

# Research Summary: Differential Item Functioning and Unidimensionality in the Pearson Test of English Academic

Hye Pae, Ph.D.  
University of Cincinnati, Ohio, USA

July 2011

## 1. Introduction

Since the Pearson Test of English Academic (PTE Academic) was designed to assess skill differences among test-takers at all points along the ability continuum, rather than to determine cutoff scores, it is important to examine the extent to which the instrument assesses what it is intended to measure (validity) as well as the extent to which the test is consistent (reliability) in measuring ELLs' academic English skills. A Rasch measurement model approach permits joint scaling of person abilities and assessment item difficulties for mapping the relationship between latent traits and responses to test items (Linacre, 2010a).

Test items should not behave differently for particular subgroups of test takers. If an item functions differently for certain groups, the item reduces the validity of the measure for that construct, and test fairness is threatened. The Rasch measurement model enables the detection of test items which are biased toward different subgroups according to construct irrelevant factors, such as ability, gender, and ethnicity, subgroups, by calculating differential test functioning (DTF) and differential item functioning (DIF) measures.

The assumptions of the Rasch model include unidimensionality (i.e., whether the items form a unitary latent trait) and local independence (i.e., the probability of the person correctly responding to an item does not depend on the other items in the test; Green, 1996; Lee, 1997). Unidimensionality and local independence are assessed using fit statistics, which report the extent to which the pattern of observed answers and the modeled expectations are evaluated in terms of item fit and person fit to the Rasch model.

In order to evaluate the psychometric properties of Form 1 of PTE Academic, two research questions were addressed in this study.

1. How does PTE Academic function in terms of the interaction of the person-gender and person-language environment (i.e. whether or not the test taker has lived in an English speaking country) with each item?
2. To what extent do the item responses of PTE Academic form a unidimensional construct according to the Rasch measurement model?

The first research question dealt with DTF and DIF so as to explore evidence for statistical between-group (i.e., gender and language environment) differences in the measurement properties at the item level. The second research question concerned the unitary latent trait underlying person ability and item difficulty produced in PTE Academic.

## 2. METHODS

### 2.1 Participants

One hundred forty ELLs took part in the study. The participants' mean age was 26.45 (SD=5.82), ranging from 17 to 46 years of age. Females accounted for 53.6% (75 examinees) and males 46.4% (65 examinees). Sixty four percent of the participants had lived in English-speaking countries.

### 2.2 Measures

The dataset used for this study was part of the larger field test administered by Pearson in 2007. Of 42 forms used for the two field tests, Form 1 of PTE Academic was analyzed in this study. It contained 86 items, including dichotomous (i.e., single point items) and polytomous (i.e., multipoint scale items) scales. Person-test reliability and item reliability of Form 1 were excellent (person  $r = .96$  and item  $r = .99$ ).

### 2.3 Procedure

Examinees who wanted to demonstrate their academic English skills for different reasons were recruited from 33 countries worldwide in 2007. Testing took place individually at test centers designated by Pearson.

### 2.4 Analysis plan

The psychometric properties of the items were analyzed utilizing the Winsteps software (Linacre, 2010b). Since the dataset comprised both dichotomous and polytomous they were analyzed utilizing the Partial Credit Model (Masters, 1982). Good-fit and misfit items were identified using infit and outfit mean-square values, and DTF and DIF were examined to test measurement invariance. The first DTF/DIF analysis was performed to examine whether the test items functioned differently by gender, and the second DTF/DIF examined whether the items favored examinees from the English-speaking setting over those from non-English-speaking countries. Unidimensionality was checked through principal component analysis (PCA), along with infit/outfit statistics and DTF/DIF.

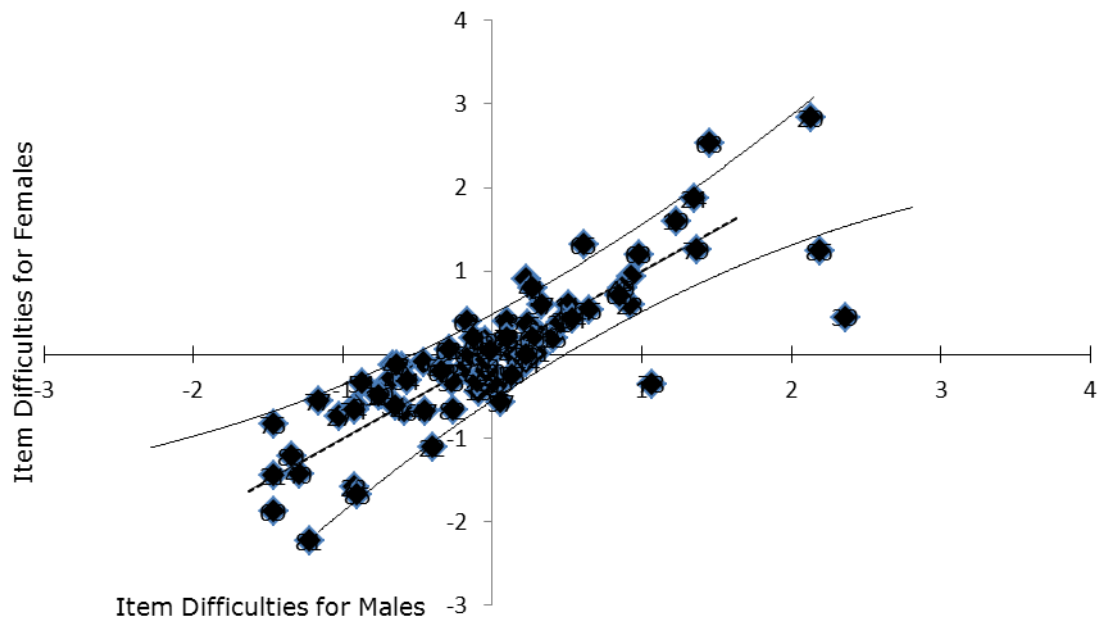
## 3. Results

### 3.1 Differential Test Functioning and Differential Item Functioning

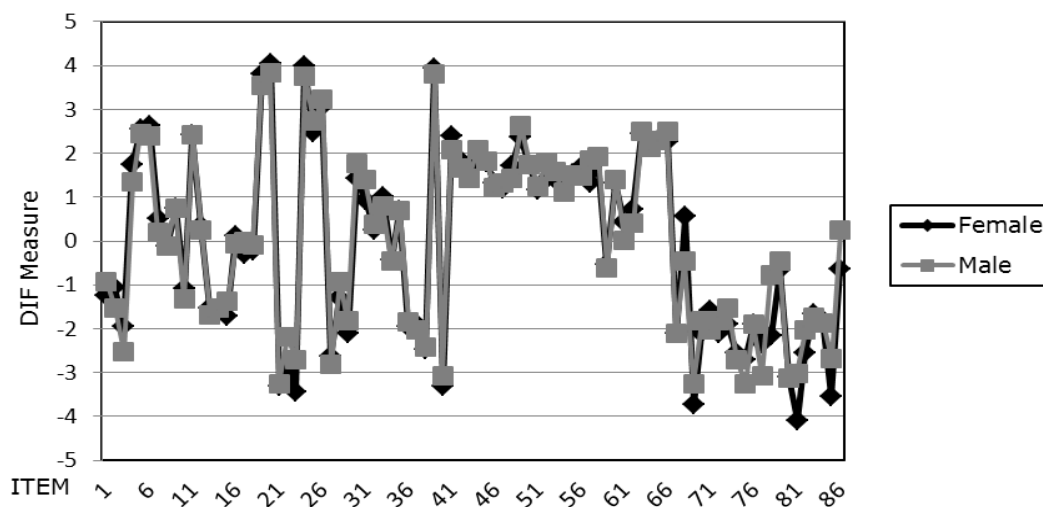
DTF investigates whether the test functions the same way for different groups of test-takers, through a comparison of how the test items function. Figure 1 displays a scatterplot for the males' and females' item difficulties. Items 20 and 68 seemed to be more difficult for females, while items 39, 78, and 86 were more difficult for the males. Although there were two outliers (items 38 and 78), it is hard to assume a test bias because the source of the outlying position is unknown. The Student's  $t$ -statistics for the items 38 and 78 were 4.91 (male measure = 2.36, female measure = 0.45;  $p < .05$ ) and 3.72 (male measure = 1.07, female measure = -0.35;  $p < 0.5$ ), respectively. The misfit-inflated standard errors were significantly small. The correlation coefficient between the males' and females' item difficulties was 0.84, and the measurement-error-removed disattenuated correlation was 0.98.

Because PTE Academic is expected to impact all ability levels in the same way across subgroups, uniform DIF analysis was run to investigate the interaction of the person-groups with each item, controlling for all other item and person measures. In order to determine whether the performance of a test item was significantly different for either gender subgroup, the difficulty was measured for the male and female subgroups.

Figure 2 shows person-item interactions with the absolute logit difficulty of each item for male and female subgroups in the same frame of reference. Although the females and males had different group sizes and mean measures, DIF computations adjusted for these differences (Linacre, 2010a). The difficulty of each item for both subgroups was remarkably similar with a few discrepancies, indicating that the test items functioned similarly for different subgroups of examinees. Given no notable DIF within the Rasch model framework, overall, PTE Academic appeared to produce DIF-free person estimates.

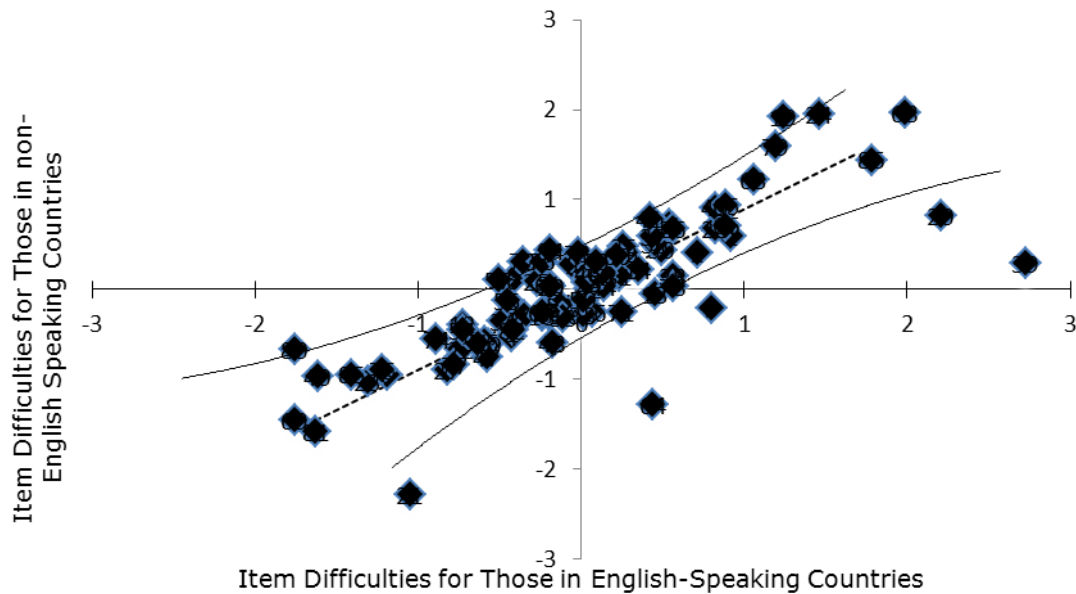


**Figure 1:** Differential Test Functioning by gender

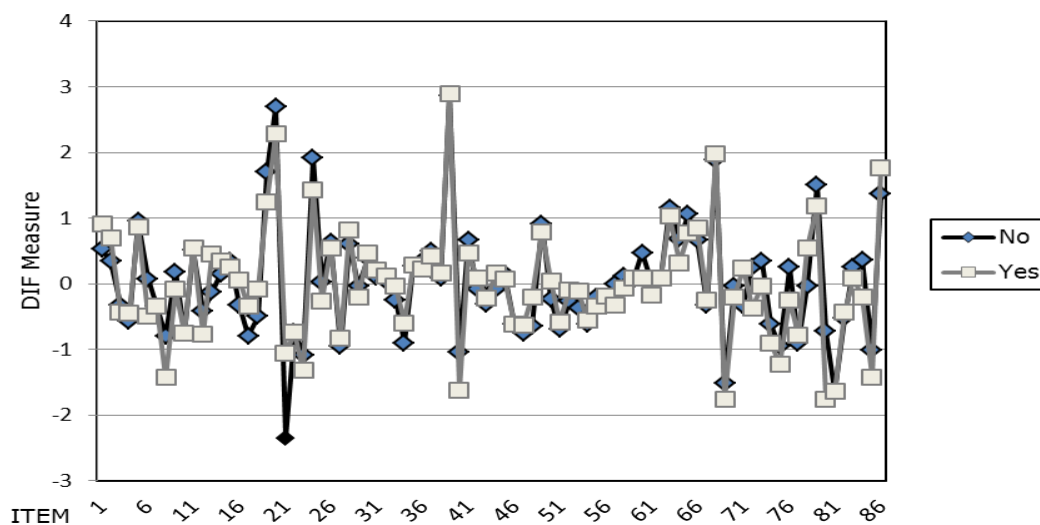


**Figure 2.** Differential Item Functioning by gender

In order to cross-validate the findings of the gender DTF and DIF, another DTF/DIF analysis was performed for learning-context subgroups (i.e., those who have lived in English-speaking countries vs. those who have not). Surprisingly, items 64, 5, 20, and 39 were more difficult for the examinees from English-speaking countries. Except these four items, the test items behaved well for the measurement model across the spoken language environment. Interestingly, item 78, which showed evidence of bias in both DTF and DIF analyses for the gender subgroups, did not show a bias for the language-setting subgroups (see Figure 3). Figure 4 displays a plot of DIF for the language subgroups. Item 21 appeared to be more difficult for individuals from non-English speaking countries than their counterparts. Item 21, which was one of the multiple-choice listening items that required a single answer, might be an item which had a potential threat to test fairness. Except item 21, the data supported the hypothesis that the difficulties of each pair of items in the two analyses were statistically not different.



**Figure 3.** Differential Test Functioning by language environment



**Figure 4.** Differential Item Functioning by language environment

### 3.2 Unidimensionality

A one-dimensional construct, equal item discrimination, and low susceptibility to guessing are the fundamental requirements of the Rasch measurement (Linacre, 2010a; Green, 1996; Sick, 2010). Given that unidimensionality is closely related to a factor structure comprising one major latent trait, principal components analysis (PCA) offers a means to evaluate the suitability of data for Rasch analysis (Camilli, Wang & Fesq, 1995).

In order to assess measurement dimensionality, PCA of the Rasch residuals was performed. The amount of the variance explained by different components in the data was 76.1% with 20.7% explained by persons and 55.4% explained by items. This indicated that a dominant first factor was present. According to Reckase (1979), the variance explained by the first factor should be greater than 20% as to be indicative of unidimensionality. The variance explained in this study exceeded the requirement of this criterion, demonstrating a unidimensional trait of the data. This indicated that the items did fit the model well with relatively good item-person targeting and wide dispersion of the items and persons. There were small amounts of unexplained variances in the components which came from the residuals (1.5%, 1.3%, 0.8%, and 0.7% for the first, second, third, and fourth contrasts, respectively). Residual factor loadings suggested that the data closely approximated the Rasch model, and there were no meaningful components beyond the primary dimension of measurement. PCA demonstrated, by and large, that there were no extraneous dimensions or sub-dimensions related to sub-skills.

Since unidimensionality is achieved by forming a single underlying pattern in a data matrix, it was assumed that local independence was met. Local independence was achieved by controlling for all the abilities so that responses to items could be independent of one another (Hambleton, Swaminathan & Rogers, 1991).

## 4. Discussion

This study examined the interaction of person abilities and item difficulties of PTE Academic. No conspicuously aberrant responses were observed. Hence, the hypothesis that Form 1 of PTE Academic data did fit the Rasch model was supported. The infit and outfit MNSQs did not excessively depart from the acceptable range, which indicated no unexpected responses and supported the measurement. The small departure suggested that misfit items might have happened due to chances or randomness predicted by the Rasch model.

DIF analysis supported a similar probability of endorsing each item category across the gender subgroups as well as the language-context subgroups. The seemingly biased items in the gender DIF did not overlap with those in the language-context DIF. Therefore, it seemed that the probabilities of correct responses were comparable, without significantly favoring one subgroup over another. Given the low degree of DIF, the hypothesis that the difficulties of each pair of items in the two subgroups were not significantly different was supported. This suggests that PTE Academic scores are relatively free of construct irrelevant variance thus supports the argument for the construct validity.

A concern about the multiplicities (Sick, 2010) of subgroup assignment suggests that keeping moderately biased items may not do harm because eliminating items with DIF by gender or language contexts does not guarantee that the test will be free of item biases by other subgroup classification.

This study contributes to the field of language testing with empirical evidence for the use and interpretation of PTE Academic scores through a detailed evaluation of the response pattern, item fit, dimensionality, and the detection of item bias. Appraising constructs and test fairness for PTE Academic is important because it is currently in

use for ELLs' academic English measurement worldwide. The appropriateness, meaningfulness, and usefulness of PTE Academic will allow decision-makers in education and business settings to make useful predictions or inferences based on PTE Academic scores.

## 5. References

- Camilli, G., Wang, M. & Fesq, J. (1995). The Effects of Dimensionality on Equating the Law School Admissions Test. *Journal of Educational Management*, 32, 1, 79-96.
- Green, K. E. (1996). Applications of the Rasch model to evaluation of survey data quality. *New Directions Evaluation*, 70, 81-92.
- Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.
- Lee, J. (1997). State activism in education reform: Applying the Rasch model to measure trends and examine policy coherence. *Educational Evaluation and Policy Analysis*, 19 (1), 29-43.
- Linacre, J. M. (2010a). A user's guide to Winsteps. Retrieved December 1, 2010 <http://www.winsteps.com/winman/index.htm?guide.htm>.
- Linacre, J. M. (2010b). *Winsteps (Version 3.70.02) [Computer Software]*. Chicago: Winsteps.com.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Pearson (2009). *Pearson Test of English Academic*. London: Pearson <http://www.pearsonpte.com/PTEACADEMIC/Pages/home.aspx>
- Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multi-Factor Tests: Results and Implications. *Journal of Educational Statistics*, 4, 207-230.
- Sick, J. (2010). Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing & Evaluation SIG Newsletter*. 14 (2), 23 – 29.