บทที่ 2

วรรณคดีที่เกี่ยวข้อง

การศึกษาการใช้จีอาร์เอ็ม และ จีพีซีเอ็ม ในการเปรียบเทียบฟังก์ชันสารสนเทศของมาตรประมาณค่าที่มีวิธีให้คะแนนที่แตกต่างกัน ผู้วิจัยได้นำเสนอเนื้อหาที่เกี่ยวข้องเป็น 5 ตอน ดังนี้

ตอนที่ 1 ความรู้ทั่วไปเกี่ยวกับแบบวัดและแบบสอบผลสัมฤทธิ์

ตอนที่ 2 การตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาค

ตอนที่ 3 มโนทัศน์เกี่ยวกับทฤษฎีการตอบสนองข้อสอบ

ตอนที่ 4 มโนทัศน์เกี่ยวกับทฤษฎี Polytomous Item Response

ตอนที่ 5 งานวิจัยที่เกี่ยวข้อง

ตอนที่ 6 โปรแกรมคอมพิวเตอร์ที่ใช้วิเคราะห์มาตรประมาณค่า

ตอนที่ 1 ความรู้ทั่วไปเกี่ยวกับแบบวัดและแบบสอบผลสัมฤทธิ์

เครื่องมือวัดผลทางการศึกษาแบ่งเป็น 2 ประเภท ที่สำคัญ คือ แบบวัดและแบบสอบ ผู้วิจัยได้เสนอความรู้ทั่วไปเกี่ยวกับแบบวัดที่เป็นประเภทมาตรประมาณค่า และแบบสอบผลสัมฤทธิ์ดังนี้

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

1.1 มาตรประมาณค่า (Rating Scale)

มาตรประมาณค่าเป็นเครื่องมือวัดผลทางการศึกษาประเภทหนึ่งประกอบด้วยข้อคำถามหรือสิ่งเร้ากับตัวเลือกที่ให้ผู้ตอบตอบสนองเพื่อที่จะใช้ประมาณค่าคุณลักษณะสิ่งใดสิ่งหนึ่งของบุคคลหรือสิ่งของ (Wiersma and Jurs, 1990) โดยในแต่ละข้อกระทงจะกำหนดตัวเลขหรือตัวอักษรเป็นช่วงตั้งแต่ 2 ช่วงขึ้นไปเพื่อใช้เป็นมาตรวัด เครื่องมือประเภทนี้จะเหมาะกับการวัดหรือประเมินคุณลักษณะที่เล็ก แคบ เฉพาะ เดี่ยว

มีความต่อเนื่องในทางบวกหรือทางลบก็ได้ (อุทุมพร จามรมาน, 2537) ซึ่งโดยทั่วไปสามารถแบ่งมาตรประมาณค่าออกได้เป็น 5 ประเภท (Guilford , 1954) ที่สำคัญคือ

ก. มาตรประมาณค่าตัวเลข (numerical rating scale) เป็นมาตรประมาณค่าที่ให้ผู้สังเกตเป็นผู้กำหนดตัวเลขที่เรียงลำดับกันให้สิ่งที่สังเกต เช่น :

ไม่พอใจ	พอใจเล็กน้อย	พอใจ	พอใจมากที่สุด
0	1	2	3

ข. มาตรประมาณค่าแบบกราฟ (graphic rating scale) เป็นเครื่องมือที่ใช้กันมาก ซึ่งแต่เดิมมาตรประมาณค่าประเภทนี้สร้างเพื่อใช้สำหรับการประเมินตนเอง เช่น :

ในการสนทนาทางสังคม	--- ---	--- ---	--- ---	--- ---	--- ---
ท่านมีพฤติกรรมอย่างไร	ช่างพูด	ชอบคุย	พูดเมื่อจำเป็น	ชอบเป็นผู้ฟัง	เงียบ

ค. มาตรประมาณค่าแบบสเกลมาตรฐาน (standard scale) มาตรประมาณค่าประเภทนี้กำหนดเกณฑ์มาตรฐานไว้เปรียบเทียบ เช่น การวัดความเป็นผู้นำจะใช้ชื่อบุคคลที่สำคัญเป็นเกณฑ์มาตรฐานแล้วพิจารณาว่าคนที่ถูกวัดมีความเป็นผู้นำมากหรือน้อยกว่าเกณฑ์ดังกล่าว ตัวอย่างมาตรประมาณค่าแบบนี้ คือ

(1) Man-to-Man Scale เป็นวิธีการที่กำหนดเกณฑ์มาตรฐานโดยการเรียงลำดับคะแนนจากการประเมินของกลุ่มบุคคลที่เกี่ยวข้องกับการวัดคุณลักษณะนั้น เช่น ให้กลุ่มนายทหารเขียนชื่อหัวหน้าระดับสูงที่ตนคิดว่ามีความเป็นผู้นำมากที่สุดมา 12 - 25 ชื่อ แล้วนำคะแนนของหัวหน้าระดับสูงมาเรียงลำดับประมาณ 5 ลำดับ และเมื่อต้องการจะวัดความเป็นผู้นำกับใครก็ให้เปรียบเทียบกับบุคคลทั้ง 5 ดังกล่าว

(2) Portrait-Matching-Scale เป็นมาตรประมาณค่าที่ใช้ศึกษากับลักษณะคน โดยให้ผู้เชี่ยวชาญประมาณ 7 คน จัดกลุ่มข้อความเกี่ยวกับคุณลักษณะที่จะวัดจากคนจำนวนมากไว้ประมาณ 10 กลุ่มข้อความ จากนั้นก็ให้ผู้ตัดสินประมาณ 48 คน ตัดสินและจัดลำดับของข้อความทั้ง 10 เพื่อหาค่าเฉลี่ยของแต่ละคุณลักษณะหรือข้อความเพื่อใช้เป็นเกณฑ์ปกติต่อไป

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ง. มาตรประมาณค่าแบบแต้มสะสม (rating by cumulated points) เป็นมาตรประมาณค่าที่ผู้ตอบเป็นผู้เลือกคุณลักษณะจากตัวเลือกหลาย ๆ ตัว และสามารถประเมินค่าได้ว่าผู้ตอบคุณลักษณะอะไรบ้างหรือส่วนใหญ่มีคุณลักษณะอย่างไร เครื่องมือประเภทนี้สามารถพบได้ 2 ลักษณะ คือ

(1) แบบ Check-List Method เป็นการประเมินคุณลักษณะโดยผู้ประเมินทำเครื่องหมาย ✓ ตรงกับคำที่แสดงคุณลักษณะแต่ละข้อในมาตรประมาณค่า

(2) แบบ The Guess Who-Technique เป็นวิธีการประเมินคุณลักษณะนักเรียนโดยใช้คำถามสั้น ๆ แล้วให้เพื่อน ๆ เขียนชื่อ ตัวอย่าง เช่น :

ใครเอ่ย

(ก) เป็นคนที่มักแกล้งเพื่อนบ่อย ๆ .....

(ข) เป็นคนที่ชอบช่วยเหลือเพื่อน ๆ .....

จ. มาตรประมาณค่าแบบตัวเลือกบังคับตอบ (forced-choice ratings) เป็นมาตรประมาณค่าที่กำหนดตัวเลือก และค่าคะแนนให้กับตัวเลือกโดยผู้ตอบจะต้องเลือกตัวใดตัวหนึ่งโดยตัวเลือกดังกล่าวอาจใช้ข้อความทางบวกหรือทางลบก็ได้ เครื่องมือประเภทนี้มีการดัดแปลงและนำมาใช้กับแบบวัดคุณลักษณะทั่วไปเป็นจำนวนมากในปัจจุบัน

มาตรประมาณค่าทั้ง 5 แบบ มีการนำมาใช้อย่างแพร่หลาย โดยมาตรประมาณค่าแบบตัวเลือกบังคับตอบที่มีลักษณะข้อความเชิงบรรยายก็เป็นเครื่องมืออีกประเภทหนึ่งที่ที่งานวิจัยส่วนใหญ่ใช้เป็นเครื่องมือวิจัย (พจนีย์ เจนพนัส, 2536)

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

1.2 แบบสอบผลสัมฤทธิ์

แบบสอบผลสัมฤทธิ์ (achievement test) เป็นเครื่องมือใช้วัดผลของการเรียนการสอน (learning outcomes) ตามวัตถุประสงค์ที่ตั้งไว้ ข้อสอบของเครื่องมือประเภทนี้แบ่งเป็น 2 ประเภท (Gronlund, 1993) ตามลักษณะการตอบ คือ ข้อสอบประเภทเลือกคำตอบ (selection type) และข้อสอบประเภทเสนอคำตอบ (supply type)

1.2.1 ข้อสอบประเภทเลือกคำตอบ ได้แก่

ก. ข้อสอบแบบหลายตัวเลือก (multiple-choice) เป็นข้อสอบที่นิยมใช้กันมาก เพราะสามารถวัดความรู้ ความเข้าใจ การนำไปใช้ วิเคราะห์ สังเคราะห์ และการประเมินค่าของนักเรียนได้ และวัดได้ครอบคลุมเนื้อหา มีส่วนประกอบที่สำคัญ 2 ส่วน คือ ข้อคำถาม และตัวเลือก ตัวอย่าง เช่น :

(1) คำในข้อใดมีความหมายต่างกัน
* ก. สังสรรค์ - รังสรรค์
ข. ประจญ - สัประยุทธ
ค. มัฆวาน - หัสนัยน์
ง. หฤทัย - ฤทัย

ข. ข้อสอบแบบถูกผิด (true-false) เป็นข้อสอบที่ผู้ตอบสามารถเลือกได้ 2 อย่าง คือ ข้อความที่กำหนดให้นั้น ถูก หรือ ผิด , ใช่ หรือ ไม่ใช่ เหมาะกับการวัดความรู้ความจำ ความเข้าใจ และการนำไปใช้ จุดอ่อนก็คือผู้สอบมีโอกาสสูงในการเดาข้อสอบได้ถูกต้อง เช่น :

ถูก *ผิด (1) ข้อสอบแบบถูก-ผิด จัดเป็นข้อสอบประเภทเสนอคำตอบ

ค. ข้อสอบแบบจับคู่ (matching) เป็นข้อสอบที่ผู้ตอบจะต้องจับคู่ระหว่างคำหรือข้อความ 2 ชุด ที่มีความสอดคล้องกัน เหมาะกับการวัดความรู้ ความจำของนักเรียน ตัวอย่าง เช่น :

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

เหตุการณ์	เกิดในปี พ.ศ.
..(ค)...1. การประกาศใช้รัฐธรรมนูญครั้งแรก	ก. 2448
..(ก) ..2. การเลิกทาส	ข. 2457
..(ง)...3. การขับไล่ 3 ทรราช	ค. 2475
	ง. 2516

1.2.2 ข้อสอบประเภทเสนอคำตอบ เป็นข้อสอบที่ผู้ตอบต้องเขียนหรือเสนอคำตอบขึ้นมาเองโดยใช้วิธีเขียนตอบ ข้อดีคือสะดวกในการเขียนข้อสอบ สามารถวัดความรู้ขั้นสูงโดยเฉพาะการวิเคราะห์ สังเคราะห์และประเมินค่า แต่จุดอ่อนคือไม่สะดวกในการตรวจให้คะแนนอาจทำให้ขาดความเป็นปรนัยในการตรวจให้คะแนนได้ ข้อสอบประเภทนี้ ได้แก่

ก. ข้อสอบแบบตอบสั้น (short answer) เป็นข้อสอบที่ผู้สอบต้องเติมคำตอบขึ้นมาเอง แต่เป็นคำตอบสั้น ๆ เหมาะกับการวัดความรู้ความจำเกี่ยวกับกฎ หรือสูตร เช่น :

(1) สูตรของกรดกำมะถันคือ .....

ข. ข้อสอบแบบความเรียง (essay test) เป็นข้อสอบที่ผู้ตอบต้องประมวลความรู้ แล้วเรียบเรียงเป็นคำตอบ เช่น :

(1) ให้ท่านอธิบายหลักและวิธีการวัดและประเมินผลในชั้นเรียนมาโดยละเอียด

1.3 การวิเคราะห์มาตรประมาณค่าและแบบสอบผลสัมฤทธิ์ที่ตรวจให้คะแนนแบบพหุวิภาค ตามแนวทฤษฎีการทดสอบแบบดั้งเดิม

1.3.1 การวิเคราะห์มาตรประมาณค่า

การวิเคราะห์มาตรประมาณค่าตามแนวทฤษฎีการทดสอบแบบดั้งเดิม มีการวิเคราะห์ที่สำคัญมีดังนี้

ก. การวิเคราะห์ค่าอำนาจจำแนก สามารถทำได้ 2 วิธี คือ

(1) โดยการคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคะแนนรายข้อกับคะแนนรวม

(2) โดยการทดสอบค่าสถิติที (t-test) ระหว่างนักเรียนที่ได้คะแนนสูงและต่ำ ซึ่งปกติจะใช้เทคนิคการแบ่งกลุ่มสูง - กลุ่มต่ำแบบ 25 เปอร์เซนต์ (McIver and Carmines, 1981 อ้างใน บุญธรรม กิจปรีดาบริสุทธิ์ ,2534)

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ข. การวิเคราะห์ความเที่ยง สามารตรวจสอบได้หลายวิธีคือ

(1) การวิเคราะห์ความเที่ยงเชิงคงที่ (stability) โดยการใช้วิธีสอบซ้ำ (test - retest) คือ ใช้แบบวัดฉบับเดิมไปใช้กับผู้สอบกลุ่มเดิมเพื่อดูความสอดคล้องของผลการวัดทั้งสองครั้ง

(2) การวิเคราะห์ความเที่ยงแบบความสอดคล้อง (consistency) สามารถทำได้โดยพิจารณาความสอดคล้องของผลการสอบจากแบบวัดคู่ขนาน (parallel test) หรือการแบ่งครึ่งฉบับ (split-half) และดูความสอดคล้องภายในโดยการคำนวณค่าสัมประสิทธิ์แอลฟ่าด้วยสูตรของครอนบาคหรือการใช้วิธีวิเคราะห์ความแปรปรวนของฮอยท์ (บุญธรรม กิจปรีดาบริสุทธิ์, 2534)

ค. การวิเคราะห์คุณภาพด้านความตรง เป็นการตรวจสอบว่าแบบวัดที่สร้างขึ้นวัดในสิ่งที่ต้องการวัดได้มากเพียงใด สามารถตรวจสอบคุณภาพได้ดังนี้

(1) การตรวจสอบความตรงเชิงเนื้อหาซึ่งส่วนมากจะให้ผู้เชี่ยวชาญเป็นผู้ตรวจสอบและตัดสินว่าข้อกระทงแต่ละข้อวัดคุณลักษณะได้ครอบคลุมในสิ่งที่ต้องการวัดหรือไม่

(2) การตรวจสอบความตรงเชิงโครงสร้าง เป็นการตรวจสอบว่าเครื่องมือ ดังกล่าววัดได้ตรงตามโครงสร้างของตัวแปรคุณลักษณะหรือไม่ สามารถทำได้โดยการวิเคราะห์องค์ประกอบ การเปรียบเทียบกับกลุ่มที่รู้แน่ชัด หรือการวิเคราะห์ความตรงเชิงจำแนก เป็นต้น

การวิเคราะห์มาตรประมาณค่าตามแนวทฤษฎีการทดสอบแบบดั้งเดิมดังกล่าวจะให้ค่าสถิติที่มักแปรเปลี่ยนไปตามกลุ่มตัวอย่างที่สอบ เช่น ถ้ากลุ่มตัวอย่างมีความเป็นเอกพันธ์ (homogeneous) ค่าอำนาจจำแนกก็มีแนวโน้มที่จะต่ำกว่าการใช้กับกลุ่มตัวอย่างที่มีความหลากหลาย (heterogeneous) เป็นต้น ดังนั้นการนำผลการวิเคราะห์ดังกล่าวไปใช้กับผู้สอบทั่วไปจึงไม่ค่อยเหมาะสมนัก ในปัจจุบันนักวัดผลจึงได้พยายามพัฒนาทฤษฎีการตอบสนองข้อสอบ (IRT) เพื่อแก้ปัญหาดังกล่าว โดยมุ่งประมาณค่าพารามิเตอร์ของผู้สอบและข้อสอบ เพื่อให้สามารถนำผลจากการวิเคราะห์ไปใช้กับกลุ่มตัวอย่างทั่วไปได้อย่างเหมาะสมมากขึ้น

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

1.3.2 การวิเคราะห์แบบสอบผลสัมฤทธิ์

การวิเคราะห์แบบสอบผลสัมฤทธิ์ที่ตรวจให้คะแนนแบบพหุวิภาค ผู้วิจัยได้เสนอการวิเคราะห์โดยยึดตามแบบการวิเคราะห์ข้อสอบประเภทความเรียง ซึ่งประยุกต์หลักการวิเคราะห์ข้อสอบประเภทเลือกตอบแบบหลายตัวเลือก (ศิริชัย กาญจนวาสี, 2535) ดังนี้

ก. การวิเคราะห์ค่าความยากและอำนาจจำแนก โดยหาค่าสัดส่วนของคะแนนนักเรียนในกลุ่มเก่ง และกลุ่มอ่อน แล้วคำนวณค่าดัชนีบงชี้ค่าความยาก และอำนาจจำแนก

ข. การวิเคราะห์ความเที่ยง สามารตรวจสอบได้โดย วิเคราะห์ความเที่ยงแบบความสอดคล้องภายใน (internal consistency) โดยพิจารณาความสอดคล้องของผลการโดยการคำนวณค่าสัมประสิทธิ์แอลฟ่า ของครอนบาค

ค. การวิเคราะห์คุณภาพด้านความตรง เป็นการตรวจสอบว่าแบบวัดที่สร้างขึ้นวัดในสิ่งที่ต้องการวัดได้มากเพียงใด สามารถตรวจสอบคุณภาพได้โดยการตรวจสอบความตรงเชิงเนื้อหาซึ่งส่วนมากจะให้ผู้เชี่ยวชาญเป็นผู้ตรวจสอบและตัดสินว่าข้อกระทงแต่ละข้อวัดคุณลักษณะได้ครอบคลุมในสิ่งที่ต้องการวัดหรือไม่

การวิเคราะห์มาตรประมาณค่าตามแนวทฤษฎีการทดสอบแบบดังเดิมดังกล่าวจะให้ค่าสถิติที่มักแปรเปลี่ยนไปตามกลุ่มตัวอย่างที่สอบ เช่น ถ้ากลุ่มตัวอย่างมีความเป็นเอกพันธ์ (homogeneous) ค่าอำนาจจำแนกก็มีแนวโน้มที่จะต่ำกว่าการใช้กับกลุ่มตัวอย่างที่มีความหลากหลาย (heterogeneous) เป็นต้น ดังนั้นการนำผลการวิเคราะห์ดังกล่าวไปใช้กับผู้สอบทั่วไปจึงไม่ค่อยเหมาะสมนัก ในปัจจุบันนักวัดผลจึงได้พยายามพัฒนาทฤษฎีการตอบสนองข้อสอบ (IRT) เพื่อแก้ปัญหาดังกล่าว โดยมุ่งประมาณค่าพารามิเตอร์ของผู้สอบและข้อสอบ เพื่อให้สามารถนำผลจากการวิเคราะห์ไปใช้กับกลุ่มตัวอย่างทั่วไปได้อย่างเหมาะสมมากขึ้น

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ตอนที่ 2 การตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาค

คำว่า "ทวิวิภาค" ตรงกับภาษาอังกฤษว่า "dichotomous" และคำว่า "พหุวิภาค" ตรงกับภาษาอังกฤษว่า "polytomous" ซึ่ง Mellenbergh (1995, cited by Weiss, 1995) ได้อธิบายรากศัพท์ของคำทั้ง 2 ไว้ว่า dichotomous มาจากภาษากรีกว่า dicho (two) + tomous (a cut) หมายถึงการแบ่งเป็น 2 ส่วน และคำว่า polytomous มาจากภาษากรีกว่า polus (many) + tomous (a cut) หมายถึง การแบ่งออกเป็นหลาย ๆ ส่วน คำทั้ง 2 มีการนำมาใช้กับการตรวจให้คะแนนในการวัดผลทางการศึกษา และเรียกแตกต่างกันไป เช่น เรียกการตรวจให้คะแนนแบบ dichotomous ว่า "การตรวจให้คะแนนแบบดั้งเดิม" "การตรวจให้คะแนนแบบทวิวิภาค" และ "การตรวจให้คะแนนแบบทวิวิภาค" ส่วนการตรวจให้คะแนนแบบ polytomous มีผู้เรียกว่า "การตรวจให้คะแนนแบบหลายค่า" และ "การตรวจให้คะแนนแบบพหุวิภาค" ในการวิจัยนี้จะเรียกการตรวจให้คะแนนแบบแรกและแบบที่สองว่า "การตรวจให้คะแนนแบบทวิวิภาค" และ "การตรวจให้คะแนนแบบพหุวิภาค" ตามลำดับ

การตรวจให้คะแนนแบบทวิวิภาค (dichotomous) เป็นวิธีที่นิยมใช้กันมาก เพราะมีความสะดวก และรวดเร็ว เหมาะกับข้อสอบประเภทเลือกคำตอบ เช่น ข้อสอบแบบหลายตัวเลือก ข้อสอบถูก-ผิด และข้อสอบแบบจับคู่ (Wiersma and Jurs, 1990) นอกจากนี้ยังพบว่าแบบวัดคุณลักษณะชนิดมาตรประมาณค่าแบบตัวเลือกบังคับตอบ (forced-choice rating) มีผู้ใช้การตรจให้คะแนนแบบนี้ด้วยเช่นกัน (บุษรินทร์ บุญรอด, 2536 ; วารี นิยมธรรม, 2536 ; นงเยาว์ พึ่งพา, 2529 ; สันติ สุขทรัพย์, 2528 ; ไสว ข้ออุดมมงคล, 2528) การตรวจให้คะแนนจะกำหนดตัวเลือกหรือคำตอบที่เหมาะสมและถูกที่สุดเป็น 1 คะแนน ส่วนคำตอบอื่น ๆ เป็น 0 คะแนน

สำหรับการตรวจให้คะแนนแบบพหุวิภาค (polytomous) เป็นวิธีที่นิยมใช้กันมากเช่นกัน วิธีนี้จะกำหนดคะแนนให้กับคำตอบของนักเรียนในแต่ละข้อไว้มากกว่า 2 ค่า รวมถึงการให้คะแนนในมาตรประมาณค่าที่ใช้วัดคุณลักษณะชนิดต่าง ๆ ด้วย ซึ่งใน

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

มาตรประมาณค่าก็จะกำหนดให้คะแนนตามลำดับขั้นของคะแนนในเครื่องมือนั้น ๆ โดยถ้าเป็นแบบวัดคุณลักษณะคะแนนแต่ละค่ามักแสดงถึงระดับของคุณลักษณะ แต่ถ้าเป็นแบบสอบผลสัมฤทธิ์การให้คะแนนแต่ละค่าแสดงถึงระดับของความสามารถของ ผู้สอบที่ตอบข้อกระทงนั้น อาจมีวิธีให้คะแนนแตกต่างกันไปเช่น การให้คะแนนตามระดับความมั่นใจในการตอบ การให้คะแนนตามความสามารถในการตัดตัวลวงของข้อกระทงนั้น และการให้คะแนนความรู้บางส่วน (partial knowledge) เป็นต้น การให้คะแนนความรู้บางส่วน เป็นวิธีที่มีผู้สนใจศึกษากันมาก (Smith, 1987, 1982; Frary, 1982; Hutchinson, 1984, Wright and Masters, 1982) ไรท์ และมาสเตอร์ส (Wright and Masters, 1982) ได้ยกตัวอย่างข้อสอบวิชาคณิตศาสตร์ที่ให้คะแนนความรู้บางส่วน เช่น :

โจทย์ $\sqrt{9.0/0.3-5} = ?$

การตรวจให้คะแนน

นักเรียนไม่ทำขั้นตอนใดเลย	**ได้ 0 คะแนน**
9.0 / 0.3 = 30	ได้ 1 คะแนน
3.0 − 5 = 25	ได้ 2 คะแนน
$\sqrt{25} = 5$	ได้ 3 คะแนน

จากตัวอย่างข้างต้นเป็นการตรวจให้คะแนนตามลำดับของความสำเร็จ การให้คะแนนในลักษณะนี้ ครูผู้สอนสามารถกำหนดเกณฑ์การให้คะแนนได้ตามความเหมาะสม สำหรับการวิจัยครั้งนี้ ผู้วิจัยใช้การตรวจให้คะแนนแบบสอบคณิตศาสตร์ตามความสามารถในการแก้โจทย์ปัญหาคณิตศาสตร์ โดยละข้อจะมี 3 ขั้นตอน คือ ความเข้าใจในวิธีการแก้โจทย์ปัญหา การคิดคำนวณ และการคำนึงถึงสภาพที่เป็นจริง นักเรียนไม่ตอบหรือตอบไม่ถูกทั้ง 3 ตอน ให้คะแนนอยู่ในลำดับขั้นที่ 1 นักเรียนที่ตอบถูก 1 ขั้นตอนให้คะแนนอยู่ในลำดับขั้นที่ 2 นักเรียนที่ตอบถูก 2 ขั้นตอน ให้คะแนนอยู่ในลำดับขั้นที่ 3 และนักเรียนที่ตอบถูกทั้ง 3 ตอน จะให้คะแนนอยู่ในลำดับขั้นที่ 4

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ตอนที่ 3 มโนทัศน์เกี่ยวกับทฤษฎีการตอบสนองข้อสอบ (IRT)

ทฤษฎีการตอบสนองข้อสอบ (item response theory: IRT) เป็นทฤษฎีทางการวัดผลที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในของบุคคลกับพฤติกรรมการตอบสนองข้อสอบแต่ละข้อว่ามีความน่าจะเป็นในการตอบถูกเพียงใด ทฤษฎีนี้ตั้งอยู่บนหลักพื้นฐานที่สำคัญ 2 ประการ (Hambleton, Swaminathan and Roger, 1991) คือ

1. ความสามารถของบุคคลในการตอบข้อสอบได้ถูกหรือผิด สามารถอธิบายได้ด้วยคุณลักษณะภายในหรือความสามารถ (latent trait or ability) ของบุคคลนั้นๆ

2. ความสัมพันธ์ระหว่างความสามารถในการตอบข้อสอบได้ถูกต้องกับความสามารถของผู้สอบที่วัดจากแบบสอบสามารถอธิบายได้ด้วยโค้งลักษณะข้อสอบ (item characteristic curve : ICC)

พัฒนาการของทฤษฎีนี้เริ่มมาจากการที่ ลอว์ลี (Lawley) ได้เสนอบทความเกี่ยวกับปัญหาการสร้างและการเลือกข้อสอบ ในปี ค.ศ.1943 (ศิริชัย กาญจนวาสี , 2535) โดยเขาได้เสนอแนวคิดในรูปของโมเดลนอร์มัลโอไจฟ์ (normal ogive model) ผู้ที่พัฒนาต่อมาในปี ค.ศ.1950 คือ ราส์ช (Rasch) นักคณิตศาสตร์ชาวเดนมาร์กซึ่งได้เสนอโมเดลราส์ช (rasch model) แบบ 1 พารามิเตอร์ โดยมีแนวคิดว่าความยากของข้อสอบ (bj) เป็นสิ่งเดียวที่มีอิทธิพลต่อการตอบสนองข้อสอบ

ในปี ค.ศ.1952 ลอร์ด (Lord) ได้เสนอฟังก์ชันนอร์มัลโอไจฟ์ แบบ 2 พารามิเตอร์ โดยเพิ่มพารามิเตอร์อำนาจจำแนก (aj) และในปี ค.ศ. 1968 เบิร์นบอม (Birnbaum) ได้พัฒนาโมเดลโลจิสติก (logistic model) เป็นฟังก์ชันที่สามารถคำนวณได้ง่ายขึ้น ในปี ค.ศ.1974 ลอร์ด (Lord) ได้เสนอโมเดลโลจิสติก แบบ 3 พารามิเตอร์ โดยเพิ่มพารามิเตอร์โอกาสการเดาข้อสอบ (ci) ในการประมาณค่าพารามิเตอร์ของผู้สอบด้วย

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ปัจจุบันทฤษฎีการตอบสนองข้อสอบได้พัฒนาไปมากโดย บาร์ตัน และ ลอร์ด (Barton and Lord) ได้เสนอโมเดลโลจิสติก แบบ 4 พารามิเตอร์ขึ้น โดยเขาเชื่อว่า นักเรียนที่มีความสามารถสูงมากไม่จำเป็นที่จะต้องตอบข้อสอบถูกเสมอไปแม้ข้อสอบจะง่ายหรือจะระมัดระวังอย่างมากแล้วก็ตามก็ยังไม่สามารถใช้สารสนเทศที่มีตอบข้อสอบได้ถูกต้องตามเฉลยของผู้ออกข้อสอบได้เสมอไป เขาเสนอแนวคิดในรูปสมการดังนี้ (Hambleton , 1982)

$$P_j(\theta)=c_j+(\gamma_j-c_j)\frac{\exp^{[Da_j(\theta-b_j)]}}{1+\exp^{[Da_j(\theta-b_j)]}}$$

เมื่อ j คือ ข้อสอบข้อที่ 1,2,3,...,n

γ_j คือ ค่าพารามิเตอร์ตัวแปรแฝงในข้อกระทง

c_j คือ ค่าพารามิเตอร์ของโอกาสการเดา

exp คือ ค่าคงที่มีค่า 2.7183

D คือ ค่าคงที่มีค่า 1.702

a_j คือ ค่าพารามิเตอร์อำนาจจำแนกข้อ j

b_j คือ ค่าพารามิเตอร์ความยากของข้อ j

โมเดลนี้ต่างจากโมเดล 3 พารามิเตอร์ เพราะ γ_j สมมติให้มีค่าต่ำกว่า 1 เล็กน้อย ดังนั้นโค้ง ICC ที่ตำแหน่งสูงสุดมีค่าต่ำกว่า 1 โมเดลนี้ให้ความสนใจศึกษาเชิงทฤษฎีเท่านั้น โดยผู้เสนอเองก็ยังไม่ได้ศึกษาในเชิงปฏิบัติหรือนำไปใช้ สำหรับการนำแนวคิดของ IRT ไปใช้กับการวิเคราะห์มาตรประมาณค่าหรือข้อสอบที่ตรวจให้คะแนนแบบพหุวิภาค (polytomous) ในปัจจุบันใช้แบบ 1 หรือ 2 พารามิเตอร์ เท่านั้น

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ตอนที่ 4 มโนทัศน์เกี่ยวกับทฤษฎี Polytomous Item Response

โมเดล IRT ที่พัฒนาในระยะแรกจะใช้กับแบบสอบที่ให้คะแนนแบบทวิวิภาค (dichotomous) เป็นส่วนใหญ่ ต่อมาได้มีผู้พัฒนาโมเดล IRT เพื่อใช้กับแบบสอบและแบบวัดทัศนคติที่ตรวจให้คะแนนแบบพหุวิภาค (polytomous) หรือลักษณะมาตรประมาณค่า (rating data) โมเดลในแนวทฤษฎีนี้ทั้งหมดเรียกว่า Polytomous Item Response Models (Muraki, 1992, 1993 ; Donoghue , 1994 ; Hambleton and Zaal, 1991) มีโมเดลที่สำคัญ คือ

1. Graded Response Model (GRM) พัฒนาโดย ซาเมจิมา (Samejima) ในปี ค.ศ.1960
2. Nominal Response Model (NRM) พัฒนาโดย บอค (Bock) ในปี ค.ศ.1972
3. Continuous Model (CM) พัฒนาโดย ซาเมจิมา ในปี ค.ศ. 1973
4. Rating Scale Model (RSM) พัฒนาโดย แอนดริช (Andrich) ปี ค.ศ.1978
5. Partial Credit Model (PCM) พัฒนาโดย มาสเตอรส์ (Masters) ปี ค.ศ.1982
6. Sucessive Interval Model(SIM) พัฒนาโดย รอส์ท (Rost) ในปี ค.ศ.1988
7. Generalized Partial Credit Model :GPCM พัฒนาโดย มูราคิ (Muraki) ในปี ค.ศ. 1992

โมเดลที่มีผู้ศึกษากันมากในปัจจุบัน ไม่เข้มงวดเกี่ยวกับข้อตกลงเบื้องต้น และสามารถใช้ได้กับแบบสอบและแบบวัดหลายลักษณะคือ GRM และ GPCM (Donoghue, 1994; De Ayala ,1994 ; Muraki , 1992, 1993 ; Reise and Yu ,1990 ; Dodd and De Ayala , 1989 ; Koch , 1983)

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ซาเมจิมา (Samejima) ได้พัฒนา GRM จากโมเดลโลจิสติก 2 พารามิเตอร์ เพื่อใช้กับแบบสอบและแบบวัดทัศนคติที่ตรวจให้คะแนนแบบพหุวิภาค (polytomous) ในขณะที่มูราคิ (Muraki) ได้พัฒนา GPCM จาก PCM ของ ไรท์ และมาสเตอรส์ (Wright and Masters, 1982) ให้สามารถใช้ได้กับแบบสอบและแบบวัดที่มีลักษณะเป็นพหุมิติ โดยรวมค่าอำนาจ

จำแนกข้อกระทงในการประมาณค่าพารามิเตอร์ของฟังก์ชันด้วย โมเดลทั้ง 2 เมื่อพิจารณาดูแล้วคล้ายคลึงกันมาก เพียงแต่มีมโนทัศน์เกี่ยวกับฟังก์ชันลักษณะเชิงปฏิบัติ (operating characteristic function : OCF) และพัฒนาการที่แตกต่างกัน รูปแบบทั่วไปของโมเดลในแนวคิดนี้พัฒนามาจากสูตร

$$P_i(u_j = 1/\theta) = \frac{\exp^{[a_j(\theta - b_j)]}}{1 + \exp^{[a_j(\theta - b_j)]}}$$

เมื่อ $P_j(u_j = 1/\theta)$ คือ ความน่าจะเป็นของคนที่ i มีความสามารถ θ สามารถตอบข้อสอบข้อ j (U=1) ได้ถูกต้อง

a_j คือ ค่าพารามิเตอร์อำนาจจำแนก

b_j คือ ค่าพารามิเตอร์ของความยากข้อสอบ

exp คือ ค่าคงที่ 2.7183

ถ้าเรามีข้อตกลงว่าข้อสอบทุกข้อมีอำนาจจำแนกแตกต่างกันโมเดลนี้ก็คือโมเดลโลจิสติกแบบ 2 พารามิเตอร์ และถ้าให้ข้อตกลงว่าข้อสอบทุกข้อมีอำนาจจำแนกเท่ากัน ค่า a_j จะเท่ากับ 1 โมเดลนี้ก็คือ Rasch's Dichotomous Response Model

ฟังก์ชันของโมเดลราส์ชดังกล่าวที่เสนอขึ้นมาจะมีความสอดคล้องกับพฤติกรรมการตอบสนองข้อสอบตามหลักทางจิตวิทยาว่าผู้ที่มีความสามารถหรือมีคุณลักษณะที่ต้องการวัดสูง มีความน่าจะเป็นในการตอบได้คะแนนที่สูงกว่าผู้ที่มีคุณลักษณะหรือความสามารถต่ำ

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

3.1 GRM (Graded Response Model)

GRM มีข้อตกลงว่าการตอบสนองข้อกระทง j ของผู้ตอบสามารถแบ่งคะแนนออกเป็น $mj+1$ ลำดับขั้น (categories) คะแนนแต่ละลำดับ (j_k) กำหนดให้เป็น 0,1,2,...,mj ตามลำดับ (Donoghue , 1994 ; Koch , 1983) เขียนสมการทั่วไปได้ดังนี้

$$P_{jk}(\theta) = \frac{\exp^{[Da_j(\theta - b_{jk})]}}{1 + \exp^{[Da_j(\theta - b_{jk})]}}$$

เมื่อ $P_{jk}(\theta)$ คือ ความน่าจะเป็นที่ผู้มีคุณลักษณะ θ จะตอบข้อกระทง j ได้คะแนน k

D คือ ค่าคงที่ของสเกล (scaling constant) เมื่อปรับโค้งฟังก์ชั่นของโมเดลโลจิสติก และโมเดล นอร์มอล ออไจฟ์ มีค่าเท่ากับ 1.70

a_j คือ ค่าอำนาจจำแนกของข้อความ j

θ คือ ความสามารถหรือคุณลักษณะภายในของผู้ตอบ

b_{jk} คือ ค่าความยากของข้อความ j ในลำดับขั้นที่ jk เมื่อ jk = 0,1, . . . ,k

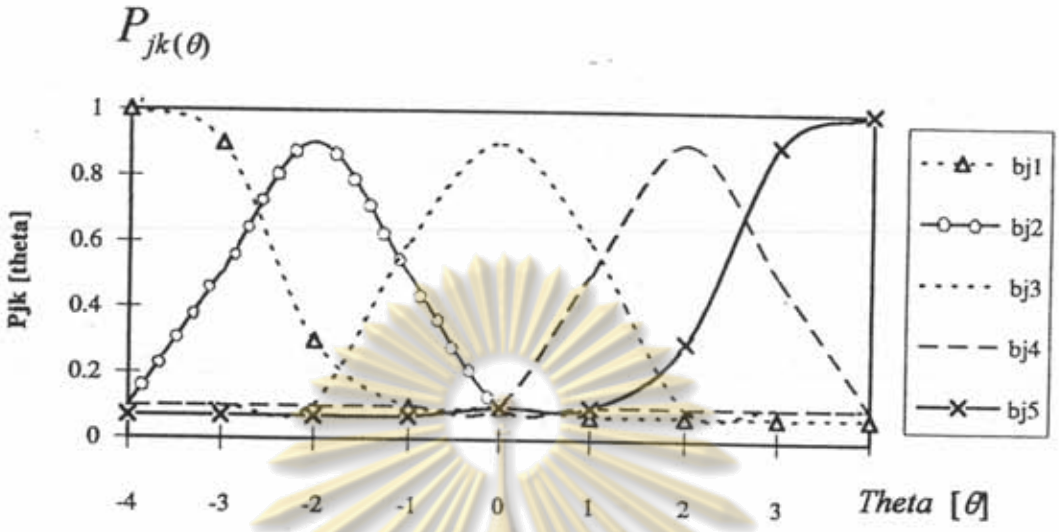
exp คือ ค่าคงที่ มีค่าเท่ากับ 2.7183

ฟังก์ชันทางคณิตศาสตร์ดังกล่าวสามารถเขียนแทนด้วยโค้งฟังก์ชันสารสนเทศได้ดังภาพที่ 1 (koch, 1983:18)

ศูนย์วิทยทรัพยากร

ในโมเดลนี้แสดงถึงความสัมพันธ์ของคนที่มีคุณลักษณะภายใน (θ) สูง มีความน่าจะเป็นในการตอบลำดับขั้นคะแนนที่สูงกว่าคนที่มีคุณลักษณะภายในต่ำ ซึ่งนักวัดผลได้มีการประยุกต์ใช้โมเดลนี้กับแบบสอบชนิดปรับเหมาะ (Adaptive Testing) (Koch and Dodd, 1989) และแบบวัดทัศนคติต่างๆ (Koch ,1983 ; Dodd ,1985)

จุฬาลงกรณ์มหาวิทยาลัย



ภาพที่ 1 โค้งฟังก์ชันสารสนเทศของ GRM

ลักษณะเฉพาะฟังก์ชันของโมเดล GRM สรุปได้ดังนี้

1. ฟังก์ชันสารสนเทศของแบบวัด (TIF) เขียนแทนด้วยสมการดังนี้

$$I(\theta) = \sum_{j=1}^{m} I_j(\theta)$$

เมื่อ I (θ) คือ ฟังก์ชันสารสนเทศของข้อกระทงตั้งแต่ข้อ 1 ถึงข้อที่ m

2. ฟังก์ชันสารสนเทศของข้อกระทง (IIF) สามารถแทนด้วยสมการดังนี้

$$I_j(\theta) = \sum_{jk=0}^{k} \frac{[P'_{jk}(\theta)]^2}{P_{jk}(\theta)}$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

เมื่อ $P'_{jk}(\theta)$ คือ ความน่าจะเป็นสำหรับคนที่มีความสามารถ θ จะตอบข้อกระทง j ได้คะแนน 1 มากกว่าคะแนน k ใดๆ

$P_{jk}(\theta)$ คือ ความน่าจะเป็นสำหรับคนที่มีความสามารถ θ จะตอบข้อกระทง j ได้คะแนน k

3. ฟังก์ชันสารสนเทศของแต่ละลำดับขั้นคะแนน (ICIFs) เขียนแทนด้วยสมการดังนี้

$$I_{jk}(\theta) = \frac{[P'_{jk}(\theta)]^2}{P_{jk}(\theta)}$$

เมื่อ $P'_{jk}(\theta)$ คือ ความน่าจะเป็นสำหรับคนที่มีความสามารถ θ จะตอบข้อกระทง j ในลำดับขั้นคะแนน ที่ 1 ได้ถูกต้อง มากกว่าคะแนน k ใดๆ

$P_{jk}(\theta)$ คือ ความน่าจะเป็นสำหรับคนที่มีความสามารถ θ จะตอบข้อกระทง j ในลำดับขั้นคะแนน k ได้ถูกต้อง

3.2 GPCM (Generalized Partial Credit Model)

GPCM (Generalized Partial Credit Model) พัฒนาจาก PCM ให้มีการตรวจให้คะแนนความรู้บางส่วนในการแก้โจทย์ปัญหาคณิตศาสตร์ โดยพิจารณาให้คะแนนตามลำดับขั้นความสำเร็จของการแก้โจทย์ปัญหาเป็นคะแนนเรียงลำดับ 0, 1, 2 , 3 (Masters, 1982) จากนั้นก็มีการนำไปใช้กับแบบวัดทัศนคติที่มีคะแนนเรียงลำดับหลายค่า (Dodd and Koch, 1989)

มูราคิ (Muraki) ได้พัฒนาโมเดลดังกล่าวมาเป็น GPCM โดยแก้ไขข้อจำกัดเกี่ยวกับความเป็นเอกมิติของเครื่องมือที่ใช้วัดคุณลักษณะ และใช้ค่าอำนาจจำแนกที่แปรเปลี่ยนไปในแต่ละข้อมารวมประมาณค่าพารามิเตอร์ด้วย โดยมีข้อตกลงว่าบุคคลที่มีความสามารถหรือมีคุณลักษณะที่ต้องการวัดสูงมีความน่าจะเป็นที่จะตอบคำตอบลำดับคะแนนที่ *K* มากกว่า *K-1* สามารถแทนด้วยสมการทั่วไปได้ดังนี้

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

$$C_{jk}(\theta) = P_{jk/k-1}(\theta) = \frac{P_{jk}(\theta)}{P_{j.k-1}(\theta) + P_{jk}(\theta)} = \frac{\exp^{[a_j(\theta)-b_{jk})]}}{1+\exp^{[a_j(\theta-b_{jk})]}}$$

เมื่อ k คือ ลำดับขั้นคะแนนที่ 2 ,3 ,...,*mj*

$$P_{jk}(\theta) = \frac{C_{jk}}{1-C_{jk}} P_{jk-1}(\theta)$$

$P_{jk}(\theta)$ คือ ความน่าจะเป็นของคนที่มีความสามารถจะตอบได้คะแนน k (กรณี k-1 ก็เช่นกัน)

$C_{jk}/(1-C_{jk})$ คือ อัตราส่วนความน่าจะเป็นของสองเงื่อนไข ซึ่งอาจแสดงในรูปสมการ $\exp^{[a_j(\theta-b_{jk})]}$

มาสเตอรส์ (Masters) เรียก b_{jk} ว่าเป็นค่าพารามิเตอร์ความยากในแต่ละลำดับขั้น เป็นจุดบนแกน θ ที่ฟังก์ชันของ $P_{j.k-1}(\theta)$และ $P_{jk}(\theta)$ ตัดกัน ซึ่งโค้งทั้ง 2 เป็นฟังก์ชันของการตอบ (item category response function : ICRFs) ใน 2 ลำดับค่าคะแนนที่ตัดกันเพียงจุดเดียวบนค่า θ ใด ๆ คือ

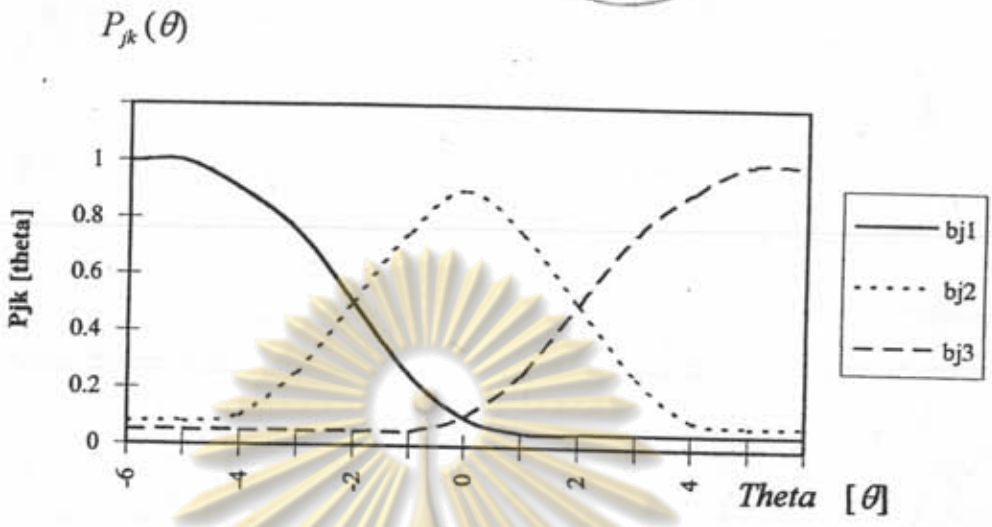
ถ้า $\theta = b_{jk}$, $P_{jk}(\theta) = P_{j.k-1}(\theta)$;

ถ้า $\theta > b_{jk}$, $P_{jk}(\theta) > P_{j.k-1}(\theta)$;

และถ้า $\theta < b_{jk}$ $P_{jk}(\theta) < P_{j.k-1}(\theta)$;

ซึ่งอยู่บนข้อตกลงว่า $a_j > 0$ และ b_{jk} ไม่จำเป็นต้องเรียงกันในข้อสอบ j เพราะค่าพารามิเตอร์จะแสดงถึงขนาดของความนาจะเป็นในการตอบ $P_{j.k-1}(\theta)$ และ$P_{jk}(\theta)$ ในลำดับที่อยู่ชิดกัน ดังภาพที่ 2

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย



ภาพที่ 2 โค้งฟังก์ชันสารสนเทศของ GPCM

การพิจารณาจุดตัดของโค้งฟังก์ชันการตอบแต่ละลำดับขั้น (item category response functions: ICRFs) ของโมเดล PCM แม้จะง่ายต่อการตีความหมายก็ตาม แต่จุดยอดของโค้งก็ไม่ใช่จะอยู่ตรงกลางของช่วงค่าคะแนนเสมอไป ซึ่งที่มาของค่าฟังก์ชันการตอบสนองข้อกระทงแต่ละลำดับขั้น ($P_{jk}(\theta)$) แสดงได้ดังนี้ (Muraki, 1992)

$$\frac{\partial}{\partial(\theta)} P_{jk}(\theta) = a_j P_{jk}(\theta)[k - \sum_{k=1}^{mj} kP_{jk}(\theta)]$$

ศูนย์วิทยทรัพยากร

โดยกำหนดสมการเป็น 0; $\sum_{k=1}^{mj} cP_j(\theta) = k$ สมการนี้จะแสดงโค้ง ICRFs ของ $P_{jk}(\theta)$ เมื่อ k = 1, 2, 3, . . . , k-1, k , k+1 , . . . , mj

จุฬาลงกรณ์มหาวิทยาลัย

พารามิเตอร์ a_j คือพารามิเตอร์อำนาจจำแนกของข้อ j โดยค่า a_j มีพิสัยที่เป็นไปได้อยู่ในช่วง -∞ ถึง + ∞ ซึ่งไม่เหมือนใน Dichotomous Model เพราะใน Polytomous Item Response Model อำนาจจำแนกในแต่ละลำดับขั้นคะแนนจะขึ้นอยู่กับผลรวมของ

I 16893700

พารามิเตอร์อำนาจจำแนกและค่าพารามิเตอร์เทรสโฮลด์ แต่ใน GPCM จะรวมค่าอำนาจจำแนกไว้ในโมเดลด้วย ซึ่งค่าอำนาจจำแนกนี้จะบอกถึงระดับการตอบในแต่ละลำดับค่าคะแนนที่เปลี่ยนไปในแต่ละข้อเช่นเดียวกับที่ค่า θ เปลี่ยนไปด้วย (Muraki ,1992)

จากแนวคิดดังกล่าวสามารถสรุปฟังก์ชันสารสนเทศที่สำคัญของ GPCM ได้ดังนี้

1. ฟังก์ชันสารสนเทศของแบบวัด (TIF) แทนด้วยสมการดังนี้

$$I(\theta) = \sum_{j=1}^{m} I_j(\theta)$$

เมื่อ $I_j(\theta)$ คือ ฟังก์ชันสารสนเทศของข้อกระทง (IIF)

2. ฟังก์ชันสารสนเทศของข้อกระทง (IIF) ซึ่งสามารถแทนด้วยสมการ

$$I_j(\theta) = D^2 a_j^{\,2} \sum_{c=1}^{mj} [T_k - \bar{T}_j(\theta)]^2 P_{jk}(\theta)$$

เมื่อ $\bar{T}_j = \sum_{k=1}^{mj} T_k P_{jk}(\theta)$

D คือ ค่าคงที่ของสเกลมีค่าเท่ากับ 1.70

a_j คือ ค่าอำนาจจำแนกของข้อ j

T_k คือ ค่าคะแนนใดๆ ใน T_j=(1,2,...,mj)

$P_{jk}(\theta)$ คือ ความน่าจะเป็นของคนที่มีความสามารถ θ จะตอบข้อ j ได้คะแนน k

k คือ ลำดับขั้นคะแนนที่ 1,2,...,k-1, k, k+1,...,mj

3. ฟังก์ชันสารสนเทศของลำดับค่าคะแนนในแต่ละข้อ (ICIFs) แทนด้วยสมการดังนี้

$$I_{jk}(\theta) = P_{jk}(\theta) I_j(\theta)$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

เมื่อ $P_{jk}(\theta)$ คือ ความน่าจะเป็นของผู้ตอบที่มีความสามารถ θ จะทำข้อสอบข้อ j ได้คะแนน k

$I_j(\theta)$ คือ ฟังก์ชันสารสนเทศข้อกระทง

ตอนที่ 5 งานวิจัยที่เกี่ยวข้อง

การศึกษาเปรียบเทียบความเหมาะสมของวิธีการตรวจให้คะแนนแบบวัดชนิดมาตรประมาณค่าที่เป็นแบบเลือกตอบยังไม่เคยมีการศึกษามาก่อน แต่มีงานวิจัยที่เกี่ยวข้องศึกษาเพื่อเปรียบเทียบความเหมาะสมของวิธีตรวจให้คะแนน แบบพหุวิภาคซึ่งการศึกษาเพื่อเปรียบเทียบวิธีตรวจให้คะแนนดังกล่าวส่วนมากจะใช้เครื่องมือที่เป็นแบบสอบวัดผลสัมฤทธิ์ (achievement test) (Smith, 1987; De-Ayala, 1993; Yamamoto and Kulick, 1992 cited by Donoghue, 1994; Wainer and Thissen, 1993; cited by Donoghue, 1994; Donoghue, 1994) มีผลการวิจัยที่ได้ข้อค้นพบเป็น 2 ประเด็นที่สำคัญ คือ

ประเด็นแรก ให้ข้อสนับสนุนว่าการตรวจให้คะแนนแบบพหุวิภาคมีความเที่ยงและ ความตรงเชิงสภาพ (concurrent validity) สูงกว่าการตรวจให้คะแนนแบบทวิวิภาค (Smith,1987) และในการเปรียบเทียบค่าฟังก์ชันสารสนเทศก็พบว่าให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค (Samejima, 1976 cited by Donoghue, 1994; Donoghue, 1994) และนอกจากนี้ยังพบว่าค่าพารามิเตอร์มีความคงที่ แม่นยำกว่าการตรวจให้คะแนนแบบทวิวิภาคด้วย (Smith, 1976 cited by Donoghue, 1994) ในส่วนของการศึกษาเกี่ยวกับการตรวจให้คะแนนความรู้บางส่วน (partial knowledge) ที่มีลักษณะการตรวจให้คะแนนแบบพหุวิภาคแบบหนึ่ง พบว่าเป็นวิธีที่มีความเที่ยงและความตรงสูงกว่าวิธีการตรวจให้คะแนนแบบทวิวิภาค (Frary and Hutchinson, 1982; Smith, 1987)

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ผลการวิจัยอีกประเด็นหนึ่งที่ศึกษากับแบบสอบวัดผลสัมฤทธิ์ ได้ข้อค้นพบว่า การตรวจให้คะแนนแบบทวิวิภาคให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบพหุวิภาค เช่น การศึกษาของ ยามาโมโต และคูลิค (Yamamoto and Kulick,1992 cited by Donoghue, 1994) เวนเนอร์ และทิชเซน (Wainer and Thissen, 1993 cited by Donoghue, 1994) อย่างไรก็ดี โดนัฟ (Donoghue, 1994) ได้ให้ข้อสังเกตว่าการศึกษาของ ยามาโมโต และคูลิก ดังกล่าวเป็นการตรวจให้คะแนนที่ไม่เหมาะสม เนื่องจากใช้แบบสอบที่สร้างขึ้นเพื่อการตรวจให้คะแนนแบบทวิวิภาคมาปรับเป็นการตรวจให้คะแนนแบบพหุวิภาค ส่วน เวนเนอร์ และทิชเซน จะมุ่งศึกษาถึงความคุ้มค่าของเวลาและการใช้จ่าย (time and expense) มากกว่าจะพิจารณาถึงความตรง (validity) และความเชื่อถือได้ (authenticity) ซึ่ง มูราคิ (Muraki, 1993) ได้ให้ข้อสังเกตว่า แบบสอบที่ตรวจให้คะแนนแบบพหุวิภาคไม่จำเป็นที่จะต้องให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาคเสมอไป แต่โดยทั่วไปจะมีโอกาสให้ค่าฟังก์ชันสารสนเทศครอบคลุมช่วงพิสัยของคุณลักษณะ (theta) มากกว่าการตรวจให้คะแนนแบบทวิวิภาค

จากงานวิจัยดังกล่าวข้างต้น เป็นการศึกษาเพื่อตอบคำถามว่าการใช้แบบสอบที่ตรวจให้คะแนนแบบทวิวิภาคกับแบบพหุวิภาค วิธีใดจะมีความเที่ยง ตรงกว่ากัน โดยมีผลการศึกษาที่ค่อนข้างชัดเจนว่าแบบสอบที่ตรวจให้คะแนนแบบพหุวิภาค ถึงแม้จะต้องเสียเวลาไปกับการตรวจให้คะแนนมากกว่าแบบสอบที่ตรวจให้คะแนนแบบทวิวิภาค แต่จะให้คุณภาพด้านความเที่ยงและความตรงที่สูงกว่าการตรวจให้คะแนนแบบทวิวิภาค ซึ่งจะเห็นได้ว่าผลการศึกษาดังกล่าวส่วนใหญ่ล้วนศึกษากับแบบสอบประเภทวัดผลสัมฤทธิ์ (achievement test) ยังไม่เคยศึกษาในแบบวัดคุณลักษณะที่มีลักษณะเป็นมาตรประมาณค่าที่มีลักษณะแบบตัวเลือกตอบอย่างจริงจัง นอกจากนี้ยังไม่เคยมีการศึกษาเพื่อเปรียบเทียบความเหมาะสมของโมเดลประเภท Polytomous Item Response ว่าโมเดลใดมีความเหมาะสมกว่ากัน ผู้วิจัยจึงเห็นว่าการศึกษาเพื่อเปรียบเทียบฟังก์ชันสารสนเทศ ของแบบวัดคุณลักษณะและแบบสอบประเภทวัดผลสัมฤทธิ์ที่มีวิธีการตรวจให้คะแนนที่แตกต่างกันและการวิเคราะห์ด้วยโมเดลที่ต่างกันน่าจะเป็นประโยชน์สำหรับการศึกษาด้านการวัดผลทางการศึกษาต่อไป

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ตอนที่ 6 โปรแกรมคอมพิวเตอร์ที่ใช้วิเคราะห์มาตรประมาณค่า

การวิเคราะห์มาตรประมาณค่าตามแนวคิดของทฤษฎีการตอบสนองข้อสอบ โดยใช้โปรแกรมคอมพิวเตอร์มี 2 แบบ คือ

5.1 การวิเคราะห์เพื่อคัดเลือกผู้ตอบที่เหมาะสม (person-fit) และข้อกระทงที่เหมาะสม (item-fit) และการวิเคราะห์ค่าน้ำหนักคะแนนที่เหมาะสม (optimal weight) ซึ่งในที่นี้จะกล่าวถึงเฉพาะการวิเคราะห์ผู้ตอบและข้อกระทงที่เหมาะสมเท่านั้น การวิเคราะห์ดังกล่าวสามารถทำได้โดยใช้โปรแกรมคอมพิวเตอร์สำเร็จรูป BIGSTEPS ที่พัฒนาโดย ไลนาเคร และไรท์ (Linacre and Wright) ในปี ค.ศ. 1994 โปรแกรมนี้สามารถใช้วิเคราะห์ ผู้ตอบที่เหมาะสมและข้อกระทงที่เหมาะสม โดยมีหลักการที่สำคัญคือ

5.1.1 การวิเคราะห์ผู้ตอบที่เหมาะสม (person-fit) เป็นการวิเคราะห์เพื่อคัดเลือกผู้ตอบที่มีแบบแผนการตอบเหมาะสมนำไปใช้ในการวิเคราะห์ต่อไป โดยมีหลักการง่ายๆ คือ ผู้ที่มีความสามารถสูงควรจะตอบข้อกระทงที่ง่ายและยากได้ถูกต้องมากกว่าผู้ที่มีความสามารถต่ำ ผู้ตอบที่ไม่เหมาะสมก็คือผู้ตอบที่มีแบบแผนการตอบที่ไม่เป็นไปตามแนวคิดนี้ โดยสามารถพิจารณาคัดเลือกบุคคลที่ไม่เหมาะสมได้จากการพิจารณาค่าสถิติ Person Infit , Person Outfit และค่า Mean Square Outfit (MNSQ) โดยผู้ตอบที่เหมาะสมควรจะมีค่าสถิติ INFIT และ OUTFIT ไม่เกิน /2/ และค่า MNSQ ไม่เกิน 1.20 (Linacre and Wrigth, 1994 :89) ผู้ตอบที่ไม่เหมาะสมอาจเกิดจากการตอบแบบเดาสุ่มหรือไม่ตั้งใจทำ หากนำผลการตอบมารวมในการวิเคราะห์ด้วยอาจทำให้การสรุปผลมีความคลาดเคลื่อนสูงได้

5.1.2 การวิเคราะห์ข้อกระทงที่เหมาะสม (item-fit) เป็นการวิเคราะห์เพื่อคัดเลือกข้อกระทงที่สามารถจำแนกคนที่มีระดับคุณลักษณะสูงและต่ำออกจากกันได้ ซึ่งมีหลักการในการคัดเลือกค่าสถิติ INFIT , OUTFIT และ MNSQ คล้ายกับการวิเคราะห์ผู้ตอบที่เหมาะสม โดยการวิเคราะห์ข้อกระทงที่เหมาะสมควรจะวิเคราะห์หลังจากได้คัด

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

คัดเลือกผู้ตอบที่ไม่เหมาะสมออกไปแล้ว เพราะถ้าหากเราเชื่อว่าเครื่องมือสร้างมา มีคุณภาพดี ความคลาดเคลื่อนจากตัวผู้ตอบน่าจะมากกว่าเครื่องมือ จึงควรคัดคนที่ไม่เหมาะสมออกก่อน

5.2 การวิเคราะห์เพื่อประมาณค่าพารามิเตอร์ของข้อกระทงและของผู้สอบ

การวิเคราะห์เพื่อประมาณค่าพารามิเตอร์ของข้อกระทงและของผู้ตอบจากการใช้แบบวัดประเภทมาตรประมาณค่า หรือแบบสอบที่ตรวจให้คะแนนแบบหลายค่า สามารถวิเคราะห์ได้โดยใช้โปรแกรมคอมพิวเตอร์ MULTILOG ซึ่งพัฒนาโดย ทิซเซน (Thissen) ในปี ค.ศ.1991 และโปรแกรม PARSCALE ที่พัฒนาโดย มูราคิ และบอค (Muraki and Bock) ในปี ค.ศ.1994 ซึ่งมีรายละเอียดโดยย่อดังนี้

5.2.1 โปรแกรม MULTILOG เป็นโปรแกรมที่ใช้สำหรับการวิเคราะห์แบบวัด หรือแบบสอบที่ตรวจให้คะแนนแบบพหุวิภาค สามารถใช้ได้กับผู้สอบและข้อกระทงจำนวนมาก เนื่องจากมีการสร้างไฟล์สำรองจำนวนมากในขณะโปรแกรมทำงาน หากมีข้อมูลมากเกินไป โปรแกรมจะเตือนและสามารถใช้คำสั่งย่อย > ESTimate BIG. ช่วยได้

โปรแกรมนี้สามารถใช้วิเคราะห์ตาม Graded Response Model (GRM), Bock Nominal Model (BNM) และมีโปรแกรมย่อยสำหรับวิเคราะห์ Logistic Model แบบ 1,2 และ 3 พารามิเตอร์สำหรับแบบสอบที่ตรวจให้คะแนนแบบทวิวิภาคได้ด้วย วิธีประมาณค่าพารามิเตอร์จะมีวิธีให้เลือก 2 วิธี คือ Marginal Maximum Likelihood (MML) และวิธี Fixed-θ Estimation ในการวิจัยครั้งนี้เลือกใช้วิธี MML ผลการวิเคราะห์จะให้ค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยากในแต่ละลำดับขั้นคะแนน (bjk) ค่าฟังก์ชันสารสนเทศรายข้อกระทง (IIF) และค่าฟังก์ชันสารสนเทศของแบบวัดทั้งฉบับ (TIF)

5.2.2 โปรแกรม PARSCALE เป็นโปรแกรมที่ใช้สำหรับการวิเคราะห์แบบวัด หรือ แบบสอบที่ตรวจให้คะแนนแบบพหุวิภาค โปรแกรมนี้สามารถใช้วิเคราะห์ตาม Graded Response Model (GRM) , Partial Credit Model (GRM) และ Generalized Partial Credit Model (GPCM) การประมาณค่าพารามิเตอร์ใช้วิธี Marginal Maximum Likelihood

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

(MML) ผลการวิเคราะห์ให้ค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยากในแต่ละลำดับขั้นคะแนน (bjk) ค่าฟังก์ชันสารสนเทศรายข้อกระทง (IIF) แต่การให้ค่าฟังก์ชันสารสนเทศเสนอเฉพาะโมเดล PCM และ GPCM เท่านั้น

การวิเคราะห์ผลในการวิจัยนี้จะเน้นการวิเคราะห์ค่าฟังก์ชันสารสนเทศจากโปรแกรม MULTILOG และ PARSCALE เป็นสำคัญ ดังนั้นผู้วิจัยจึงได้สรุปลักษณะและการนำโปรแกรมคอมพิวเตอร์ทั้งดังกล่าวไปใช้ได้ดังตารางที่ 2

ตารางที่ 2 ลักษณะของโปรแกรม MULTILOG และ PARSCALE

โปรแกรมคอมพิวเตอร์	การตรวจให้คะแนน	โมเดลที่สามารถวิเคราะห์	ผลจากการวิเคราะห์					
			ค่าพารามิเตอร์				ค่าฟังก์ชันสารสนเทศ	
			a_j	b_j	c_j	θ	IIF	TIF
1. MULTILOG (version 6.0)	1.1 ทวิวิภาค	Logistic1 PL	-	/	-	/	/	/
		2 PL	/	/	-	/	/	/
		3 PL	/	/	/	/	/	/
	1.2 พหุวิภาค	GRM	/	/	-	/	/	/
		GPCM	-	-	-	-	-	-
2.PARSCALE (version 2.2)	2.1 ทวิวิภาค	Logistic 1 PL	-	-	-	-	-	-
		2 PL	-	-	-	-	-	-
		3 PL	-	-	-	-	-	-
	2.2 พหุวิภาค	GRM	/	/	-	/	-	-
		GPCM	/	/	-	/	/	/

หมายเหตุ

/ หมายถึง ผลจากการวิเคราะห์ ที่โปรแกรมคอมพิวเตอร์สามารถทำได้

- หมายถึง ผลจากการวิเคราะห์ ที่โปรแกรมคอมพิวเตอร์ไม่สามารถทำได้

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

จากตารางที่ 2 จะเห็นว่าโปรแกรม MULTILOG สามารถวิเคราะห์ข้อมูลการตรวจให้คะแนนทั้งแบบทวิวิภาคและแบบพหุวิภาค คือสามารถวิเคราะห์ด้วยโมเดลโลจิสติก 1,2 และ 3 พารามิเตอร์ และ GRM ซึ่งการวิเคราะห์ดังกล่าวจะให้ค่าพารามิเตอร์ของผู้ตอบ ของข้อกระทง และค่าฟังก์ชันสารสนเทศของข้อกระทงและรวมทั้งฉบับ ส่วนโปรแกรม PARSCALE ไม่มีโปรแกรมย่อยสำหรับการวิเคราะห์การตรวจให้คะแนนแบบทวิวิภาค แต่สามารถวิเคราะห์การตรวจให้คะแนนแบบพหุวิภาคได้ทั้ง GRM และ GPCM แต่การวิเคราะห์ GRM จะไม่แสดงผลของค่าฟังก์ชันสารสนเทศ ดังนั้นในการวิจัยครั้งนี้ผู้วิจัยจึงจำเป็นต้องวิเคราะห์ค่าฟังก์ชันสารสนเทศของแบบวัดคุณลักษณะและแบบสอบ ซึ่งใช้วิธีการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคด้วยโปรแกรมคอมพิวเตอร์ MULTILOG และ PARSCALE

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย