

บทที่ 2

แนวความคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

2.1 แบบทดสอบปรับระดับขึ้น

แบบทดสอบปรับระดับขึ้น (Stradaptive Test) หรือมีชื่อเต็มเป็นภาษาอังกฤษว่า The Stratified Adaptive Test เป็นแบบทดสอบที่ใช้วัดหาระดับความสามารถของผู้สอบ โดยคะแนนที่ได้จากแบบทดสอบนี้มีความถูกต้องแน่นอนและมีความคงที่ แม้ว่าจะใช้ทดสอบข้ามกลุ่มผู้สอบที่ระดับความสามารถแตกต่างกันในช่วงกว้างก็ตาม แบบทดสอบปรับระดับขึ้นนี้เป็นรูปแบบหนึ่งของการทดสอบแบบเทเลอร์ กล่าวคือ ในขณะที่กำลังดำเนินการทดสอบจะมีการปรับระดับความยากของแบบทดสอบหรือมีการเลือกข้อสอบให้มีความยากเหมาะสมกับระดับความสามารถของผู้สอบตลอดเวลาโดยใช้กฎเกณฑ์ในการแยกทาง (Branching Rule) เพื่อนำผู้สอบไปทำข้อสอบที่เหมาะสมกับระดับความสามารถของผู้สอบ

การทดสอบแบบปรับระดับขึ้นนี้ ได้พัฒนามาจากแบบทดสอบ 2 ชนิดคือ Peaked Ability Test และ Binet's Testing Strategy (Weiss, 1973)

Peaked Ability Test คือแบบทดสอบความสามารถชนิดหนึ่ง ที่ข้อสอบทุกข้อมีระดับความยากใกล้เคียงกัน ในทางทฤษฎีนั้นข้อสอบแต่ละข้อของแบบทดสอบดังกล่าวควรมีระดับความยากเท่ากันหมดทุกข้อ ดังนั้นระดับความยากของข้อสอบในแบบทดสอบชนิดนี้จึงไม่มีความแปรปรวน (Variance) แต่ในทางปฏิบัตินั้น การทำข้อสอบในแบบทดสอบให้มีความยากเท่ากันหมดทั้งฉบับเป็นเรื่องที่ทำได้ยาก การสร้าง Peaked Ability Test จึงทำได้เพียงให้ข้อสอบมีความยากที่ไม่แตกต่างกันมากนัก นั่นคือทำให้ความยากของข้อสอบมีความแปรปรวนต่ำมาก ๆ เท่านั้น

Lord (อ้างถึงใน Weiss, 1973) กล่าวว่า Peaked Ability Test จะให้ประสิทธิภาพในการวัดดีกว่า Adaptive Test ที่มีจำนวนข้อสอบเท่ากัน แต่ประสิทธิภาพการวัดจะน้อยกว่า Adaptive Test เมื่อระดับความยากของ Peaked Ability Test ไม่สอดคล้องกับระดับความสามารถของผู้สอบ ดังนั้นการทดสอบโดยใช้ Peaked Ability Test จึงมีจุดอ่อนอยู่ที่การประมาณระดับความสามารถของผู้สอบก่อนที่จะเริ่มดำเนินการทดสอบ ซึ่งอาจมีการประเมินผิดพลาด ทำให้ไม่ได้สารสนเทศอะไรมากนักเกี่ยวกับความสามารถของผู้สอบและอาจมีผลต่อภาวะจิตใจของผู้สอบด้วย

Binet's Testing Strategy เป็นกระบวนการทดสอบที่ใช้วัดระดับสติปัญญา วัดสมรรถภาพทางสมองของมนุษย์ ซึ่งแบบทดสอบชนิดนี้ประกอบไปด้วยชุดของแบบทดสอบย่อย ๆ และในแต่ละชุดย่อย ๆ นั้น ก็จะมีระดับความยากที่พอเหมาะกับผู้สอบในระดับอายุต่าง ๆ กัน เช่นแบบทดสอบชุดใดก็ตามที่กลุ่มผู้สอบอายุ 8 ปี สามารถตอบถูกประมาณร้อยละ 50 ก็หมายความว่า แบบทดสอบชุดนั้น มีระดับความยากพอเหมาะกับกลุ่มผู้สอบอายุ 8 ปี แบบทดสอบชุดนี้จะมีลักษณะเช่นเดียวกับแบบทดสอบ Peaked Test คือข้อสอบแต่ละข้อในแบบทดสอบมีระดับความยากใกล้เคียงกัน และเนื่องจากว่าแบบทดสอบนี้ใช้วัดสมรรถภาพทางสมองหรือสติปัญญาของผู้สอบ จึงเรียกแบบทดสอบย่อยหรือแบบทดสอบที่เหมาะสมกับระดับอายุของผู้สอบนั้น ๆ ว่า Mental Age Test ดังนั้นแบบทดสอบของ Binet จึงประกอบไปด้วยชุดของข้อสอบย่อย ๆ ที่มีอายุสมองสอดคล้องกับระดับความสามารถทางสมองหรือสติปัญญาของผู้สอบส่วนมากในระดับอายุต่าง ๆ และเรียงลำดับชุดของข้อสอบตามลำดับอายุสมองจากน้อยไปมาก

ลักษณะกระบวนการทดสอบอย่างคร่าว ๆ ของแบบทดสอบ Binet นั้นจะดำเนินการดังนี้ คือให้พิจารณาอายุจริงของผู้สอบก่อน เช่นถ้าผู้สอบมีอายุจริง 8 ปี 1 เดือน ก็จัด Peaked Test ที่เหมาะกับผู้ที่อายุสมองเท่ากับ 8.0 ให้ผู้สอบนั้นทำ แต่ถ้ามีสารสนเทศก่อนสอบว่าผู้สอบนั้นมีสมรรถภาพทางสมองมากกว่าอายุจริง ก็อาจจะให้ผู้สอบเริ่มทำ Peaked Test ที่เหมาะกับผู้ที่อายุสมองสูงกว่า เช่น เท่ากับ 9.0 ก็ได้ หรือถ้าทราบมาก่อนว่าผู้สอบมีสติปัญญาต่ำก็อาจจะให้ผู้สอบเริ่มทำ Peaked Test ที่เหมาะกับผู้ที่อายุสมองต่ำกว่า 8.0 ก็ได้ เช่น อายุสมองเท่ากับ 6.5 เป็นต้น จากนั้นให้ผู้สอบทำ Peaked Test ที่ลดอายุสมองต่ำลงมาเรื่อย ๆ จนพบ Peaked Test ที่ผู้สอบสามารถตอบถูกทั้งหมด เราจะเรียก Peaked Test นั้นว่า ฐานอายุ (Basal Age) ของผู้สอบ แล้วให้ผู้สอบทำ Peaked Test ที่เพิ่มอายุสมองสูงขึ้นเรื่อย ๆ จนพบ Peaked Test ที่ผู้สอบตอบผิดทั้งหมด เรียก Peaked Test นั้นว่า เพดานอายุ (Ceiling Age) ของผู้สอบ ซึ่งทั้ง ฐานอายุ และเพดานอายุนั้น จะไม่ให้สารสนเทศเกี่ยวกับสมรรถภาพทางสมองของผู้สอบเลย แต่ผลการตอบแบบทดสอบที่อยู่ระหว่างฐานอายุและเพดานอายุของผู้สอบนั้นจะให้สารสนเทศที่มีความถูกต้องแน่นอน ในการวัดระดับสติปัญญาของผู้สอบแต่ละคน ซึ่งการประมาณระดับสติปัญญา โดยวิธีนี้จะเชื่อถือได้มากแค่ไหนก็ขึ้นอยู่กับจำนวน Peaked Test ที่อยู่ระหว่างฐานอายุและเพดานอายุของผู้สอบแต่ละคน ถ้ามีจำนวน Peaked Test น้อยการประมาณระดับสติปัญญาของผู้สอบก็จะเชื่อถือได้มาก แต่ถ้ามีจำนวน Peaked Test มากการประมาณระดับสติปัญญาของผู้สอบก็จะเชื่อถือได้น้อย และยังแสดงถึงความสามารถทางสมองที่ไม่คงเส้นคงวาของผู้สอบอีกด้วย

กระบวนการดำเนินการสอบแบบปรับระดับขึ้น ก็มีลักษณะคล้ายคลึงกับการทดสอบ Binet ซึ่งเริ่มจากกลุ่มของข้อสอบขนาดใหญ่ ที่แบ่งกันเป็นชั้น ๆ ตามระดับความยากของข้อสอบ แต่ละชั้นจะ

ประกอบไปด้วยกลุ่มของข้อสอบที่มีระดับความยากใกล้เคียงกัน ผู้สอบแต่ละคนจะเริ่มทำข้อสอบที่มีความยากซึ่งประมาณแล้วว่าสอดคล้องกับระดับความสามารถของผู้สอบ จากนั้นก็ใช้ผลการตอบข้อสอบข้อแรกและกฎการแยกทางที่กำหนดขึ้น ดำเนินการทดสอบไปเรื่อย ๆ จนกระทั่งพบ ข้อที่ง่ายที่สุด หรือ ข้อที่ยากที่สุดของผู้สอบ ซึ่งเป็นข้อที่ผู้สอบตอบคำถามถูกต้อง (หรือตอบถูกด้วยสัดส่วนตั้งแต่ 0.80 ขึ้นไป) หรือตอบคำถามผิดหมด (หรือตอบถูกด้วยสัดส่วนตั้งแต่ 0.20 ลงมา) ทุกข้อที่ได้รับตามลำดับ หรือพบเกณฑ์ในการยุติการสอบที่กำหนดไว้ การทดสอบของผู้สอบคนนั้นก็จะหยุดลง ผลการตอบข้อสอบที่อยู่ระหว่าง ข้อที่ง่ายที่สุดและข้อที่ยากที่สุดของผู้สอบ จะให้สารสนเทศมากที่สุดเกี่ยวกับระดับความสามารถของผู้สอบ ซึ่งจากลักษณะการเริ่มทำข้อสอบของผู้สอบแต่ละคน ลักษณะการแยกทางไปทำข้อสอบข้อต่อไป โดยการพิจารณาผลการตอบข้อสอบแต่ละข้อก่อน และเกณฑ์ในการยุติการทดสอบดังกล่าวของการทดสอบแบบปรับระดับขึ้นนั้น ทำให้ผู้สอบแต่ละคนได้รับข้อสอบที่ไม่เหมือนกัน จำนวนข้อสอบไม่เท่ากัน และเวลาที่ใช้ในการทดสอบก็ไม่เท่ากันด้วย แต่ชุดของข้อสอบที่ผู้สอบได้รับนั้นจะมีความยากที่เหมาะสมกับความสามารถของผู้สอบ ซึ่งลักษณะของการทดสอบแบบปรับระดับขึ้นแต่ละอย่างนี้จะได้อธิบายในรายละเอียดต่อไป

2.1.1 โครงสร้างของกลุ่มข้อสอบ (Item Pool Structure)

แบบทดสอบปรับระดับขึ้น จะต้องมียุทธศาสตร์ของข้อสอบขนาดใหญ่ ที่แบ่งกันเป็นชั้น ๆ โดยใช้ระดับความยากของข้อสอบเป็นเกณฑ์แบ่ง และข้อสอบทุกข้อในกลุ่มนั้นจะต้องใช้วัดความสามารถด้านเดียวกันของผู้สอบ ขั้นตอนในการสร้างกลุ่มของข้อสอบ มีดังนี้

2.1.1.1 จะต้องมีข้อสอบจำนวนมากที่ใช้วัดความสามารถด้านเดียวกัน และนำไปทดลองใช้กับกลุ่มตัวอย่างขนาดใหญ่ ซึ่งเป็นตัวแทนที่ดีของประชากร กลุ่มตัวอย่างมีขนาดใหญ่มากเท่าไรก็จะทำให้ค่าพารามิเตอร์ของข้อสอบ (ค่าความยากและค่าอำนาจการจำแนก) มีความคงที่มากขึ้นเท่านั้น

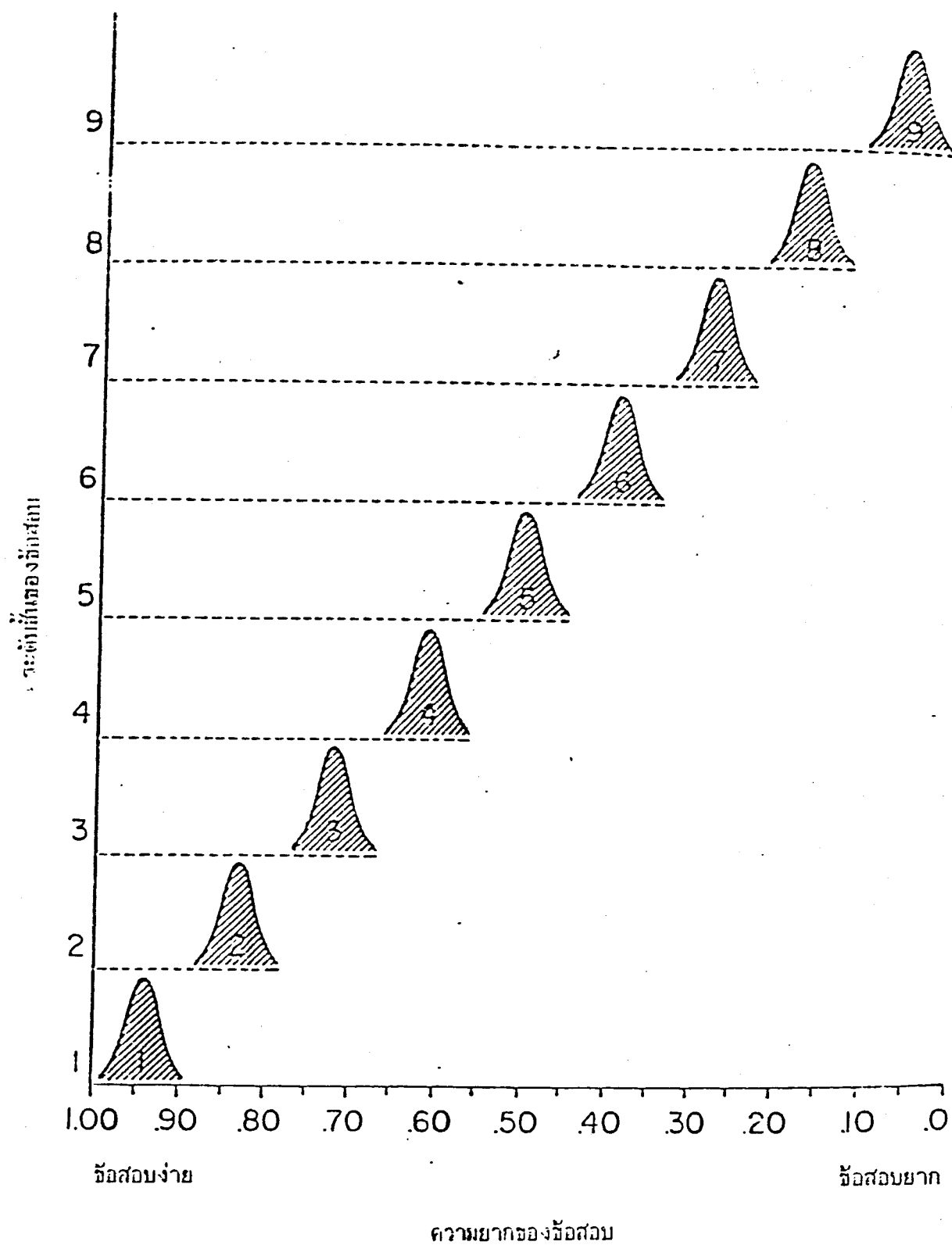
2.1.1.2 นำผลการทดลองสอบกับกลุ่มตัวอย่างนั้นไปทำการวิเคราะห์เพื่อหาค่าพารามิเตอร์ของข้อสอบ ซึ่งอาจจะใช้วิธีการแบบดั้งเดิม คือ เทคนิค 27 % หรือจะใช้ทฤษฎีคุณลักษณะแฝง (Latent Trait Theory) ก็ได้ แต่ถ้าใช้ทฤษฎีคุณลักษณะแฝง จะให้การประมาณค่าพารามิเตอร์ของข้อสอบที่คงที่กว่า เพราะอธิบายความหมายของค่าพารามิเตอร์ต่าง ๆ ด้วยทฤษฎีความน่าจะเป็นไปได้ของผู้สอบในระดับความสามารถต่าง ๆ แต่การวิเคราะห์แบบดั้งเดิมนั้น อธิบายความหมายของค่าพารามิเตอร์ในรูปของสัดส่วน เช่นค่าความยาก คือสัดส่วนของผู้สอบที่ตอบถูกกับผู้สอบทั้งหมด ซึ่งค่าพารามิเตอร์ดังกล่าวจะแปรเปลี่ยนไปตามลักษณะของกลุ่มตัวอย่าง

2.1.1.3 จัดกลุ่มของข้อสอบโดยให้แต่ละกลุ่มมีระดับความยากใกล้เคียงกันและเรียงลำดับกลุ่มข้อสอบเป็นชั้น ๆ (Strata) ตามระดับความยากของข้อสอบ จากชั้นที่ง่ายที่สุดไปหาชั้นที่ยากที่สุด โดยแต่ละชั้นจะมีระดับความยากที่ไม่มีส่วนคาบเกี่ยวกัน (No-Overlap) ระหว่างชั้น จำนวนชั้นของข้อสอบจะต้องมีมากพอที่จะครอบคลุมช่วงระดับความสามารถของผู้สอบ ซึ่งจำนวนชั้นอย่างต่ำควรจะมีประมาณ 9 หรือ 10 ชั้น จำนวนข้อสอบในแต่ละชั้นจะแปรผันไปตามขนาดของกลุ่มข้อสอบ และระดับความยากของข้อสอบ และในแต่ละชั้นก็ไม่จำเป็นต้องมีจำนวนข้อสอบเท่ากัน แต่อย่างน้อยควรมี 10 ข้อ และอย่างมากควรมีประมาณ 50 ข้อ (Weiss, 1973) โดยเฉพาะชั้นที่มีระดับความยากปานกลาง และชั้นที่มีระดับความยากสูงควรมีจำนวนข้อสอบมาก นอกจากนี้ในการศึกษาเกี่ยวกับแบบทดสอบปรับระดับชั้นบางครั้งอาจมีการกำหนดให้แต่ละชั้นมีจำนวนข้อสอบเท่ากันก็ได้ เช่นการศึกษาของ Vale and Weiss (1973)

2.1.1.4 ในแต่ละชั้นให้จัดข้อสอบเรียงลำดับกันตามค่าอำนาจจำแนกจากค่ามากไปหาน้อย ข้อสอบจะต้องสามารถจำแนกผู้สอบได้เป็นอย่างดี นั่นคือข้อสอบแต่ละข้อควรมีค่าอำนาจจำแนกไม่ต่ำกว่า 0.30 (Weiss, 1973) แต่ในการศึกษาบางครั้งอาจกำหนดให้แบบทดสอบปรับระดับชั้นทั้งฉบับมีค่าอำนาจจำแนกเท่ากันก็ได้ (Vale and Weiss, 1973)

ผลจากการบวนการสร้างโครงสร้างของกลุ่มข้อสอบดังกล่าวอาจแสดงได้ดังภาพที่ 2 ซึ่งเป็นภาพสมมติของกลุ่มข้อสอบแบบปรับระดับชั้น วิเคราะห์ค่าพารามิเตอร์โดยใช้วิธีการแบบดั้งเดิมโดยที่แบบทดสอบประกอบไปด้วยกลุ่มข้อสอบย่อย ๆ 9 ชั้น ในแต่ละชั้นมีข้อสอบที่มีความยากใกล้เคียงกัน และในชั้นหนึ่ง ๆ จะมีความยากโดยเฉลี่ยเรียงลำดับจากชั้นที่ง่ายไปหาชั้นยาก ตัวอย่างจากภาพที่ 2 ชั้นที่ 1 จะมีความยากของข้อสอบโดยเฉลี่ยเท่ากับ 0.94 แต่ช่วงความยากของข้อสอบทั้งหมดในชั้นนี้จะมีค่าความยากตั้งแต่ 0.99 ถึง 0.89 และในชั้นที่ 2 ซึ่งมีความยากมากขึ้นไปอีก คือมีความยากโดยเฉลี่ยเท่ากับ 0.83 แต่ช่วงความยากของข้อสอบทั้งหมดในชั้นที่ 2 มีค่าตั้งแต่ 0.88 ถึง 0.78 สำหรับชั้นที่ 9 ช่วงความยากของข้อสอบทั้งหมดมีค่าตั้งแต่ 0.01 ถึง 0.11 และมีค่าความยากของข้อสอบโดยเฉลี่ยในชั้นนี้เท่ากับ 0.06 และจากแผนภาพที่แสดงให้เห็นนั้น การแจกแจงของข้อสอบตามความยากจะไม่มีส่วนที่คาบเกี่ยวกันระหว่างชั้น

ส่วนในตารางที่ 1 นั้น เป็นโครงสร้างของกลุ่มข้อสอบ ที่ประมาณค่าพารามิเตอร์ของข้อสอบโดยใช้ทฤษฎีคุณลักษณะแฝง โครงสร้างของกลุ่มข้อสอบแบบปรับระดับชั้นตามตัวอย่างในตารางนี้เป็นแบบทดสอบวัดความสามารถทางภาษา (Verbal Ability) จำนวน 229 ข้อ จัดเป็นกลุ่ม ๆ 9 กลุ่ม โดยให้แต่ละกลุ่มมีความยากของข้อสอบใกล้เคียงกัน เรียงกันเป็นชั้น ๆ ตามระดับความยากจากน้อยไปมาก จำนวนข้อสอบในแต่ละชั้นมีตั้งแต่ 10 ข้อ คือในชั้นที่ 9 ไปจนถึง 36 ข้อ ดังในชั้นที่



ภาพที่ 2 แสดงการจัดกลุ่มข้อสอบตามค่าความยากของข้อสอบ ที่อธิบายความยากของข้อสอบ ด้วยสัดส่วนของการตอบถูก (P) (จาก Weiss, 1973)

ตารางที่ 1 แสดงค่าความยาก (b) และค่าอำนาจจำแนก (a) ของข้อสอบที่วิเคราะห์โดยใช้
Latent Trait Theory และจัดรูปแบบข้อสอบตามลักษณะของแบบทดสอบ
ปรับระดับขึ้น (จาก Weiss, 1973)

	(easy)				Stratum								(difficult)						
	1	2	3	4	5	6	7	8	9										
Item Difficulties	-2.39	-1.64	-1.01	-0.34	0.33	.98	1.63	2.31	3.11										
H ₁	-2.98	-2.32	-1.63	-1.00	-0.28	.74	1.00	1.65	2.32										
L ₀	-2.65	-1.92	-1.29	-0.63	.02	.65	1.33	2.01	2.62										
Mean																			
No. of Items	35	36	36	30	25	19	23	15	10										
Item Number Within Stratum																			
	h	n	h	n	h	n	h	n	h	n	h	n	h	n	h	n	h	n	h
1	-2.42	3.00	-1.97	1.76	-1.51	1.40	-.70	1.82	-.05	1.31	.73	.98	1.07	.72	1.89	.83	2.45	.84	
2	-2.42	3.00	-1.78	1.54	-1.23	1.35	-.73	.72	.14	1.07	.74	.91	1.49	.62	2.03	.64	2.47	.48	
3	-2.45	3.00	-2.22	1.52	-1.08	1.23	-.52	.86	-.13	.98	.65	.77	1.33	.60	1.93	.57	2.61	.43	
4	-2.45	3.00	-1.68	1.46	-1.33	1.16	-.68	.86	.15	.97	.79	.70	1.54	.58	2.31	.54	2.86	.42	
5	-2.72	3.00	-1.87	1.43	-1.34	1.02	-.59	.83	-.08	.91	.79	.67	1.11	.56	1.79	.50	2.35	.42	
6	-2.72	3.00	-1.92	1.23	-1.10	.99	-.75	.82	.16	.86	.49	.56	1.40	.55	2.04	.49	2.67	.42	
7	-2.72	3.00	-1.08	1.14	-1.42	.92	-.57	.77	-.21	.86	.42	.55	1.17	.52	1.79	.49	2.32	.38	
8	-2.66	1.79	-2.17	1.10	-1.21	.91	-.85	.75	-.25	.86	.98	.52	1.30	.52	1.88	.45	2.37	.34	
9	-2.54	1.59	-1.64	1.08	-1.06	.89	-.47	.71	.21	.86	.37	.50	1.38	.51	2.07	.43	2.11	.33	
10	-2.81	1.48	-2.22	1.07	-1.34	.89	-.40	.68	.16	.83	.46	.49	1.44	.49	2.13	.42	2.50	.31	
11	-2.46	1.29	-1.67	1.02	-1.31	.87	-.90	.67	-.23	.81	.46	.48	1.31	.44	2.31	.40			
12	-2.78	1.26	-1.71	.99	-1.10	.77	-1.00	.67	.30	.78	.65	.48	1.25	.43	1.65	.39			
13	-2.47	1.16	-2.26	.98	-1.55	.77	-.69	.66	.08	.76	.78	.45	1.00	.42	1.82	.36			
14	-2.43	1.01	-2.21	.96	-1.07	.76	-.81	.66	-.28	.75	.71	.44	1.00	.40	2.26	.35			
15	-2.86	1.01	-1.66	.93	-1.43	.76	-.56	.66	.24	.66	.65	.43	1.26	.39	2.18	.34			
16	-2.94	.96	-1.65	.92	-1.40	.75	-.58	.66	.33	.53	.62	.41	1.36	.37					
17	-2.83	.94	-1.65	.82	-1.15	.73	-.84	.65	-.23	.43	.83	.37	1.24	.36					
18	-2.74	.93	-2.32	.80	-1.42	.71	-.85	.64	.09	.43	.75	.37	1.60	.35					
19	-2.89	.91	-1.80	.77	-1.63	.71	-.41	.27	.15	.42	.92	.37	1.21	.35					
20	-2.54	.88	-1.80	.76	-1.47	.67	-.94	.60	-.09	.41			1.47	.34					
21	-2.55	.79	-1.93	.74	-1.60	.66	-.41	.59	-.26	.40			1.61	.34					
22	-2.81	.74	-2.28	.70	-1.33	.62	-.89	.53	.08	.39			1.63	.32					
23	-2.50	.68	-1.83	.66	-1.04	.58	-.52	.48	.09	.37			1.36	.31					
24	-2.82	.67	-1.74	.63	-1.17	.57	-.58	.48	-.04	.35									
25	-2.54	.67	-1.70	.59	-1.27	.56	-.39	.40	.12	.32									
26	-2.50	.66	-2.19	.56	-1.07	.56	-.36	.40											
27	-2.51	.64	-1.89	.52	-1.02	.54	-.58	.40											
28	-2.39	.62	-2.20	.50	-1.01	.52	-.38	.38											
29	-2.58	.57	-1.71	.47	-1.31	.52	-.34	.32											
30	-2.98	.56	-2.21	.44	-1.30	.52	-.67	.30											
31	-2.73	.52	-2.08	.42	-1.19	.52													
32	-2.77	.50	-1.80	.42	-1.57	.49													
33	-2.68	.48	-1.82	.42	-1.26	.44													
34	-2.56	.44	-2.12	.41	-1.59	.38													
35	-2.95	.41	-1.92	.32	-1.35	.34													
36			-1.84	.31	-1.08	.32													

2 และ 3 ข้อสอบแต่ละชั้นดังกล่าวถูกเลือกมาจากกลุ่มข้อสอบขนาดใหญ่ ซึ่งทุกข้อใช้วัดความสามารถด้านเดียวกัน จำนวน 500 ข้อ และผ่านการวิเคราะห์ประมาณค่าพารามิเตอร์ของข้อสอบ ตามรูปแบบของ Normal Ogive มีค่าอำนาจจำแนกและค่าความยากแทนด้วยสัญลักษณ์ a และ b ตามลำดับ วิธีการประมาณค่าพารามิเตอร์ของข้อสอบดังกล่าว เสนอโดย Lord และ Novick (1968) ค่าความยากของข้อสอบส่วนใหญ่จะมีช่วงตั้งแต่ -3.00 ไปจนถึง 3.00 แบ่งช่วงความยากของข้อสอบทั้งหมดออกเป็น 9 ช่วงย่อย ๆ มีขนาดความกว้างของแต่ละช่วงเท่า ๆ กัน ส่วนค่าอำนาจจำแนกจะใช้ตั้งแต่ 0.30 ขึ้นไป แต่สำหรับตัวอย่างนี้จะยกเว้นในชั้นที่ 9 ซึ่งต้องการให้มีข้อสอบ 10 ข้อ จึงยอมให้บางข้อมีค่าอำนาจจำแนกต่ำกว่า 0.30 คือมีค่าเท่ากับ 0.21

กลุ่มข้อสอบทั้ง 9 ชั้นในตารางที่ 1 นั้น มีค่าความยากโดยเฉลี่ยของแต่ละชั้นตั้งแต่ -2.65 ถึง 2.62 ชั้นที่มีข้อสอบยากที่สุดคือชั้นที่ 9 มีข้อสอบทั้งหมด 10 ข้อ ความยากโดยเฉลี่ยเท่ากับ 2.62 และความยากของข้อสอบในชั้นนี้มีช่วงตั้งแต่ 2.32 ถึง 3.11 สำหรับในชั้นที่ 8 ซึ่งเป็นชั้นที่มีความยากรองลงมา คือมีความยากเฉลี่ยเท่ากับ 2.01 มีข้อสอบจำนวน 15 ข้อ ความยากของข้อสอบทั้งหมดในชั้นที่ 8 มีค่าตั้งแต่ 1.65 ถึง 2.31 ในแต่ละชั้นจะเรียงลำดับข้อสอบตามค่าอำนาจจำแนกจากมากไปน้อย เช่นในชั้นที่ 9 ข้อสอบข้อที่ 1 มีค่า $a = 0.84$ และข้อต่อ ๆ ไปจะมีค่าอำนาจจำแนกลดลง จนถึงข้อที่ 10 ซึ่งมีค่า $a = 0.21$

2.1.2 วิธีดำเนินการทดสอบ

การทดสอบแบบปรับระดับขึ้น เป็นรูปแบบหนึ่งของการทดสอบเทเลอร์ คือมีการปรับระดับความยากของแบบทดสอบให้เหมาะสมกับระดับความสามารถของผู้สอบแต่ละคน ซึ่งเป็นหลักการใหญ่ของการทดสอบแบบเทเลอร์ ในแต่ละรูปแบบของการทดสอบแบบเทเลอร์ก็จะมีรายละเอียดในการดำเนินการทดสอบแตกต่างกันไป สำหรับการทดสอบปรับระดับขึ้นก็มีรายละเอียดในการดำเนินการสอบที่เป็นรูปแบบเฉพาะของตนเองดังนี้

2.1.2.1 การเริ่มต้นเข้าทำข้อสอบ (Entry Point)

การดำเนินการสอบของแบบทดสอบปรับระดับขึ้น มีการเริ่มต้นเข้าทำข้อสอบที่แปลกไปจากวิธีการของแบบทดสอบเทเลอร์ชนิดอื่น ๆ คือผู้สอบแต่ละคนจะเริ่มต้นเข้าทำข้อสอบข้อแรกที่ระดับขึ้นแตกต่างกัน ตามระดับความสามารถของผู้สอบ แต่ก็ไม่จำเป็นมากนักสำหรับการเริ่มต้นที่แตกต่างกันนี้ คือผู้สอบทุกคนอาจจะเริ่มต้นทำข้อสอบข้อเดียวกันก็ได้ แต่การเริ่มต้นทำข้อสอบในชั้นที่เหมาะสมกับระดับความสามารถของผู้สอบแต่ละคนนั้น มีข้ออย่างน้อย 2 ประการคือ (Weiss, 1973)

ประการที่ 1 การเริ่มต้นทำข้อสอบในขั้นที่แตกต่างกันตามระดับความสามารถของแต่ละบุคคลนั้น จะช่วยประหยัดเวลาในการทดสอบ และใช้ข้อสอบจำนวนน้อย คือผู้สอบไม่ต้องไปเสียเวลาทำข้อสอบที่ยากเกินไปหรือง่ายเกินไปสำหรับตนเอง ซึ่งข้อสอบเหล่านี้จะไม่ให้สารสนเทศเกี่ยวกับความสามารถของผู้สอบ

ประการที่ 2 ทำให้ผู้สอบเกิดแรงจูงใจอย่างต่อเนื่อง มีความตั้งใจทำข้อสอบและแสดงความสามารถที่แท้จริงออกมาอย่างเต็มที่ ผลการทดสอบครั้งนี้ก็จะได้สารสนเทศเกี่ยวกับความสามารถของผู้สอบอย่างมาก

ข้อดีทั้ง 2 ประการดังกล่าวนี้จะปรากฏออกมาอย่างน่าพอใจก็ต่อเมื่อผู้สอบได้ เริ่มต้นทำข้อสอบในขั้นที่มีความยากสอดคล้องกับระดับความสามารถของผู้สอบ แหล่งข้อมูลที่เราใช้ในการประมาณหาจุดเริ่มต้นเข้าทำข้อสอบมีดังนี้

แหล่งที่ 1 ใช้ผลการเรียนเดิมหรือระดับผลการเรียนเฉลี่ยของวิชาต่าง ๆ ที่สอดคล้องกับวิชาในแบบทดสอบปรับระดับขึ้น เป็นข้อมูลในการประมาณจุดเริ่มต้นในการเข้าทำข้อสอบ

แหล่งที่ 2 ใช้แบบทดสอบดั้งเดิม ให้ผู้สอบทุกคนทำก่อน แล้วนำผลการตอบแบบทดสอบนั้นไปประมาณจุดเริ่มต้นเข้าทำข้อสอบในแบบทดสอบปรับระดับขึ้น

ในภาพที่ 3 แสดงแนวทางในการประมาณจุดเริ่มต้นเข้าทำข้อสอบในแบบทดสอบปรับระดับขึ้น 2 แนวทาง สำหรับการทดสอบแบบปรับระดับขึ้นที่ใช้วัดความสามารถทางภาษา ตามโครงสร้างของกลุ่มข้อสอบในตารางที่ 1 ส่วนหนึ่งของภาพที่ 3 เป็นการประมาณจุดเริ่มต้นของการเข้าทำข้อสอบแบบปรับระดับขึ้น ที่ใช้ระดับผลการเรียนเฉลี่ยของวิชาที่เกี่ยวข้องกับความสามารถทางภาษา ซึ่งแบ่งช่วงของระดับผลการเรียนเฉลี่ยออกเป็น 9 ช่วง เท่ากับจำนวนขั้นในแบบทดสอบปรับระดับขึ้น ถ้าผู้สอบคนใดมีระดับผลการเรียนเฉลี่ยอยู่ในช่วง 3.76 ถึง 4.00 จะเริ่มทำข้อสอบข้อแรกในขั้นที่ 9 หรือถ้าผู้สอบคนใดมีระดับผลการเรียนเฉลี่ยอยู่ในช่วง 2.51 ถึง 2.75 ก็จะเริ่มทำข้อสอบข้อแรกในขั้นที่ 4 เป็นต้น

ส่วนล่างของภาพที่ 3 แสดงความแตกต่างของการเริ่มเข้าทำข้อสอบอีกแบบหนึ่ง โดยใช้ผู้สอบตอบแบบทดสอบชนิดดั้งเดิมฉบับเดียวกันหมด ผลการทดสอบของผู้สอบแต่ละคนจะใช้ในการประมาณจุดเริ่มต้นเข้าทำข้อสอบ

2.1.2.2 การแยกทางไปทำข้อสอบข้อต่อไป (Branching)

ในการดำเนินการทดสอบแบบปรับระดับขึ้น การแยกทางไปทำข้อสอบข้อต่อไปหลังจากตอบคำถามของข้อสอบข้อหนึ่ง ๆ เสร็จแล้วนั้น นับเป็นกฎเกณฑ์เฉพาะอย่างของการทดสอบแบบนี้ ซึ่งมีความสำคัญและจำเป็นมาก เพราะวิธีการนี้จะใช้เป็นกระบวนการค้นหาระดับความสามารถที่แท้จริง

การประมาณจุดเริ่มต้นในการเข้าทำข้อสอบโดยใช้ระดับผลการเรียนเฉลี่ยของวิชาที่เกี่ยวข้อง

	ช่วงของระดับผลการเรียนเฉลี่ย	ระดับชั้นที่ เริ่มต้นทำข้อสอบ
1.	3.76 - 4.00	9
2.	3.51 - 3.75	8
3.	3.26 - 3.50	7
4.	3.01 - 3.25	6
5.	2.76 - 3.00	5
6.	2.51 - 2.75	4
7.	2.26 - 2.50	3
8.	2.01 - 2.25	2
9.	2.00 หรือต่ำกว่า	1

การประมาณจุดเริ่มต้นเข้าทำข้อสอบจากการตอบแบบทดสอบดั้งเดิม

ถ้าตอบข้อสอบถูก :	1 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 1
	2 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 2
	3 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 3
	4 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 4
	5 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 5
	6 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 6
	7 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 7
	8 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 8
	9 ใน 10 ข้อ จุดเริ่มต้นจะอยู่ในชั้นที่ 9

ภาพที่ 3 การประมาณจุดเริ่มต้นเข้าทำข้อสอบในแบบทดสอบปรับระดับชั้น
(ดัดแปลงจาก Weiss, 1973)

ของผู้สอบ หรือ เป็นการปรับระดับความยากของชุดข้อสอบที่ผู้สอบได้รับ ให้เหมาะสมกับระดับความสามารถของผู้สอบตนเอง ลักษณะของกฎการแยกทางของแบบทดสอบปรับระดับขึ้นคือ เมื่อผู้สอบตอบข้อสอบข้อแรกของเขาถูกต้อง เขาก็จะได้ข้อสอบข้อต่อไปที่มีความยากขึ้นซึ่งอยู่ในชั้นที่สูงถัดขึ้นไป แต่ถ้าเขาตอบข้อสอบข้อแรกผิด ข้อสอบข้อต่อไปของเขาจะง่ายลง ซึ่งอยู่ในชั้นต่ำลงมา และจะใช้วิธีการแยกทางทำนองนี้ในการปรับความยากของชุดข้อสอบ ให้เหมาะสมกับระดับความสามารถของผู้สอบไปเรื่อย ๆ จนกว่าจะถึงเกณฑ์ในการหยุดทำข้อสอบ

กฎการแยกทางอย่างง่าย ๆ คือ ขึ้นหนึ่งชั้น/ลงหนึ่งชั้น (Up-one/Down-one) การแยกทางในลักษณะนี้ ผู้สอบจะถูกปรับระดับขึ้นของข้อสอบที่จะได้รับขึ้นหรือลงครั้งละ 1 ชั้นเท่านั้น เช่น สมมติว่า ผู้สอบกำลังทำข้อสอบในชั้นที่ 5 อยู่ และถ้าผู้สอบตอบข้อสอบที่กำลังทำอยู่ได้ถูกต้อง ก็จะได้รับข้อสอบในชั้นที่ 6 เป็นข้อต่อไป แต่ถ้าตอบผิดข้อสอบที่จะได้รับต่อไปจะอยู่ในชั้นที่ 4 เป็นต้น ซึ่งวิธีการนี้เหมาะสำหรับแบบทดสอบปรับระดับขึ้นที่มีช่วงความยากระหว่างชั้นกว้าง แต่ถ้ามีช่วงความยากของข้อสอบระหว่างชั้นแคบและต้องการความเร็วในการค้นหา ชั้นต่ำสุดและชั้นสูงสุดของผู้สอบ ก็อาจใช้กฎการแยกทางแบบ ขึ้นสองชั้น/ลงสองชั้น (Up-two/Down-two)

สำหรับผู้สอบบางคนที่มีความสามารถตรงกับชั้นสูงสุดหรือชั้นต่ำสุดของแบบทดสอบปรับระดับขึ้น หรือมีความสามารถใกล้เคียงกับชั้นสูงสุดหรือชั้นต่ำสุดของแบบทดสอบ ซึ่งไม่สามารถนำข้อสอบที่มีค่าความยากที่สูงกว่าหรือต่ำกว่าค่าความยากในแบบทดสอบมาให้ผู้สอบทำต่อไป ในกรณีนี้ผู้สอบทำข้อสอบในชั้นที่ 9 ถูก หรือทำผิดในชั้นที่ 1 ในกรณีเช่นนี้ จะต้องดำเนินการทดสอบต่อไป โดยให้ผู้สอบทำข้อสอบในระดับชั้นเดิมคือชั้นสูงสุดหรือชั้นต่ำสุดของแบบทดสอบนั้น ๆ ต่อไป

2.1.2.3 การยุติการทดสอบ (Termination)

ลักษณะเด่นของการทดสอบแบบปรับระดับขึ้นอีกอย่างหนึ่งคือ การยุติการทดสอบของผู้สอบแต่ละคน เมื่อเปรียบเทียบความแตกต่างกับการทดสอบแบบเทเลอร์อื่น ๆ เช่น การทดสอบแบบยูลีวีสองชั้น การทดสอบแบบรูปปริมาตร และการทดสอบแบบเฟลิกซ์เลเวล ซึ่งแบบทดสอบเหล่านี้กำหนดจำนวนข้อสอบที่คงที่ไว้ก่อน และดำเนินการทดสอบต่อไปจนกว่าจะหมดข้อสอบ แต่การทดสอบแบบปรับระดับขึ้นนี้ผู้สอบแต่ละคนไม่จำเป็นต้องได้ข้อสอบจำนวนเท่ากัน สำหรับ Bayesian Adaptive Testing Strategy ของ Urry (1969, 1970 อ้างถึงใน Weiss, 1973) และ Maximum Likelihood Testing Strategy ของ Urry (1970, อ้างถึงใน Weiss, 1973) ก็ยอมให้ผู้สอบแต่ละคนได้รับจำนวนข้อสอบที่แตกต่างกันเช่นกัน แต่ก็ยังมีข้อตกลงเกี่ยวกับรูปร่างการกระจายของระดับความสามารถของกลุ่มผู้สอบ ซึ่งต้องเป็นไปตามสมมติฐานบางอย่าง และผู้ดำเนินการสอบต้องมีความชำนาญในการคำนวณทางคณิตศาสตร์ แต่การทดสอบแบบปรับระดับขึ้นนี้ จำนวนข้อสอบของแต่ละคนที่ได้รับแตกต่างกัน ไม่ต้องมีข้อตกลงเกี่ยวกับรูปร่างการกระจายของระดับความ

สามารถของผู้สอบ และไม่ต้องอาศัยการคำนวณที่ยุ่งยาก

จากลักษณะเดิ ดังกล่าวข้างต้น กระบวนการทดสอบแบบปรับระดับขึ้น ก็จะรักษารั้งของข้อสอบที่เหมาะสมที่สุดสำหรับผู้สอบแต่ละคน โดยใช้กฎการแยกทางดังกล่าว ขึ้นของข้อสอบเหล่านี้จะให้การสแกนที่สูงที่สุดเกี่ยวกับระดับความสามารถของผู้สอบ ขึ้นของข้อสอบที่เหมาะสมกับผู้สอบคนใดก็หมายความว่าผู้สอบคนนั้นสามารถตอบข้อสอบเหล่านี้ได้ประมาณร้อยละ 50 ของข้อสอบที่ได้รับ (Weiss, 1973) ส่วนขึ้นของข้อสอบที่ไม่เหมาะสม ก็คือขึ้นที่ผู้สอบตอบถูกหรือตอบผิดหมด (หรือสัดส่วนในการตอบถูกน้อยกว่า 0.20 โดยทำข้อสอบในขั้นนั้นแล้วไม่ต่ำกว่า 5 ข้อ)

สำหรับเกณฑ์ในการยุติการทดสอบนั้น ผู้ดำเนินการสอบอาจกำหนดขึ้นเองได้ ซึ่งมีอยู่ 2 แนวทาง คือ

1) โดยการกำหนดจำนวนข้อสอบที่ผู้สอบได้รับใน ขึ้นสูงสุดของผู้สอบ เช่น ถ้ากำหนดไว้ 5 ข้อ เมื่อผู้สอบตอบผิดหมดทั้ง 5 ข้อ ก็ให้หยุดทำข้อสอบ

2) โดยการกำหนดสัดส่วนในการตอบถูกในแต่ละขั้นที่ทำข้อสอบ และจะต้องกำหนดจำนวนข้อสอบในแต่ละขั้นนั้นด้วยว่าจะต้องทำข้อสอบอย่างน้อยกี่ข้อ เช่นถ้ากำหนดสัดส่วนในการตอบถูกของแต่ละขั้นไม่ต่ำกว่า 0.20 และต้องทำข้อสอบอย่างน้อย 5 ข้อในแต่ละขั้น (ทั้งนี้ต้องไม่รวม ขึ้นสูงสุดและ ขึ้นต่ำสุดของผู้สอบ) การทดสอบก็จะเสร็จสิ้นลง (Weiss, 1973)

2.1.2.4 การให้คะแนน (Scoring)

การทดสอบแบบปรับระดับขึ้นนั้น เหนายามที่จะปรับระดับความยากของชุดข้อสอบให้เหมาะสมกับระดับความสามารถของผู้สอบ ดังนั้นจำนวนข้อสอบที่ตอบถูกจึงคิดเป็นคะแนนของผู้สอบแต่ละคน เช่นเดียวกับการให้คะแนนของแบบทดสอบชนิดดั้งเดิมนั้น จึงไม่เหมาะสมกับการทดสอบแบบปรับระดับขึ้น เพราะผู้สอบแต่ละคนได้รับข้อสอบไม่เหมือนกัน จำนวนข้อสอบที่ได้รับของแต่ละคนก็ไม่เท่ากัน บางคนใช้ข้อสอบเพียง 11 หรือ 12 ข้อ ก็ยุติการทดสอบได้ ในขณะที่บางคนอาจใช้ข้อสอบ 30 หรือ 40 ข้อ จึงจะไปถึงเกณฑ์ในการยุติการสอบ

วิธีการให้คะแนนที่เหมาะสมกับการทดสอบแบบปรับระดับขึ้นวิธีหนึ่ง ก็คือใช้สัดส่วนระหว่างข้อสอบที่ตอบถูกกับข้อสอบทั้งหมดที่ทำไปเป็นคะแนนของผู้สอบ แต่อย่างไรก็ตามการให้คะแนนผู้สอบด้วยวิธีนี้ คะแนนที่ได้ก็ไม่ได้ให้สารสนเทศเกี่ยวกับระดับความสามารถของผู้สอบมากนัก คือไม่แตกต่างจากการให้คะแนนโดยนับจำนวนข้อสอบที่ทำถูกต้องตามการทดสอบแบบดั้งเดิม

วิธีการให้คะแนนในการทดสอบแบบปรับระดับขึ้นอีกวิธีหนึ่งซึ่งน่าจะเหมาะสมกว่า คือการให้คะแนนที่อาศัยพื้นฐานจากค่าความยากของข้อสอบ ซึ่งคะแนนความสามารถของผู้สอบอาจประมาณได้จากวิธีต่อไปนี้

- 1) ใช้ค่าความยากของข้อสอบที่ยากที่สุด ที่ผู้สอบตอบถูก

- 2) ใช้ค่าความยากของข้อสอบข้อที่ $N+1$ (N คือจำนวนข้อสอบที่ผู้สอบได้ทำ) ถ้ามีการสอบต่อ
- 3) ใช้ค่าความยากของข้อสอบที่ยากที่สุด ที่ผู้สอบตอบถูกในชั้นที่ต่ำกว่า ชั้นสูงสุดของผู้สอบ
- 4) ใช้ค่าความยากเฉลี่ยของชั้นข้อสอบที่สูงที่สุดที่ผู้สอบตอบถูกอย่างน้อย 1 ข้อ ในชั้นนั้น
- 5) ใช้ค่าความยากเฉลี่ยของชั้นข้อสอบที่มีข้อสอบข้อที่ $N+1$ อยู่ (N คือจำนวนข้อสอบที่ผู้สอบได้ทำ)
- 6) ใช้ค่าความยากเฉลี่ยของชั้นข้อสอบที่สูงที่สุดที่ผู้สอบตอบถูกแต่ต่ำกว่า ชั้นสูงสุดของผู้สอบ
- 7) คะแนนความสามารถของผู้สอบหาได้จากสูตร

$$A = \bar{D}_{c-1} + S \cdot P_{c-1} - 0.50$$

เมื่อ $\bar{D}_{c-1}$ = ค่าความยากเฉลี่ยของชั้นที่ $c-1$ เมื่อ c คือ ชั้นสูงสุดของผู้สอบ

P_{c-1} = สัดส่วนการตอบถูกของผู้สอบในชั้นที่ $c-1$

$S = \bar{D}_c - \bar{D}_{c-1}$ ถ้า $P_{c-1} > 0.50$
หรือ $S = \bar{D}_{c-1} - \bar{D}_{c-2}$ ถ้า $P_{c-1} < 0.50$

- 8) ใช้ค่าความยากเฉลี่ยของข้อสอบที่ตอบถูกทั้งหมด
- 9) ใช้ค่าความยากเฉลี่ยของข้อสอบที่ตอบถูกที่อยู่ระหว่าง ชั้นต่ำสุดและชั้นสูงสุดของผู้สอบ
- 10) ใช้ค่าความยากเฉลี่ยของข้อสอบที่ตอบถูกในชั้นที่สูงที่สุดของผู้สอบ ซึ่งต่ำกว่าชั้นสูงสุดของผู้สอบ และวิธีนี้เป็นวิธีที่มีความยุติธรรมและมีความคงที่สูงในการประมาณค่าความสามารถของผู้สอบ (Weiss, 1973)

วิธีการให้คะแนนแบบต่าง ๆ ที่เสนอขึ้นมานี้ ทำให้การวิจัยที่เกี่ยวกับการทดสอบแบบปรับระดับขึ้นมีความน่าสนใจมากขึ้น วิธีการให้คะแนนอาจแปรเปลี่ยนไปตามลักษณะทางจิตวิทยา เช่น ความคงทนรูปร่างการแจ่มแจ้งของคะแนนหรือความสัมพันธ์กับคะแนนจากการทดสอบแบบอื่น ๆ หรือวิธีการให้คะแนนอาจแปรเปลี่ยนไปตามความตรง (Validity) และ/หรือ การใช้งาน (Utility) ซึ่งบางวิธีอาจดีที่สุดในการทำนายกับเกณฑ์ภายนอก หรือใช้ได้ดีในสถานการณ์ต่างๆ นอกจากนี้วิธีการให้คะแนนอาจแปรเปลี่ยนไปตามลักษณะของการวิจัย เช่น การวิจัยเชิงประจักษ์ การวิจัยในสถานการณ์จำลอง หรือการศึกษาในเชิงทฤษฎี ซึ่งผลการวิจัยจะเป็นตัวกำหนดวิธีการให้คะแนนที่ดีที่สุดสำหรับสถานการณ์ต่าง ๆ

2.1.2.5 ความคงที่ในการประมาณค่าความสามารถ

การประมาณระดับความสามารถของผู้สอบ โดยใช้กระบวนการทดสอบแบบปรับระดับขึ้น ผลของการประมาณระดับความสามารถจะมีความคงที่มากแค่ไหนนั้นอาจสังเกตได้จากผลการตอบของผู้สอบที่ถูกบันทึกไว้ นั่นคือถ้าระดับความสามารถของผู้สอบมีความคงเส้นคงวา ผลการตอบข้อสอบก็จะปรากฏออกมาให้เห็นว่าจำนวนชิ้นของข้อสอบที่ใช้จะมีเพียง 2 หรือ 3 ชิ้นเท่านั้น และจำนวนข้อสอบที่ใช้ก็มีน้อยด้วย แต่ถ้าระดับความสามารถของผู้สอบไม่คงที่ในขณะที่ทำการทดสอบอยู่นั้นจำนวนชิ้นของข้อสอบที่ถูกใช้จะมีประมาณ 5 ชิ้น หรือมากกว่า ทำให้การประมาณค่าความสามารถของผู้สอบขาดความกระชับแน่นอนไปด้วย

เพื่อความเข้าใจในการแปลผลการทดสอบดีขึ้น จึงขอเสนอตัวอย่างการบันทึกผลการทดสอบตามแผนภาพต่อไปนี้ ซึ่งเป็นผลจากการทดสอบความสามารถทางภาษา โดยใช้แบบทดสอบปรับระดับขึ้นของ Weiss (1973) ซึ่งมีโครงสร้างของกลุ่มข้อสอบดังตารางที่ 1 และสารสนเทศที่ใช้ในการกำหนดการเริ่มเข้าทำข้อสอบนั้น ใช้ระดับผลการเรียนเฉลี่ยของวิชาที่เกี่ยวข้อง ดังภาพที่ 3 ใช้กฎการแยกทางแบบ ขึ้นหนึ่งขั้น/ลงหนึ่งขั้น เกณฑ์ในการเสร็จสิ้นการทดสอบที่ใช้คือ จะหยุดการดำเนินการสอบเมื่อพบ ชิ้นสูงสุดของผู้สอบ โดยขั้นที่เป็นชิ้นสูงสุดของผู้สอบนั้น จะต้องมีส่วนในการตอบถูกน้อยกว่า 0.20 เมื่อทำข้อสอบในขั้นไม่ต่ำกว่า 5 ข้อ หรือจะหยุดการดำเนินการสอบเมื่อข้อสอบในขั้นที่จะทำต่อไปหมดแล้ว

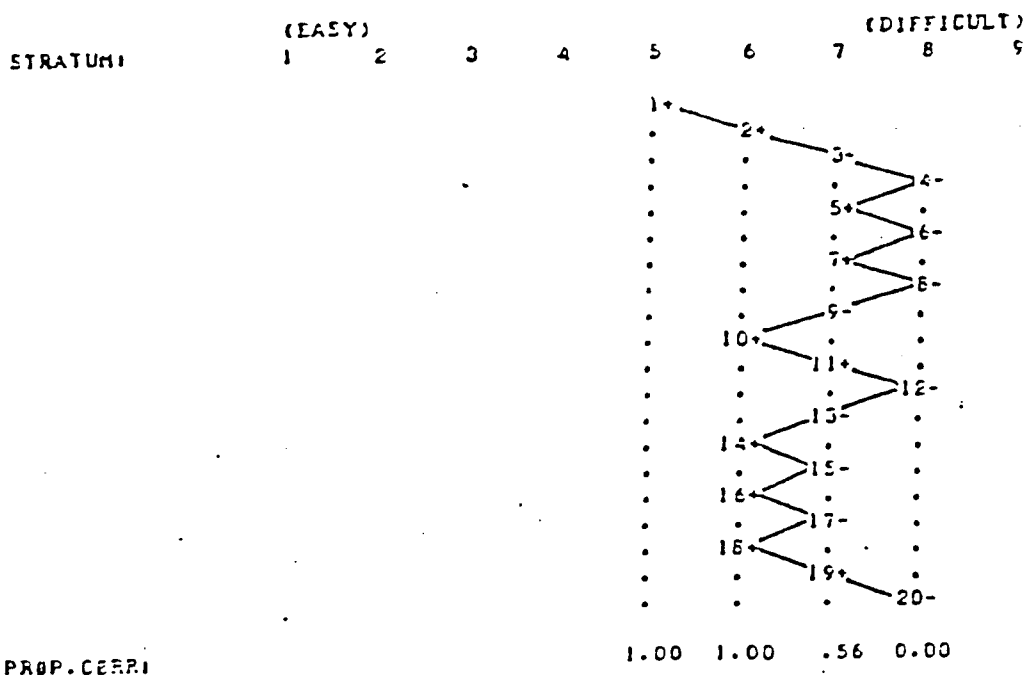
ในภาพที่ 4 เป็นผลการบันทึกการตอบข้อสอบในแบบทดสอบปรับระดับขึ้น และการให้คะแนนด้วยวิธีการต่างๆ 10 วิธีที่กล่าวไปแล้ว ซึ่งในภาพนี้เป็นบันทึกผลการตอบของ William เขาถูกกำหนดให้เริ่มทำข้อสอบในขั้นที่ 5 เครื่องหมาย + หลังข้อที่ทำแสดงว่าเขาตอบข้อสอบนั้นถูก ดังนั้นข้อสอบข้อต่อไปที่ได้รับจะอยู่ในขั้นที่สูงขึ้น คือข้อสอบจะยากขึ้นนั่นเอง ส่วนเครื่องหมาย - หลังข้อสอบที่ทำหมายถึง เขาตอบข้อสอบข้อนั้นผิด ดังนั้นข้อสอบข้อต่อไปจะอยู่ในขั้นต่ำลงมาคือข้อสอบจะง่ายลง จากภาพที่ 4 แสดงให้เห็นว่าขั้นต่ำสุดของเขายู่ในขั้นที่ 6 ชิ้นสูงสุดของเขายู่ในขั้นที่ 8 ซึ่งมีช่วงกว้างเพียง 3 ชิ้นแสดงว่า ในการทำข้อสอบครั้งนี้ของ William เขาแสดงความสามารถที่คงที่พอสมควร แต่การประมาณจุดเริ่มต้นเข้าทำข้อสอบนั้นต่ำกว่าระดับความสามารถที่ประมาณได้จากแบบทดสอบปรับระดับขึ้น ซึ่งทำให้เขาเสียเวลาไปบ้างเล็กน้อย

ส่วนในภาพที่ 5 เป็นการบันทึกผลการตอบของ Carol จากแผนภาพแสดงให้เห็นว่า การประมาณจุดเริ่มต้นในการเข้าทำข้อสอบของเธอสูงกว่าระดับความสามารถที่ประมาณได้จาก แบบทดสอบปรับระดับขึ้น คือจุดเริ่มต้นทำข้อสอบอยู่ในขั้นที่ 6 แต่ช่วงคะแนนความสามารถของเธออยู่ระหว่างขั้นที่ 1 กับ 3 ซึ่งมีเพียง 3 ชิ้น ก็ยังแสดงว่าระดับความสามารถของ Carol คงที่พอสมควร แต่อยู่ในระดับต่ำ และผลจากการที่ประมาณจุดเริ่มต้นทำข้อสอบสูงเกินไป ทำให้เธอเสียเวลากับการทำข้อสอบที่ยากเกินไปซึ่งไม่ให้สารสนเทศอะไรมากนัก

REPORT ON STRADAPTIVE TEST

NAME: WILLIAM W.

DATE TESTED: 73/07/12

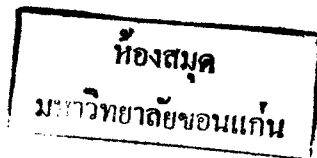


TOTAL PROPORTION CORRECT= .550

SCORES ON STRADAPTIVE TEST

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 1.49
2. DIFFICULTY OF THE N+1 TH ITEM= 1.44
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 1.49
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= 1.33
5. DIFFICULTY OF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 1.33
7. INTERPOLATED STRATUM DIFFICULTY= 1.37
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= .55
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA = 1.23
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= 1.23

ภาพที่ 4 แสดงรายงานผลการตอบข้อสอบของ William ในการทดสอบแบบปรับระดับขึ้น (จาก Weiss, 1973)



REPORT 2N STRADAPTIVE TEST

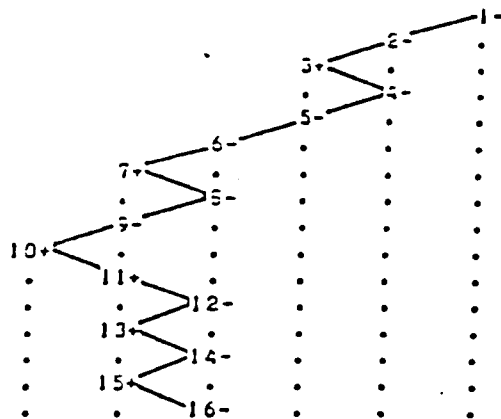
NAME: CAROL C.

DATE TESTED: 73/07/12

STRATUM: (EASY) 1 2 3 4 5 6 (DIFFICULT: 7 8 9)

W
LB
0263
B1

91 51581



PROP. CORRECT: 1.00 .80 0.00 .50 0.00 0.00
TOTAL PROPORTION CORRECT= .375

SCORES ON STRADAPTIVE TEST

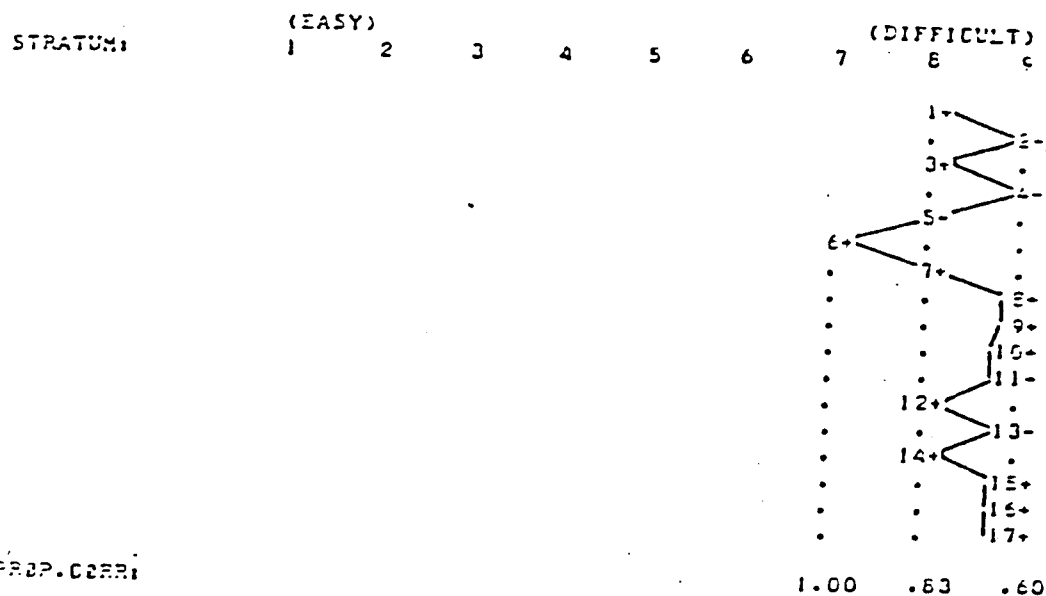
1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= -.70
2. DIFFICULTY OF THE N+1 TH ITEM= -1.92
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= -1.65
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= -.63
5. DIFFICULTY OF THE N+1 TH STRATUM= -1.92
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= -1.92
7. INTERPOLATED STRATUM DIFFICULTY= -1.73
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= -1.51
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA = -1.94
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= -1.94

ภาพที่ 5 แสดงรายงานผลการตอบข้อสอบของ Carol ในการทดสอบแบบปรับระดับขึ้น (จาก Weiss, 1973)

REPORT ON STEADAPTIVE TEST

NAME: NANCY N.

DATE TESTED: 73/04/09



TOTAL PROPORTION CORRECT= .706

SCORES ON STEADAPTIVE TEST

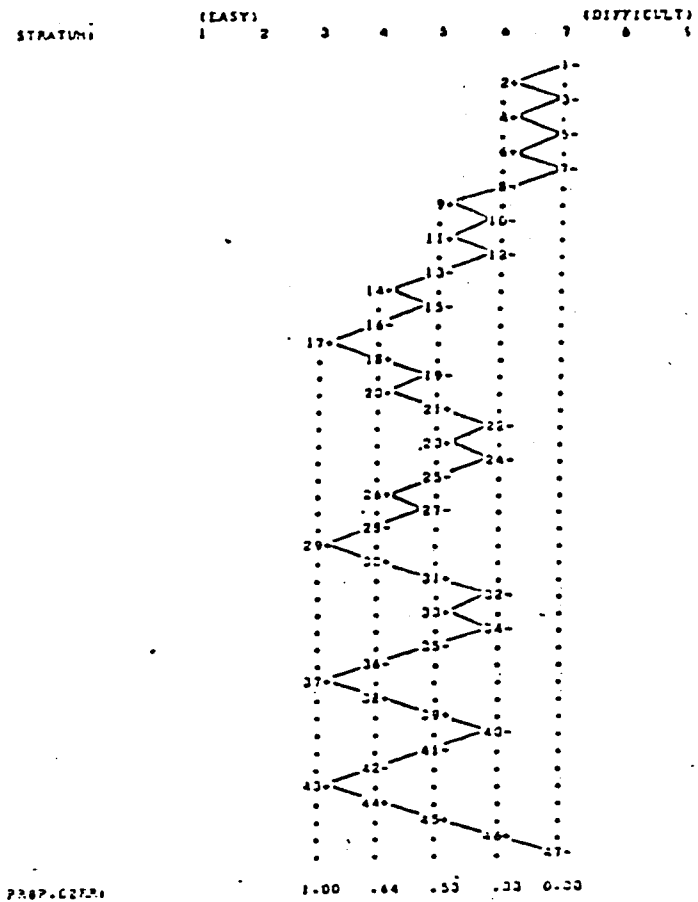
1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 3.11
2. DIFFICULTY OF THE N+1 TH ITEM= 1
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 3.11
4. DIFFICULTY OF HIGHEST STRATUM
WITH A CORRECT ANSWER= 2.62
5. DIFFICULTY OF THE N+1 TH STRATUM= 3.27
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 2.62
7. INTERPOLATED STRATUM DIFFICULTY= 2.69
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= 2.24
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN
CEILING AND BASAL STRATA = 2.35
10. MEAN DIFFICULTY OF ITEMS CORRECT
AT HIGHEST NON-CHANCE STRATUM= 2.63

ภาพที่ 6 แสดงรายงานผลการตอบข้อสอบของ Nancy ในการทดสอบแบบปรับระดับขึ้น
(จาก Weiss, 1973)

REPORT ON STRADAPTIVE TEST

NAME: DIXIE D.

DATE TESTED: 73/04/09



PROP. CORRECT

TOTAL PROPORTION CORRECT= .459

SCORES ON STRADAPTIVE TEST

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= .73
2. DIFFICULTY OF THE N-1 TH ITEM= .78
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= .73
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= .63
5. DIFFICULTY OF THE N-1 TH STRATUM= .63
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= .45
7. INTERPOLATED STRATUM DIFFICULTY= .54
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= -.30
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= -.09
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= .59

ภาพที่ 7 แสดงรายงานผลการตอบข้อสอบของ Dixie ในการทดสอบแบบปรับระดับขึ้น
(จาก Weiss, 1973)

ในภาพที่ 6 เป็นการบันทึกผลการตอบของ Nancy ซึ่งไม่สามารถบอกได้ว่ากำหนดจุดเริ่มต้นเข้าทำข้อสอบได้เหมาะสมหรือไม่ เพราะในขั้นที่ 9 เป็นขั้นที่มีความยากที่สุดของแบบทดสอบ เธอก็ยังตอบถูก แต่ข้อสอบที่ยากมากกว่าขั้นที่ 9 นี้ไม่มี เธอจึงต้องทำข้อสอบข้อต่อไปในขั้นเดิม จนกว่าจะถึงเกณฑ์ในการหยุดทำข้อสอบ และอาจกล่าวได้ว่าแบบทดสอบชุดนี้ง่ายเกินไปสำหรับ Nancy

และในภาพที่ 7 เป็นการบันทึกผลการตอบของ Dixie แสดงให้เห็นว่าความสามารถของ เขาในขณะที่ทดสอบไม่คงเส้นคงวา จึงต้องใช้จำนวนข้อสอบถึง 47 ข้อ จาก 5 ชั้น จึงจะถึงเกณฑ์ในการยุติการสอบและคะแนนความสามารถที่ประมาณได้ก็ขาดความกระชับแน่นอนไป คือยังมีความคลาดเคลื่อนอยู่มาก

จากตัวอย่างการบันทึกผลการตอบของแบบทดสอบปรับระดับขึ้นข้างต้นนี้ จะเห็นว่าผลจากการทดสอบด้วยวิธีนี้ทำให้ได้สารสนเทศจากการทดสอบอย่างมากมาย มีความถูกต้องแน่นอน และเนื่องจากการใช้ระบบคอมพิวเตอร์มาช่วยในการดำเนินการทดสอบ ทำให้มีความสะดวกรวดเร็วและรู้ผลการตอบได้ทันที จะดำเนินการทดสอบเป็นกลุ่มหรือเป็นรายบุคคลก็ได้ ไม่จำเป็นต้องสอบกันและเวลาเดียวกันก็ได้ และควรจะให้ผู้สอบนั้นพร้อมที่จะสอบมากที่สุด เพื่อให้การวินิจฉัยความสามารถของผู้สอบนั้นมีความเที่ยงสูง และข้อสอบแต่ละข้อนี้จะปรากฏบนจอภาพ ไม่มีการพิมพ์แผ่นกระดาษ ผู้สอบแต่ละคนก็ทำชุดของข้อสอบที่ไม่เหมือนกัน จึงไม่มีโอกาสที่ข้อสอบจะรั่วไหลออกมากระทบกระเทือนต่อความเที่ยงของแบบทดสอบได้เลย

2.2 สถิติของการวิจัยที่เกี่ยวกับแบบทดสอบเทเลอร์

Weiss และ Betz (อ้างถึงใน ปรีชา, 2526) ได้จำแนกสถิติของการวิจัยที่เกี่ยวกับการสอบแบบเทเลอร์ ออกเป็น 3 ชนิดดังนี้

2.2.1 การศึกษาเชิงประจักษ์นิยม (Empirical Study)

เป็นการศึกษาที่ใช้ผู้สอบและกลุ่มของข้อสอบจริง มีวิธีการดำเนินการทดสอบที่เฉพาะ ซึ่งประกอบด้วย การดำเนินการสอบที่ใช้วิธีให้ผู้สอบเขียนตอบ และการใช้เครื่องคอมพิวเตอร์

ข้อสรุปที่ได้จากการศึกษาเชิงประจักษ์นิยมต้องพิจารณาด้วยความระมัดระวัง ทั้งนี้เพราะมีข้อจำกัดในเรื่องคุณลักษณะของผู้สอบ ความเหมาะสมของกลุ่มข้อสอบที่ใช้ในการศึกษา และปฏิสัมพันธ์ (Interaction) ของตัวผู้สอบกับวิธีการดำเนินการทดสอบ

2.2.2 การศึกษาในสภาพการณ์จำลอง (Simulation Study) มี 2 ชนิด คือ

ชนิดที่ 1 การใช้ข้อมูลจริง (Real Data) เป็นการศึกษาที่ใช้ข้อมูลจากผลการทดสอบที่ใช้

แบบทดสอบดั้งเดิม ซึ่งข้อมูลเหล่านี้ประกอบด้วยผลการตอบข้อสอบแต่ละข้อของผู้สอบแต่ละคน คะแนนรวมและรวมทั้งข้อมูลเกี่ยวกับความยากและค่าอำนาจจำแนกของข้อสอบ ในการจำลองการทดสอบแบบทดสอบเทเลอร์กับกลุ่มผู้สอบนั้นอาจจะใช้วิธีการใดวิธีการหนึ่ง หรือหลายวิธีการของการทดสอบแบบเทเลอร์ กับกระดาษคำตอบของผู้สอบแต่ละคนและมีการให้คะแนนกับกระดาษคำตอบของผู้สอบแต่ละคนใหม่ โดยถือเสมือนว่าแบบทดสอบได้รับการดำเนินการทดสอบแบบเทเลอร์

ชนิดที่ 2 การศึกษาในสถานการณ์จำลองแบบมอนติคาร์โล (Monticarlo Simulation Study) เป็นการศึกษาที่ประกอบด้วยการสร้างกลุ่มของข้อสอบในเชิงทฤษฎีในลักษณะที่เป็นการตรวจสอบคุณลักษณะที่เฉพาะและถูกต้องแน่นอนเกี่ยวกับค่าความยากและค่าอำนาจจำแนกของข้อสอบ ตลอดจนความสามารถของผู้สอบ แล้วใช้ทฤษฎีของโค้งคุณลักษณะของข้อสอบ (Item Characteristic Curve Theory, ICC) และตัวเลขสุ่มที่ได้จากเครื่องคอมพิวเตอร์ (Computer Generated Random Numbers) ตลอดจนเมทริกซ์ของคำตอบของข้อสอบอีกทั้งคะแนนรวมและระดับความสามารถของผู้สอบ มาสร้างเป็นกลุ่มของข้อสอบโดยใช้วิธีการทดสอบแบบเทเลอร์ที่เฉพาะเจาะจงกับผู้สอบที่จะทำการศึกษา แต่ไม่ให้เกี่ยวกับผลกระทบที่เกิดจากการทดสอบเทเลอร์ที่เกิดขึ้นกับผู้สอบ

2.2.3 การศึกษาเชิงทฤษฎี (Theoretical Study)

เป็นการศึกษาที่ตั้งอยู่บนพื้นฐานของทฤษฎีโค้งคุณลักษณะของข้อสอบ ซึ่งต้องมีค่า พารามิเตอร์ที่คงที่ การศึกษาในลักษณะนี้เป็นการศึกษาเพื่อตรวจสอบตลอดช่วงความสามารถที่ต่อเนื่องของผู้สอบมากกว่าที่จะศึกษาในช่วงความสามารถที่เฉพาะเจาะจงในช่วงใดช่วงหนึ่ง สำหรับกลุ่มของข้อสอบในเชิงทฤษฎีที่จะให้ผลการวัดในระดับสูงสุดโดยไม่มีข้อจำกัดเกี่ยวกับจำนวนผู้สอบในสถานการณ์จำลอง ผลการศึกษาที่ได้จากวิธีนี้ จะอยู่ในลักษณะของการหาคำตอบของชุดสมการทางคณิตศาสตร์ที่คิดขึ้นโดย Lord (อ้างถึงใน Weiss, 1973), Lord และ Novick (1968) และคนอื่น ๆ ซึ่งเป็นการสร้างโค้งการกระจายของคะแนนผลการสอบจากโค้งลักษณะของข้อสอบที่สมมุติขึ้น สำหรับผู้สอบที่มีความสามารถผันแปรไปภายใต้วิธีการทดสอบที่เฉพาะ การศึกษาในลักษณะนี้ไม่มีประโยชน์เพราะทำได้รวดเร็วในการประเมินวิธีการทดสอบแบบต่างๆของแบบทดสอบเทเลอร์ ตลอดจนค่าพารามิเตอร์ของข้อสอบ แต่ไม่สามารถรายงานถึงผลกระทบที่เกิดจากการทดสอบแบบเทเลอร์ในสถานการณ์จริงได้

2.3 เกณฑ์ที่ใช้ในการประเมินแบบทดสอบเทเลอร์

Sympson (อ้างถึงใน ปรีชา, 2526) ได้ให้ข้อคิดเกี่ยวกับการประเมินกระบวนการทดสอบไว้ว่า ไม่ควรจะทำในลักษณะรวม ๆ แต่จะประเมินผลกระทบแต่ละอย่างในองค์ประกอบใดองค์ประกอบหนึ่ง การประเมินองค์ประกอบเหล่านี้จะเป็นการเปรียบเทียบกระบวนการทดสอบ 2

กระบวนการที่มีความแตกต่างกันเพียงองค์ประกอบเดียวเท่านั้น ถ้าพบว่ากระบวนการทดสอบหนึ่งดีกว่าอีกกระบวนการหนึ่งแล้ว ก็จะสามารถได้ว่าองค์ประกอบที่แตกต่างกันนั้น เป็นตัวการทำให้กระบวนการทดสอบแตกต่างกัน แต่โดยสภาพทั่ว ๆ ไปแล้วการประเมินกระบวนการทดสอบในแง่ขององค์ประกอบดังกล่าวข้างต้นมักจะเป็นไปไม่ได้ เนื่องจากข้อจำกัดหลายประการ เช่น วิธีการเลือกข้อสอบบางแบบ ไม่อาจจะใช้วิธีดำเนินการสอบโดยวิธีอื่นได้นอกจากจะใช้เครื่องคอมพิวเตอร์ หรือการให้คะแนนทางวิธีเลือกข้อสอบไปลักษณะเดียวกันตลอดกระบวนการทดสอบ ดังนั้นจึงจำเป็นต้องทำการเปรียบเทียบกระบวนการทดสอบซึ่งแตกต่างกันในสององค์ประกอบหรือมากกว่า

Sympson (อ้างถึงใน ปรีชา, 2526) ได้จำแนกประเภทของเกณฑ์ที่ใช้ในการประเมินกระบวนการทดสอบ ซึ่งไม่จำกัดเฉพาะกับแบบทดสอบเทเลอร์มีสาระสำคัญดังนี้

2.3.1 เกณฑ์เชิงความเป็นไปได้ (Possibility Criteria)

จะเป็นการพิจารณาถึงคุณลักษณะของคะแนนที่ได้จากกระบวนการทดสอบว่า ได้แสดงให้เห็นถึงลักษณะที่ต้องการวัดเพียงใด เกณฑ์เชิงความเป็นไปได้นี้ประกอบด้วย ความเที่ยง ชนิดของสัมประสิทธิ์ความคงที่ (Stability Coefficient) สัมประสิทธิ์ความคงที่ภายใน (Internal Consistency Coefficient) และค่าสหสัมพันธ์จากแบบทดสอบคู่ขนาน (Alternate Form Correlation) ค่าสหสัมพันธ์กับแบบทดสอบอื่นๆ ตลอดจนค่าสหสัมพันธ์กับตัวแปรอื่นๆ ที่นอกเหนือจากคะแนนที่ได้จากการทดสอบ รวมทั้งลักษณะการกระจายของคะแนนผลการสอบในกลุ่มผู้สอบกลุ่มใดกลุ่มหนึ่ง ความแตกต่างระหว่างการกระจายของคะแนนที่ได้จากผู้สอบต่างกลุ่มกัน และการใช้วิธีการทางสถิติ หรือกราฟสำหรับการประเมินความสอดคล้องกัน (Goodness-of-fit) กับรูปแบบตามทฤษฎีที่มีอยู่เดิม

ในการทดสอบแบบเทเลอร์ ผู้สอบแต่ละคนจะได้ตอบข้อสอบที่แตกต่างกัน ดังนั้นจึงไม่สามารถคำนวณค่าสัมประสิทธิ์ความคงที่ภายในได้ จึงจำเป็นต้องใช้สหสัมพันธ์ชนิดแบบทดสอบคู่ขนานในการประมาณค่าความเที่ยงของคะแนนที่ได้จากแบบทดสอบ

2.3.2 เกณฑ์เชิงทฤษฎี (Theoretical Criteria)

จะใช้สำหรับตรวจสอบว่ากระบวนการทดสอบหนึ่งดีกว่าอีกกระบวนการหนึ่งหรือไม่ เมื่อทราบแต่แรกแล้วว่ากระบวนการทั้งสองน่าจะมีความตรงเชิงโครงสร้างเท่ากัน เกณฑ์เชิงทฤษฎีประกอบด้วย การกระจายของค่าประมาณคุณลักษณะแฝง โค้งของสารสนเทศ (Information Curve) โค้งประสิทธิภาพสัมพัทธ์ (Relative Efficiency Curve) โค้งความลำเอียง (Bias Curve) โค้งความคลาดเคลื่อนมาตรฐานของการวัด ตลอดจนการคงสภาพ (Robustness) ในลักษณะต่างๆ ซึ่งได้แก่ความคลาดเคลื่อนในการประมาณค่า บารามิเตอร์ของข้อสอบหรือโค้งคุณลักษณะของข้อสอบที่

ได้จากกระบวนการทดสอบที่แตกต่างกัน เป็นต้น

การใช้เกณฑ์เชิงทฤษฎีนี้ ต้องมีการกำหนดรูปแบบของโดเมนลักษณะของข้อสอบที่เฉพาะเจาะจงขึ้น และต้องทราบค่าจริงของค่าพารามิเตอร์ของข้อสอบตั้งแต่แรก แล้วใช้วิธีการทางคณิตศาสตร์ หรือใช้คอมพิวเตอร์ในการคำนวณ เกณฑ์เชิงทฤษฎีเหล่านี้ ไม่สามารถใช้กับการศึกษาในสถานจริงได้ ทั้งนี้เพราะเราไม่ทราบสถานภาพจริงของผู้สอบแต่ละคนมาตามตรรกะลักษณะแฝง

2.3.3 เกณฑ์เชิงจิต-สังคม (Psycho-Social Criteria)

เป็นการศึกษาถึงผลกระทบที่เกิดจากกระบวนการทดสอบ ที่มีต่อคุณลักษณะทางจิตวิทยาในตัวผู้สอบ เช่น การจูงใจ (Motivation) และความวิตกกังวล (Anxiety) ในการทดสอบ ตลอดจนผลกระทบที่มีต่อสังคม เช่น การทดสอบแบบเทเลอร์จะทำให้เกิดความไม่เข้าใจกันไม่ศรัทธาไป เพราะการที่ผู้สอบแต่ละคนได้ตอบข้อสอบที่แตกต่างกัน อาจทำให้เกิดความสงสัยว่าจะเปรียบเทียบความสามารถระหว่างผู้สอบได้อย่างไรในเมื่อผู้สอบเหล่านี้ ไม่ได้ตอบข้อสอบที่เหมือนกัน

2.3.4 เกณฑ์เชิงค่าใช้จ่าย (Cost Criteria)

เป็นการประเมินค่าใช้จ่ายในระบบการสร้างแบบทดสอบ การสื่อสารกับตัวผู้สอบด้วยเครื่องคอมพิวเตอร์ หรือมูลค่าวัสดุที่ต้องใช้ในระบบการทดสอบ ค่าใช้จ่ายในการจัดสร้างและการหาเกณฑ์ปกติของกลุ่มข้อสอบที่มีขนาดตามที่กระบวนการนั้นต้องการ อีกทั้งจุดอ่อนในเรื่องที่ผู้สอบอาจจะกำเริบพลาดในระหว่างการทำข้อสอบ หรือจุดอ่อนในเรื่องการให้คะแนนที่ผิดพลาด ตลอดจนจุดอ่อนในการเสียเวลาในระบบการถาม-ตอบ (Delivery System) การขาดเอกสารหรือพยานยืนยันเวลาที่สูญเสียไปในการดำเนินการทดสอบ รวมทั้งการให้คะแนนและการตีความหมายของคะแนน

Weiss และ Betz (อ้างถึงใน ปรีชา, 2526) ได้กล่าวถึงการประเมินแบบทดสอบเทเลอร์ โดยเฉพาะในลักษณะที่สัมพันธ์กับชนิดของการวิจัย มีสาระสำคัญดังนี้

1) การประเมินประสิทธิภาพของแบบทดสอบเทเลอร์ ส่วนใหญ่จะเป็นการเปรียบเทียบกับแบบทดสอบดั้งเดิม โดยใช้เกณฑ์เชิงปฏิบัติซึ่งประกอบด้วยเวลาที่ใช้ในการดำเนินการสอบ ค่าใช้จ่ายในการดำเนินการสอบและความยุ่งยากซับซ้อนในการดำเนินการสอบ

2) การศึกษาบางครั้งก็ใช้ค่าสหสัมพันธ์ของคะแนนที่ได้จากแบบทดสอบเทเลอร์กับคะแนนจากแบบทดสอบดั้งเดิมเป็นเกณฑ์ในการประเมิน การศึกษาในลักษณะนี้โดยปกติจะทราบคะแนนจากแบบทดสอบดั้งเดิมตั้งแต่เริ่มต้น ข้อสอบจากแบบทดสอบเทเลอร์ที่สร้างขึ้นนี้จะถูกนำไปดำเนินการสอบกับผู้สอบกลุ่มเดิม ในสถานการณ์จริงหรือไม่ก็ในสถานการณ์จำลอง เพื่อให้ได้คะแนนผลการสอบออกมาจุดมุ่งหมายในการศึกษาลักษณะนี้ ส่วนใหญ่จะเป็นการกำหนดว่าแบบจำลองการทดสอบชนิดเทเลอร์

แบบใดที่สามารถประมาณคะแนนรวมในแบบทดสอบตั้งเดิมซึ่งข้อสอบในแบบทดสอบนั้นถูกเลือกไปสร้างเป็นแบบทดสอบเทเลอว์ได้ดีที่สุด

3) การศึกษาเชิงทฤษฎี จะใช้ค่าสหสัมพันธ์ของคะแนนที่ได้จากแบบทดสอบกับความสามารถของผู้สอบเป็นเกณฑ์ในการประเมิน เกณฑ์การประเมินของการศึกษาในสถานการณ์จำลองแบบมอนติ-คาร์โล คือค่าสหสัมพันธ์ระหว่างความสามารถของผู้สอบกับค่าประมาณความสามารถที่ประมาณจากรูปแบบของคำตอบของผู้สอบแต่ละคนในสถานการณ์การทดสอบที่สมมุติขึ้น การศึกษาในลักษณะนี้ ผู้กระทำการศึกษาจะสนใจในเรื่องความตรงของแบบจำลองการทดสอบว่าคะแนนที่ได้มาสามารถทำนายความสามารถของผู้สอบที่สมมุติขึ้นได้มากเพียงใด

4) ในการศึกษาบางครั้งอาจใช้ฟังก์ชันของสารสนเทศ (Birnbbaum, อ้างถึงใน ปรีชา, 2526) เป็นเกณฑ์ในการประเมินแบบจำลองของการทดสอบซึ่งเกณฑ์นี้จะเป็นการเปรียบเทียบแบบจำลองการทดสอบตั้งแต่สองแบบขึ้นไป ในทอมปริมาณของสารสนเทศที่ได้จากแบบจำลองการทดสอบที่ระดับต่างๆ ตลอดช่วงความสามารถของผู้สอบ (Underlying Ability Continuum) ฟังก์ชันสารสนเทศที่มีการตีความหมายแตกต่างกันใน 3 ลักษณะคือ ค่าของสารสนเทศจะบอกให้ทราบถึงประสิทธิภาพสัมพัทธ์ของจำนวนข้อสอบในแบบทดสอบต่างๆ ที่จะส่งผลให้แบบทดสอบเหล่านั้นมีระดับความเที่ยงเท่ากัน ตลอดจนสัมพัทธ์ภาพของความคลาดเคลื่อนมาตรฐานของการวัด (Relative Error of Measurement) และสัมพัทธ์ภาพของความสามารถ (Relative Capability) ของแบบจำลองการทดสอบต่างๆ ที่จะจำแนกผู้สอบตามระดับความสามารถ

2.4 การประมาณค่าพารามิเตอร์ของข้อสอบและการประมาณระดับความสามารถของผู้สอบตามทฤษฎีคุณลักษณะแฝง (Latent Trait Theory)

ทฤษฎีคุณลักษณะแฝง เป็นทฤษฎีที่อธิบายความสัมพันธ์ระหว่างปริมาณของการกระทำหรือคะแนนของผู้สอบกับปริมาณความสามารถของผู้สอบ (สำเริง, 2527)

ถ้าให้ P แทน ผลการสอบ (Performance)
 θ แทน ความสามารถ (Ability หรือ Trait)
 f แทน ความสัมพันธ์ (Function)

ดังนั้นสมการแสดงความสัมพันธ์ระหว่างผลการสอบและความสามารถของผู้สอบคือ

$$P = f(\theta) \dots\dots\dots (1)$$

ซึ่งเป็นสมการที่อยู่ในรูปทั่วไป ยังไม่จำเพาะเจาะจง การที่จะนิยามทฤษฎีนี้ จำต้องกำหนดข้อตกลงบางอย่าง เพื่อสร้างสมการที่เหมาะสมสำหรับแสดงความสัมพันธ์ระหว่าง P และ θ

2.4.1 ข้อตกลงเบื้องต้นของทฤษฎีคุณลักษณะแฝง

ข้อตกลงทั่ว ๆ ไป เพื่อพัฒนาความสัมพันธ์ระหว่าง P และ θ มีดังนี้

2.4.1.1 แบบทดสอบที่ใช้ในการวัดความสามารถจะต้องเป็นแบบทดสอบที่วัดความสามารถด้านเดียว เพื่อให้ง่ายต่อการรวมคะแนนและการแปลความหมาย แบบทดสอบใด ๆ ที่ใช้วัดความสามารถด้านเดียวนั้น อาจใช้การวิเคราะห์ตัวประกอบ (Factor Analysis) ตรวจสอบได้ โดยการกระทำตามขั้นตอนต่อไปนี้

ขั้นที่ 1 แปลงผลการตอบข้อสอบของผู้สอบในแต่ละข้อให้เป็นคะแนน 0/1 (ถ้าตอบผิดให้คะแนนเป็น 0 ตอบถูกให้คะแนนเป็น 1)

ขั้นที่ 2 คำนวณหาค่าสหสัมพันธ์ระหว่างข้อแบบ Tetrachoric แล้วจัดให้อยู่ในรูป Matrix ของความสัมพันธ์

ขั้นที่ 3 แทนที่สหสัมพันธ์ในแนว Diagonal (1.00) ด้วยค่าสหสัมพันธ์ที่มีค่ามากที่สุดของแต่ละแถว

ขั้นที่ 4 วิเคราะห์ตัวประกอบ โดยใช้โมเดล Principal Component

เมื่อทำการวิเคราะห์ตัวประกอบเสร็จแล้ว จะตรวจสอบว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียวหรือไม่ โดยพิจารณาจากตัวบ่งชี้ต่อไปนี้

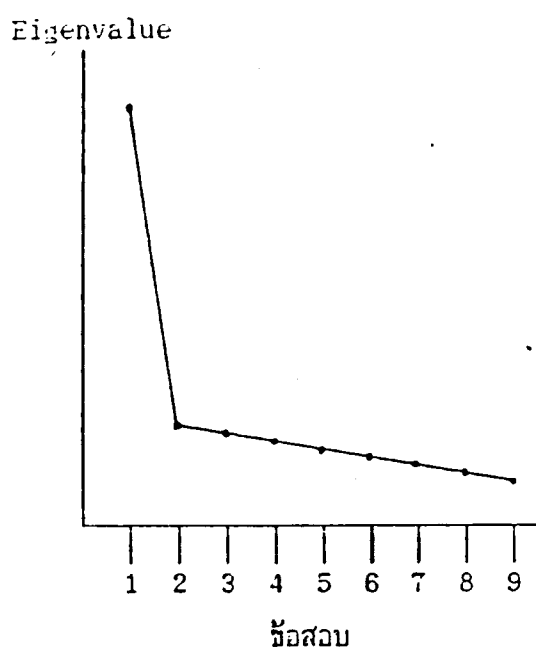
(1) The Eigenvalue Test นำค่า Eigenvalue ของแต่ละตัวประกอบไปเขียนกราฟ ถ้าพบว่ากราฟมีลักษณะดังภาพที่ 8 ก. ก็แสดงว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียว

(2) The Random Baseline Test ทำการสุ่มตัวอย่างข้อมูลมาเพียงบางส่วน แล้วดำเนินการวิเคราะห์ตัวประกอบตามขั้นตอนที่กล่าวไว้ข้างต้น แล้วนำค่า Eigenvalue ของแต่ละตัวประกอบที่ได้ใหม่นั้นไปเขียนกราฟแบบเดียวกันกับข้อ (1) ถ้ากราฟที่ได้มีลักษณะดังในภาพที่ 8 ก. ก็แสดงว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียว

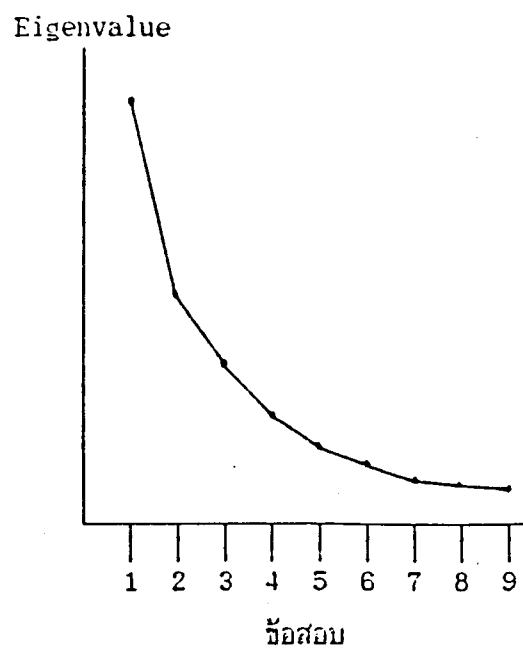
(3) The Biserial Test คำนวณหาค่าสหสัมพันธ์แบบ Biserial ระหว่างข้อสอบแต่ละข้อกับคะแนนรวม แล้วนำค่าสหสัมพันธ์ที่ได้ไปหาค่าสหสัมพันธ์กับค่า Factor Loading ของข้อสอบแต่ละข้อที่มีต่อ Factor แรก ถ้าค่าสหสัมพันธ์ที่ได้ในครั้งหลังนี้มีค่าสูงตั้งแต่ 0.80 ขึ้นไปก็แสดงว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียว

(4) The Factor Loading Test พิจารณาค่า Factor Loading ของข้อสอบแต่ละข้อที่มีต่อ Factor แรก ถ้าพบว่าแต่ละค่าของ Factor Loading มีค่าสูงอย่างมีนัยสำคัญและมีเครื่องหมายเหมือนกัน (+ หรือ -) แสดงว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียว

(5) The Congruence Test แบ่งผู้สอบออกเป็น 2 กลุ่ม (เช่น กลุ่มผู้ชาย/กลุ่มผู้หญิง) ทำการวิเคราะห์ตัวประกอบของแบบทดสอบในแต่ละกลุ่ม แล้วคำนวณหาค่าสัมประสิทธิ์ความสอดคล้อง (C_{AB}) ของ Factor Loading ใน Factor แรก ระหว่าง 2 กลุ่มโดยใช้สูตร

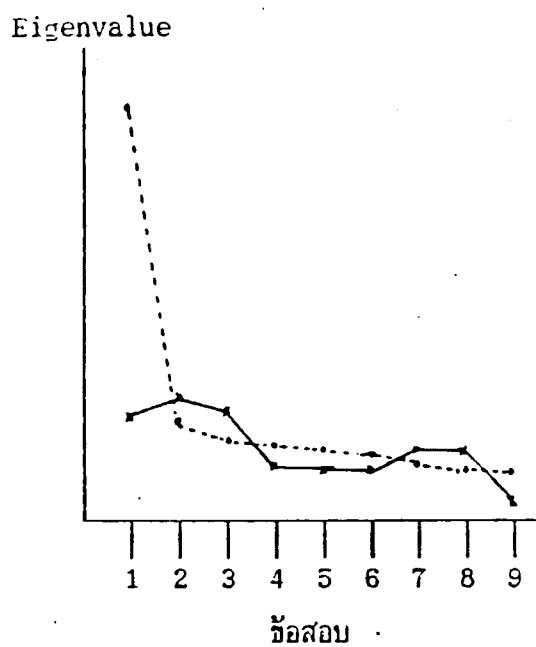


ก. UNIDIMENSIONAL

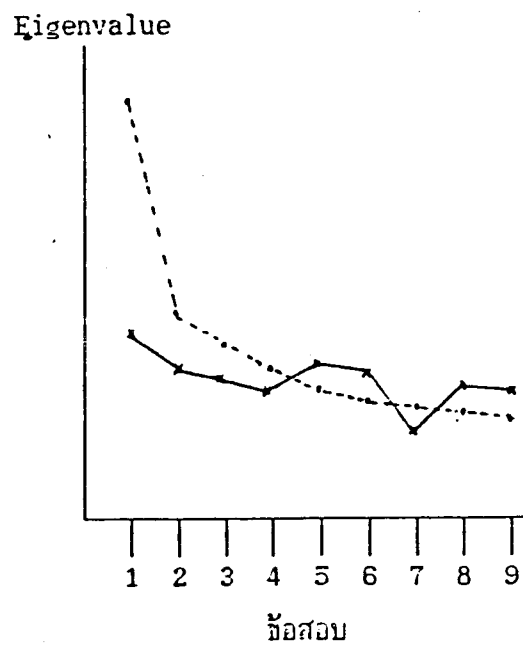


ข. NONUNIDIMENSIONAL

พิจารณาจากกราฟของ Eigenvalue



ค. UNIDIMENSIONAL



ง. NONUNIDIMENSIONAL

พิจารณาจาก Random Baseline

ภาพที่ 8 กราฟแสดงค่า Eigenvalue กับตัวประกอบ (ข้อสอบ) สำหรับการพิจารณาความเป็น Unidimensionality ของแบบทดสอบ (จาก Warm, 1978)

$$C_{AB} = \sqrt{\frac{\sum_{i=1}^n (L_{ia} - L_{ib})^2}{n}} \dots\dots\dots (2)$$

โดยที่ L_{ia} = ค่า Factor Loading ของข้อสอบข้อที่ i ใน Factor แรก ในกลุ่ม A

L_{ib} = ค่า Factor Loading ของข้อสอบข้อที่ i ใน Factor แรก ในกลุ่ม B

n = จำนวนข้อสอบในแบบทดสอบ

C_{AB} = สัมประสิทธิ์ความสอดคล้องระหว่างกลุ่ม A และ B

ถ้าพบว่าค่า C_{AB} เข้าใกล้ 0 แสดงว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียว

(6) The Communalities Test โดยการคำนวณค่า G จากสูตร

$$G = \frac{\sum \frac{r_{ij}}{\sqrt{h_i h_j}}}{n}, i > j \dots\dots\dots (3)$$

โดยที่ r_{ij} = ค่าสหสัมพันธ์แบบ interitem tetrachoric

h_i = ค่าความแปรปรวนร่วมของข้อสอบที่ i

n = จำนวนของค่าสหสัมพันธ์

ถ้าพบว่าค่า G เข้าใกล้ 1.00 แสดงว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียว

(7) The Part/Whole Test นำแบบทดสอบทั้งฉบับ (Whole) เช่น สมมติว่ามีข้อสอบ 20 ข้อ ไปทำการวิเคราะห์หาค่าพารามิเตอร์ของข้อสอบแต่ละข้อ แล้วแบ่งข้อสอบบางส่วน (Part) สมมติว่าแบ่งมาเพียง 10 ข้อ ไปทำการวิเคราะห์หาค่าพารามิเตอร์ของข้อสอบแต่ละข้อเช่นเดียวกัน แล้วนำค่าพารามิเตอร์ทั้ง 2 ชุดของข้อสอบทั้ง 10 ข้อ มาหาค่าสหสัมพันธ์กัน ถ้าพบว่ามีความสัมพันธ์กันสูง แสดงว่าแบบทดสอบ 10 ข้อนี้ใช้วัดความสามารถด้านเดียว

(8) The Vector Frequency Test สมมติว่าความสามารถของกลุ่มผู้สอบ (๑) มีการกระจายแบบปกติ และข้อสอบแต่ละข้อในแบบทดสอบนั้นทราบค่าพารามิเตอร์แล้ว ถ้าแบบแผนการตอบ (Patterns) อย่างคาดหวัง และแบบแผนการตอบจริงของผู้สอบ ไม่แตกต่างกันอย่างมีนัยสำคัญ (ใช้ Chi-Square ทดสอบ) แสดงว่าแบบทดสอบนั้นใช้วัดความสามารถด้านเดียว

2.4.1.2 โอกาสที่ผู้สอบตอบคำถามข้อหนึ่งถูกต้องนั้นไม่เกี่ยวข้องกับการตอบคำถามข้ออื่น ๆ ในแบบทดสอบเรียกข้อตกลงนี้ว่า Local Independence

ถ้าให้ U_g แทน ผลการตอบคำถามข้อที่ g ของผู้สอบคนหนึ่ง

$$g = 1, 2, 3, \dots, n$$

โดยให้คะแนนแบบ ตอบถูกให้ 1 คะแนน และตอบผิดให้ 0 คะแนน

P_g แทน โอกาสที่ผู้สอบตอบคำถามข้อ g ถูกต้อง

$$Q_g = 1 - P_g$$

จากข้อตกลงเกี่ยวกับ Local Independence จะได้ว่า

$$\text{Prob} \{ U_1 = U_1, U_2 = U_2, \dots, U_n = U_n \}$$

$$= \prod_{g=1}^n P_g^{u_g} Q_g^{1-u_g} \dots\dots\dots (4)$$

สมการนี้อ่านว่า โอกาสของการตอบแบบทดสอบได้ถูกต้องทั้งหมดทุกข้อ คือ ผลคูณ (π) ของ โอกาสของการตอบคำถามแต่ละข้อได้ถูกต้อง

จากข้อตกลงดังกล่าวนี้ ทำให้สามารถสร้างสมการแสดงความถี่ของคะแนนของกลุ่มผู้สอบที่มี ความสามารถ θ ได้ดังนี้

$$f(x/\theta) = \sum_{g=1}^n \pi P_g^{u_g} Q_g^{1-u_g} \dots\dots\dots (5)$$

เมื่อ x เป็นคะแนนของผู้สอบซึ่งมีค่าตั้งแต่ 0 ถึง n

การทดสอบว่าข้อสอบแต่ละข้อเป็นไปตามข้อตกลงของ Local Independence หรือไม่นั้น ให้ตรวจสอบดูแบบแผนการตอบข้อสอบของผู้สอบที่มีความสามารถใกล้เคียงกันว่าข้อสอบแต่ละข้อมีลักษณะ การตอบแตกต่างกันหรือไม่ ซึ่งสามารถตรวจสอบได้จากการคำนวณค่าไคสแควร์ เพื่อทดสอบความเป็น อิสระของการตอบข้อสอบเหล่านี้ (ลำเริง, 2527) และในการทดสอบต้องให้เวลาผู้สอบได้ทำข้อ สอบอย่างเต็มที่เพื่อให้ผู้สอบแสดงความสามารถที่แท้จริงออกมา (อุปัทมภ์, 2528) โดยที่โอกาสของ การตอบข้อสอบข้อหนึ่ง ๆ ถูกของผู้สอบคนหนึ่งนั้น ต้องไม่ขึ้นอยู่กับความแข็งแรงความสามารถของกลุ่ม ประชากร นั้นคือรูปร่างเส้นโค้งคุณลักษณะของข้อสอบ ไม่ได้ขึ้นอยู่กับ การกระจายของความสามารถ ของกลุ่มผู้สอบ

และถ้าเป็นไปตามข้อตกลงดังกล่าวข้างต้น การวิเคราะห์ข้อสอบโดยอาศัยทฤษฎีคุณลักษณะ แฝง ก็จะทำให้ข้อดีหรือข้อได้เปรียบดังนี้

- 1) การประมาณค่าความสามารถของผู้สอบแต่ละคน จะขึ้นอยู่กับคำตอบแบบทดสอบที่ถูกเลือกมาจากกลุ่มข้อสอบขนาดใหญ่ ที่วัดความสามารถด้านเดียวกัน
- 2) ค่าพารามิเตอร์ของข้อสอบจะไม่แปรผันไปตามกลุ่มตัวอย่างผู้สอบ
- 3) ให้การวัดประมาณค่าความสามารถที่ถูกต้องแน่นอนในแต่ละระดับความสามารถ ความคลาดเคลื่อนมาตรฐานของการวัดของแต่ละบุคคลสามารถประมาณได้

2.4.2 รูปแบบแสดงความสัมพันธ์ระหว่างผลการสอบและความสามารถของผู้สอบ ตามทฤษฎีคุณลักษณะแฝง

2.4.2.1 Normal-Ogive Model

ฟังก์ชันทางคณิตศาสตร์ของรูปแบบนี้คือ

$$P_g(\theta) = \int_{-\infty}^u [1/(2\pi)^{1/2}] e^{-t^2/2} dt \quad \dots\dots\dots (6)$$

เมื่อ $u = a_g(\theta - b_g)$
 $P_g(\theta) =$ โอกาสที่ผู้ที่มีความสามารถ θ จะตอบข้อสอบข้อ g ได้ถูกต้อง
 $b_g =$ ค่าความยากของข้อสอบข้อ g
 $a_g =$ ค่าอำนาจจำแนกของข้อสอบข้อ g

2.4.2.2 The Logistic Model

มีรูปแบบฟังก์ชันทางคณิตศาสตร์โดยทั่วไปดังนี้

$$p_g(\theta) = 1/(1 + e^{-u}) \quad \dots\dots\dots (7)$$

ซึ่งความหมายของ u , a_g และ b_g ก็เหมือนกับที่เคยให้ความหมายไว้แล้วในรูปแบบของ Normal-Ogive และใน The Logistic Model นี้ยังแบ่งออกเป็น 3 รูปแบบย่อย ๆ ตามจำนวนของตัวพารามิเตอร์ของข้อสอบ

ในข้อสอบประเภทเลือกตอบนั้น แต่ละข้อจะมีตัวเลือกตอบ 4-5 ตัวเลือก ผู้สอบคนหนึ่งที่อาจเลือกตอบอย่างสุ่มหรือเดาคำตอบที่ถูกได้โดยมีโอกาสดังกล่าวถึง 1/4 หรือ 1/5 นั้นแสดงว่า แม้แต่ผู้สอบที่มีความสามารถต่ำ ๆ ก็ยังมีโอกาสตอบถูกหรือกล่าวอีกนัยหนึ่งคือมีการเดาเกิดขึ้น ดังนั้น Lower Asymptote ของเส้นโค้งคุณลักษณะของข้อสอบจะไม่เป็นศูนย์ พารามิเตอร์ของการเดา (Guessing

Parameter จะแทนด้วย c) ย่อมมีความสำคัญที่จะต้องพิจารณาด้วย ดังนั้น Normal-Ogive Model จึงไม่เหมาะสมสำหรับเหตุการณ์นี้ แม้ว่าเราอาจจะดัดแปลงให้รูปแบบนี้ใช้กับข้อสอบที่มี 3 พารามิเตอร์ได้ แต่ก็ไม่สะดวกเท่ากับการใช้ Logistic Model ซึ่งมีอยู่ 3 รูปแบบดังต่อไปนี้

1) Three-Parameter Logistic Model

มีฟังก์ชันทางคณิตศาสตร์ดังนี้

$$p_g(\theta) = c_g + (1 - c_g) \frac{e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}} \dots\dots\dots (8)$$

เมื่อ c_g = โอกาสที่ผู้มีความสามารถต่ำมากๆ จะตอบข้อสอบข้อที่ g ได้ถูกโดยการเดา

b_g = ค่าความยากของข้อสอบที่ g ซึ่งส่วนใหญ่จะมีค่าอยู่ระหว่าง -3.00 ถึง 3.00 (ความยากของข้อสอบแต่ละข้อมีค่าที่อาจเป็นไปได้ตั้งแต่ $-\alpha$ ถึง α) ค่า b_g ที่เข้าใกล้ -3.00 แสดงว่าข้อสอบข้อนั้นง่าย และค่า b_g ที่เข้าใกล้ 3.00 แสดงว่าข้อสอบนั้นยาก

a_g = ค่าอำนาจจำแนกของข้อสอบที่ g ซึ่งมีค่าตั้งแต่ $-\alpha$ ถึง α แต่ในการทดสอบความสามารถและเบสิกสมฤทธิ์ เราไม่สนใจค่า a_g ที่เป็นลบ และจะใช้ค่า a_g ตั้งแต่ 0.20 ขึ้นไป ซึ่งถือว่ามีความอำนาจจำแนกสูง

$$D = 1.7$$

2) Two-Parameter Logistic Model

รูปแบบที่มี 2 พารามิเตอร์นี้ เราสามารถดัดแปลงมาจากรูปแบบที่มี 3 พารามิเตอร์ได้ โดยการให้ค่า c_g เป็นศูนย์ คือไม่คิดว่าผู้สอบจะตอบคำถามโดยการเดา ดังนั้นรูปแบบนี้จึงมีฟังก์ชันทางคณิตศาสตร์เป็นดังนี้

$$p_g(\theta) = \frac{e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}} \dots\dots\dots (9)$$

3) One-Parameter Model (Rasch Model)

Rasch Model ถูกพัฒนาขึ้นมาโดยไม่เกี่ยวข้องกับ Latent Trait Model อื่นๆ แต่อย่างไรก็ตาม รูปแบบนี้ก็สามารถดัดแปลงได้จาก Two-Parameter Logistic Model โดยสมมติว่าค่าอำนาจจำแนกของข้อสอบทุกข้อเท่ากัน ฟังก์ชันทางคณิตศาสตร์จึงเป็นดังนี้

$$p_x(\theta) = \frac{e^{Dx(\theta-b_x)}}{1 + e^{Dx(\theta-b_x)}} \dots\dots\dots (10)$$

เมื่อ $\bar{x}$ = ค่าอำนาจจำแนกของข้อสอบทุกข้อ ซึ่งมีค่าเท่ากัน

จากจำนวนของพารามิเตอร์ที่แตกต่างกันในแต่ละรูปแบบ ทำให้ได้คุณลักษณะของข้อสอบนั้นแตกต่างกันไปด้วยดังในภาพที่ 9

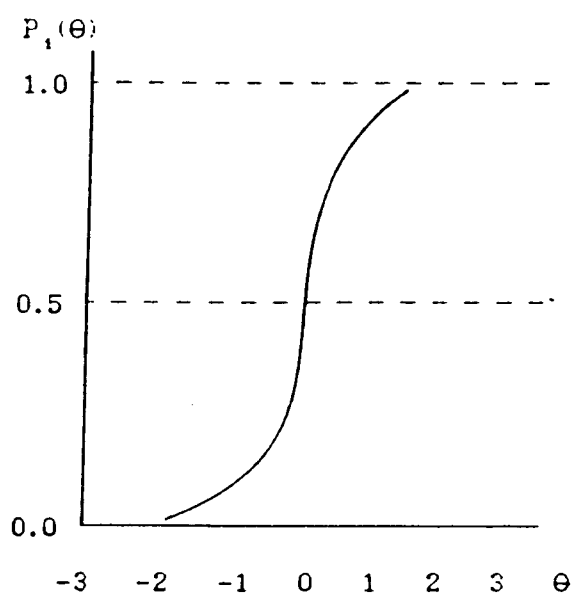
ในการกำหนดเส้นโค้งคุณลักษณะของข้อสอบนั้น ขั้นแรกเราต้องประมาณค่าพารามิเตอร์ของข้อสอบ $[p(\theta), \theta]$ และพารามิเตอร์ของข้อสอบก่อน (a_x, b_x, c_x) สำหรับ Three-Parameter Logistic Model นั้น พารามิเตอร์ทั้ง 2 ชนิด ประมาณได้โดยใช้วิธีการที่เรียกว่า Maximum Likelihood Technique และในปัจจุบันนี้ได้มีการพัฒนาโปรแกรมคอมพิวเตอร์ที่ชื่อ LOGIST เพื่อประมาณค่าพารามิเตอร์ของผู้สอบและของข้อสอบ โดยจะต้องมีจำนวนข้อสอบและจำนวนผู้สอบมากๆ ค่าพารามิเตอร์ที่ประมาณได้จึงจะมีความคงที่

2.5 งานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับระดับขึ้น

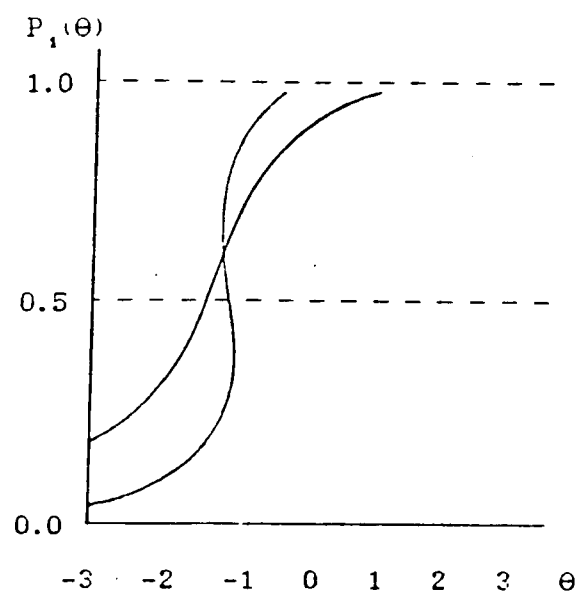
การศึกษาวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับระดับขึ้น ยังไม่เป็นที่แพร่หลายนัก โดยเฉพาะในประเทศไทยยังไม่มีใครศึกษาวิจัยกันเลย เนื่องจากมีความยุ่งยากในการสร้างแบบทดสอบและวิธีการดำเนินการทดสอบ ซึ่งเหมาะที่จะใช้กับระบบคอมพิวเตอร์เท่านั้น แต่จากการศึกษาเอกสารและผลงานวิจัยของต่างประเทศ ได้ข้อสรุปว่าการทดสอบแบบปรับระดับขึ้น มีประสิทธิภาพในการวัดผลมากกว่าการทดสอบแบบเขียนตอบ เช่น

Vale และ Weiss (1975) ได้ศึกษาเปรียบเทียบประสิทธิภาพของแบบทดสอบปรับระดับขึ้น 2 ฉบับกับแบบทดสอบดั้งเดิม โดยศึกษาแบบใช้สถานการณ์จำลอง ใช้โปรแกรมคอมพิวเตอร์สร้างชุดคำตอบแบบทดสอบทั้ง 3 ฉบับ จำนวน 1000 ชุด มีการตรวจให้คะแนนที่แตกต่างกัน 2-3 วิธี มีการเปลี่ยนแปลงค่าอำนาจจำแนกของข้อสอบ ความยาวของแบบทดสอบและระดับความสามารถของผู้สอบ การประเมินคุณภาพของแบบทดสอบ ประเมินจากค่าสหสัมพันธ์ระหว่างระดับความสามารถที่วัดได้จากแบบทดสอบกับสารสนเทศเกี่ยวกับระดับความสามารถที่ไว้ก่อนสอบ และประสิทธิภาพด้านความแม่นยำในการวัดของแบบทดสอบแต่ละฉบับ ผลการศึกษพบว่า

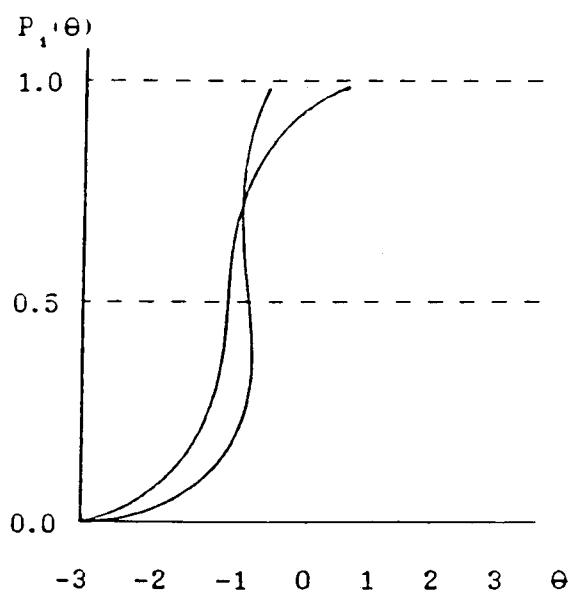
- 1) เมื่อค่าอำนาจจำแนกของข้อสอบเพิ่มขึ้นจนเข้าใกล้ 1.00 ค่าสหสัมพันธ์ระหว่างคะแนนของแบบทดสอบดั้งเดิมกับระดับความสามารถของผู้สอบจะลดลง
- 2) แบบทดสอบดั้งเดิม เมื่อเพิ่มค่าอำนาจจำแนกของข้อสอบขึ้น คะแนนที่ได้จะให้ความแม่นยำในการวัดน้อยลง



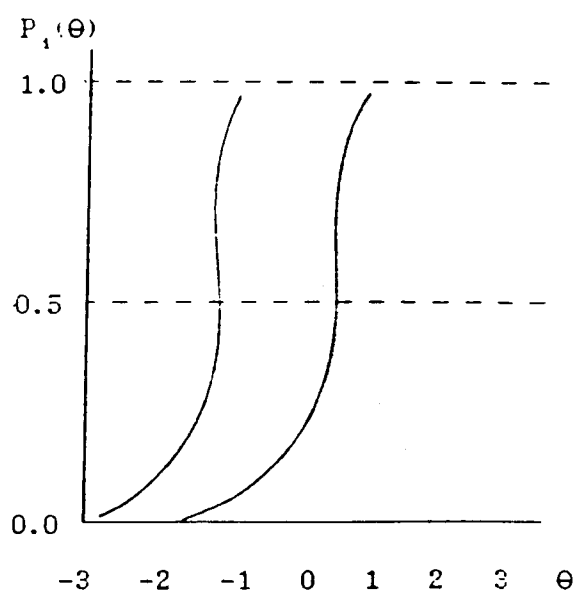
ก. Normal Ogive Model



ข. Three-parameter logistic Model



ค. Two-parameter logistic Model



ง. One-parameter logistic Model

ภาพที่ 9 แสดงได้งคุณลักษณะของข้อสอบ ในแต่ละรูปแบบ

- 3) แต่ละแบบที่ได้จากแบบทดสอบปรับระดับขึ้นจะไม่มีลักษณะดังกล่าวทั้ง 2 อย่าง
- 4) เมื่อค่าอำนาจจำแนกของข้อสอบสูงขึ้น ค่าสหสัมพันธ์ระหว่างระดับความสามารถกับคะแนนที่ได้จากแบบทดสอบปรับระดับขึ้น จะสูงกว่าค่าสหสัมพันธ์ระหว่างระดับความสามารถกับคะแนนที่ได้จากแบบทดสอบดั้งเดิม
- 5) เมื่อมีจำนวนข้อสอบ เท่ากันและค่าอำนาจจำแนกเท่ากัน แบบทดสอบปรับระดับขึ้นจะให้สารสนเทศเกี่ยวกับผู้สอบและความแม่นยำในการวัดมากกว่าแบบทดสอบดั้งเดิม
- 6) ในการทดสอบแบบปรับระดับขึ้น โดยให้ผู้สอบ เริ่ม เข้าทำข้อสอบตามระดับความสามารถของผู้สอบ จะให้คะแนนที่มีคุณสมบัติด้านความแม่นยำในการวัดมากกว่าคะแนนที่ได้จาก เริ่มทำข้อสอบในจุดเดียวกันทุกครั้ง
- 7) การตรวจให้คะแนนของแบบทดสอบปรับระดับขึ้น โดยใช้วิธีของเบย์ (Bayesian Scoring Technique) จะทำให้ได้คะแนนที่มีประสิทธิภาพในการวัดที่ดี

ซึ่งจะเห็นว่า การทดสอบแบบปรับระดับขึ้นมีข้อดีข้อได้เปรียบมากกว่าการทดสอบแบบดั้งเดิม จึงจะนำแบบทดสอบใหม่เข้ามาประยุกต์ใช้กับการศึกษา ดังที่ Weiss และ Kingsbury (1984) ได้ศึกษาผลงานวิจัยต่าง ๆ ที่เกี่ยวกับการนำการทดสอบแบบปรับระดับขึ้นไปประยุกต์ใช้ในการช่วยแก้ปัญหาต่าง ๆ ในทางการศึกษา และได้ข้อสรุปว่าการทดสอบแบบนี้ให้ผลดีต่อการเรียนการสอนในเรื่องต่อไปนี้คือ

- 1) ใช้วินิจฉัยว่าผู้สอบมีความรอบรู้ในส่วใดของเนื้อหาวิชาที่เรียนไป
- 2) ใช้ในการพิจารณาให้หรือผลการเรียนแก่ผู้เรียนได้
- 3) ใช้ในการประมาณการเปลี่ยนแปลงของระดับผลสัมฤทธิ์ของผู้เรียนได้