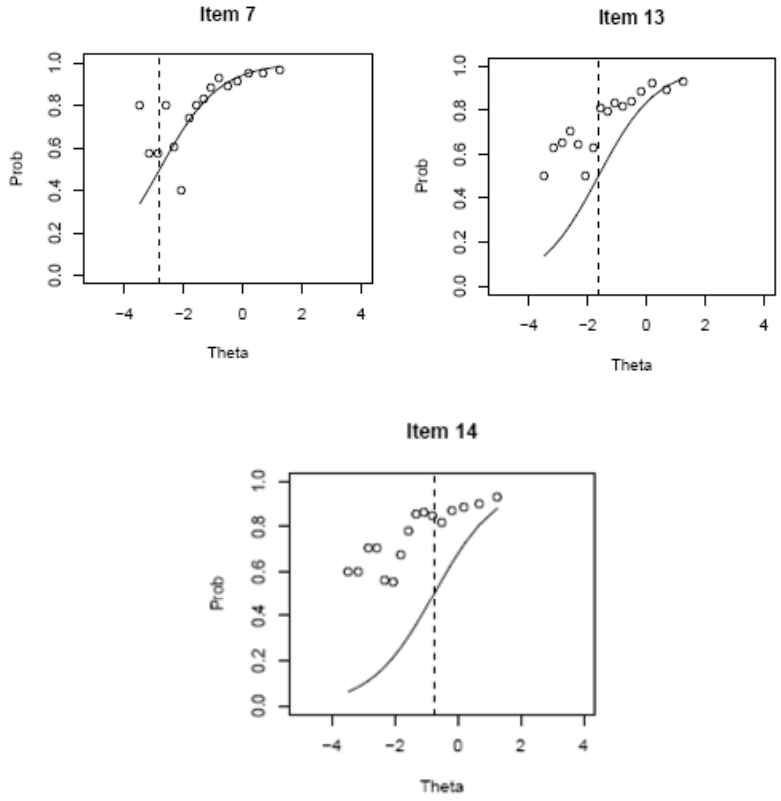


***Figure 1.*** **Item plots for listening.**

*Note.* Circles represent subjective probabilities (confidence); solid lines represent objective probabilities (accuracy); and the vertical dotted lines represents item difficulties.

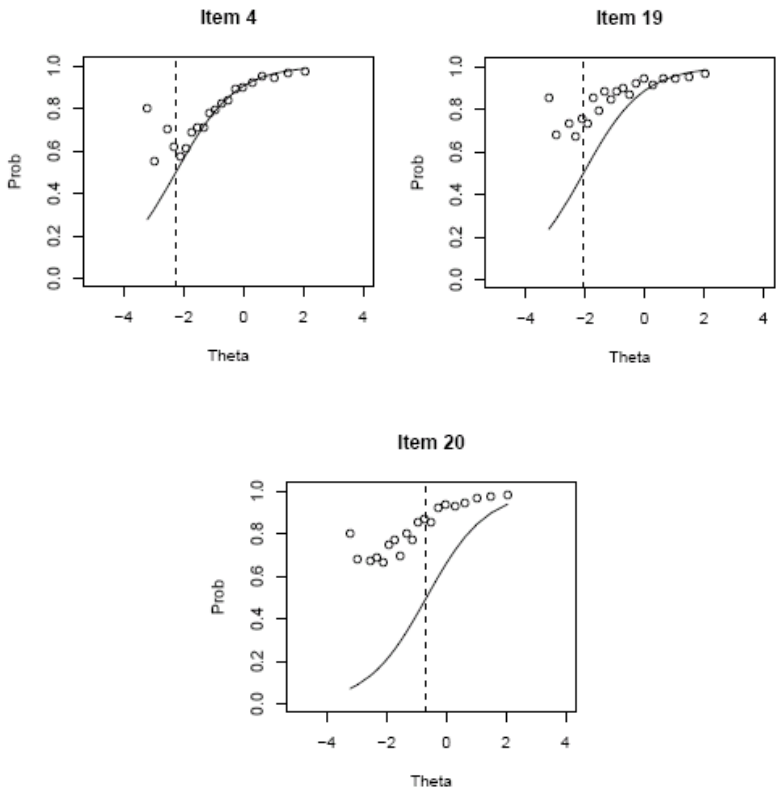


***Figure 2.*** **Item plots for reading.**

*Note.* Circles represent subjective probabilities (confidence); solid lines represent objective probabilities (accuracy); and the vertical dotted lines represents item difficulties.

Figure 2 shows the three example items from the Reading section. Item 4 shows a small IDI (IDI = 0.004), Item 19 has a medium IDI (IDI = 0.061), and Item 20 has a large IDI (IDI = 0.257). The rest of the Reading section items with medium or large IDIs showed overconfidence patterns that were very similar to those seen in Items 19 and 20. In general, as people's ability on the accuracy dimension increased, the confidence ratings (expressed in subjective probability) came closer to their objective performance (shown by objective probability).

Figure 3 illustrates the change in overconfidence as a function of ability on the accuracy dimension at the test level. The plots were constructed using the objective expected score $\sum_i P_i$ (Equation 13) and the subjective expected score $\sum_i P_i^*$ for a given $\theta_{ac}$. Overall, both the Listening and Reading sections showed pronounced patterns of overconfidence across all ability levels on the accuracy dimension. However, the participants' subjective predictions (i.e.,

confidence) became closer to their actual performance level as their ability level increased. One may claim that better prediction of participants' performance is associated with improved precision of their predictions (i.e., reduction of variability in their confidence rating). Others may argue that the better predication of higher-ability people could be due to the ceiling effect in the confidence rating scale that is bounded between 0 and 1 in probability.

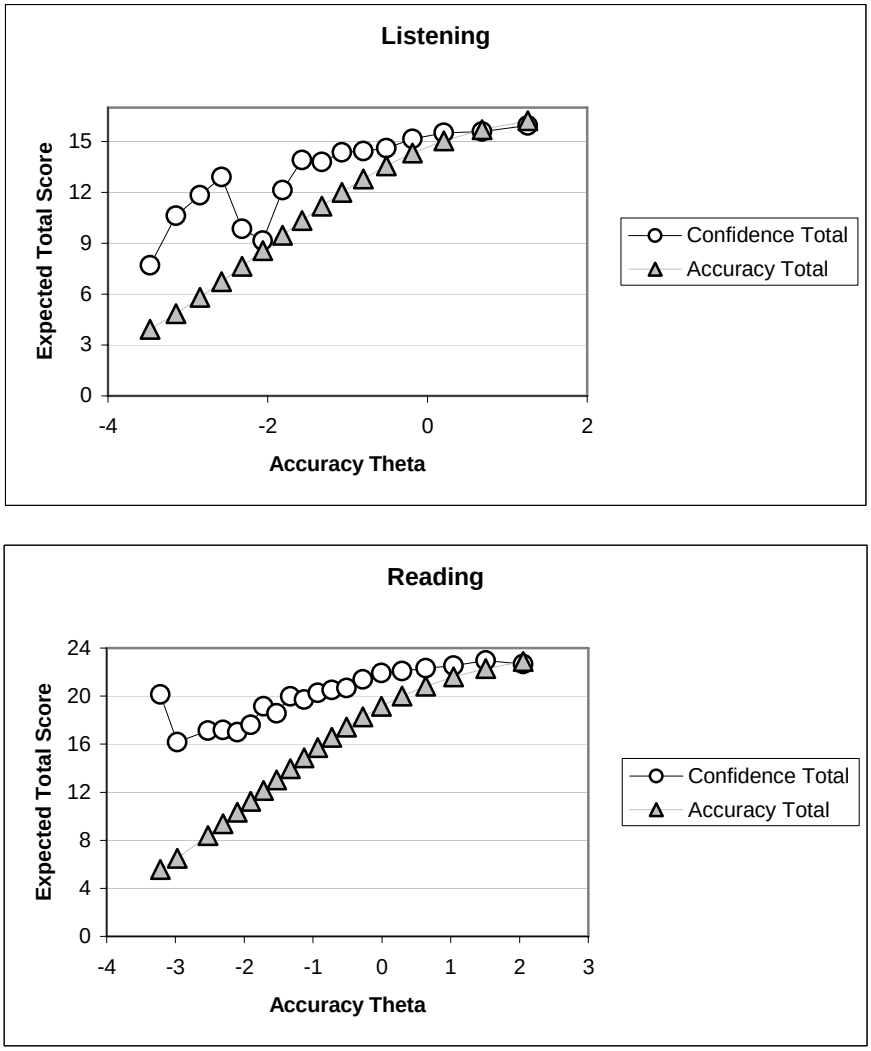


***Figure 3.*** **Confidence and accuracy at the test score level.**

Figure 4 shows the average standard deviation (SD) for the subjective probability conditional on $\theta_{ac.}$ The average SD for the subjective probability decreased as the ability on the accuracy dimension increased.

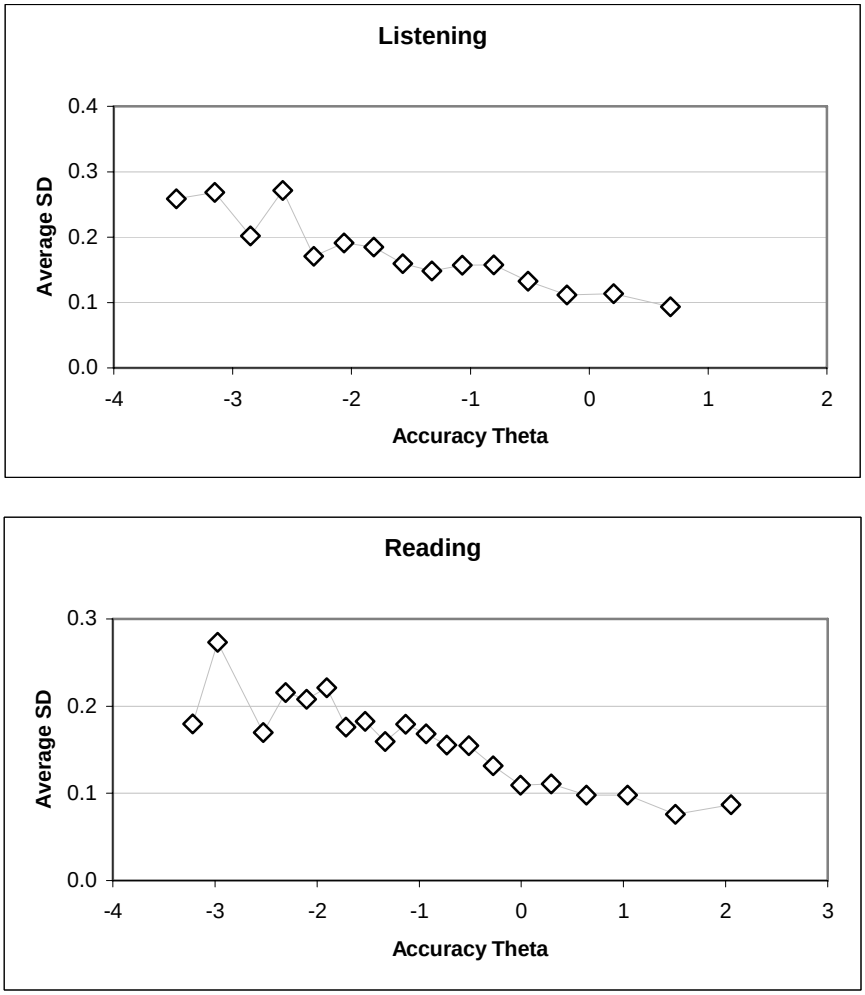


***Figure 4.*** **Average standard deviation plot for subjective probability.**

The plots in Figures 1 through 4 were constructed using $P_i^*$ (subjective probabilities on the *i*th item) and $\sum_i P_i^*$ (subjective expected test scores). Because of their nonparametric nature, different sample sizes at each level of $\theta_{ac}$ could affect the stability of $P_i^*$ and $\sum_i P_i^*$. To examine the degree of heterogeneity in the stability for $P_i^*$ and $\sum_i P_i^*$, the frequencies at each $\theta_{ac}$ were calculated, which are shown in Table 2. Both the Listening and Reading sections were easy for this group of participants, and average item difficulties were −2.09 and −1.70. Thus, higher frequencies are observed at the higher end of the accuracy scale, resulting in more stability for SD, $P_i^*$, and $\sum_i P_i^*$ at the higher-ability range on the accuracy dimension. However, the ceiling

effect in the confidence rating scale used in the present study, as addressed before, could be again contributing to this reduced variability at the higher end of ability.

**Table 2**

***Frequency Distributions for Ability Estimate on Accuracy***

| Listening | | | Reading | | |
|---|---|---|---|---|---|
| Number correct score | Ability estimate on accuracy | Frequency | Number correct score | Ability estimate on accuracy | Frequency |
| 2 | −3.47 | 1 | 4 | −3.22 | 2 |
| 3 | −3.15 | 7 | 5 | −2.97 | 4 |
| 4 | −2.85 | 4 | 7 | −2.53 | 7 |
| 5 | −2.58 | 9 | 8 | −2.31 | 11 |
| 6 | −2.32 | 5 | 9 | −2.10 | 16 |
| 7 | −2.06 | 2 | 10 | −1.91 | 13 |
| 8 | −1.81 | 16 | 11 | −1.72 | 22 |
| 9 | −1.57 | 18 | 12 | −1.53 | 24 |
| 10 | −1.32 | 25 | 13 | −1.33 | 35 |
| 11 | −1.07 | 38 | 14 | −1.13 | 34 |
| 12 | −0.80 | 50 | 15 | −0.93 | 41 |
| 13 | −0.51 | 71 | 16 | −0.73 | 36 |
| 14 | −0.19 | 127 | 17 | −0.51 | 42 |
| 15 | 0.21 | 143 | 18 | −0.27 | 59 |
| 16 | 0.68 | 186 | 19 | −0.01 | 69 |
| 17 | 1.26 | 118 | 20 | 0.29 | 86 |
| | | | 21 | 0.64 | 91 |
| | | | 22 | 1.04 | 103 |
| | | | 23 | 1.51 | 99 |
| | | | 24 | 2.05 | 26 |
| | Total *N* | 820 | | Total *N* | 820 |

***3. Are People's Confidence Levels Related to the Difficulty Level of a Specific Task?***

The Rasch item difficulty estimated from the unidimensional Rasch calibration for the accuracy dimension was compared with the item-level confidence measure (IDI). The correlation between the IDI and the item difficulties was 0.92 for the Listening section and 0.93 for the Reading section, showing a strong relationship between the item difficulty and item-level overconfidence. However, these relationships were rather curvilinear: As the item difficulty increased, the overconfidence increased in a nonlinear fashion. To find a trend in the nonlinear relationship, simple linear and polynomial regression models were fitted, and the model

comparisons were conducted using a general *F*-test in a forward selection manner with α level of 0.05 (see, for example, Sen & Srivastava, 1990, for *F*-test details). A lower-degree regression with respect to the predictor (difficulty) was compared to a one-degree higher polynomial regression. The model comparison continued consecutively from the simple linear model until the statistical test showed nonsignificance. We found that the second-degree polynomial model provided the best trend-line function describing the relationship between the item difficulty and the IDI for both the Listening and Reading sections. Figure 5 shows the plots where the second-degree polynomial line was overlaid with the coefficient of determination ($R^2$) and the regression equations estimated. The increment in $R^2$ was increased by 10% and 11% when the linear relationship was replaced by the second-degree polynomial. These $R^2$ changes were statistically significant. The third-degree polynomial relationship did not allow rejection of the second-degree polynomial.

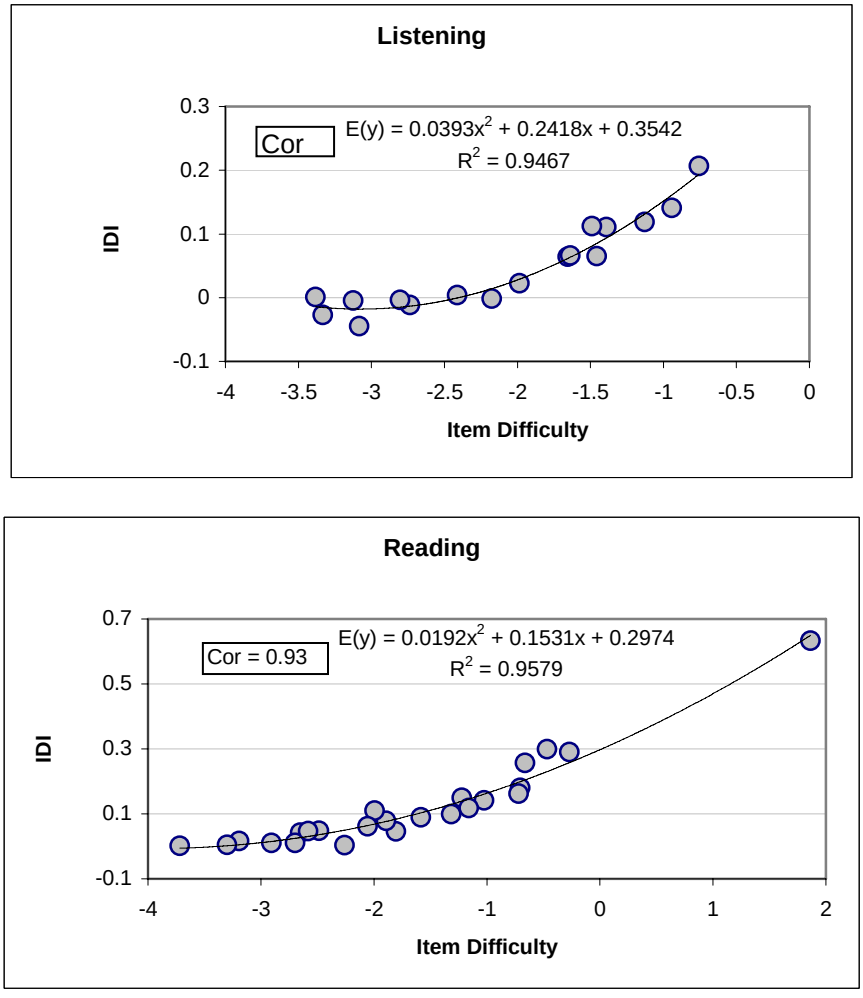


***Figure 5.*** **Relationship between item difficulty and item discrepancy index (IDI).**

### *4. What Is the Strength of Association Between People's Accuracy and Confidence Levels?*

Table 3 shows the variance, covariance, and correlation between accuracy and confidence scores estimated from the multidimensional Rasch analysis. The correlations between accuracy and confidence were 0.67 and 0.72 for the Listening and Reading sections, respectively. Confidence alone explained 45% of the variation in the accuracy dimension for the Listening section and 52% for the Reading section.

**Table 3**

***Latent Correlation Between Confidence and Accuracy Dimensions***

| | Listening | | Reading | |
|---|---|---|---|---|
| | Accuracy | Confidence | Accuracy | Confidence |
| Accuracy | 1.286 | 0.805 | 1.628 | 1.019 |
| Confidence | 0.670 | 1.124 | 0.720 | 1.231 |

*Note.* Values below the diagonal are correlations, and values above are covariances.

## Discussion

The IRT Rasch model-based approach allowed us to examine under/overconfidence at any desired level of ability at the item or test levels. Investigations could be conducted at certain ability points or intervals, which can be defined as a local level of confidence, or across all ability ranges. Furthermore, confidence can be examined at an individual item level, over a subset of items, or across the overall test. The local under/overconfidence indices at the item and test levels (Equations 6, 17, and 20) can also be calculated with a modification for an interval of interest in the latent dimension $\theta$. Their mathematical expressions are shown in Appendix B. The graphical representations, such as Figures 2 and 3, illustrate the gaps between accuracy and confidence, making it easy to diagnose magnitudes of confidence at points on the accuracy dimension and to examine the pattern of the relationship between confidence and accuracy.

The proposed IDI, TDI, DP, LIDI, LTDI, and LDP are robust against outliers in that they are weighted summary indices. These indices are calculated not only by the gap between confidence and accuracy, but also by their relative frequency distributions. For example, a contribution to the calculation of these indices for a large gap with a fairly small relative frequency would be minimized by the weighting factor, relative frequency. Summary indices

more sensitive to outliers can be obtained by omitting the term $w(\theta_{ac})$ in Equations 6, 17, 20, B1, B2, and B3 ,which would produce unweighted summary measures. The comparison between the unweighted and weighted indices would indicate the extent to which a relatively small number of people appear to have large gaps.

The weight $w(\theta_{ac})$ in Equations 6, 17, 20, B1, B2, and B3 makes use of the individual $\theta_{ac}$ distribution. Instead, another choice of weight is to use the population weight. In the MML estimation of the Rasch models, $\theta_{ac}$ was assumed to follow a normal distribution in this study. Thus, the population weight $w(\theta_{ac})$ is

$$\text{Population weight } w(\theta_{ac}^{*}) = \frac{1}{\sigma}\phi\left(\frac{\theta_{ac}^{*} - \mu}{\sigma}\right), \tag{25}$$

where $\phi(\cdot)$ is the standard normal density function, $\mu$ is the mean of the normal distribution, $\sigma$ is the standard deviation of the normal distribution, and $\theta_{ac}^{*}$ is the quadrature point used in the MML estimation.[3] The use of a population weight may be more attractive when the population weights are estimated as model parameters in a semi-parametric IRT MML approach (see, for example, Mislevy, 1984 for nonparametric modeling of the population distribution).

Although the IDI, TDI, and DP are useful summary measures, they have limitations. In particular, when confidence and accuracy are crossing in such a way that positive and negative gaps are essentially offsetting in the calculation of the difference between the two, the graphical representation should be used together with the IDI, TDI and DP. We can also quantify such a crossing trend by modifying Equations 6, 17, B1, and B2. Instead of using the difference between accuracy and confidence, the absolute difference (or the square difference) can be used. For example, $(P_i^{*} - P_i)$ can be replaced with $\left|P_i^{*} - P_i\right|$ (or $\left(P_i^{*} - P\right)^2$) and $\sum_i P_i^{*} - \sum_i P_i$ with $\left|\sum_i P_i^{*} - \sum_i P_i\right|$ (or $\left(\sum_i P_i^{*} - \sum_i P_i\right)^2$) prior to the summation over $\theta_{ac,}$ when the absolute difference is employed. A crossing pattern between confidence and accuracy is likely to be observed when there is a small IDI or TDI with large values of this modified version of IDI or TDI. Note that this modified version of IDI or TDI cannot provide directional information (i.e., overconfidence or

underconfidence) as our initially proposed IDI, TDI, and DP. Again, graphical displays with these summary indices can show local under/overconfidence at the item or test levels.

## Summary

This study employed the IRT analyses for the Reading and Listening sections of an English language proficiency test. One of the main purposes in this paper was to show the usefulness of IRT modeling in studying the relationship between confidence and accuracy. In our analyses, we treated confidence and accuracy as two different constructs. Based on the unidimensional Rasch model calibration, we proposed effect-size indices at the item and test levels that are potentially useful in measuring underconfidence or overconfidence. In addition, the use of the multidimensional Rasch modeling allowed us to investigate the strength of association between these two latent dimensions. It should be noted that the procedures and proposed over/underconfidence indices (e.g., IDI, TDI, and DP) in this study are applicable in general to whatever IRT models are adopted for analysis.

The IRT approach was also useful in providing answers for the substantive research questions that we had. We showed that the participants were overconfident in general: They tended to overestimate their performance on both the Listening and Reading sections of the English language proficiency test. Their overconfidence was negatively related to their ability levels and was positively related to the item difficulty. The strength of association between the accuracy and confidence latent dimensions was moderate: The confidence dimension alone explained 45% (Listening) and 52% (Reading) of the variability in the accuracy dimension.

There are a couple of cautionary statements that we want to make. One is that the overconfidence pattern observed in this study may not be salient in other cognitive subject tests (Juslin & Olsson, 1997). Especially, the pattern of the overconfidence could be different under a different testing environment, such as computerized adaptive testing where the item difficulty level is optimized for the examinee ability to calculate his or her ability level efficiently.[4] Another caution that should be made is that the decrease of the gap between accuracy and confidence at the high end of ability may be due to the subjective rating that was given on a probability unit with the maximum boundary of unity. If we were to gather the subjective rating on confidence into a scale set with no maximum boundary, we may be able to examine whether the better prediction of higher-ability participants has to do with the form of the subjective rating scale.

Although this study proposed potentially useful measures for under/overconfidence, these measures lack statistical significance testing procedures, which is one of the limitations of the current study. In addition, the thresholds for under/overconfidence classification at the test level were not developed in the present study. A simple classification for TDI can be made by the following rule:

$$\text{Small or negligible: test length}*\nu_M \leq |\text{TDI}|,$$

$$\text{Medium: test length}*\nu_M < |\text{TDI}| \leq \text{testlength}*\nu_L, \text{ and}$$

$$\text{Large: } |\text{TDI}| > \text{ test length}*\nu_L,$$

where $\nu_M$ and $\nu_L$ are the desired minimum expected threshold values of an item to be qualified as a medium or large TDI. If we use a rather conservative approach of $\nu_M$ = 0.05 (for medium) and $\nu_L$ = 0.10 (for large), the thresholds are 0.85 (17*0.05) and 1.7 (17*0.1) for the Listening section and 1.2 (24*0.05) and 2.4 (24*0.1) for the Reading section. We see that the TDI classification is small for the Listening section and large for the Reading section. The use of 0.05 and 0.10 is considered conservative because we can expect the IDIs to be at least medium or large to be flagged as medium or large TDI. This means that it may not capture the amplification effect arising from, say, many small positive IDIs whose effects could be accumulated to be relatively large TDI values. Both the classification of TDI and the development of the statistical significance testing described in this section warrant more research.

We also want to mention that this study did not account for a potential serial correlation impact from the use of the same stimulus. Item responses for the accuracy and confidence ratings were gathered under the same item stem, which may cause additional dependency between the accuracy and confidence dimensions in the data. Also, Reading and Listening items that share common stimuli (e.g., the same reading passage) may have extra data dependency as well, which is commonly known as a *testlet effect*. Modeling these potential clustering effects caused by the use of the same stimulus in a set of items can be explored in future studies.

## References


Adams, R. J., & Khoo, S. T. (1996). *Quest.* Melbourne, Australia: Australia Council for Educational Research.

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

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

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.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometirka, 46*, 443–459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Crawford, J., & Stankov, L. (1996). Age differences in the realism of confidence judgments: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual Differences, 6*, 84–103.

Dempster, A. P, Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* [series B]*, 39*, 1–38.

Dorans, N., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.35–66). Hillsdale, NJ: Lawrence Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fischer, G. H. (1989).Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3–26.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 15–38). New York: Springer-Verlag.

Hakstian, A. R., & Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures. *Journal of Educational Measurement. 12*(4), 231–239.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 53–68). New York: Springer-Verlag.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organisational Behavior and Human Decision Processes*, *57*, 226–246.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 10,* 344–366.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologia, 77*, 217–273.

Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, &A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases*. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 39-52). New York: Springer-Verlag.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237–255.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49–57.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago: University of Chicago Press.

Sen, A., & Srivastava, M. (1990). *Regression analysis: Theory, methods, and applications.* New York: Springer-Verlag.

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21*, 971–986.

Stankov, L., & Lee, J. (2007). *Confidence and cognitive test performance* (ETS Research Rep. No. RR-07-03). Princeton, NJ: ETS.

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

Wright, D. B., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, D. B., & Stone, M. H. (1979). *Best test design*. Chicago: University of Chicago.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models*. Unpublished master's thesis, University of Melbourne, Australia.

Wu, M. L., Adams, R. J., & Wilson. M. R. (1998). *ConQuest: Generalized item response modeling software.* Australia: Australian Council on Educational Research.

Yen, W. M., & Fitzpatrick, A R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.111–154). Westport, CT: Praeger.

## Notes

[1] The term *bias* is typically associated with statistical differential item function (DIF) techniques and used as a term for test fairness in educational testing context. Readers should not be confused by this.

[2] For the multidimenssional Rasch model calibration, the strict monotonicity in both $\theta_{ac}$ and $\theta_{co}$ with normality are assumed, but the minimal assumptions to generate item responses for confidence do not need to specify normality, the shape of IRF, and strict monotonicity in $\theta_{ac}$.

[3] Gauss-Hermite quadratures are used in the evaluation of an integral in the marginal likelihood during the MML estimation. ConQuest provides empirical Bayes solutions in the Rasch calibration for the normal distribution mean and variance. In the actual estimation, only the variance of the normal distribution is estimated, while the mean is fixed as 0 because of the identifiability issue. The number of quadrature points can be user-specified values (or default values can be used) in ConQuest.

[4] In addition, the results could be very different with samples from different populations.

## Appendix A

## Item Difficulty Estimate and Item Fit Statistic

**Table A1**

***Listening***

| Item no. | Difficulty | *SE* | WMNSQ |
|---|---|---|---|
| 1 | −1.129 | 0.086 | 1.05 |
| 2 | 2.175 | 0.108 | 1.04 |
| 3 | −2.412 | 0.116 | 0.92 |
| 4 | −1.656 | 0.095 | 0.96 |
| 5 | −1.986 | 0.103 | 0.96 |
| 6 | −2.736 | 0.128 | 1.01 |
| 7 | −2.803 | 0.131 | 1.02 |
| 8 | −1.457 | 0.091 | 1.09 |
| 9 | −0.942 | 0.084 | 1.02 |
| 10 | −1.391 | 0.090 | 1.11 |
| 11 | −3.084 | 0.144 | 0.97 |
| 12 | −3.332 | 0.157 | 0.94 |
| 13 | −1.638 | 0.095 | 1.00 |
| 14 | −0.758 | 0.082 | 1.12 |
| 15 | −1.490 | 0.092 | 0.98 |
| 16 | −3.126 | 0.146 | 0.98 |
| 17 | −3.383 | 0.160 | 0.84 |

*Note.* WMNSQ = weighted mean squared.

**Table A2**

***Reading***

| Item no. | Difficulty | *SE* | WMNSQ |
|---|---|---|---|
| 1 | −2.484 | 0.116 | 1.01 |
| 2 | −3.718 | 0.176 | 0.98 |
| 3 | −1.222 | 0.088 | 0.98 |
| 4 | −2.258 | 0.109 | 0.88 |
| 5 | −2.653 | 0.121 | 0.94 |
| 6 | −0.704 | 0.083 | 0.97 |
| 7 | −3.191 | 0.144 | 1.04 |
| 8 | −0.467 | 0.082 | 1.04 |
| 9 | −1.803 | 0.098 | 0.89 |
| 10 | −1.891 | 0.100 | 0.84 |
| 11 | −1.316 | 0.089 | 1.07 |
| 12 | −2.907 | 0.131 | 0.94 |
| 13 | −1.993 | 0.102 | 0.87 |
| 14 | −0.718 | 0.083 | 0.98 |
| 15 | −1.025 | 0.086 | 1.02 |
| 16 | −2.697 | 0.123 | 0.99 |
| 17 | −1.584 | 0.094 | 0.91 |
| 18 | −0.269 | 0.081 | 1.05 |
| 19 | −2.056 | 0.103 | 1.03 |
| 20 | −0.663 | 0.083 | 0.90 |
| 21 | −3.299 | 0.150 | 0.90 |
| 22 | −2.580 | 0.119 | 0.94 |
| 23 | −1.160 | 0.087 | 1.00 |
| 24 | 1.864 | 0.098 | 1.31 |

*Note.* WMNSQ = weighted mean squared.

## Appendix B

## Local Item-Level Discrepancy Index, Local Test-Level Discrepancy Index, and Local Discrepancy Percentage

$$\text{Local IDI (LIDI)} = \sum_{\theta_{ac} \in [\theta_S, \theta_L]} (P_i^* - P_i) w(\theta_{ac}), \tag{B1}$$

$$\text{Local TDI (LTDI)} = \sum_{\theta_{ac} \in [\theta_S, \theta_L]} [\sum_i P_i^* - \sum_i P_i] w(\theta_{ac}), \text{ and} \tag{B2}$$

$$\text{Local DP (LDP)} = \frac{\sum_{\theta_{ac} \in [\theta_S, \theta_L]} [\sum_i P_i^* - \sum_i P_i] w(\theta_{ac})}{\text{Test Length}} \times 100, \tag{B3}$$

where $\theta_S$ and $\theta_L$ are the smallest and the largest boundary of interest in $\theta$.