行政院國家科學委員會專題研究計畫 成果報告

科學課堂中的性別刻板印象威脅對女學生科學學習之影響
研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 96-2522-S-110-001-
執行期間：96年11月01日至97年10月31日
執行單位：國立中山大學教育研究所

計畫主持人：鄭英耀
共同主持人：鍾素香
計畫參與人員：碩士班研究生-兼任助理人員：陳秋蟬
碩士班研究生-兼任助理人員：張逸柔
碩士班研究生-兼任助理人員：蔡學斌
博士班研究生-兼任助理人員：劉昆夏
博士班研究生-兼任助理人員：陳宜伶

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫涉及專利或其他智慧財產權，1年後可公開查詢

中華民國 97年12月31日

行政院國家科學委員會補助專題研究計畫 ■ 成果報告
□期中進度報告

科學課堂中的性別刻板印象威脅對女學生科學學習之影響

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 96－2522－S－110－001
執行期間：　96　年　11　月　1　日至　97　年　10　月　31　日

計畫主持人：鄭英耀　　國立中山大學教育研究所
共同主持人：鍾素香　　國立中山大學教育研究所
計畫參與人員：劉昆夏、陳秋蟬、張逸柔、陳宜伶、蔡學斌

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、
列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立中山大學教育研究所

中　華　民　國　97　年　12　月　30　日

科學課堂中的性別刻板印象威脅對女學生科學學習之影響

Influence of Gender Stereotype Threat on Female Students' Learning in Science Classes

計畫編號：NSC 96-2522-S-110-001-

執行期限：96 年 11 月 1 日至 97 年 10 月 31 日

主持人：鄭英耀 國立中山大學教育研究所

共同主持人：鍾素香 國立中山大學教育研究所

研究助理：劉昆夏、陳秋蟬、張逸柔、陳宜伶、蔡學斌

中文摘要

本研究旨在探討科學課堂中不同性別的師生互動行為對學生的性別刻板印象威脅之影響，以及科學性別刻板印象威脅、科學能力知覺、科學價值期望、科學成就目標和科學成就表現之關係。首先，本研究依據性別刻板印象威脅理論及相關研究，發展「科學性別刻板印象量表」和「科學領域認同量表」。研究工具包括「科學價值期望量表」、「科學能力知覺量表」、「科學成就目標量表」。其次，以立意取樣選取高雄市11 所國中22 個班級進行問卷施測。研究結果發現：被男老師教的男學生之科學領域認同的分數顯著高於被男老師教的女學生。被男老師教的女學生之科學性別刻板印象顯著高於被女老師教的男學生。科學能力知覺和科學價值有正相關，且兩者皆可正向的預測趨向表現目標、趨向精熟目標。科學能力知覺可以直接預測學生的學習成就。

關鍵詞：性別刻板印象威脅、性別差異、科學學習、師生互動

Abstract

The study aims to explore the gender influence of teacher-student interaction on students' gender stereotype threat in science classes and the relationship among gender stereotype threat, perceptions of science competence, science values, achievement goals and science learning achievement. Based on gender stereotype threat theory and related studies, the questionnaire of "the Gender Stereotype of Science Inventory" and "Science Identification Inventory" were developed. The instruments include "Expectancies-Values Questionnaire," "Perception of Scientific Competence Scale," and "The Achievement Goal Questionnaire." Data were collected from 22 science class students of 11 junior high schools. Results showed: male-male group (male teacher and male student) had stronger science identification than male-female group (male teacher and female student); the male-female group showed significantly higher gender stereotype belief in sciences than the female-male group. Findings supported the proposed model in which perceptions of science competence was positively correlated with science values. Both perceptions of science competence and science values positively predicted performance-approach goals, mastery-approach goals. Perceptions of science competence directly impacted science achievement.

Keywords: gender stereotype threat, gender difference, science learning, teacher-student interaction

一、研究背景

大規模調查的科學成就測驗發現，男、女學生在科學表現上存有差異 (Martin, Mullis, Gonzalez, & Chrostowski, 2004; OECD, 2007)。根據 Steele(1997)的「刻板印象威脅」（gender stereotype threat）理論，學生的學業表現可能會受他們的性別刻板印象信念影響 (Kiefer & Sekaquaptewa, 2007; Spencer, Steele, & Quinn, 1999)。針對課堂中的師生互動行為，Martin 和 Marsh(2005)提出性別刻板印象模式(Gender-stereotypic model)，他們認為男學生被男老師教時，表現比較好；女學生被女老師教時，表現比較好。Simth(2006)所提出的刻板印象任務投入歷程模式認為，刻板印象威脅會導致個體採取逃避表現目標取向，進而對於成就表現產生負面的影響。然而，她並未說明為何刻板印象威脅會導致個體採取逃避表現目標取向。經深入探究有關動機信念、價值與目標的相關理論（Bandura, 1997; Covington, 2000; Eccles & Wigfield, 2002; Fryer & Elliot, 2007; Moller & Elliot, 2006; Weiner, 1992）發現，Jacquelynne S. Eccles 的研究團隊自 1983 年起即長期投入性別與成就的相關研究，他們所發展與建構的價值期望理論（Eccles, Adler, Futterman, Goff, Kaczala, Meece, & Midgley, 1983; Simpkins, Davis-Kean, & Eccles, 2006; Wigfield & Eccles, 2000）似乎可以較完整的解釋性別刻板印象對成就表現的影響歷程。依據 Eccles 等人所建構的價值期望模式（Expectancy-Value Model, 如圖 1），學生在選擇是否投入某種任務時，會同時考量成功的可能性與任務的價值性。其中成功期望會受學生所訂定的目標和自我概念基模（包括自我與社會認同、個人能力的自我概念等）影響。而自我概念基模的形成原因則來自學生對於性別角色與特定任務、活動的刻板印象的知覺。換句話說，當學生知覺到學習環境中存在刻板印象訊息，且任務具有難度時，依據刻板印象威脅理論（Steele, 1997; Steele & Aronson, 1995），此時學生內在既有的刻板印象可能會被激發，進而經由成功期望評估後，再決定是否投入學習。此外，個人主觀任務價值判斷（包括興趣-快樂價值、完成任務的重要性和實用價值等），亦是影響學生選擇與投入學習的另一關鍵因素。

二、研究目的

基於上述，本研究試圖探究科學課堂中不同性別的師生互動行為對學生的性別刻板印象威脅之影響，以及科學性別刻板印象威脅、科學能力知覺、科學價值期望、科學成就目標和科學成就表現之關係。具體而言，本研究的目的為：

（一）發展一份具良好信效度的「科學性別刻板印象威脅量表」，供本研究以及教育與心理領育之研究使用。

（二）比較不同性別的師生配對，學生知覺科學性別刻板印象之差異。

（三）探討科學性別刻板印象威脅、學生科學能力知覺、科學價值期望、科學成就目標和科學成就表現之關係。

三、研究方法與結果

本計畫針對上述三個研究目的，分別進行研究，研究結果已撰寫論文，投稿至國際學術研討會，分別摘述如下：

Study 1: Development of the Gender Stereotype of Science Inventory and the Science Identification Inventory

Theoretical Framework

Stereotype threat refers to being at risk of confirming, as self-characteristic, a negative stereotype about one's group (Steele & Aronson, 1995, p. 797). A negative stereotype threat must be self-relevant. How threatening this recognition becomes depends on the person's identification with the stereotype-relevant domain (Steele, 1997).The essential components for stereotype threat include identification with a domain, self-relevant stereotype belief, and stereotype salience of situations.

Three facets were considered in the development of the GSSI and SII: (a) theoretical constructs of stereotype threat (Steele, 1997; Steele & Aronson, 1995), (b) taxonomy of educational objectives of affective domain (Krathwohl, Bloom, & Masia, 1964), and (c) procedure of constructing measures (Wilson, 2005).

There were three components in the first facet: identification with science, gender stereotype belief of science, and gender stereotype salience of classroom situations. In the second facet, the taxonomy of educational objectives of affective domain (e.g., awareness, willingness to receive, satisfaction in response, acceptance of a value, preference for a value, and commitment) were adopted to generate items. In the third facet, the four procedure of constructing measures, construct map, items design, outcome space, and measurement model, were applied to connect item generation, analyses, and evaluation.

Method

Item Generation and Revision

Four graduate students were recruited to generate draft items. Four experts, including two professors of social psychology, one professor of science education, and one professor of educational and psychological measurement were asked to check the correspondence between the definition of theoretical constructs and the content of items, to verify the order of the affective level of each item based on the taxonomy of educational objectives of affective domain, and to provide additional feedback for enhancing content validity of the items. From these procedures, 15 items were obtained. The GSSI had two subscales, *Stereotype Belief* and *Stereotype Situation*, each with five 5-point Likert-type items. The SII also had five 5-point Likert-type items.

Participants and Procedure

Two samples of students participated in test development, one for item revision and the other for validation. Sample 1 consisted of 295 (143 boys and 152 girls) eighth graders from five middle schools in Taiwan. They completed the drafts of the GSSI and SII. Sample 2 consisted of 604 (303 boys and 301 girls) eighth graders from 11 middle schools in Taiwan. In addition to the revised GSSI and SII, they completed the subscales *Interest in Science* and *Perceptions of Science Importance* of the Expectancies-Values Inventory (EVI; Simpkins, Davis-Kean, & Eccles, 2006), and the subscales *Self-efficacy* and *Perceived Difficulty* of the Subjective Competence Inventory (SCI, Cherng, 2006).

Analysis

There were two phases in data analysis. The first phase of inventory construction consisted of internal consistency analysis and model-data fit analysis with the Rasch rating scale model (RSM; Andrich, 1978) using the computer program ConQuest (Wu, Adams, & Wilson, 2007). The second phase of inventory examination consisted of Rasch analysis, confirmatory factor analysis and multigroup analysis using the computer program AMOS (Arbuckle, 2006). Finally, criterion-related validity analysis was conducted to provide additional evidence of construct validity.

Results

Phase 1

For Sample 1 of students, the Cronbach alpha coefficient was .88 and .77 for the subscales of Stereotype Belief and Stereotype Situation, and .88 for the SII. In the Rasch analysis, all items had a good model-data fit. However, its difficulty order did not match very well the expected order by experts. We then decided to randomize item order and administered the tests to Sample 2 of students. A better model-data fit and better item order matches were obtained.

Phase 2

Person-item Match

The thresholds (difficulties) of the five items in the SII were between -1.00 and 0.62 logits ($M = 0$), which matched the students' latent trait levels (ranging from -3.47 to 3.52 logits, $M = 0.27$) fairly well. The thresholds of the five items in the subscale Stereotype Belief were between -0.38 and 0.91 logits ($M = 0$), which were higher than the students' latent trait levels (ranging from -3.04 to 2.28 logits, $M = -1.12$). Thus, it was relatively difficult for the students to agree with the items. Likewise, the thresholds of the five items in the subscale Stereotype Situation were between -0.41 and 0.42 logits ($M = 0$), which were also higher than the students' latent trait levels (ranging from -2.67 to 1.57 logits, $M = -1.45$). Obviously, easier items in the subscales of Stereotype Belief and Stereotype Situation should be developed in the future to match this kind of samples.

Confirmatory Factory Analysis

According to stereotype threat theory, a three-factor structure of stereotype threat was tested. The fit indices (GFI = .94, AGFI = .92, SRMR = .05, RMSEA = .06, TLI = .94, and CFI = .95) revealed a good model-data fit. In addition, the reliability measures of factor loading (ranging from .50 to .82) and composite reliability (ranging from .77 to .87) were all in an acceptable range. Overall, our data supported the three-factor structure of stereotype threat.

Multigroup Analysis

Table 1 summarizes the results of multigroup (boys and girls) confirmatory factor analysis. A ΔCFI value smaller than or equal to -0.01 is indicative of significant drop in model fit (Cheung & Rensvold, 2002). Smaller values of AIC and BCC were better fit than larger values (Arbuckle, 2006). Overall, Hypothesis 2 of equal loadings across genders had the best model-data fit.

Table 1. Goodness-of-fit statistics for test of invariance across genders

Hypothesis Model	χ^2	*df*	$\chi^2/_{df}$	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA	AIC	BCC
H1: Base model	405.23	174	2.33	—	—	.937	—	.047	537.23	544.64
H2: Equal loadings	416.11	186	2.24	10.88	12	.937	.000	.045	524.11	530.17
H3: Equal loadings, factor covariances	433.51	192	2.26	28.28	18	.934	-.003	.046	529.51	534.90
H4: Equal loadings, factor covariances, measurement residuals	505.04	207	2.44	99.81^{***}	33	.918	-.016	.049	571.04	574.74

Note. Boys: $N = 303$; Girls: $N = 301$; *** $p < .001$.

Criterion-related Validity

The relationships among Science Identification, Science Interest, Science Importance, and Self-efficacy are shown in Table 2. As expected, Science Identification was positively correlated with these three criteria. In addition, girls' Stereotype Belief was positively correlated with Stereotype Situation ($r = .55$) and Perceived Difficulty ($r = .15$). These results provided additional evidences of criterion-related validity for the GSSI and SII.

Table 2. Relationships among Science Identification, Science Interest, Science Importance, and Self-efficacy

	Science Identification	Science Interest	Science Importance
Science Interest	.74**	—	
Science Importance	.66**	.65**	—
Self-efficacy	.70**	.64**	.49**

Note. $N = 604$; $^{**}p < .01$.

Conclusion

In this study, we successfully develop the GSSI and SII to measure gender stereotype threat in learning sciences. The development integrates theoretical construct of stereotype threat, affective level of items, and construct mapping. Classical testing theory analyses support reliability, construct validity, and criterion-related validity for the two tests. Rasch analyses demonstrate adequate model-data fit and provide person-item match and item difficulty order, suggesting the affective levels are adequate, which is an important step forward understanding students' attitude in learning sciences. Multigroup confirmatory factor analysis supports equal loadings across genders.

Study 2: Relationship between Teachers' and Students' Genders and Students' Stereotype in Sciences

Method

Participants and Procedures

Participants were 640 junior high school students (325 boys and 315 girls). According to the principle, one school out of the 11 districts in Kaohsiung, these participants were from 22 classes taught by 22 science teachers (11 men and 11 women). Participants received the Science Gender Stereotype Inventory (SGSI), measuring individuals' awareness of gender stereotype threat in learning sciences, and the Science Identification Inventory (SII), measuring individuals' attitude in learning sciences. A 2 by 2 teacher-student matched-pair analysis (including, male-male matched group, male-female matched group, female-male matched group, and female-female matched group) was conducted in the study.

Instrument Development

Stereotype threat refers to being at risk of confirming, as self-characteristic, a negative stereotype about one's group (Steele & Aronson, 1995, p.797). A negative stereotype threat must be self-relevant. How threatening this recognition becomes depends on the person's identification with the stereotype-relevant domain (Steele, 1997). The essential components for stereotype threat include identification with a domain, self-relevant stereotype belief, and stereotype salience of situations.

Items of the Science Gender Stereotype Inventory and the Science Identification were initially generated through brainstorming by 4 graduate students. The Science Gender Stereotype Inventory (SGSI) included two subscales: Gender Stereotype Belief Subscale and Gender Stereotype Situation Subscale, measuring individuals' awareness of gender stereotype threat in learning sciences. The Science Identification Inventory (SII) measured individuals' attitude in learning sciences. Each SGSI subscale and SII had 5 items using a 5-point Likert-type scale with responses ranging from *strongly disagree* (1) to *strongly agree* (5). Four experts, including two professors of social psychology, one professor of science education, and one professor of educational and psychological measurement were invited to check the draft items. Their work included checking the correspondence between the definition of theoretical constructs and the content of the items, examining the order of the affective level of each item based on the taxonomy of educational objectives of affective domain, and providing additional feedback for enhancing content validity of the items. After the process, 15 items were obtained. Examples of SSI ranked from the lowest to the highest by the affective levels were: "I am willing to learn physical science in class" (willingness to receive), "It is joyful for me to learn physical science" (satisfaction in response), "It is important for me to learn physical science" (acceptance of a value), “I would like to learn subjects related to physical science” (preference for a value), and “I will invest more time in learning physical science”(commitment). After the process, fifteen items were obtained.

Confirmatory Factory Analysis

According to stereotype threat theory, a three-factor structure was employed to test the components of gender stereotype threat. The fit indices (GFI = .94, AGFI = .92, SRMR = .05, RMSEA = .06, TLI = .94, and CFI = .95) revealed a good model-data fit. In addition, factor loading (ranging from .50 to .82) and composite reliability (ranging from .77 to .87) were all in an acceptable range. Overall, our data supported the three-factor structure of stereotype threat.

Data analysis

Data were analyzed with one-way ANOVA to compare the differences between gender matched groups, and Cohen’s (1977) statistic index *d* was employed to estimate the effect size of experimental treatment.

Result

Science identification of teacher-student gender matched groups

Table 1 indicated the means and standard deviations of students’ science identification for different teacher-student matched groups. One-way ANOVA showed there was a significant difference among the four groups in the science identification ($F_{3,\,636}$=3.96, $p < .01$, η^2=.0183). Further post hoc analysis indicated science identification of male-male matched group was significantly higher than that of male-female matched group (d = 0.35). That is, male students exhibited higher science identification than female students when they were taught by male science teachers.

Table 1: Science Identification of Teacher-student Gender Matched Groups

Gender Matched Groups	*M*	*SD*	*N*
male-male	16.70	4.54	184
male-female	15.16	4.20	177
female-male	15.47	5.04	141
female-female	15.68	4.03	138
Total	15.78	4.49	640

Science gender stereotype of teacher-student gender matched groups

Based on the Gender Stereotype Belief Subscale, the means and standard deviations of students’ science gender stereotype belief for different teacher-student matched groups were shown in Table 2. Analyzed through one-way ANOVA, the four groups showed a significant difference in science gender stereotype belief ($F_{3,\,636}$=2.63, $p < .05$, η^2=.0123). Further post hoc analysis displayed that science gender stereotype belief of male-female matched group was significantly higher than that of female-male matched group (d = 0.31); namely, female students taught by male teachers had higher science gender stereotype belief than male students taught by female teachers.

Table 2: Science Gender Stereotype Belief of Teacher-student Gender Matched Groups

Gender Matched Groups	*M*	*SD*	*n*
male-male	10.80	4.56	184
male-female	11.61	4.88	177
female-male	10.19	4.22	141
female-female	11.18	4.68	138
Total	10.97	4.62	640

Based on Gender Stereotype Situation Subscale, Table 3 showed the means and standard deviations of students’ perceived gender stereotype from teacher-student interaction of different teacher-student gender matched groups. One-way ANOVA showed there was no significant difference of perceived gender stereotype from teacher-student interaction among the four gender matched groups ($F_{3,\,636}$=.95, $p > .05$). That is, no matter male students or female students taught by male teachers or female teachers, students’ perceived gender stereotype from teacher-student interaction was not significantly different.

Table 3: Perception of Gender Stereotype from Teacher-student Interaction of Different Teacher-student Gender Matched Groups

Gender Matched Groups	*M*	*SD*	*n*
male-male	9.33	3.66	184
male-female	8.95	3.70	177
female-male	9.45	3.51	141
female-female	8.86	3.54	138
Total	9.15	3.61	640

Conclusion and Discussion

In this study, we investigated whether students' perception of stereotype threat would be different due to gender matching of teachers and students. The results indicated there were differences among teacher-student gender matched groups. Firstly, male students' exhibited stronger science identification than female students when they were taught by male science teachers. Secondly, the belief of gender stereotype would be higher when female students were taught by male science teachers than when male students were taught by female science teachers. Thirdly, there was no significant difference of perceived gender stereotype from teacher-student interaction among the four gender matched groups.

Study 3: A Predictive Model for Science Achievement: Integrating Expectancy-Value Theory and Achievement Goal Framework

Theoretical Framework

On the basis of the expectancy-value theory and achievement goal framework, the study proposed a hypothetic model to predict students' science achievement, as shown in Figure 1. Specifically, the following relationships were posited: (a) perceptions of science competence will positively relate to science values and will directly impact science achievement. (b) Both perceptions of science competence and science values will positively predict performance-approach and mastery-approach. (c) Performance-approach will positively impact science achievement.

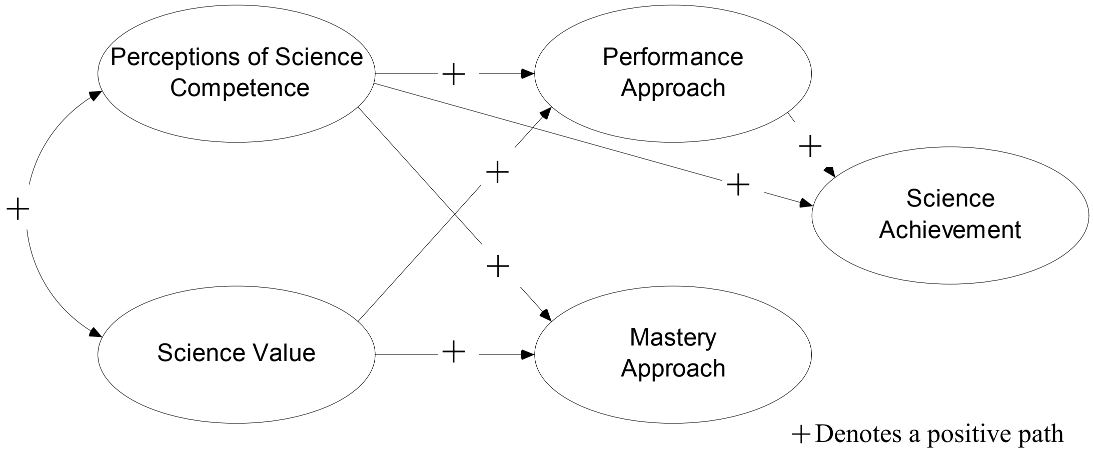


Figure 1. Hypothetic model for predicting students' science achievement

Method

Subjects and Procedures

Subjects in the study were 604 (303 boys and 301 girls) eighth graders selected form 11 middle schools in Taiwan. These students completed the PSC (Cherng, 2006), designed to measure perceptions of science competence and the SV (Simpkins, Davis-Kean, & Eccles, 2006), designed to measure science values in the second week of the second semester of 2007. After the final term exam of the second semester, these students were administered the AG (Elliot & McGregor, 2001), designed to measure performance- approach and mastery-approach and the science achievement test (SAT).

Instruments

Perceptions of Science Competence (PSC): Cherng's (2006) *Subjective Ability Perception Scale* was used to assess students' perceptions of science competence, including 3 sub-scales—"Perception of Difficulty," "Self-efficacy," and "Expectation of Success." Each subscale had 5 items and had Cronbach's alpha of .96, .95, and .96 respectively.

Science Value (SV): Simpkins, Davis-Kean, and Eccles's (2006) *Expectancy-Value Questionnaire* was used to assess students' science values, including 2 subscales— "Interest in Science" and "Perception of Science Importance". Each subscale had 2 items and had Cronbach's alpha of .71 and .84 respectively.

Achievement Goals (AG): Elliot and McGregor's (2001) *Achievement Goal Questionnaire* was used to assess students' achievement goals, including 2 subscales— "Performance- approach Goals" and "Mastery-approach Goals". Each subscale had 3 items and had Cronbach's alpha of .92 and .87 respectively.

Science Achievement Test (SAT): This test was developed by a team of teachers from Compulsory Education Advisory Group of Kaohsiung Municipal Department of Education based on science-related unit themes and cognitive dimensions in Curriculum Guidelines of Science and Technology in Taiwan. There are totally 30 multiple-choice questions with average difficulty .51, average discrimination .48, and reliability .85.

Analysis

There were two phases in data analysis. The first phase of instruments confirmation tested the factor structure of the PSC, SV, and AG using confirmatory factor analysis. The second phase of model examination tested the structure of hypothesized model and structure invariance using multigroup analysis.

Results

Instruments Confirmation

To confirm the factor structure of the PSC, SV, and AG, 6 fit indices were considered: the goodness-of-fit index (GFI), the adjusted GFI (AGFI), the standardized root mean square residual (SRMR), the root mean square error of approximation (RMSEA), the Tucker-Lewis index (TLI), and the comparative fit index (CFI). The GFI, AGFI, TLI, and CFI values greater than .90 suggest an acceptable model fit (Hu & Bentler, 1999; Marsh & Hau, 1996). The SRMR value below .08 and the RMSEA value below .06 indicate a good model fit (Hu & Bentler, 1999). In PSI, consisted of 3 subscales, the fit indices (GFI= .98, AGFI= .96, SRMR= .02, RMSEA= .05, TLI=.99, and CFI= .99) indicated that the three- factor model well fit the data. In SV, consisted of 2 subscales, the fit indices (GFI= .99, AGFI= .98, SRMR= .01, RMSEA= .06, TLI=.99, and CFI= .99) indicated that the two- factor model well fit the data. In AG, consisted of 2 subscales, the fit indices (GFI= .99, AGFI= .97, SRMR= .02, RMSEA= .06, TLI=.99, and CFI= .99) indicated that the two- factor model well fit the data. Additionally, the reliability of the three instruments was established by examining the factor loading, composite reliability and variance extracted. Generally, composite reliability scores of above 0.6, and variance extracted scores above 0.5 are deemed acceptable (Bagozzi & Kimmel, 1995). As Table 1 showed, all reliability measures showed acceptable levels. Overall, these results implied support for the factor structure and reliability of the three instruments.

Table 1 Reliability of three instruments

Instrument	Factor Loading	Composite Reliability	Variance Extracted
Perceptions of Science			
Self-efficacy	.84, .88, .87	.90	.75
Expectation of Success	.91, .91, .83	.92	.78
Perception of Difficulty	.68, .75, .83	.80	.57
Science Values			
Interest in Science	.82, .86	.83	.71
Perception of Science	.77, .84	.79	.65
Achievement Goals			
Performance-approach	.78, .90, .87	.89	.73
Mastery-approach	.82, .83, .83	.87	.68

Note. *N*=604.

Model Examination

We began by testing a single group structure model with all subjects. Results showed that the hypothesized model displayed a good fit to the data (GFI= .94, AGFI= .90, SRMR = .04, RMSEA= .08, TLI=.95, and CFI= .96), excepted that χ^2 statistic was statistically significant (χ^2(47, *N* = 604) = 231.25, p< .000) which could due to sample size sensitivity (Bollen & Long, 1993). All of the hypothesized direct paths were statistically significant showed in Figure 2. As anticipated, perceptions of science competence were positively related to science values and directly impacted science achievement. Both perceptions of science competence and science values positively predicted performance-approach and mastery-approach. In addition, performance-approach positively impacted science achievement.

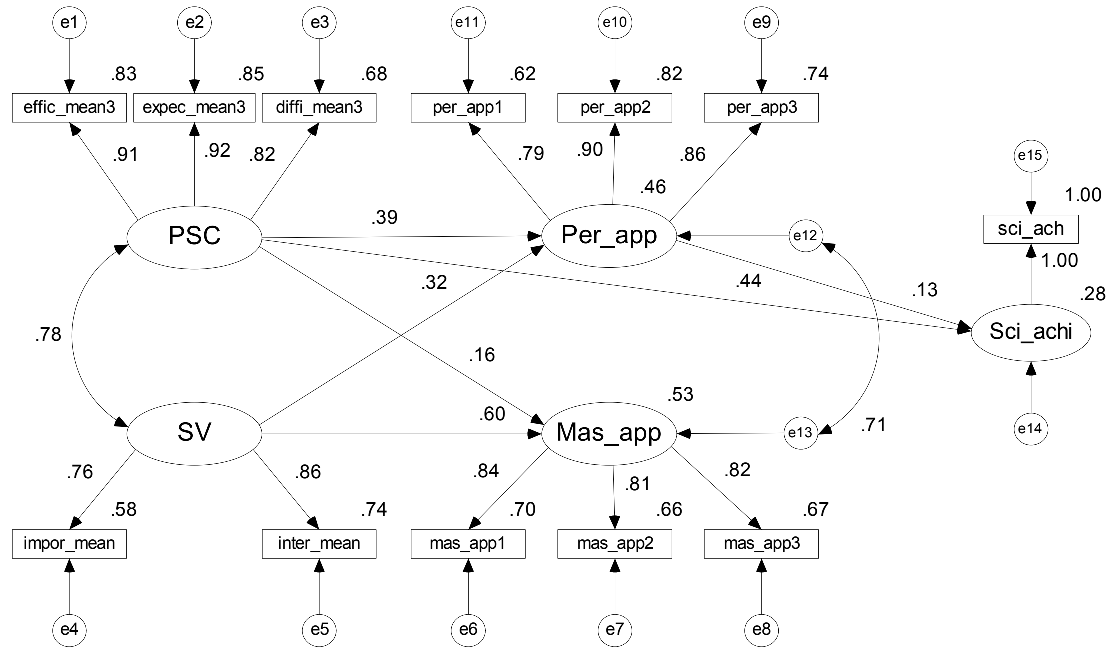


Figure 2. The standardized solution of the Predictive Model for Science Achievement.

Furthermore, multigroup analysis was used to examine whether the structure model was invariant across genders. Five model invariance hypotheses were subsequently formulated for testing these parameter estimates observed in this study. The results of these invariance tests are provided in Table 2. In addition to the χ^2 statistic and RMSEA, the change in the CFI (ΔCFI), the Akaike's information criterion (AIC), and Browne and Cudeck's criterion (BCC) were adopted for model comparison. A ΔCFI value smaller than or equal to -0.01 is indicative of significant drop in model fit (Cheung & Rensvold, 2002). Smaller values of AIC and BCC were better fit than larger values (Arbuckle, 2006). Overall, Hypothesis 4 of equal loadings, structural weights, and structural covariances across genders had the best model-data fit.

Table 2 Goodness-of-fit statistics for test of invariance across genders

Hypothesis Model	χ^2	*df*	$\chi^2/_{df}$	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA	AIC	BCC
H1: Base model	309.24	94	3.29	—	—	.958	—	.062	433.24	438.84
H2: Equal loadings	307.98	97	3.13	7.03	7	.959	.000	.060	426.27	431.23
H3: Equal loadings, structural weights	326.47	105	3.06	17.91	13	.957	-.001	.058	425.15	429.57
H4: Equal loadings, structural weights structural covariances	331.58	108	3.02	23.00	16	.956	-.001	.058	424.24	428.39
H5: Equal loadings, structural weights structural covariances structural residuals	331.58	108	3.00	32.56*	20	.956	-.001	.058	425.80	429.59

Note. Boys: *N*=303; Girls: *N*=301; **p*< .05.

Conclusion

The present study was designed to examine a proposed model for predicting students' science achievement based on expectancy-value theory and achievement goal framework. The results indicated that perceptions of competence were positively related to science values and directly impact science achievement which are consistent with previous research (Bouffard, Marcoux, Vezeau, & Bordeleau, 2003; Harackiewicz, Barron, & Elliot, 1998; Leondari & Gialamas, 2002). In addition, findings indicated that science values both positively predicted performance-approach and mastery-approach which are also consistent with previous researches that students with mastery-approach goals often hope to develop their ability and have greater interest in learning tasks (Elliot & Dweck, 1988; Grant & Dweck, 2003; Harackiewicz, Barron, Tauer, & Elliot, 2002).

四、計畫成果自評

本計畫執行成果已撰寫成四篇論文，其中三篇已發表在國際學術研討會，另一篇也即將投稿至IMPS 2009，未來亦將改寫投稿至SSCI學術期刊。

1. **Cheng, Y. Y.,** Liu, K. S, & Chen, Y. L. (2008, Nov). *Rasch* a*nalysis of item quality in teacher-made science achievement test*. Paper presented at 2008 Asia-Pacific Education Research Association Conference. National Institute of Education, Singapore.
2. Chen, Y. L., **Cheng, Y. Y**., Liu, K. S., & Chang, Y. R. (2008, Dec). *Relationship between teachers' and students' gender and students' stereotype in science*. Paper presented at the World Association of Lesson Studies International Conference 2008. The Hong Kong Institute of Education, Hong Kong.
3. **Cheng, Y. Y.,** Liu, K. S., & Chen, Y. L. (2009, Mar). *A predictive model for science achievement: Integrating expectancy-value theory and achievement goal framework*. Paper will be presented at 2008 AERA Annual Meeting. New York, NY.
4. **Cheng, Y. Y.,** Liu, K. S., Wang, W. C. & Chung, S. H. (to be submitted) *Development of the Gender Stereotype of Science Inventory and the Science Identification Inventory*. Paper will be submitted at The 16th International Meeting of the Psychometric Society. St John's College, Cambridge, July 20-24, 2009.

五、參考文獻

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

Arbuckle, J. L. (2006). *Amos 7.0 user's guide*. Spring House, PA: Amos Development Corporation.

Bagozzi, R. P., & Kimmel, S. K. (1995). A comparison of leading theories for the prediction of goal directed behaviors. *British Journal of Social Psychology, 34*, 437–461.

Bettinger, E. P., & Long, B. T. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review*, *95*, 2, 152-157.

Bollen, K. A., & Long, S. J. (1993). *Testing structural equation models*. Thousand Oaks, CA: Sage.

Bouffard, T., Marcoux, M. F., Vezeau, C., & Bordeleau, L. (2003). Changes in self-perceptions of competence and intrinsic motivation among elementary schoolchildren. *British Journal of Educational Psychology, 73*, 171-186.

Cherng, B. L. (2006). Students' perception of subjective competence and their use of avoidance strategies. *Journal of Taiwan Normal University: Education, 51*(2), 1-24.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Christidou, V. (2006). Greek students' science-related interests and experiences: Gender differences and correlations. *International Journal of Science Education, 20*(10), 1181-1199.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought process. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed. pp. 255-296). New York: Macmillan.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised ed.). New York: Academic Press.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., & Meece, J. L. (1983). Expectancies, values and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75– 146). San Francisco: Freeman.

Elliot, A. J., & McGregor, H. A. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501-519.

Elliot, A. J., McGregor, H., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Experimental Social Psychology, 91,* 549–563.

Elliott, E.S., & Dweck, C.S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*, 5-12.

Grant, H. & Dweck, C.S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology, 85*, 541-553.

Greenfiled, T.A. (1997). Gender-and grade-level differences in science interest and participation. *Journal of Research in Science Teaching*, *81*, 259–276.

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94,* 638–645.

Harackiewicz, J.M., Barron, K.E., & Elliot, A.J. (1998). Rethinking achievement goals: When are they adaptive for college students and why? *Educational Psychologist, 33,* 1-21.

Harackiewicz, J.M., Barron, K.E., Tauer, J. M., & Elliot, A.J. (2002). Predicting success in college: A longitudinal study of achievement goal and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94*(3), 562-575.

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics*, *15*(1), 37-53.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55.

Kahle, J. B. (1990). Real students take chemistry and physics: gender issues. In: K. Tobin, J. B. Kahle & B. J. Fraser (Eds), *Windows into science classrooms*: *Problems associated with higher-level cognitive learning.* New York, Falmer Press.

Kelly, A. (1988). Gender differences in teacher–pupil interactions: A meta-analytic review. *Research in Education*, *39*, 1–23.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *Journal of Experimental Social Psychology, 43*, 825-832.

Kit, K. A., Tuokko, H. A., & Mateer, C. A. (2008). A review of the stereotype threat literature and its application in a neurological population. *Neuropsychology Review, 18*, 132-148.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives handbook II: Affective domain*. New York: David McKay Co.

Leondari, A., & Gialamas, V. (2002). Implicit therries, goal orientations, and perceived competence: Impact on students' achievement behavior. *Psychology in the Schools, 39*(3), 279-291.

Marsh, H. W., & Hau, K. T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64*, 364-390.

Marsh, H. W., Martin, A. J., & Cheng, J. H. S. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology, 100*(1), 78–95.

Martin, A. J., & Marsh, H. W. (2005). Motivating boys and motivating girls: Does teacher gender really make a difference? *Australian Journal of Education, 49,* 320–334.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report.* Chestnut Hill, MA: Boston College.

Meece, J. L., Glienke, B. B., & Burg S. (2006). Gender and motivation. *Journal of School Psychology, 44*, 351-373.

Moller, A.C., & Elliot, A.J. (2006). The 2 x 2 achievement goal framework: An overview of empirical research. In A. Mittel (Ed.), *Focus on educational psychology* (pp. 307-326). NY: Nova Science Publishers, Inc.

Nelson, R. M., & DeBacker, T. K. (2008). Achievement motivation in adolescents: The role of peer climate and best friends. *The Journal of Experimental Education, 76*(2), 170-189.

OECD (2007). *PISA 2006: Science competencies for tomorrow's world: Executive summary*. Paris: OECD.

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.)

Sadker, M. & Sadker, D. (1994). *Failing at fairness: How America's schools cheat girls.* New York, Scribner's.

She, H. C. & Barrow, L. H. (1997). Gifted elementary students' interactions with female and male scientists in a biochemistry enrichment program. *Journal of Elementary Science Education*, *9*, 45–66.

She, H. C. (2000). The interplay of a biology teacher's beliefs, teaching practices and gender-based student–teacher classroom interaction. *Educational Research*, *42*, 28–39.

She, H. C. (2001). Different gender students' participation in the high-and low-achieving middle school questioning-orientated biology classrooms in Taiwan. *Research in Science & Technological Education, 19*(2)*, 147-158.*

Simpkins, S. D., Davis-Kean, P. E., & Eccels, J. S. (2006). Math and Science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology, 42*(1), 70-83.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4-28.

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Steele, C. M. (1997). A threat in the air: How stereotype shape intellectual identity and performance. *American Psychologist, 52*(6), 613-629.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

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

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ConQuest*〔Computer software and manual〕.
Camberwell, Victoria, Australia: Australian Council for Educational Research.

出席國際學術會議心得報告

計畫編號	96-2522-S-110-001-
計畫名稱	科學課堂中的性別刻板印象威脅對女學生科學學習之影響
出國人員姓名 服務機關及職稱	鄭英耀 中山大學教育研究所教授
會議時間地點	National Institute of Education, Singapore.
會議名稱	2008 Asia-Pacific Education Research Association Conference.
發表論文題目	1. Rasch analysis of item quality in teacher-made science achievement test 2. Longitudinal effects of educational expectations and achievement attributions on middle school students' academic achievements

一、參加會議經過

亞太教育研究學會國際研討會是專為教育研究者、實踐者和政策制定者而設的學術盛會，旨在透過亞太的觀點，分享及整合亞太地區及其他國際區域在教育研究、實踐和政策制定上的各種經驗及成果，以促進教育發展。

為了提升本校教育研究所學生的國際視野和學術研究品質，我積極鼓勵與指導研究生將研究論文投稿至此研討會，並於今年6月接獲主辦單位審核通過，同意在該研討會發表論文，與國際學者分享研究成果。

今年（2008 年）的亞太教育研究學會國際研討會，於 11 月 26 日至 11 月 28 日假新加坡國立教育學院舉辦，約計有 1200 位，來自 39 個國家的教育工作者與會。26 日上午 9 點開幕式後，首先由英國教育學者 Julian Elliott 發表演說：Western Influences on the East, Eastern Influences on the West: Lessons for the East and West。緊接著由聯合國教科文組織（UNESCO）的教育公共福利聯盟（Commonwealth Consortium for Education）的主席 Colin Power 發表演說：Asia and the Pacific in 2020: Scenarios for Education Research。兩場演說之後，即開始進行各場次的論文發表，3 天的研討會，總計有 8 場的 Keynote Addresses 和 654 篇論文發表。會議期間，我除了發表研究論文，和與會學者經驗交流外，亦聆聽多場其他學者的論文發表以及座談會。我和研究生的兩篇論文發表時間安排在 26 日下午，主題分別為：

(1) "Analysis of Item Quality in Teacher-made Science Achievement Test"，此篇論文係以 Rasch Model 分析學校教師自編的科學成就測驗，並從學生實際的作答反應，展示各種有用的教學回饋訊息，以作為教師教學與評量反省之參考。

(2) "Longitudinal effects of educational expectations and achievement attributions on adolescents' academic achievements"，此篇論文係以「台灣教育長期追蹤資料庫」（Taiwan Education Panel Survey, TEPS）的實徵資料，運用階層線性模式(hierarchical linear modeling, HLM)縱

貫資料分析方法，檢驗教育期望與成就歸因對中學生學業成就成長模式之影響。

二、與會心得

（1）補助與支持進行跨文化議題的教育研究

Julian Elliott於Keynote Address中，比較了東西方的教育觀點，以他過去對俄國的研究與長期的觀察，發現俄國的家長和老師對學生的要求較高，有低估學生能力的傾向；而以英美為主的西方國家的家長和教師，卻有高估學生能力的傾向。自主自動的學習，導致西方教育多年來呈現平庸(mediocrity)的教育。他指出未來東方的教育應向西方學習"創新、勇於冒險、面對挑戰、解決問題"；西方的教育應向東方學習"鼓勵學生努力投入學習、教師應對學生有高期望、高要求、家長必須重視孩子的教育"等觀點，開啟個人學術視野與觸角。

就台灣而言，我們的家長和學校老師和俄國是相接近的，對學生普遍存有較高的期望和高要求，台灣未來的教育如何在創新自主和勤勉精進之間取得平衡，值得思考。此外，各種教育理論是否適用於不同文化？其共同點與差異點為何？這些有關文化議題的教育研究，應是教育部或國科會近期可列入重點支持的研究議題。

（2）鼓勵與支持研究生參與國際學術研討會

藉由參與國際學術研討會，研究生除了可認識世界各國的學者、聆聽當前各國最新的研究議題外，經由親身以英文發表研究成果，並與國外學者討論的經驗，確實有助於提升台灣研究生的學術熱忱與國際觀，對於台灣研究人才的培育有莫大的助益。建議教育部和本校能增加補助研究生出國參加研討會的經費，鼓勵與支持研究生參與國際學術研討會，增廣國際視野。

（3）挹注研究經費支持教育和教學研究

新加坡教育部高級政務部長傅海燕於本次研討會宣布，新加坡政府今年將補助近 1 億新幣（約新台幣 22 億元）給新加坡國立教育學院的"教學法與實踐研究中心"，作為未來五年的教育研究經費，著重於科技融入教學、創新教學和種子教學計畫。建議教育部可參考新加坡模式，增加補助教育研究的經費，一方面可作為教育政策推動的研究基礎，他方面可提升台灣教育研究的水準和學術研究的國際競爭力。

（4）成立台灣教育研究協會接軌國際學術社群

此次亞太教育研究協會年度會議，國際上 30 個教育研究組織的代表，包括 American Educational Research Association (AERA)、Asia-Pacific Educational Research Association (APERA)、British Educational Research Association (BERA)、European Educational Research Association (EERA)、Australian Association for Research in Education (AARE)、Korean Educational Research Association (KERA)、Japanese Educational Research Association (JERA)、Educational Research Association of Singapore (ERAS)…等也達成共識，正籌劃成立世界教育研究協會（World Educational Research Association, WERA），以聯繫各國教育研究組織，增進教育研究的國際交流網絡。建議台灣也應該成立 Taiwan Educational Research Association (TERA)，成為 WERA 的一員，與國際教育學術研究接軌。

Rasch Analysis of Item Quality in Teacher-made Science Achievement Test

Ying-Yao Cheng   Kun-Shia Liu   Yi-Ling Chen
National Sun Yat-sen University, Kaohsiung, Taiwan

Abstract

Since 1994's 410 Education Reform Movement, how to advance the quality of school education has become the core value of Taiwan education reform. Teaching assessment is much related to quality of school teaching. This study used Rasch model to analyze item quality of teacher-made achievement test to provide information of students' real learning behavior for teachers in reflecting on their teaching and assessment. 489 Taiwan subjects in grades 9 (251 males and 238 females) completed a teacher-made science achievement test. The results indicated (1) there were differences between teachers' expectation and students' responses in the difficulty of test items; (2) some test items showed gender differential functioning. The study further suggests that compared with classical test theory, Rash model is a more effective way for schools to offer teachers more abundant teaching feedback for enhancing teaching quality.

Keywords: classroom assessment, diagnostic assessment, Rasch measurement

Introduction

Classroom assessment is closely related to quality of teaching and learning. Every component in instructional system, including goal-setting, diagnosis of entering behavior, and adoption of instructional procedure, is closely connected to classroom assessment. Through classroom assessment, teachers can know whether the expected goals are achieved, decide competency level for students as their entering point, and then adjust their teaching progress and methods. Hence, a good classroom assessment not only provide feedback information for teachers, but also can integrate the whole teaching process together to exert the best teaching and learning effects. In traditional assessments, teachers often conduct the procedure of assessing students' learning through teacher-made tests which are mainly based on teaching contents, using the total scores of a test to interpret learning effects, and reviewing tests one item by one item. This kind of assessments provides quite limited information for improving learning. Compared with traditional assessment, Rasch model (Rasch, 1960) of item response theory (IRT) has the advantages of providing objective and interval scales, and presenting abundant useful reference information for learning diagnosis and teaching improvement. Based on these, the study aimed to use Rasch model to analyze a teacher-made science achievement test and to offer teachers various useful feedback information from students' authentic response for reflecting on their teaching and assessment.

Rasch Model

Rasch model (Rasch, 1960) is an assessment model proposed by George Rasch, a Danish mathematician, with the aim to obtain an objective and interval scale from subjects' responses. Rasch contended that every subject's response on each test item can be modeled through two parameters — the ability of the person tested and the difficulty of test items. In Rasch model, what influences the response of a person *n* on item *i* can be divided into two parameters, the ability of the person tested θ_n and the difficulty of test item δ_i. The formula is in the following:

$$\log(P_{ni1}/P_{ni0}) = \theta_n - \delta_i \qquad (1)$$

where P_{ni1} and P_{ni0} denote the probability of scoring 1 point (correct response) and 0 point (incorrect response) for person *n* respectively. From formula (1), person *n*'s probability of correct response for test item *i* as:

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

Because θ and δ belong to the same measurement unit, when θ_n is greater than δ_i, person *n*'s probability of correct response for item *i* will be greater than 0.5; when θ_n is smaller than δ_i, the probability will be smaller than 0.5; when θ_n is equal to δ_i, the probability will be equal to 0.5. As the item difficulty is fixed, the probability of correct response increases with a person's ability. When ability approximates infinity, the probability of giving correct response approximates 1; when ability approximates infinitesimal, the probability of correct response approximates 0. Due to the objective and interval attributes of Rasch model (Wang, 2004a), misfit subjects or items can be picked out through examining model-data fit to provide reference information for ability diagnosis and item modification.

Method

Subjects and Procedure

Subjects in the study were 489 ninth graders (251 males and 238 females) selected form 5 middle schools (2 classes per school) in Taiwan. A science achievement test was conducted two weeks after school's secondly exam in the second semester of 2007 school year. Participants completed all the test items.

Instrument

The science achievement test was compiled by a group of middle school science teachers from Kaohsiung city. The development of the test items was based on taxonomy of educational objectives of cognitive domain and the attributes of science literacy listed in the Grade 1-9 Science and Technology Curriculum Guidelines. Proportion of all kinds of test items and proportion of item difficulty were also considered. There were totally 50 multiple-choice questions related to four attributes, including "The Nature of Science," "The Development of Science and Technology Knowledge," "Science Process Skills," and "The Development of Processing Intelligence." Table 1 indicates the checklist for each test item on educational objectives of cognitive domain and the attributes of science literacy. Table 2 indicates the items difficulties what teachers considered on each educational objective of cognitive domain.

[Insert Table 1 and 2 here]

Analysis

The study employed Rasch model (Rasch, 1960) to analyze students' answering response, using ConQuest (Wu, Adams, & Wilson, 2007) and ConstructMap (Kennedy, Wilson, & Draney, 2008) as the analyzing software. The criteria of model-data fit adopted 0.6~1.4 Infit MNSQ (Wright & Linacre, 1994). Equal-mean-difficulty method was used for differential item functioning analysis (DIF). Cohen's (1977) effect size statistic was adopted for analyzing the two-group students' scoring differences with the criteria 0.5 (medium), and 0.8 (high).

Results

DIF Analysis

Table 3 shows the estimates and fit statistic for overall item difficulties and the difference of item difficulties between males and females. All the 50 items had an Infit MNSQ with the critical range (0.6, 1.4), indicating that they all had a reasonably good fit with each group and that the item parameter estimates could be directly compared over groups for evidence of DIF. "Male-Female Difficulty" column in Table 3 refers to a value deducting the item difficulty for females from item difficulty for males. If the value is positive, the item is more difficult for males than for females; if it is negative, the item is more difficult for females than for males. Even though some of the 50 items difficulties were found significantly different between genders, most of the differences were smaller than 0.5 logits. Based on Cohen's (1977) standard for effect size, the ratio of the values of Male-Female Difficulty to the standard deviation of all students' ability (*SD* = 1.28 logits) could be regarded as effect size of DIF between genders. When the Male-Female Difficulty reaches 0.5 time (0.64 logits), and 0.8 time (1.02 logits) of the *SD*, the effect size of DIF between genders are medium, and high respectively. Therefore, there were at least five items exhibited medium DIF between genders. Item 2, Item 20, and Item 40 are more difficult for females than males, whereas Item 36, and Item 37 are more difficult for males than females. Take Item 40 for example. After comparing males' and females' expected scores on Item 40, based on same ability level of gender groups, we found that the expected scores for males are higher than those for females, indicating that Item 40 is advantageous to males(see Figure 1). Taking a look at Item 40 (see Figure 2), it showed that this item needs the number change of chromosomes in Diagram 2 to judge which part of the structure undergoes the task. Compared with male students, Taiwan female students of junior high schools tend to have mental barriers when looking at the picture of male genitals (Diagram 1). This may disturb them when answering this item, so they are less likely to perform well on this item.

[Insert Table 3, Figure 1 and 2 here]

Wright Map

As far as the whole test items are concerned, the distribution of the difficulty of all test items fits the distribution of students' ability (Figure 3). The test can effectively discriminate students with middle level ability. However, Figure 3 indicates that teachers' understandings of item difficulty are somewhat different from students' real responses. What was considered difficult by teachers (for example, I20, I36, I41 and I46) was revealed to be easy in the result of the real data analysis, whereas what was considered easy by teachers (for example, I3, I10, I27, I33, and I47) was regarded as difficult in the result.

[Insert Figure 3 here]

Diagnostic Map

Figure 4 showed the Diagnostic Map (the Kidmap) for Student 130 (ability = -0.02 logit, Infit MNSQ = 1.05). In the Diagnostic Map option in the ConstructMap software, the item difficulties are displayed vertically in logit scale, with the easiest items at the bottom of the map and the most difficult item at the top. The items located on the left side of the map are those on which this particular student was successful, and the items on the right side of the map are those that the student did not complete successfully. The horizontal "XXX" in the center column indicates the respondent's location, or proficiency level, on the same vertical scale. The two horizontal dashed lines are located one standard deviation above and below the respondent's location to delineate the range of expected responses. When a person is located near an item's difficulty, that person has an approximately 50% chance of success. When the person is above the item's difficulty, the chance of success is greater than 50%, while when the person is below the item's difficulty, the chance of success will be less than 50%.

The Diagnostic Map is divided into 4 quadrants. The bottom left area indicates an area in which the item difficulty is lower than the student's ability and the student answered it correctly, namely, "Easier Achieved." The top right denotes the item difficulty is higher than the student's ability and the student answered it incorrectly, namely, "Harder Not Achieved." Hence, these are areas of fit, where the person's responses match the Rasch model's expectations. The top left area indicates that the item difficult is higher than the student's ability; whereas, the student answered it correctly, namely, "Harder Achieved." The bottom right denotes the item difficulty is lower than student's ability; however, the student didn't answer it correctly, namely, "Easier Not Achieved." Hence, these are areas of misfit.

Take the diagnostic map of student 130 for example. There are five "Easier Not Achieved" items (I14, I16, I34, I38, and I46) that the student was expected (probability greater than 0.5) to have achieved given his ability estimate. These are surprising responses. A teacher should take a closer look at why that occurred, and whether other students exhibited similar patterns on that item. There are seven "Harder Achieved" items (I4, I9, I10, I27, I45, I47, and I48) that the student was not expected (probability less than 0.5) to have achieved. At this time, a teacher should further understand whether the student had the required knowledge for answering these questions or he did them correctly because of other reasons (for example, reading the questions before the test and memorizing the answers, just lucky guesses, or cheating, etc.) to make sure the student learns the required knowledge of these questions. As to the "Harder Not Achieved" items, a teacher should teach these items in the future to enhance student' ability.

[Insert Figure 4 here]

Conclusion and Discussion

The study aimed to use Rasch model to analyze the quality of a teacher-made science achievement test and offered various important feedback figures for teachers to reflect on their teaching and assessment. It was found that there was some mismatch between teachers' expectations of item difficulty and the item difficulties obtained from real data analysis. The teachers considered that memory-required items were simple, but students' real responses didn't follow the teachers' expectation. Therefore, when teaching the memory-related knowledge, teachers should still pay attention to students' comprehension to help students store these concepts in their long-term memory. Furthermore, the 4 quadrants on the diagnostic map of Rasch analysis diagnosed students' various responses and offered teachers reference for conducting make-up teaching and advanced teaching.

In addition, from the analysis of these items, we found several items showed gender differential item functioning. Some items were beneficial to males and some were beneficial to females. This study also used examples to explain possible reasons for gender DIF. Further, the analyzing software used in this study was equal-mean-difficulty method of ConQuest; its presumption was that the average difficulty of male beneficial items would be equal to that of female beneficial items. Therefore, the result would appear that some items would be beneficial to males and some would be advantageous to females. However, if the real situation doesn't follow the presumption, estimation bias will occur. Anchor item methods (Shih & Wang, in press, Wang, 2004b) were suggested for obtaining more accurate estimates.

Besides, every component in instructional system, including goal-setting, diagnosis of entering behavior, and adoption of instructional methods, is closely connected to classroom assessment. A good assessment system not only helps teachers to know whether students achieve the teaching goals, but also assists teachers in diagnosing students' learning problems and analyzing the quality of test items. Recently, the Berkeley Evaluation and Assessment Research (BEAR) Center has been involved in the development of an assessment system, which call the BEAR Assessment System (Wilson, 2008; Wilson & Carstensen, 2005). The system consists of four principles, each associated with a practical "building block" including the construct map, the items design, the outcome space, and the measurement model (Wilson, 2005) as well as an integrative activity that can take on different aspects under different circumstances. We suggest teachers adopt the procedure of BEAR Assessment System in real classroom assessment in the future and develop instruments that combine assessment and diagnosis to enhance teaching and learning quality.

References

Cohen, J. (1977). *Statistical power analysis for the behavioral science* (revised ed.). New York: Acadimic Press.

Kenney, C. A., Wilson, M., & Draney, K. (2008). *ConstructMap* (computer program). BEAR Center: UC Berkeley, CA.

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.)

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

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

Wang, W.-C. (2004b). 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.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum.

Wilson, M. (2008). Cognitive diagnosis using item response models. *Journal of Psychology, 216*(2), 74-88.

Wilson, M., & Carstensen, C. (2005). Assessment to improve learning in mathematics: The BEAR assessment system. *Journal of Educational Research and Development, 1*(3), 27-50.

Wright, B. D., & Linacre, J. M. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8*: 2, 350.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ConQuest*〔Computer software and manual〕. Camberwell, Victoria, Australia: Australian Council for Educational Research.

Table 1. *Checklist of Biology Test Items on Unit Theme and Cognitive Dimensions*

Attributes of Science Literacy	Educational Objectives of Cognitive Domain			
	Knowledge	Comprehension	Application	Analysis
The Nature of Science	7	9, 12, 13, 34, 44	46	15, 48
The Development of Science and Technology Knowledge	47	11, 19, 21, 22, 23, 24, 25, 28, 31, 33, 37, 38, 50	2, 18, 27, 30,41	26, 39
Science Process Skills	45, 49	16, 29, 36	32, 35	17, 40, 42, 43
The Development of Processing Intelligence	1	3, 8, 10	6, 14, 20	4, 5

Table 2. *Items Difficulties What Teachers Considered on Each Unit Theme.*

Attributes of Science Literacy	Item Difficulty		
	Easy	Medium	Hard
The Nature of Science	7, 15	9, 12, 13, 34, 44	46, 48
The Development of Science and Technology Knowledge	2, 11, 19, 26, 27, 33, 47	22, 24, 25, 28, 30, 31, 37, 38	18, 21, 23, 39, 41, 50
Science Process Skills		16, 17, 29, 32, 40, 43, 45, 49	35, 36, 42
The Development of Processing Intelligence	3, 10	1, 5, 6, 8, 14	4, 20

Table 3 *Estimates and Fit for Overall Item Difficulties and the Difference of Item Difficulties between Male and Female*

Item	Overall Difficulty	Male-Female	Infit MNSQ
I1	-0.68	0.028	0.89
I2	-0.65	**-0.72**	1.31
I3	0.39	-0.47	1.08
I4	0.96	-0.19	1.24
I5	-0.13	0.23	0.75
I6	0.05	-0.01	0.82
I7	-2.38	0.25	0.92
I8	-0.77	0.21	0.97
I9	1.58	-0.49	1.18
I10	0.33	0.39	1.03
I11	-0.11	-0.05	1.02
I12	1.47	0.07	0.99
I13	-0.06	-0.14	0.85
I14	-0.39	0.01	0.98
I15	-0.9	0.29	0.81
I16	-0.81	0.14	0.92
I17	-0.97	-0.03	0.84
I18	0.7	-0.25	0.93
I19	-1.11	-0.07	0.97
I20	-0.55	**-0.78**	0.97
I21	0.21	-0.20	1.01
I22	1.05	-0.51	1.11
I23	0.17	-0.11	1.12
I24	0.46	0.10	1.04
I25	-0.02	0.09	1.02
I26	-0.12	0.41	0.82
I27	1.05	0.04	1.26
I28	1.15	-0.42	1.2
I29	0.41	-0.29	0.83
I30	0.41	0.19	1.17
I31	1.02	-0.07	1.32
I32	0.26	-0.38	0.88
I33	1.29	-0.38	0.99
I34	-1.47	0.19	0.85
I35	0.56	-0.05	1.03
I36	-1.7	**1.05**	0.87
I37	-1.55	**0.76**	0.91
I38	-0.45	0.23	1.05
I39	1.05	-0.28	1.01
I40	-0.02	**-0.85**	0.92
I41	-0.34	0.15	0.93
I42	0.19	0.19	0.82
I43	-1.07	0.07	0.95
I44	1.23	0.02	0.99
I45	0.43	0.63	0.97
I46	-0.56	-0.20	0.97
I47	1.08	0.27	1.34
I48	0.27	0.26	1.18
I49	-1.14	0.50	0.89
I50	0.19	0.15	0.94

Note：Boldface value exhibited DIF between genders. I2, I20, I40 are more difficult for females than males, whereas I36, I37 are more difficult for males than females.

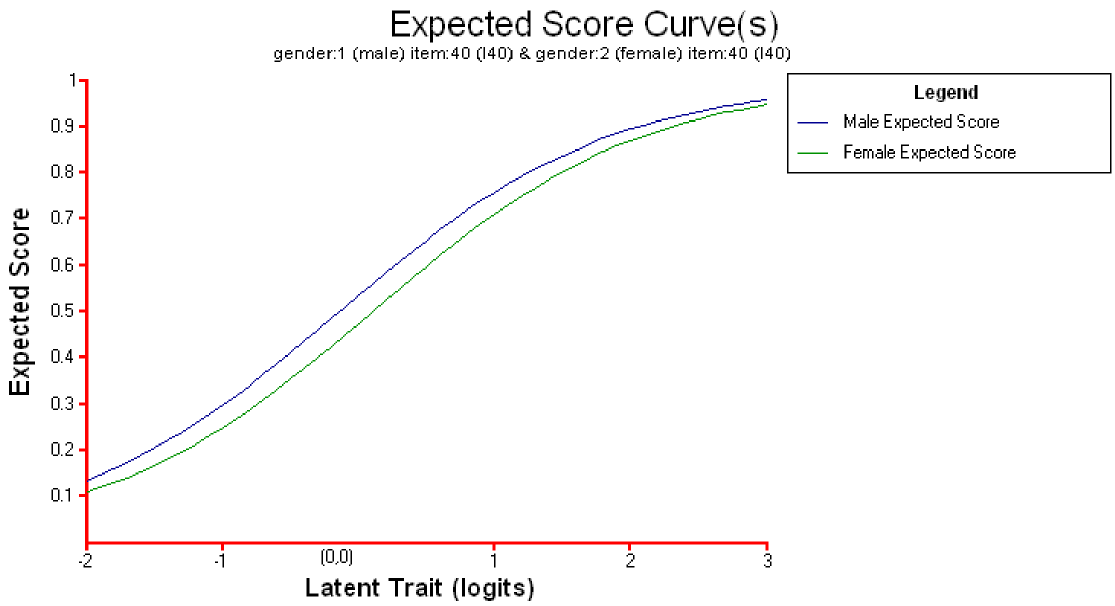


Figure 1. The Expected Scores for Males and Females on Item 40

I40: Diagram 1 is a picture of male genitals and Diagram 2 is the changing of chromosome numbers during the process of cell division. Which part of Diagram 1 can undergo the process of division in Diagram 2?

(A)A *(B)B* *(C)C* *(D)D*

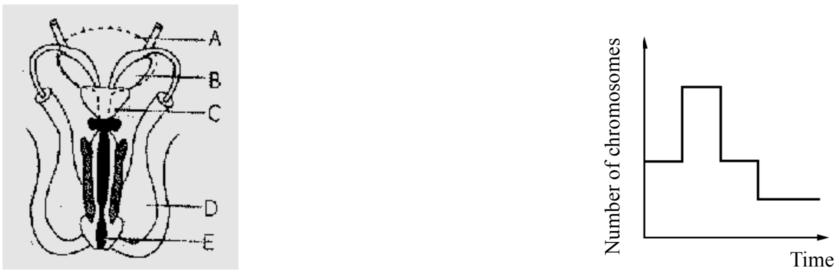
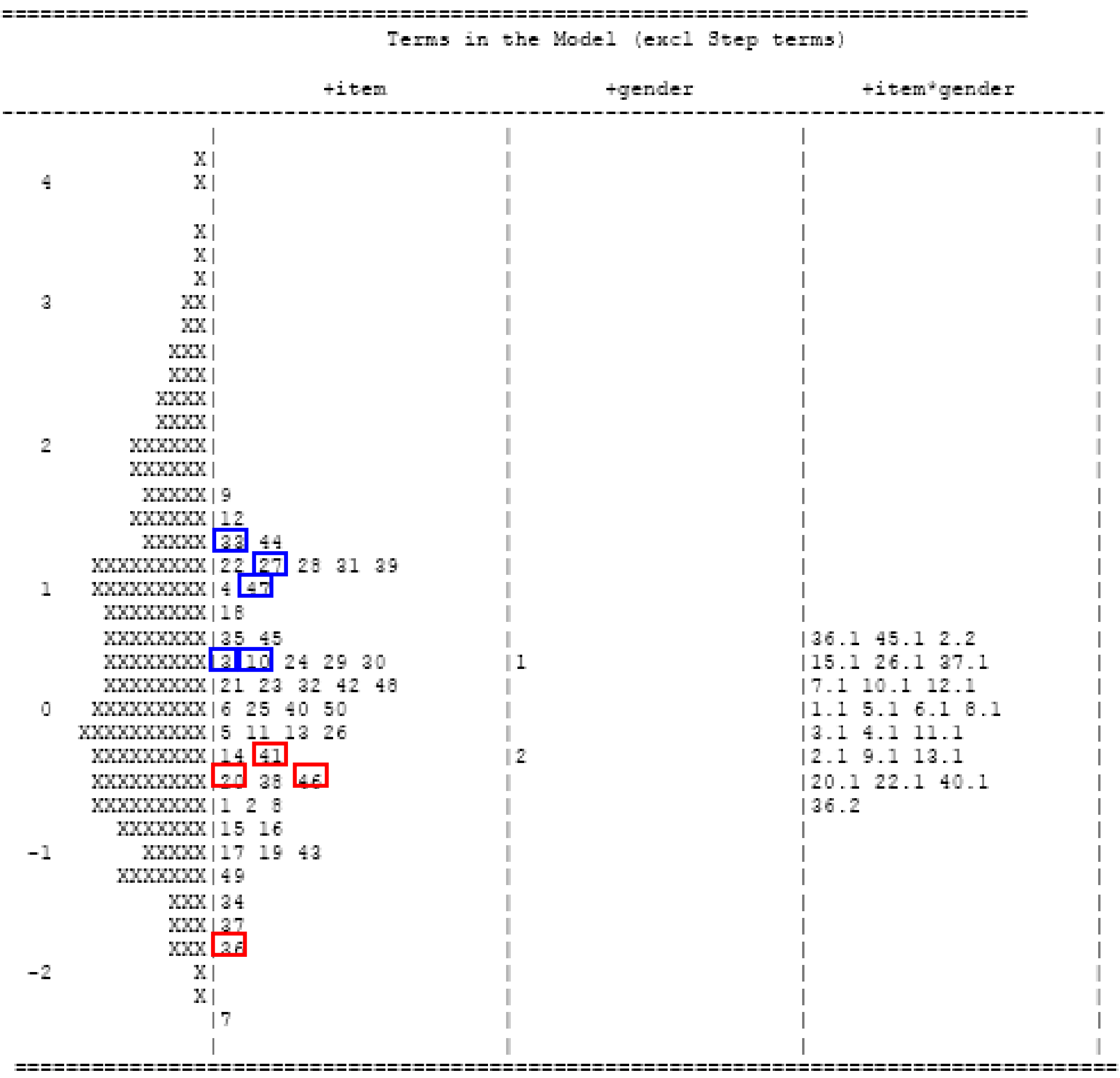


Diagram 1

Diagram 2

Figure 2. The content on Item 40

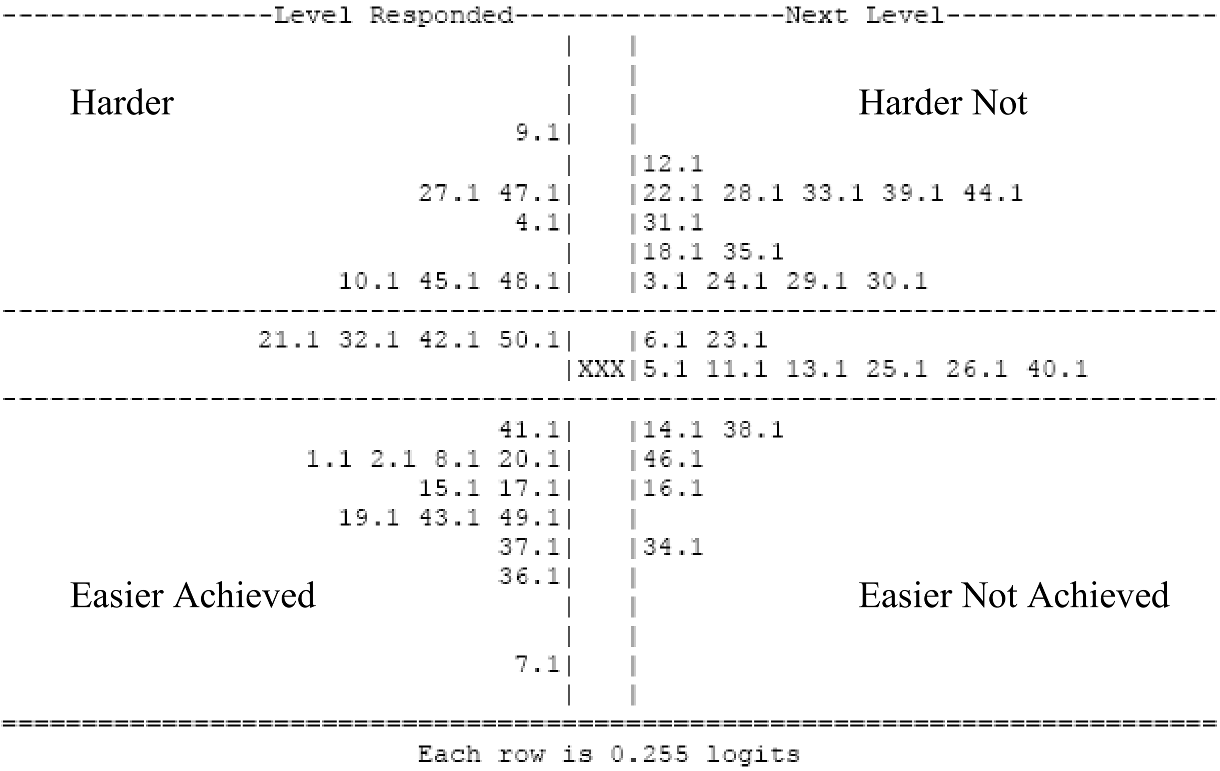


```
================================================================================
                              Terms in the Model (excl Step terms)

                         +item                 +gender             +item*gender
--------------------------------------------------------------------------------------
                |                      |                      |                      |
               X|                      |                      |                      |
   4           X|                      |                      |                      |
                |                      |                      |                      |
               X|                      |                      |                      |
               X|                      |                      |                      |
               X|                      |                      |                      |
   3          XX|                      |                      |                      |
              XX|                      |                      |                      |
             XXX|                      |                      |                      |
             XXX|                      |                      |                      |
            XXXX|                      |                      |                      |
            XXXX|                      |                      |                      |
   2      XXXXXX|                      |                      |                      |
          XXXXXX|                      |                      |                      |
           XXXXX|9                     |                      |                      |
          XXXXXX|12                    |                      |                      |
           XXXXX|33 44                 |                      |                      |
       XXXXXXXXX|22 27 28 31 39        |                      |                      |
   1   XXXXXXXXX|4 47                  |                      |                      |
        XXXXXXXX|18                    |                      |                      |
        XXXXXXXX|35 45                 |                      |36.1 45.1 2.2         |
        XXXXXXXX|3 10 24 29 30         |1                     |15.1 26.1 37.1        |
        XXXXXXXX|21 23 32 42 48        |                      |7.1 10.1 12.1         |
   0   XXXXXXXXX|6 25 40 50            |                      |1.1 5.1 6.1 8.1       |
      XXXXXXXXXX|5 11 13 26            |                      |3.1 4.1 11.1          |
       XXXXXXXXX|14 41                 |2                     |2.1 9.1 13.1          |
       XXXXXXXXX|20 38 46              |                      |20.1 22.1 40.1        |
       XXXXXXXXX|1 2 8                 |                      |36.2                  |
         XXXXXXX|15 16                 |                      |                      |
  -1       XXXXX|17 19 43              |                      |                      |
         XXXXXXX|49                    |                      |                      |
             XXX|34                    |                      |                      |
             XXX|37                    |                      |                      |
             XXX|36                    |                      |                      |
  -2           X|                      |                      |                      |
               X|                      |                      |                      |
                |7                     |                      |                      |
                |                      |                      |                      |
 ======================================================================================
```

Note. Red: originally considered “difficult items,” including 20, 36, 41, and 46；Blue: originally considered “easy items,” including 3, 10, 27, 33, and 47

Figure 3. The Distribution of Students’ Abilities and Item Difficulties



```
-----------------Level Responded-----------------Next Level-----------------
                                   |   |
                                   |   |
    Harder                         |   |                  Harder Not
                                9.1|   |
                                   |   |12.1
                          27.1 47.1|   |22.1 28.1 33.1 39.1 44.1
                                4.1|   |31.1
                                   |   |18.1 35.1
                     10.1 45.1 48.1|   |3.1 24.1 29.1 30.1
----------------------------------------------------------------------------
                21.1 32.1 42.1 50.1|   |6.1 23.1
                                   |XXX|5.1 11.1 13.1 25.1 26.1 40.1
----------------------------------------------------------------------------
                               41.1|   |14.1 38.1
                   1.1 2.1 8.1 20.1|   |46.1
                          15.1 17.1|   |16.1
                     19.1 43.1 49.1|   |
                               37.1|   |34.1
                               36.1|   |
    Easier Achieved                |   |                  Easier Not Achieved
                                   |   |
                                7.1|   |
                                   |   |
============================================================================
                          Each row is 0.255 logits
```

Figure 4. The Diagnostic Map for Student 130