

ทฤษฎีแนวคิดและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ศึกษาทฤษฎีเกี่ยวข้องกับการศึกษาค้นคว้า โดยศึกษาเนื้อหาเป็นอันดับดังนี้

1. ทฤษฎีการทดสอบดั้งเดิมกับทฤษฎีการตอบคำถาม
2. ทฤษฎีการตอบข้อคำถาม
 - 2.1 ประวัติความเป็นมาของทฤษฎีการตอบข้อคำถาม
 - 2.2 โมเดลต่าง ๆ ในทฤษฎีการตอบข้อคำถาม
 - 2.3 ข้อตกลงเบื้องต้นของทฤษฎีการตอบข้อคำถาม
 - 2.4 ลักษณะเด่นของทฤษฎีการตอบข้อคำถาม
 - 2.5 พิสัยของค่าพารามิเตอร์
 - 2.6 คุณภาพของการทดสอบ
3. ทฤษฎีการตอบข้อคำถามกับการทดสอบเทเลอร์ (Tailored Testing)
4. การสร้างกลุ่มข้อสอบ
5. การทดสอบเทเลอร์ (Tailored Testing)
 - 5.1 การทดสอบดั้งเดิมกับการทดสอบเทเลอร์
 - 5.2 วิธีการของการทดสอบเทเลอร์
 - 5.3 รูปแบบของการทดสอบเทเลอร์
 - 5.4 การประมาณค่าความสามารถและความคลาดเคลื่อนในการทดสอบเทเลอร์
 - 5.5 การทดสอบเทเลอร์โดยวิธีเบส์
6. การทดสอบอิงเกณฑ์

1. ทฤษฎีทดสอบดั้งเดิมกับทฤษฎีการตอบข้อสอบ

ทฤษฎีการทดสอบดั้งเดิม (Classical Test Theory) ซึ่งปรากฏในหนังสือ Mental Test Theory นั้น (Gullikson, 1950) ได้มีการนำมาใช้เป็นหลักในวิเคราะห์คุณภาพของข้อคำถามเกี่ยวกับค่าความยากง่ายและค่าอำนาจจำแนก ตลอดจนใช้แปลความหมายผลการสอบเพื่อวินิจฉัยความสามารถของผู้สอบแต่ละคน การวัดผลที่ยึดถือทฤษฎีการทดสอบดั้งเดิมนี้มีปัญหาสำคัญ 3 ประการคือ (Hambleton and others, 1978 อ้างถึงใน สำเริง บุญเรือง-รัตน์ 2529 หน้า 41-42)

1. คุณสมบัติของข้อคำถามของแบบทดสอบไม่คงที่แปรเปลี่ยนไปตามกลุ่มตัวอย่าง
2. ในการเปรียบเทียบความสามารถของผู้สอบต่าง ๆ นั้นจะต้องใช้แบบทดสอบที่มีข้อคำถามเหมือนกันหรือเป็นแบบทดสอบคู่ขนาน แต่การทดสอบบางครั้งนั้นเราเลือกข้อสอบมาจำนวนหนึ่งเพื่อวัด "คุณลักษณะใดคุณลักษณะหนึ่ง" ซึ่งกรณีนี้ผู้สอบสองคนก็ไม่จำเป็นต้องตอบกลุ่มคำถามเหมือนกันและเพราะว่าข้อสอบทั้งสองต่างก็วัดคุณสมบัติอันเดียวกันนี้ แต่อาจมีความยากง่ายไม่เท่ากัน คนหนึ่งอาจได้คำถามที่ง่าย แต่อีกคนหนึ่งอาจได้คำถามที่ยาก ฉะนั้นคะแนนของสองคนนี้ก็ไม่สามารถนำมาจัดอันดับเพื่อเปรียบเทียบความสามารถกันได้

3. ข้อมูลที่ได้จากการทดสอบตามมาตรฐานดั้งเดิมนั้น ไม่สามารถบอกได้ว่าผู้สอบมีความสามารถทำอะไรได้หรือมีข้อบกพร่องตอนใด เนื่องจากผลการสอบไปเปรียบเทียบกับกลุ่มผู้สอบนั่นเอง

ปัญหาทั้ง 3 ประการนี้ ทฤษฎีการทดสอบดั้งเดิมไม่สามารถเสนอแนวทางแก้ไขได้ นักวัดผลจึงได้เสนอทฤษฎีใหม่ เพื่อแก้ปัญหาดังกล่าวนั้นคือ ทฤษฎีเลแทนท์เรท (Latent Trait Theory) ซึ่งได้มีการพัฒนาทฤษฎีนี้กันอยู่หลายโมเดล ทฤษฎีหนึ่งซึ่งเป็นที่นิยมใช้ในทฤษฎีเลแทนท์เรท คือ ทฤษฎีการตอบข้อคำถาม (Item Response Theory)

2. ทฤษฎีการตอบข้อคำถาม

2.1 ประวัติความเป็นมาของทฤษฎีการตอบข้อคำถาม

ผู้ที่ริเริ่มให้หลักการเบื้องต้นของทฤษฎีคุณลักษณะแฝง (Latent Trait Theory) คือ เฟอ์กูสัน (Ferguson, 1942) และลอเลย์ (Lawley; 1943) ซึ่งทฤษฎีการตอบข้อคำถาม (Item Response Theory) ก็เป็นส่วนหนึ่งของทฤษฎีคุณลักษณะแฝง ผู้ศึกษาในระยะเริ่มแรกที่อยู่ในแนวความคิดอันเดียวกันนี้ได้แก่ บรอกเดน (Brogden, 1946) ทักเกอร์ (Tucker, 1946) แครอลล์ (Carroll, 1950) และครอนบัคกับวอร์ริง (Cronbach and Worring, 1952)

ในปี ค.ศ.1952 ลอร์ด (Lord) ได้เสนอ วิทยานิพนธ์ในระดับปริญญาเอกของเขาซึ่งในวิทยานิพนธ์นั้นเสนอ โมเดลหรือทฤษฎีการตอบข้อคำถาม เขาเรียกว่าทฤษฎีลักษณะโค้งข้อสอบ (Item Characteristic Curve Theory: ICC) ลอร์ด กล่าวว่า พฤติกรรมการตอบสนองข้อสอบแต่ละข้อของผู้สอบ สามารถอธิบายได้ด้วยโมเดลทางคณิตศาสตร์ คือ โค้ง-ลักษณะของข้อสอบแต่ละข้อมีลักษณะเป็นโค้งปกติสะสม หรือ ที่เรียกว่าโมเดลโค้งปกติสะสม (Normal Ogive Model) ซึ่งโมเดลนี้กล่าวถึงพารามิเตอร์ 2 ตัว คือค่าความยากและค่าอำนาจจำแนก แต่เนื่องจากโมเดลนี้มีความยุ่งยากในการคำนวณ จึงทำให้ลอร์ดใช้เวลาในการพิสูจน์ความเหมาะสมและความเป็นไปได้ของทฤษฎีเป็นเวลา 10 ปี

ในปี ค.ศ.1960 ราสค์ (Goerg Rasch) ได้เสนอโมเดลความเป็นอิสระจากกลุ่มตัวอย่าง พารามิเตอร์ตัวเดียว คือ ค่าความยาก ซึ่งเรียกว่าราสค์โมเดล (Rasch Model) ซึ่งได้รับความสนใจอย่างมากและทำให้เกิดมีการวิจัยในระยะต่อมาอย่างกว้างขวาง ผู้นำคนสำคัญในสหรัฐอเมริกาคือ เบนจามิน ไรท์ (Benjamin Wright) แห่งมหาวิทยาลัยชิคาโก

ในปี ค.ศ. 1965 ลอร์ด (Lord) ได้ข้อสรุปอันหนึ่งจากการใช้กลุ่มตัวอย่างกว่าแสนคน ผลจากการศึกษา แสดงให้เห็นว่าปัญหาที่ทำให้งานศึกษาของเขาต้องชะงักไปเป็นเวลานานนั้น ไม่ได้เป็นปัญหาแต่อย่างใด และทฤษฎีการตอบข้อคำถามเหมาะสมกับข้อสอบแบบเลือกตอบ เขาจึงเริ่มศึกษาทฤษฎีการตอบข้อคำถามอีกครั้ง

ในปี ค.ศ. 1968 ลอร์ด และโนวิก (Lord and Novick) ได้พิมพ์หนังสือ *Statistical Theories of Mental Test Scores* ในหนังสือเล่มนี้บทที่ 17-20 เขียนโดย เบิร์นบอม (Birnbau) ซึ่งเป็นนักสถิติที่มีชื่อเสียงมาก ได้แสดงให้เห็นถึงรายละเอียดทางคณิตศาสตร์ของพารามิเตอร์ 2 ตัว และพารามิเตอร์ 3 ตัว ของโมเดลโค้งปกติสะสม (Normal Ogive) และโมเดลโลจิสติก (Logistic Model)

โมเดลโลจิสติก ที่เบิร์นบอมเสนอโดยใช้พารามิเตอร์ 2 ตัว คือค่าความยากและค่าอำนาจจำแนก ซึ่งเป็นโมเดลที่ง่ายกว่าของลอร์ด จึงทำให้เป็นที่นิยมใช้กันแพร่หลายและมีการพัฒนาขึ้นเรื่อย ๆ จนสามารถดัดแปลงใช้ได้กับพารามิเตอร์ตัวเดียวและพารามิเตอร์สามตัว

ในปี ค.ศ. 1977 ลอร์ด (Lord) ได้เปลี่ยนชื่อโมเดลของเขาจากทฤษฎีโค้งลักษณะข้อคำถาม (Item Characteristic Curve Theory) เป็นทฤษฎีการตอบข้อคำถาม (Item Response Theory)

2.2 โมเดลต่าง ๆ ในทฤษฎีการตอบข้อคำถาม

ทฤษฎีการตอบข้อคำถามได้มีการพัฒนาให้เกิดโมเดลเฉพาะขึ้น โดยแต่ละโมเดลจะแตกต่างกันตามฟังก์ชันคณิตศาสตร์ และจำนวนพารามิเตอร์ ที่ใช้อธิบายโค้งลักษณะข้อสอบของแต่ละโมเดล เช่น โมเดลโค้งปกติสะสม (The Normal Ogive Model) โมเดลเส้นตรง (The Linear Model) โมเดลโลจิสติก (The Logistic Model) ซึ่งจะขอกกล่าวเฉพาะโมเดลโลจิสติกซึ่งมีโมเดลย่อยอยู่ 3 โมเดล คือ (Hambleton and Cook; 1977: 81-82)

2.2.1 โมเดลที่ใช้พารามิเตอร์ตัวเดียว (One-parameter Logistic Model : Rasch Model)

ราชได้พัฒนาโมเดลของเขาขึ้นมา โดยไม่ได้ศึกษาโมเดลทางคุณลักษณะแฝง (Latent Trait) มาก่อนเลย เรียกว่า ราชโมเดล (Rasch Model) เมื่อได้ตีพิมพ์เสนอผลการศึกษานี้ของเขาออกมา ก็ปรากฏว่า สอดคล้องกับโมเดลคุณลักษณะแฝง (Latent Trait) ในกรณีพารามิเตอร์ตัวเดียวหรือราชโมเดลอาจเป็นกรณีของ โมเดลพารา

มีเตอร์สองตัว (Two-parameter Logistic Model) ของเบิร์นบอม เมื่อที่ทุกข้อมีค่า a_u เท่ากัน ดังสมการ

$$P_u(\theta) = \frac{\exp(Da_u(\theta - b_u))}{1 + \exp(Da_u(\theta - b_u))}$$

โดยถือว่าค่าอำนาจจำแนกของข้อสอบ (a_u) เท่ากับค่าอำนาจจำแนกเฉลี่ย (1) และไม่มีการเดาได้ถูก ($C = 0$) (Hambleton, 1980:75) โดยเหตุที่ Da_u เป็นค่าคงที่ ไรท์และสโตน (Wright and Stone, 1979) จึงนิยมเขียนเป็น

$$P_u(\theta) = \frac{\exp(\theta - b_u)}{1 + \exp(\theta - b_u)}$$

เมื่อ θ คือ ระดับความสามารถที่แท้จริงที่คำนวณจากคะแนนจริงแต่ปรับหน่วยให้เป็นมาตรฐาน

$P_u(\theta)$ คือ ค่าความน่าจะเป็นที่ผู้สอบที่มีความสามารถ θ จะทำข้อสอบ u ได้ถูกต้อง

a_u คือ ค่าอำนาจจำแนกที่เป็นสัดส่วนโดยตรงกับค่าความชันของโค้ง θ จุดเปลี่ยนโค้งหรือจุดที่ชันที่สุด

b_u คือ ค่าความยากที่แสดงระดับความสามารถที่แท้จริงที่สุด โค้งชันที่สุดหรือในกรณีที่ไม่มีการเดาค่า b_u คือ θ จุดความน่าจะเป็น .50

D คือ ค่าคงที่มีค่า 1.7

$\exp$ คือ ค่า 2.7182818

2.2.2 โมเดลที่ใช้พารามิเตอร์สองตัว (Two-Parameter Model)

เบิร์ตบอม ได้เสนอโมเดลโลจิสติกที่มีค่าพารามิเตอร์ 2 ตัว คือ ค่าความยาก และค่าอำนาจจำแนก ในปี ค.ศ.1968 มีสมการดังนี้

$$P_{\theta}(\theta) = \frac{\exp(Da_{\theta}(\theta - b_{\theta}))}{1 + \exp(Da_{\theta}(\theta - b_{\theta}))}$$

2.2.3 โมเดลที่ใช้พารามิเตอร์สามตัว (Three-Parameter Model)

โมเดลนี้ดัดแปลงจากโมเดลที่ใช้พารามิเตอร์สองตัวโดยพารามิเตอร์ที่เพิ่มขึ้น คือ สัมประสิทธิ์การเดา (C_{θ}) มีสมการดังนี้

$$P_{\theta}(\theta) = C_{\theta} + (1 - C_{\theta}) \frac{\exp(Da_{\theta}(\theta - b_{\theta}))}{1 + \exp(Da_{\theta}(\theta - b_{\theta}))}$$

เมื่อ C_{θ} คือ ค่าความน่าจะเป็นที่ผู้สอบมีความสามารถต่ำมาก มีโอกาสจะทำข้อนี้ได้ถูกต้อง

2.3 ข้อตกลงเบื้องต้นของทฤษฎีการตอบข้อคำถาม

2.3.1 ความเป็นอิสระที่ระดับความสามารถเดียวกัน (Local Independence)

ข้อตกลงนี้หมายความว่า โอกาสของผู้สอบคนหนึ่ง จะตอบข้อสอบได้ถูกต้องนั้น ไม่ได้มีผลมาจากคำตอบของข้อสอบข้ออื่นในแบบทดสอบชุดเดียวกัน รวมทั้งหมายความว่า ข้อสอบไม่มีความสัมพันธ์กับความแตกต่างของผู้สอบ ที่มีระดับความสามารถเดียวกัน ซึ่งสามารถทดสอบด้วยสถิติคือ (Warm, 1978: 100)

$$r_{gh|0} = 0, \quad g \neq h$$

เมื่อ $r_{gh|0}$ คือ สหสัมพันธ์แบบเตตระคอร์ริเครหว่างข้อสอบข้อ g กับ h สำหรับผู้สอบที่มีระดับความสามารถเดียวกันอย่างแท้จริง

การใช้สถิติในการทดสอบนี้ คือ 1) จำเป็นต้องใช้กลุ่มผู้สอบมาก 2) ค่าสหสัมพันธ์ต้องไม่แตกต่างจาก 0 อย่างมีนัยสำคัญ การใช้สถิติวิธีนี้มีความยุ่งยาก 2 ประการ

1) ต้องหาสหสัมพันธ์ที่ทุกระดับความสามารถ (θ) ซึ่งคงเป็นไปได้ยากเพราะค่าระดับความสามารถ (θ) มีจำนวนมาก

2) จำเป็นต้องหาค่าสหสัมพันธ์ของข้อสอบทุกคู่ ซึ่งจำนวนคู่ข้อสอบที่จะต้องหาค่าสหสัมพันธ์เตตระคอร์ริคที่แต่ละระดับความสามารถคือ $n(n-1)/2$ เมื่อ n คือจำนวนข้อสอบในแบบทดสอบ ดังนั้นถ้ามีข้อสอบใน 50 ข้อ จะต้องหาค่าสหสัมพันธ์ทั้งหมด 1225 ครั้ง ที่ระดับความสามารถเดียว ถ้าใช้ระดับความสามารถ (θ) ทั้งหมด 10 ค่าก็หมายความว่าเราจะต้องหาสหสัมพันธ์ 12,250 ครั้ง วิธีการที่ง่ายกว่าวิธีแรกก็คือการคำนวณโดยสูตร

$$r_{gh|0} = \frac{r_{gh} - r_{g0}r_{h0}}{\sqrt{1 - r_{g0}^2} \sqrt{1 - r_{h0}^2}}$$

เมื่อ r_{gh} คือ สหสัมพันธ์แบบเตตระคอร์ริคของผู้สอบทั้งหมดระหว่างข้อสอบข้อ g และ h ($g \neq h$)

r_{g0} คือ สหสัมพันธ์แบบไบซีเรียลระหว่างข้อสอบ g กับ θ

ผลจากการคำนวณค่า $r_{gh|0}$ ต้องไม่ต่างจาก 0 อย่างมีนัยสำคัญ

ลอร์ด (Lord อ้างถึงใน สำเนา บุญเรืองรัตน์ 2529 หน้า 46) เสนอแนะว่าให้ตรวจสอบแผนการตอบข้อคำถามของผู้สอบ ซึ่งมีความสามารถปาน ๆ กันว่าข้อสอบแต่ละข้อ เช่น ข้อ 1 และข้อ 2 ข้อ 2 และข้อ 3 นั้นลักษณะการตอบแตกต่างกันหรือไม่ ซึ่งตรวจสอบได้จากการคำนวณค่าไคสแควร์เพื่อทดสอบความเป็นอิสระของการตอบคำถามเหล่านั้นได้

จากวิธีการทดสอบที่กล่าวมาแล้วข้างต้น การทดสอบความเป็นอิสระที่ระดับความสามารถเดียวกัน (Local Independence) มีความยุ่งยากมาก ดังนั้น โดยปกติแล้วจึงไม่มีการทดสอบกัน (Warm, 1978: 101)

2.3.2 ความเป็นมิติเดียว (Unidimensionality)

ข้อตกลงเรื่องนี้มีควมสำคัญมากที่สุด ความเป็นมิติเดียว(Unidimensional) หมายถึง ข้อสอบวัดเนื้อหาความรู้หรือความสามารถเดียวเท่านั้น (Warm, 1978: 101) หรือข้อสอบแต่ละข้อในแบบทดสอบจะต้องวัดในความสามารถหรือคุณลักษณะเดียวกัน หรือมีความเป็นเอกพันธ์ (สำเนา บุญเรืองรัตน์ 2525 หน้า 3-6) โดยทั่วไปแล้ว มักมีข้อตกลงเกี่ยวกับมิติที่จะวัดใช้เป็นเพียงมิติเดียว (Unidimensional) ก็เพราะทำให้ผู้สร้างแบบทดสอบรวมคำถามที่มุ่งวัดความสามารถอย่างใดอย่างหนึ่งไว้ในฉบับเดียวกัน และรวมคะแนนจากการตอบได้ง่าย พร้อมกับแปลความหมายของคะแนนก็สะดวกขึ้น แบบทดสอบที่มีลักษณะเป็นมิติเดียวจะมีลักษณะดังนี้ สมมติว่าแบบทดสอบฉบับหนึ่งมี n ข้อและใช้วัดผู้สอบ r กลุ่ม และสมมติอีกว่าแบบทดสอบนี้วัดกลุ่มบุคคลที่มีความสามารถเท่ากับ θ ถ้านำคะแนนจากการสอบทั้ง r กลุ่มมาแจกแจงความถี่แล้วคะแนนมีลักษณะการแจกแจงอย่างเดียวกัน ก็ถือได้ว่าแบบทดสอบนั้นมีลักษณะเป็นมิติเดียว แต่ถ้าการกระจายของคะแนนมีลักษณะแตกต่างกันไปก็แสดงว่าแบบทดสอบนั้นคงวัดสิ่งอื่น ๆ อีกด้วย ไม่ได้วัดความสามารถเดียวเท่านั้น

หลักที่ถือปฏิบัติกันอย่างหนึ่ง คือ แบบทดสอบที่ดูแล้วน่าเชื่อได้ว่าเป็นมิติเดียว (Unidimensional) ก็ถือว่าเป็นมิติเดียว (Unidimensional) ตัวอย่างแบบทดสอบวัดความสามารถ เช่น แบบทดสอบการใช้ภาษา (Verbal) ตัวเลข (Numerical) มิติสัมพันธ์ (Spatial)

หลักที่ถือปฏิบัติอีกอย่างหนึ่งคือ ข้อสอบในแบบทดสอบวัดในความรู้ที่มีเนื้อหา
 เกี่ยวพันกันเล็กน้อย ก็นำถือว่าข้อสอบเป็นมิติเดียว (Unidimensional) ตัวอย่างที่
 ดีที่สุดในหลักปฏิบัตินี้คือ จากการศึกษาเบจา ไวสส์ และคิงสเบอวี (Bejar, Weiss and
 Kingsbury cited in Warm, 1978: 102) ซึ่งศึกษากับแบบทดสอบชีววิทยาเบื้องต้นในระดับ
 วิทยาลัยฉบับหนึ่ง ซึ่งกระบวนวิชานี้ประกอบด้วยเนื้อหา 3 ส่วนคือ เคมี (Chemistry)
 เซลล์ (Cell) และพลังงาน (Energy) ซึ่งแบบทดสอบฉบับเดียวมีเนื้อหาทั้ง 3 เรื่องก็คงยัง
 เป็นมิติเดียว (Unidimensional) อยู่

ความเป็นมิติเดียว (Unidimensional) จากที่กล่าวมานี้เป็นสิ่งที่แปลก
 มาก ความจริงแล้วสรุปได้จากข้อสอบที่ใช้ในการทดสอบมีความรู้ส่วนหนึ่ง เกี่ยวพันกันเหมือน
 กับว่าคุณสมบัติหรือปฏิกิริยาเคมีมีความจำเป็นที่จะทำความเข้าใจเรื่องเซลล์ และเนื้อหาเรื่อง
 เซลล์ก็มีความสำคัญที่จะทำให้เข้าใจเรื่องพลังงาน

หลักที่ถือปฏิบัติโดยการนิยามไว้นี้บางครั้งก็เกิดความคลาดเคลื่อน จึงนิยม
 ใช้วิธีการเชิงประจักษ์ คือใช้สถิติทดสอบ การทดสอบโดยใช้สถิตินี้ไม่มีวิธีใดดีที่สุด สมบูรณ์ที่สุด
 เหตุผลก็เพราะว่า มีวิธีการทดสอบอยู่หลายวิธี ซึ่งส่วนใหญ่ใช้วิธีวิเคราะห์องค์ประกอบ
 (Factor Analysis) ซึ่งใช้สหสัมพันธ์เตตระคอร์ริคาสสหสัมพันธ์ภายในข้อสอบ (Interitem
 Tetrachoric Correlations) การทดสอบโดยใช้การวิเคราะห์องค์ประกอบเท่าที่ปฏิบัติมีอยู่
 6 วิธี คือ (Warm, 1978: 103) 1) The Eigenvalue Test 2) The Random
 Baseline Test 3) The Biserial Test 4) The Factor Loading Test 5)
 The Congrevnce Test 6) The Commuality Test กล่าวโดยสรุปการทดสอบโดยวิธี
 การวิเคราะห์องค์ประกอบ (Factor Analysis) คือ ถ้าแบบทดสอบไม่เป็นมิติเดียว ข้อ
 สอบในแบบทดสอบจะแบ่งออกเป็นกลุ่ม ๆ ตามองค์ประกอบ (Factors) แบบทดสอบใดมี
 ลักษณะนี้เรียกว่ามีหลายมิติ (Multi-Dimensional)(Hambleton and Traub cited in
 Hambleton and Cook, 1977) กรีนและคณะ (Green and others, 1984: 349)
 กล่าวว่า เป็นไปได้ที่ข้อมูลผลการตอบข้อสอบเหมาะสม (Fit) กับโมเดลของทฤษฎีการตอบข้อคำถาม

นั้นไม่ไวต่อการเสียความเป็นมิติเดียว (Johnes, 1980 cited in Green and others, 1984: 349)

ลอร์ด (Lord, 1980 อ้างถึงใน ผจงจิต อินทสุวรรณ 2530 หน้า 25) กล่าวว่าถ้าแบบทดสอบนั้นเป็น Unidimensional เพียงพอที่จะบอกได้ว่าแบบทดสอบนั้นเป็นไปตามข้อตกลงเรื่อง Local Independence เสมอ แต่แบบทดสอบเป็นไปตามข้อตกลงของ Local Independence บอกไม่ได้ว่าแบบทดสอบนั้นมีความเป็น Unidimensional (Warm, 1978: 107)

2.2.3 ข้อตกลงเกี่ยวกับ ลักษณะของ โค้งจากผลการตอบข้อสอบแต่ละข้อ (Item Characteristic Curve)

โอกาสของการตอบข้อสอบถูกนั้น ไม่ขึ้นอยู่กับความสามารถของกลุ่มประชากร นั่นคือ รูปร่างของเส้นโค้ง ของข้อคำถามไม่ขึ้นอยู่กับการกระจายของความสามารถกลุ่มผู้สอบ ดังนั้นลักษณะของคำถามและคุณสมบัติข้อคำถามไม่ขึ้นอยู่กับลักษณะของกลุ่มผู้สอบ จึงเป็นข้อคำถามที่มีคุณสมบัติคงที่ตลอดเวลา

ในโมเดลราซค์ นอกจากจะมีข้อตกลงในทฤษฎีการตอบข้อคำถาม (IRT) แล้ว ยังมีข้อตกลงเพิ่มขึ้นดังนี้ (อวยพร วิบูลย์กาญจน์ 2526 อ้างถึงใน ชวลิต โพธิ์นคร 2528 หน้า 42) 1) คะแนนเป็นแบบถูก 1 ผิด 0 2) ข้อสอบแต่ละข้อมีอำนาจจำแนกเท่ากันในการคำนวณ จึงให้เป็น 1 และไม่มีการเดาเกิดขึ้น

2.4 ลักษณะเด่นของทฤษฎีการตอบข้อคำถาม

สงบ ลักษณะ (2525 หน้า 49) กล่าวว่า ถ้าข้อตกลงเบื้องต้นของการสอบตรงตามเงื่อนไขของทฤษฎี เราสามารถคำนวณค่าพารามิเตอร์ได้เท่ากับโมเดล ผลลัพธ์ที่ได้รับคือ

1. ค่าพารามิเตอร์ของผู้สอบ คือค่าความยาก ค่าอำนาจจำแนก และค่าสัมประสิทธิ์การเดา จะเป็นค่าที่ไม่เปลี่ยนแปลงไปตามกลุ่มผู้สอบ ไม่ว่าจะนำไปสอบกับผู้ใดก็ตาม

2. เมื่อทราบลักษณะข้อสอบแต่ละข้อของผู้สอบคนใด เราจะสามารถคำนวณหาค่าความสามารถที่แท้จริงของบุคคลนั้นได้ ค่าความสามารถที่แท้จริงนี้มีความสัมพันธ์โดยตรงกับคะแนนจริง

การคำนวณค่าความสามารถที่แท้จริงอาจใช้ข้อสอบข้อใดก็ได้ที่วัดในสิ่งเดียวกัน ลักษณะเช่นนี้ถือว่าเป็นลักษณะของความเป็นอิสระของข้อสอบ

2.5 นิสยของค่าพารามิเตอร์

ในทฤษฎีการตอบข้อคำถามค่าพารามิเตอร์ของข้อสอบ (Item Parameter) ได้แก่ ค่าความยาก (b_u) ค่าอำนาจจำแนก (a_u) ค่าสัมประสิทธิ์การเดา (C_u) และค่าพารามิเตอร์ของผู้สอบ (Examinee Parameter) ได้แก่ ระดับความสามารถของผู้สอบ (θ) นิสยค่าพารามิเตอร์ต่าง ๆ เหล่านี้มีดังนี้ (Hambleton and Cook, 1977 : 81)

ค่าความยากของข้อสอบ (b_u) มีค่าตั้งแต่ $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติจะมีค่าอยู่ระหว่าง -2 ถึง +2 ค่า -2 แสดงว่าข้อสอบนั้นง่ายมาก และค่า +2 แสดงว่าข้อสอบข้อนั้นยากมาก

ค่าอำนาจจำแนกข้อสอบ (a_u) มีค่าตั้งแต่ $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติจะมีค่าตั้งแต่ 0 ถึง +2 เพราะค่าอำนาจจำแนก (a_u) เป็นลบแสดงว่าข้อสอบไม่ดีใช้ไม่ได้จึงตัดข้อสอบข้อนั้นทิ้งไป ค่า 0 แสดงว่าข้อสอบไม่มีอำนาจจำแนกหรือมีอำนาจจำแนกต่ำ ค่า +2 แสดงว่าข้อสอบมีอำนาจจำแนกสูง

ระดับความสามารถของผู้สอบ (θ) มีค่าอยู่ระหว่าง -3 ถึง +3 ค่า -3 แสดงว่ามีความสามารถต่ำ และค่า +3 แสดงว่าความสามารถสูง

2.6 คุณภาพของการทดสอบ

สำหรับการวิเคราะห์คุณภาพของแบบทดสอบ หรือการพิจารณาความแน่นอนของการประมาณค่าความสามารถที่แท้จริงนั้น ตามทฤษฎีการตอบข้อคำถาม จะใช้การพิจารณาจากค่าสาระประโยชน์ของแบบทดสอบ (Test Information Function) ซึ่งได้จากผลรวมของค่าสาระประโยชน์ของข้อสอบแต่ละข้อ (Item Information Function) นั้นเอง (Birnbau, 1968 อ้างถึงใน ผจงจิต อินทสุวรรณ 2525 หน้า 64) ฟังก์ชันค่าสาระประโยชน์ของแบบทดสอบเป็นส่วนกลับกำลังสองของความยาวช่วงความเชื่อมั่น (Confidence Interval) ที่เป็นผลมาจากการประมาณค่าความสามารถของผู้สอบจากการทำข้อสอบของเขา โค้งค่าสาระประโยชน์

ของแบบทดสอบเป็นเครื่องแสดงถึงความถูกต้องแน่นอน (precision) ของค่าที่ประมาณได้
แสดงในรูปสมการได้

$$I(\theta) = \sum I(\theta, u)$$

ค่าสาระประโยชน์ของข้อสอบ ในแต่ละค่าความสามารถมีความสัมพันธ์กับความชัน
(Slope) ถ้ามีความชัน (Slope) มาก ค่าสาระประโยชน์ก็จะมากด้วย ความชัน (Slope)
ของฟังก์ชันการตอบข้อคำถามหาได้โดย (Warm, 1978: 65)

$$P' = \frac{1.7a(1-c)\exp(1.7a(\theta-b))}{[1 + \exp(1.7a(\theta-b))]^2}$$

เมื่อ a , b และ c คือ ค่าพารามิเตอร์ข้อสอบ

P' คือ ความชัน

อย่างไรก็ตาม การวัดความชันนี้ไม่สมบูรณ์เสียทีเดียวในการอธิบายสาระประโยชน์ของ
ข้อสอบ ดังนั้นจึงใช้วิธีการหาค่าสาระประโยชน์ของข้อสอบโดยตรงจากสูตร

$$I(\theta, u) = \frac{P^2}{P(\theta)Q(\theta)} = \frac{(1.7a)^2(1-c)}{[c+\exp(1.7a(\theta-b))][1+\exp(-1.7a(\theta-b))]^2}$$

เมื่อ $I(\theta, u)$ คือ ค่าสาระประโยชน์ของข้อสอบแต่ละข้อ

$P(\theta)$ คือ ความน่าจะเป็นที่ผู้สอบที่มีความสามารถ θ จะตอบข้อสอบข้อนั้น
ถูก

$Q(\theta)$ คือ $1-P(\theta)$

P' คือ ความชันของโค้งลักษณะข้อสอบที่ระดับความสามารถ θ

จากสมการแสดงให้เห็นว่า ฟังก์ชันค่าสาระประโยชน์ของข้อสอบแต่ละข้อขึ้นอยู่กับความชันของโค้งลักษณะข้อสอบ (ICC) และความแปรปรวนของคะแนนจากข้อสอบแต่ละระดับของความสามารถยิ่งโค้งลักษณะข้อสอบชันเท่าไร และความแปรปรวนเล็กลงเท่าไร โค้งแสดงสาระประโยชน์ของข้อสอบก็จะยิ่งสูงขึ้น ความสูงของโค้งค่าสาระประโยชน์ของข้อสอบที่ระดับหนึ่งของความสามารถ เป็นเครื่องวัดโดยตรงของความมีประโยชน์ของข้อคำถามนั้น ในการวัดความสามารถที่ระดับนั้นได้แน่นอน (ผจงจิต อินทสุวรรณ 2525 หน้า 64)

ในกรณีที่เป็นโมเดลพารามิเตอร์เดียว (ราชคโมเดล) ก็จะหาสาระประโยชน์ของข้อสอบได้จากสูตร (McKinley and Reckase, n.d. : 4)

$$I(\theta, u) = \frac{\exp(-(\theta-b))}{[1+\exp(-\theta-b)]^2}$$

ซึ่งการหาค่าของสาระประโยชน์ของข้อสอบในโมเดลราชคนี้ ไรท์และสโตน (Wright and Stone, 1979 : 73) ได้ทำตารางสำเร็จไว้ในหนังสือ Best Test Design

3. ทฤษฎีการตอบข้อคำถามกับการทดสอบเทเลอร์

ทฤษฎีการทดสอบแบบเดิม (Classical Test Theory) ไม่เหมาะสมกับการทดสอบที่จัดข้อสอบให้เหมาะสมกับระดับความสามารถของผู้สอบ (Adaptive Testing) หรือการทดสอบแบบเทเลอร์ (Tailored Test) ทฤษฎีการทดสอบแบบเดิมดัชนีที่จะเป็นตัวชี้ถึงความเชื่อมั่น ความเที่ยงตรง และคุณภาพของข้อสอบนั้นจะสัมพันธ์กับผู้สอบทั้งหมดซึ่งสอบแบบทดสอบชุดเดียวกัน แต่การทดสอบที่จัดข้อสอบให้เหมาะสมกับระดับความสามารถ ผู้สอบแต่ละคนจะได้รับข้อทดสอบ

แตกต่างกัน ทฤษฎีที่เหมาะสมกับการทดสอบที่จัดข้อสอบให้เหมาะสมกับระดับความสามารถ ก็คือ ทฤษฎีคุณลักษณะแฝง (Latent Trait Theory) ซึ่งปรากฏในตำราของ ลอร์ดและโนวิก (Lord and Novick, 1968) และเมื่อปี ค.ศ. 1980 ลอร์ดก็ได้อธิบายให้เหตุผลทฤษฎีนี้อย่าง สมบูรณ์ ซึ่งในปัจจุบันเรียกว่าทฤษฎีการตอบข้อคำถาม (Green and others, 1984:348)

ไวสส์และคิงส์เบอรี (Weiss and Kingsbury, 1984:362) ได้กล่าวถึงองค์ ประกอบของการทดสอบแบบเดิฟท์พีหรือการทดสอบเทเลอร์

1. โมเดลทฤษฎีการตอบข้อคำถาม (Item Response Model) ซึ่งทฤษฎีการตอบข้อ คำถามที่แพร่หลายในขณะนี้ การทดสอบแบบเดิฟท์พีที่ใช้ในคอมพิวเตอร์ (CAT) สามารถจะเลือก ใช้ได้ทั้งโมเดลโลจิสติก พารามิเตอร์ตัวเดียว พารามิเตอร์สองตัว และพารามิเตอร์สามตัว หรือโมเดลนอร์มอลอวโจว (Normal Ogive Models) การเลือกโมเดลที่จะใช้ขึ้นอยู่กับธรรมชาติของข้อสอบที่ใช้ (ตอบอิสระหรือแบบเลือกตอบ) และผลการตอบเหมาะกับโมเดลที่เลือก

2. กลุ่มข้อสอบ (Item Pool) เป็นสิ่งที่จำเป็นสำหรับการทดสอบแบบอะเดิฟท์พี คือ จะต้องมียุุ่มข้อสอบซึ่งค่าพารามิเตอร์ของข้อสอบประมาณโดยทฤษฎีการตอบข้อคำถาม ไม่มีข้อ กำหนดที่เฉพาะในเรื่องขนาดของยุ่มข้อสอบ ซึ่งยุ่มข้อสอบประมาณ 100 ข้อ ก็สามารถที่จะ ให้ผลที่ดีได้ โดยที่ค่าความยากของข้อสอบจะต้องกระจายเต็มมาตรวัดคุณลักษณะแฝง การวัดจะมี ประสิทธิภาพมากที่สุดจากการใช้ข้อที่มีอำนาจจำแนกสูง

3. การเริ่มต้นสอบ (Entry Level) ในการทดสอบแบบอะเดิฟท์พี หรือเทเลอร์ เป็นไปได้ที่นักเรียนจะได้สอบข้อสอบที่มีระดับความยากแตกต่างกัน เช่น นักเรียนที่มีระดับผล สัมฤทธิ์สูงก็จะสามารถเริ่มทำข้อสอบที่มีความยาก เพราะระดับความยากของข้อสอบจะวัดระดับ ความสามารถที่สูงขึ้น ๆ ตามลำดับ ความคลาดเคลื่อนของการเริ่มต้นแบบจะไม่มีผลกระทบต่อ ผลสรุป แต่การเริ่มต้นที่มีความเที่ยงตรงจะช่วยลดจำนวนของข้อสอบที่จะใช้วัด

4. กฎการเลือกข้อสอบ (Item Selection Rule) กระบวนการคัดเลือกข้อสอบที่มี ประสิทธิภาพนั้นมี 2 วิธี คือ 1) ค่าสาระประโยชน์สูงสุด (Maximum Information) (Weiss, 1982) และ 2) วิธีเบย์ส์ (Bayesian) (Owen, 1969, 1975) วิธีการเลือก

ข้อสอบแบบค่าสาระประโยชน์สูงสุด เป็นการเลือกข้อสอบที่ค่าค่าสาระประโยชน์ของข้อสอบใกล้เคียงกับระดับความสามารถขั้นสุดท้าย (๑) ที่ประมาณได้ และการเลือกข้อสอบโดยวิธีเบส์เป็นการเลือกข้อสอบที่มีค่าความแปรปรวนต่ำสุด ณ ระดับความสามารถ (๑) ที่ประมาณได้

5. วิธีการให้คะแนน (Scoring Method) ระดับความสามารถแฝงของแต่ละคนสามารถที่จะประมาณได้จากวิธีความเป็นไปได้สูงสุด (Maximum Likelihood) หรือวิธีเบส์ (Bayesian Method) (Bejar and Wiss, 1979) โดยวิธีเบส์ (Owen, 1969, 1975) เป็นการกำหนดระดับความสามารถ ณ จุดหนึ่ง จากผลการตอบข้อสอบเพียงข้อเดียว ซึ่งวิธีความเป็นไปได้สูงสุด (Maximum Likelihood) ไม่สามารถจะกำหนดได้ แต่วิธีเบส์ค่าประมาณความสามารถ (๑) จะใช้การถดถอย (regression) จากค่าประมาณความสามารถที่ได้มาก่อน (prior ๑) ซึ่งผลของการถดถอยนั้นจะทำให้ใช้แบบทดสอบที่สั้นลง

วิธีการทั้งสองจะประมาณค่า ความแปรปรวน และค่าความคลาดเคลื่อน ซึ่งจะใช้สร้างช่วงของความเชื่อมั่น (confidence interval ซึ่งเรียกว่า credibility interval ในวิธีเบส์) ของการประมาณความสามารถของนักเรียน ซึ่งช่วงของความเชื่อมั่นนี้จะช่วยให้ทราบถึงระดับความเที่ยงตรงของค่าความสามารถของนักเรียนแต่ละคนที่ประมาณได้ เมื่อช่วงของความแปรปรวนแคบลงก็จะทำให้การประมาณค่าความสามารถมีความเที่ยงตรงสูงขึ้น ดังนั้นช่วงของความเชื่อมั่นนี้จะ เป็นประโยชน์มากต่อการทดสอบเทเลอร์

6. เกณฑ์การสิ้นสุด (Termination Criterion) คุณลักษณะที่สำคัญของการทดสอบเทเลอร์ที่ใช้ในคอมพิวเตอร์ (CAT) นั้น จะทำข้อสอบเท่าที่จำเป็นของผู้สอบแต่ละคน ซึ่งการทดสอบจะกระทำต่อเนื่องจนถึงระดับที่ถือเป็นเกณฑ์ว่ามีความเที่ยงตรง ดังนั้น ผู้สอบแต่ละคนก็จะมีจุดสิ้นสุดการทดสอบที่แตกต่างกัน

4. การสร้างกลุ่มข้อสอบ (Item Pool)

สิ่งจำเป็นในการทดสอบเทเลอร์ (Tailored Test) จะต้องมีการสร้างกลุ่มข้อสอบซึ่งประมาณค่าด้วยทฤษฎีการตอบข้อคำถาม ถึงแม้ว่าจะสามารถทำได้โดยไม่ต้องใช้ทฤษฎีตอบข้อคำถาม

แต่พื้นฐานของทฤษฎีตอบข้อคำถาม (IRT) ก็มีข้อดีเหนือกว่าวิธีอื่น ๆ ข้อดีประการหนึ่งก็คือการทำ
ให้ค่าความยากของข้อสอบ และค่าประมาณความสามารถของผู้สอบอยู่ในมาตราวัดอันเดียวกัน
ดังนั้น จึงเป็นไปได้ที่จะคัดเลือกข้อสอบที่มีความยากเหมาะสมกับแต่ละบุคคล (Garrison and
Baumgarten, 1986:25) การทดสอบเทเลอร์ต้องใช้ข้อสอบจำนวนมาก จึงจำเป็นต้องมี
กลุ่มข้อสอบ (Item Pool)

ในการสร้างกลุ่มข้อสอบ การที่จะให้กลุ่มผู้สอบต้องสอบข้อสอบชุดเดียวเป็นจำนวน
ร้อย ๆ ข้อนั้น เป็นสิ่งที่ยากและไม่เป็นที่นิยมกระทำกัน ไวท์และสโตน (Wright and Stone
1979:98-106) ได้เสนอวิธีการสร้างกลุ่มข้อสอบโดยใช้การเชื่อมโยงแบบทดสอบ (Linking
Test) กล่าวคือ จะใช้ข้อสอบร่วมเพื่อเชื่อมโยงข้อสอบสองฉบับเข้าด้วยกัน สมมติว่าต้องการ
สร้างกลุ่มข้อสอบที่มีจำนวนข้อสอบ 100 ข้อ เราก็ต้องสร้างแบบทดสอบขึ้น 2 ฉบับ ดังนี้คือ
(Warm, 1978 : 114)

1. แบบทดสอบฉบับที่ 1 ประกอบด้วยข้อสอบข้อที่ 1-60
2. แบบทดสอบฉบับที่ 2 ประกอบด้วยข้อสอบข้อที่ 40-100

ข้อสอบแต่ละฉบับมีจำนวน 60 ข้อ มีข้อสอบจำนวน 20 ข้อ ที่เป็นข้อสอบร่วมในแบบ
ทดสอบทั้งสอง นำแบบทดสอบทั้งสองไปสอบกับนักเรียน 2 กลุ่ม (กลุ่มละ 1 ฉบับ) แล้วนำผล
การตอบของแบบทดสอบ 2 ฉบับ ไปวิเคราะห์แยกกัน เมื่อต้องการที่จะทำให้ข้อสอบทั้ง 100 ข้อ
อยู่ในมาตราวัดเดียวกัน (Same Scale) ก็จะกระทำโดย

1. หาค่าเฉลี่ยและค่าความคลาดเคลื่อนมาตรฐานของค่าความยากเฉพาะส่วน
ที่เป็นข้อสอบร่วม 20 ข้อ ของแต่ละฉบับ
2. จัดกระทำค่าพารามิเตอร์ของข้อสอบทั้งหมดให้เป็นข้อสอบชุดเดียวกัน โดยใช้สมการ (Warm, 1978: 113-115)

$$b_1 = \left[\frac{SD_{b_1}}{SD_{b_2}} \right] b_2 + \left[\bar{b}_1 - \left[\frac{SD_{b_1}}{SD_{b_2}} \right] \bar{b}_2 \right] \quad (4.1)$$

$$\theta_1 = \left[\frac{SD_{b1}}{SD_{b2}} \right] \theta_2 + \left[\bar{b}_1 - \left[\frac{SD_{b1}}{SD_{b2}} \right] \bar{b}_2 \right] \quad (4.2)$$

เมื่อ b_1 คือ ค่าความยากข้อสอบของกลุ่มผู้สอบกลุ่ม 1

b_2 คือ ค่าความยากข้อสอบของกลุ่มผู้สอบกลุ่ม 2

$\bar{b}_1$ และ SD_{b1} คือ ค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของค่าความยากข้อสอบของกลุ่มผู้สอบกลุ่ม 1

$\bar{b}_2$ และ SD_{b2} คือ ค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของค่าความยากข้อสอบของกลุ่มผู้สอบกลุ่ม 2

$$a_1 = a_2 \left[\frac{SD_{b2}}{SD_{b1}} \right] \quad (4.3)$$

เมื่อ a_2 และ b_2 คือ ค่าอำนาจจำแนกและค่าความยากบนสเกลเดิม

a_1 และ b_1 คือ ค่าอำนาจจำแนกและค่าความยากบนสเกลใหม่

สำหรับค่าโอกาสการเดา (c-values) จะเท่าเดิมเพราะค่าเหล่านี้อยู่บนแกน P(0) ของฟังก์ชันการตอบข้อคำถาม (IRF) ดังนั้นค่าโอกาสการเดาจึงไม่ต้องเปลี่ยนแปลง (Warm, 1979:115) การให้ข้อสอบร่วม (Anchor Items) เพื่อที่จะเชื่อมโยงข้อสอบให้อยู่ในมาตรวัด (scale) เดียวกันต้องมีข้อสอบร่วม (Anchor Items) อย่างน้อย 17 ข้อ ถ้ามากกว่า 17 ข้อก็จะยิ่งดี (Warm, 1979:115)

บางครั้งแบบทดสอบ 2 ฉบับแตกต่างกันโดยไม่มีข้อสอบร่วม (Common Items) ไปสอบกับกลุ่มผู้สอบ 2 กลุ่ม คนละเวลากัน และมีผู้สอบส่วนหนึ่งได้สอบแบบทดสอบทั้งสองฉบับ เรียกว่ากลุ่มผู้สอบร่วม (Anchor Persons) ในลักษณะนี้สามารถทำได้ 2 วิธี คือ

วิธีที่ 1 ถ้าจำนวนผู้สอบเพียงพอทำโดยรวมกระดาษคำตอบของกลุ่มผู้สอบร่วม (Anchor Persons) ในการสอบแบบทดสอบทั้งสองฉบับเข้าด้วยกัน เหมือนกับกระทำให้แบบทดสอบทั้งสองฉบับเป็นข้อสอบยาว 1 ฉบับ ค่าความสามารถ (θ - scales) ของแบบทดสอบทั้งสองสามารถที่จะปรับสเกลให้รวมอยู่ในสเกลเดียวกันโดยใช้สมการที่ 4.2

วิธีที่ 2 แยกวิเคราะห์เป็นรายฉบับ จากนั้นใช้ค่าความสามารถ (θ - scales) 2 ค่า ของกลุ่มผู้สอบร่วม คำนวณหาค่าเฉลี่ยและค่าความคลาดเคลื่อนมาตรฐานของแต่ละแบบทดสอบแล้วใช้สมการที่ 4.4 เป็นตัวปรับ (rescale) ค่าความสามารถของผู้สอบทั้งหมด

$$\theta_1 = \left[\frac{SD_{\theta_1}}{SD_{\theta_2}} \right] \theta_2 + \left[\bar{\theta}_1 - \left[\frac{SD_{\theta_1}}{SD_{\theta_2}} \right] \bar{\theta}_2 \right] \quad (4.4)$$

ใช้สมการที่ 4.5 เป็นตัวปรับ (rescale) ค่าความยาก (b-values)

$$\theta_1 = \left[\frac{SD_{\theta_1}}{SD_{\theta_2}} \right] b_2 + \left[\bar{\theta}_1 - \left[\frac{SD_{\theta_1}}{SD_{\theta_2}} \right] \bar{\theta}_2 \right] \quad (4.5)$$

ใช้สมการที่ 4.6 เป็นตัวปรับ (rescale) ค่าอำนาจจำแนก (a-values)

$$a_1 = a_2 \left[\frac{SD_{\theta_2}}{SD_{\theta_1}} \right] \quad (4.6)$$

และค่าโอกาสการเดาก็ยังใช้สเกลเดิมอยู่

การใช้ค่าความสามารถ (θ) ในการปรับสเกลมีข้อดลลงว่า ค่าความสามารถ (θ) ไม่เปลี่ยนแปลงในระหว่างการดำเนินการสอบของแบบทดสอบทั้งสอง ในบางครั้งถ้าไม่มีกลุ่มผู้สอบ

ร่วม (Anchor Persons) ที่ทำแบบทดสอบทั้งสองชุด ก็ยังสามารถทำได้โดยสร้างแบบทดสอบฉบับที่ 3 ให้เป็นข้อทดสอบร่วม (Anchor Items) โยงกับแบบทดสอบทั้ง 2 ฉบับแรก

5. แบบทดสอบเทเลอร์ (Tailored Test)

5.1 การทดสอบแบบดั้งเดิมกับการทดสอบเทเลอร์

การทดสอบแบบดั้งเดิมที่ใช้ทดสอบกับผู้สอบเป็นกลุ่มนั้น จะกำหนดให้ผู้สอบที่ซึ่งมีความแตกต่างกันอยู่แล้วนั้นสอบข้อสอบทุกข้อ เมื่อข้อทดสอบยากเกินไปสำหรับผู้สอบอีกกลุ่มหนึ่ง ข้อสอบก็จะไม่สร้างความท้าทายให้ผู้สอบแสดงความสามารถสูงสุดออกมา ซึ่งอาจจะสร้างความเบื่อหน่ายให้กับผู้สอบ ในลักษณะนี้แบบทดสอบที่ยากมากหรือง่ายมาก นำไปสู่การตัดสินใจที่คลาดเคลื่อน (Weiss, 1974:1) ข้อจำกัดอีกประการหนึ่งคือการกำหนดเวลาให้ทำแบบทดสอบเมื่อผู้สอบเป็นกลุ่ม ซึ่งเป็นข้อบังคับขั้นแรกของการทดสอบแบบดั้งเดิมจะเป็นพื้นฐานนำไปสู่ความคลาดเคลื่อนของคะแนนสอบด้วย ถ้าเวลาที่กำหนดไว้ในการทำแบบทดสอบเหมาะสมกับบางส่วนของกลุ่มผู้สอบ ก็จะทำให้เขาได้รับคะแนนสูง ในทางตรงกันข้ามเวลาที่กำหนดไม่เหมาะสมกับผู้สอบอีกส่วนหนึ่ง ก็จะทำให้เขาได้คะแนนต่ำ ดังนั้นการกำหนดเวลาก็จะมีผลต่อความเที่ยงตรงในการทดสอบ และจะนำไปสู่การไม่รู้ค่าความคลาดเคลื่อนของการทดสอบ (Weiss, 1974:1-2) ถึงแม้ว่าการทดสอบเป็นกลุ่มจะถูกออกแบบไว้เพื่อกำจัดความไม่ยุติธรรมหรือผลกระทบจากการผู้ดำเนินการสอบ แต่ความมีปฏิสัมพันธ์กันโดยตรงระหว่างผู้ดำเนินการสอบกับผู้สอบก็แสดงให้เห็นว่ายังมีอยู่ด้วยเงื่อนไขด้านการดำเนินการสอบ (Weiss and Weiss, 1973 cited in Weiss, 1974:2) ดังนั้นความแตกต่างกันของผู้ดำเนินการสอบจึงเป็นพื้นฐานนำไปสู่ความคลาดเคลื่อนของคะแนนของผู้สอบแต่ละคน

จากเหตุผลเหล่านี้ ควรจะใช้การทดสอบวิธีอื่นที่จะให้ความเชื่อมั่น และความเที่ยงตรงสูงกว่าการทดสอบแบบดั้งเดิม (Weiss and Betz, 1973, cited in Weiss 1974:2) ปัญหา^{นี้} สามารถแก้ได้อย่างเหมาะสมโดยใช้การทดสอบเทเลอร์ (Tailored Test) ในความหมายเดียวกับแบบทดสอบอะแดปทีฟ (Adaptive Test) การทดสอบเทเลอร์มุ่งที่จะเลือกข้อสอบที่มีความยากเหมาะสมกับระดับความสามารถของผู้สอบแต่ละคน

5.2 วิธีการของการทดสอบเทเลอร์ (Tailored Test)

ลอร์ด (Lord, 1980 อ้างถึงในสุธรรม จันทน์หอม 2529 หน้า 11) กล่าวว่า การใช้แบบทดสอบเทเลอร์ ก็คือ การคัดเลือกข้อสอบที่เหมาะสมเพื่อใช้สอบและคิดคะแนนเป็นคราว ๆ ไป คะแนนจากการทำแบบทดสอบที่ผ่านมา จะเป็นเครื่องชี้แนะการเลือกแบบทดสอบฉบับต่อ ๆ ไปจนเป็นที่พอใจว่าสามารถคำนวณค่าคะแนนจริงได้แล้วจึงยุติการสอบ ซึ่งกรูจเตอร์ (Grujter, 1980 อ้างถึงในสุธรรม จันทน์หอม 2529 หน้า 11) ก็กล่าวไว้ในทำนองเดียวกันว่า การวัดผลด้วยแบบทดสอบเทเลอร์นั้นชุดของคำถามจะถูกเลือกให้เหมาะสมกับระดับความสามารถของผู้สอบแล้ว คะแนนจากการทำแบบทดสอบที่ผ่านมาจะถูกใช้ในการคัดเลือกแบบทดสอบฉบับต่อไปให้ผู้สอบทำตามระดับความสามารถ ผลที่ได้จากการทดสอบจึงเป็นค่าที่ใกล้เคียงกับสมรรถภาพที่แฝงอยู่ในตัวผู้สอบนั้นจริง ๆ

ไวสส์และคิงส์เบอรี (Weiss and Kingsbury, 1984:361) กล่าวว่า การทดสอบเทเลอร์ เป็นวิธีการดำเนินการสอบอย่างหนึ่งซึ่งข้อสอบที่เลือกให้สอบนั้นขึ้นอยู่กับผลการตอบของแต่ละคน ผลการตอบซึ่งเป็นคะแนนจากการสอบข้อสอบ 1 ข้อ จะถูกประมาณระดับคุณลักษณะแฝงของผู้สอบ เพื่อจะคัดเลือกข้อใหม่ต่อไป โดยวิธีการของการทดสอบเทเลอร์ การทดสอบเสนอให้ทำเฉพาะข้อสอบที่เหมาะสมกับระดับความสามารถที่ประมาณได้จากผลตอบเท่านั้น โดยทั่วไปแล้วการเริ่มต้นสอบจะเริ่มจากข้อสอบที่มีความยากระดับกลาง ๆ ของกลุ่มข้อสอบ ถ้าตอบถูกก็จะทำข้อสอบข้อต่อไปที่มีความยากมากขึ้น แต่ถ้าตอบผิดข้อต่อไปก็จะง่ายลง ซึ่งผลการตอบแต่ละข้อจะถูกประมาณค่าความสามารถของผู้สอบ (Green and others, 1984:347)

จากที่กล่าวมา จะเห็นได้ว่าโดยวิธีการดำเนินการทดสอบเทเลอร์นั้นจะมี 2 วิธีการคือ

- 1) การจัดข้อสอบเป็นฉบับ ไว้เป็นชุด ๆ โดยใช้ค่าพารามิเตอร์ของข้อสอบให้มีค่าความยากหลาย ๆ ระดับ เพื่อจะได้แยกให้ผู้สอบได้สอบตามความสามารถของผู้สอบหลังจากการสอบฉบับแรก ดังนั้นผู้สอบจะต้องสอบทีละฉบับ
- 2) ผู้สอบจะสอบทีละข้อ ผลจากการตอบในแต่ละข้อจะถูกประมาณความสามารถเพื่อจะคัดเลือกข้อสอบที่เหมาะสมกับความสามารถต่อไป ซึ่ง ไวสส์และคิงส์เบอรี (Weiss and Kingsbury, 1984:361) กล่าวว่าโดยวิธีการทดสอบแบบแรก

นั่นเป็น วิธีที่หลีกเลี่ยงความยุ่งยากในการประมาณระดับคุณลักษณะภายหลังจากการตอบข้อสอบที่ละเอียด แต่การทดสอบเทเลอร์ในลักษณะที่สองนั้น มีความยุ่งยากในการประมาณค่าความสามารถ แต่เมื่อคอมพิวเตอร์ได้มีพัฒนาและเข้ามาเกี่ยวข้องกับการทดสอบมากขึ้น จึงได้มีการประยุกต์การทดสอบเทเลอร์ในลักษณะที่สองเข้าไปใช้กับคอมพิวเตอร์ (Urry, 1977) ซึ่งทำให้มีความนิยมการทดสอบในลักษณะที่สองมากขึ้น แต่อย่างไรก็ตามโดยวิธีการทดสอบแบบเทเลอร์ในลักษณะแรกนั้นดูจะเหมาะสมที่จะนำมาพัฒนาใช้กับสภาพการวัดผลของไทยเราในขณะนี้ได้เหมาะสมมากกว่า

5.3 รูปแบบของการทดสอบเทเลอร์ (Tailored Test)

การทดสอบเทเลอร์มีผู้เสนอไว้หลายแบบซึ่ง ไวสส์ (Weiss, 1974) ได้สรุปวิธีการสร้างและดำเนินการสอบเป็น 2 วิธีใหญ่ ๆ คือ

1. วิธีสองชั้น (Two-Stage Strategies)
2. วิธีหลายชั้น (Multi-Stage Strategies)

2.1 แบบการแยกคงที่ (Fixed-Branching Models)

2.1.1 แบบรูปปิรามิด (Pyramid Model)

2.1.2 แบบทดสอบเฟล็กซ์ิเลเวล (Flexilevel Test)

2.1.3 แบบทดสอบสแตรดาทีฟ (Stradative Test)

2.2 แบบการแยกทางแปรผัน (Variable Branching Models)

2.2.1 วิธีเบส์ (Bayesian Strategies)

2.2.2 วิธีความเป็นไปได้สูงสุด (Maximum Likelihood)

ในการวิจัยครั้งนี้มีการทดสอบเทเลอร์ทั้ง 2 รูปแบบใหญ่ที่นำมาใช้ในการวิจัย คือทั้งวิธีสองชั้น และวิธีหลายชั้นวิธีเบส์ (Bayesian Strategies) สำหรับรูปแบบในลักษณะวิธีสองชั้นนั้นผู้วิจัยได้เลือกรูปแบบที่เสนอโดย ฟิชเชอร์และเพนเดิล (Fischer and Pendl, 1980)

และการทดสอบเทเลอร์ที่ไรท์และสโตน (Wright and Stone, 1977:151) เสนอ ดังนั้น เพื่อความเข้าใจที่ชัดเจนจึงขอลำดับการทดสอบเทเลอร์วิธีสองขั้นก่อน

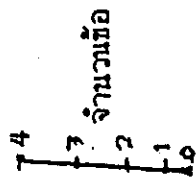
5.3.1 การทดสอบเทเลอร์วิธีสองขั้น (Two-Stage Strategies)

การทดสอบเทเลอร์แบบนี้เป็นแบบง่ายที่สุด คือ วิธีแบบนี้โดยปกติแล้ววิธีนี้จะประกอบด้วยแบบทดสอบนำร่อง(Routing Test) และแบบทดสอบวัดผล(Measurement Test) แบบทดสอบนำร่องจะเป็นแบบทดสอบวัดความสามารถกว้าง ๆ ประกอบไปด้วยข้อสอบที่แตกต่างกันไปด้วยค่าความยาก ซึ่งจะมีตั้งแต่ข้อที่ง่ายที่สุด ไปจนถึงข้อที่ยากที่สุด โดยที่ข้อสอบนั้นจะวัดในสมรรถภาพเดียวกัน ข้อสอบทั้งหมดในชุดนี้จะกระจายระดับความยากเพื่อให้ครอบคลุมระดับความสามารถ ซึ่งแบบทดสอบนำร่องออกแบบไว้สำหรับประมาณค่าความสามารถขั้นเริ่มแรกของผู้สอบแต่ละคน ผลคะแนนที่ได้จากแบบทดสอบนำร่องของผู้สอบแต่ละคน จะเป็นเครื่องชี้เพื่อแยกไปสอบแบบทดสอบวัดผล (Measurement Test) แบบทดสอบวัดผลแต่ละฉบับก็คือแบบทดสอบที่วัดในสมรรถภาพเดียวกัน ซึ่งแตกต่างกันไปตามระดับความยากและกำหนดให้วัดความสามารถที่ละเอียดและแคบมากกว่าแบบทดสอบนำร่อง ดังปรากฏในภาพ 1

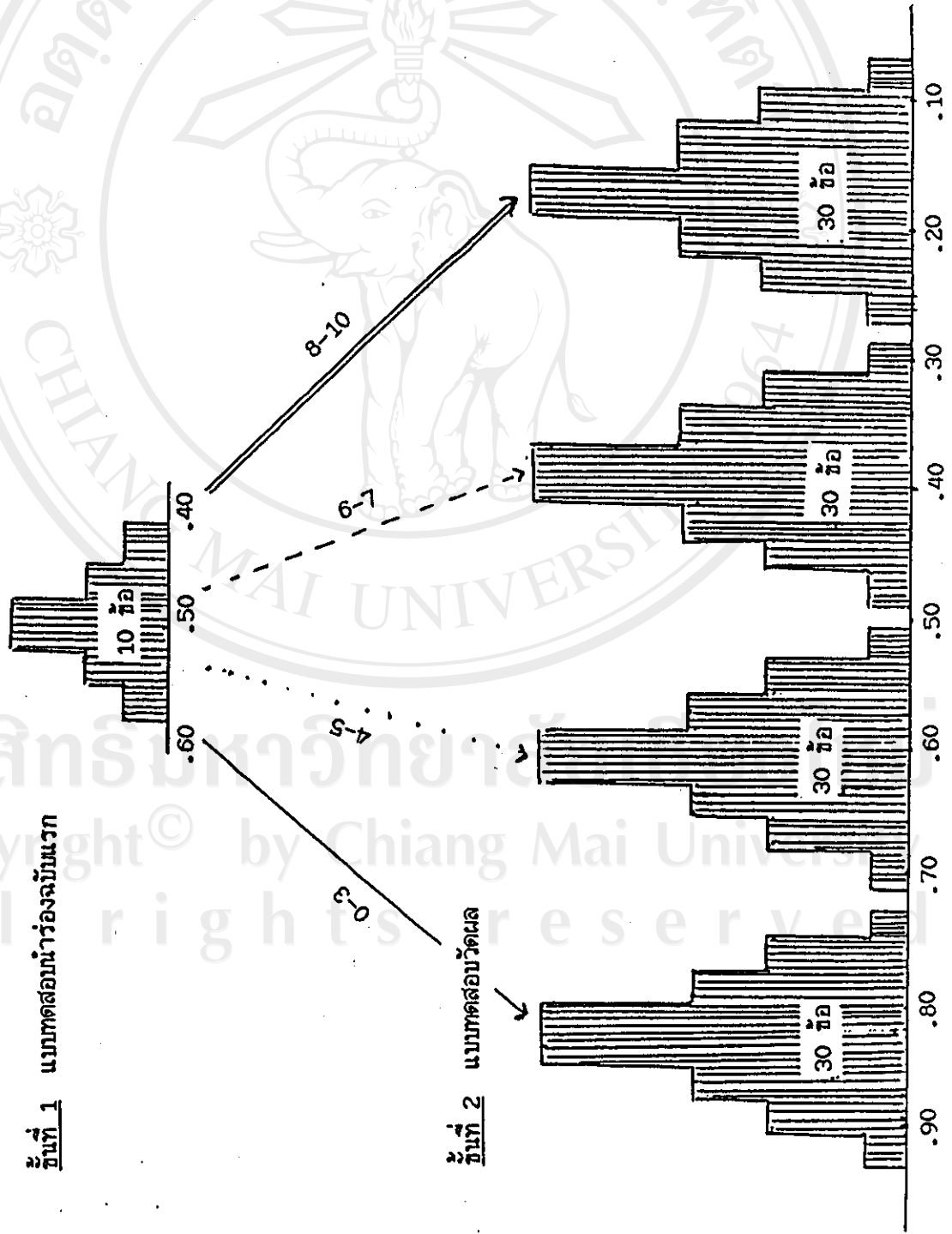
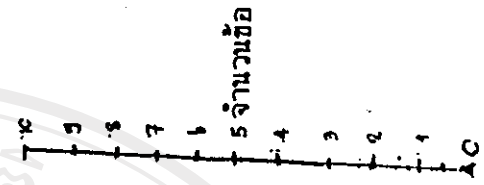
ภาพ 1 แสดงการจัดแบบทดสอบแบบเลอว์ริสสองขั้นที่ข้อสอบในแบบทดสอบนำร่องมีความยาก

อยู่ในระดับปานกลาง

ขั้นที่ 1 แบบทดสอบนำร่องฉบับแรก



ขั้นที่ 2 แบบทดสอบวัดผล



ข้อสอบง่าย

ค่าความยาก/ความสามารถ

ข้อสอบยาก

(ความสามารถต่ำ)

(สัดส่วนการตอบถูก)

(ความสามารถสูง)

ในภาพ 1 แสดงการทดสอบเทเลอรั่ววิธีสองชั้นที่ใช้แบบทดสอบนำร่องเพียงฉบับเดียว แบบทดสอบในลักษณะนี้ศึกษาโดยแองกอฟฟ์และฮัดเดิลสตัน (Angoff and Huddleston, 1958) เบทซ์และไวส์ (Betz and Weiss, 1973, 1974) ลอร์ด (1971) และวูด (Wood, 1971) (Weiss, 1974:5) ในภาพ 1 นี้ แบบทดสอบนำร่องประกอบด้วยข้อสอบจำนวน 10 ข้อ โดยใช้ค่าความยากอยู่ในช่วง .45 - .55 ถ้าให้คะแนนแต่ละข้อแบบถูก 1 ผิด 0 คะแนน จะแปรเปลี่ยนอยู่ระหว่าง 1 - 10 คะแนน การแยกหรือการนำร่องเพื่อจะไปสอบแบบทดสอบ ในชั้นที่สองขึ้นอยู่กับผลการตอบในชั้นแรกของแบบทดสอบนำร่อง ผู้สอบที่ได้คะแนน 0, 1, 2 หรือ 3 ในแบบทดสอบนำร่องก็จะถือว่าเป็นผู้มีระดับความสามารถต่ำสุด ก็จะแยกไปทำแบบทดสอบตามเส้นลูกศร ซึ่งเป็นแบบทดสอบที่มีค่าความยากน้อยสุดที่ข้อสอบในฉบับทั้งหมด 30 ข้อ ซึ่งเป็นกลุ่มข้อสอบที่กระจายอยู่รอบ ๆ ค่าความยาก .85 ในทำนองเดียวกันผู้ที่สอบได้คะแนน 4 หรือ 5 ในแบบทดสอบนำร่องก็จะแยกไปสอบแบบทดสอบวัดผลซึ่งมีค่าความยากมากขึ้น ด้วยกลุ่มข้อสอบที่มีค่าความยากกระจายอยู่รอบ ๆ .63 ทางแยกของคะแนน 6 หรือ 7 ก็จะ ไปสอบแบบทดสอบวัดผลที่มีความยากมากขึ้น (คิดเฉลี่ย $p = .37$) ขณะที่ผู้สอบได้คะแนน 8, 9 หรือ 10 ของแบบทดสอบนำร่องก็จะแยกไปสอบแบบทดสอบวัดผลที่มีความยากมากที่สุด ซึ่งค่าเฉลี่ยเพียง .15

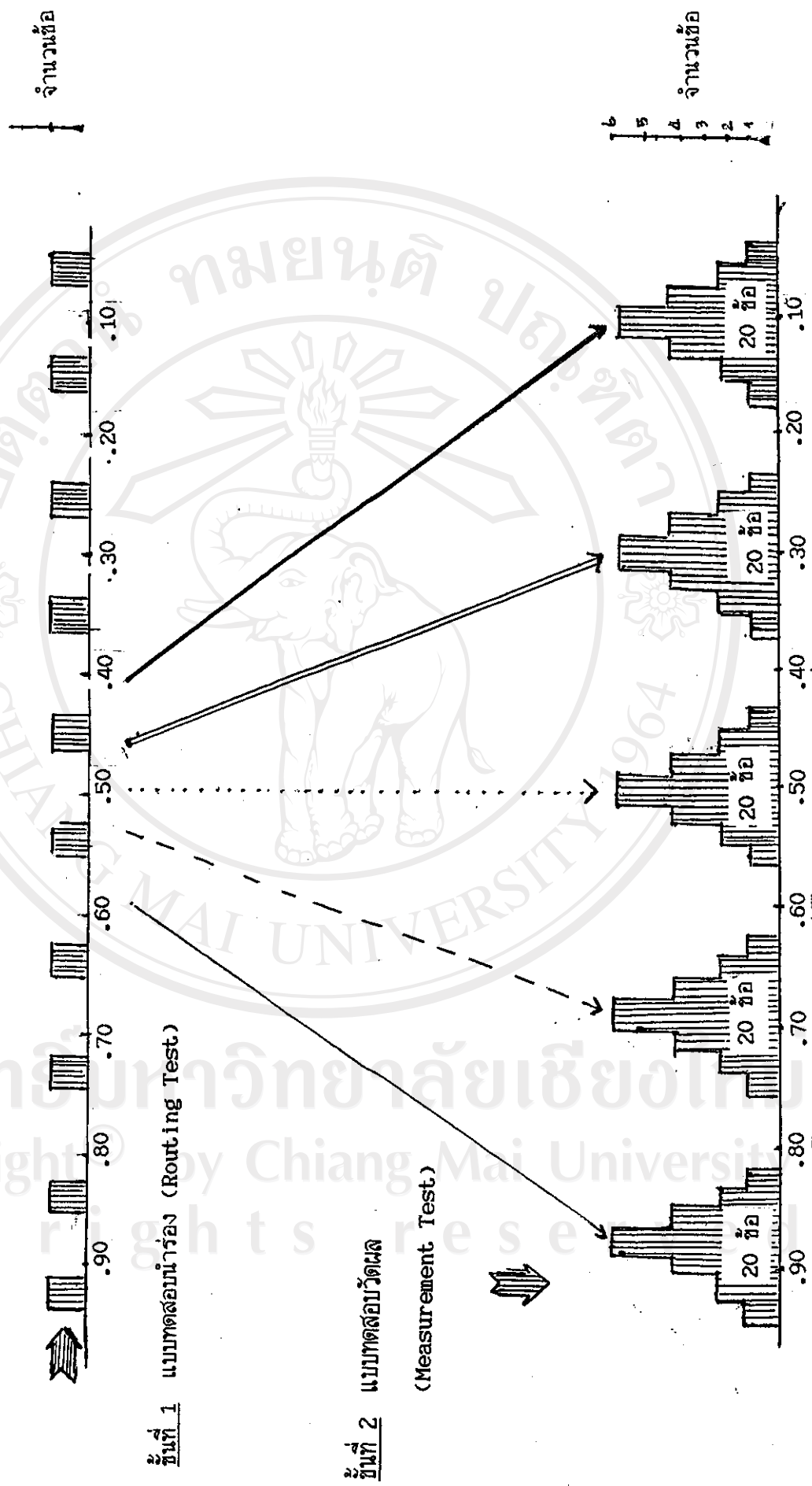
การประหยัดเวลาที่ต้องใช้ในการทดสอบเทเลอรั่ววิธีสองชั้น เห็นได้จากภาพ 1 ถ้าผู้สอบแต่ละคนทดสอบข้อสอบทั้งหมดในภาพ 1 ผู้สอบแต่ละคนจะต้องทำข้อสอบทั้งหมด 130 ข้อ (10 ข้อจากแบบทดสอบนำร่อง และแบบทดสอบวัดผล 4 ฉบับ ๆ ละ 30 ข้อ) การใช้วิธีสองชั้น ผู้สอบแต่ละคน จะทำข้อสอบเพียง 40 ข้อ ถ้าแบบทดสอบนำร่องแยกผู้สอบได้อย่างเหมาะสม ข้อสอบ 40 ข้อ ที่จะทดสอบแต่ละคนอย่างน้อยที่สุด 30 ข้อ (ของแบบทดสอบวัดผล) จะเป็นค่าความยากที่เกือบจะนับได้ว่าเหมาะสมกับระดับความสามารถของผู้สอบแต่ละคน ซึ่งการทำให้เหมาะสมกับระดับความสามารถของผู้สอบแต่ละคน จะเป็นการเพิ่มแรงจูงใจในการทำข้อสอบ อันจะให้ผลดีต่อความเชื่อมั่นและความเที่ยงตรงของการทดสอบ (Weiss and Betz, 1973 cited in Weiss, 1974:5) เบทซ์ และไวส์ (Betz and Weiss, 1974 cited in

Weiss, 1974:5) ได้แสดงให้เห็นว่า การทดสอบเทเลอรั่ววิธีสองชั้นให้ค่าความเชื่อมั่น (การสอบซ้ำ) สูงกว่าแบบทดสอบดั้งเดิมที่มีความยาวเหมือนกัน

แนวทางอีกอย่างหนึ่งที่เป็นข้อเสนอของการทดสอบเทเลอรั่ววิธีสองชั้น ดังภาพ 2 ใน หน้า 39 แสดงให้เห็นว่าข้อสอบในแบบทดสอบนำร่องจะกระจายไปตามระดับความยาก แทนที่จะกระจายอยู่รอบ ๆ ค่าเฉลี่ย ในแต่ละระดับความยากก็จะมีข้อกระทงเพียงข้อเดียวกระจายไปทุกช่วงความยาก การนำร่องเพื่อแยกไปสอบแบบทดสอบวัดผลก็ยังขึ้นอยู่กับผลคะแนนของแบบทดสอบนำร่อง ในภาพ 2 จะมีแบบทดสอบวัดผล 5 ฉบับ ซึ่งวัดไว้อย่างต่อเนื่องตามช่วงของความยาก ในแบบทดสอบวัดผล 1 ฉบับ ก็จะมีลักษณะเหมือนกับแบบทดสอบนำร่องในภาพ 1 การใช้แบบทดสอบสองชั้น ซึ่งข้อสอบในแบบทดสอบนำร่องกระจายตามระดับความยากนี้ศึกษาโดย เคลียร์, ลินน์ และ ร็อค (Cleary, Rinn and Rock, 1968) และ ลินน์, ร็อค และ เคลียร์ (Rinn, Rock and Cleary, 1969 cited in Weiss, 1974:7)

จากรายงานการศึกษาของลินน์ และเคลียร์ (Rinn and Cleary, 1969 cited in Weiss, 1974:7) โดยให้การเปลี่ยนแปลงหลายวิธีการบนพื้นฐานของการทดสอบเทเลอรั่ววิธีสองชั้น ซึ่งผลการศึกษานำไปสู่การเปลี่ยนแปลงในการทดสอบนำร่อง ซึ่งการเปลี่ยนแปลงอันหนึ่งก็คือ การใช้แบบทดสอบนำร่องสองครั้ง ดังภาพ 3 ในหน้า 40 แบบทดสอบนำร่องชุดแรกประกอบด้วยข้อสอบ 10 ข้อ คะแนนจากแบบทดสอบจะแยกเป็น 2 พวก เพื่อแยกไปสอบในแบบทดสอบนำร่องในชุดที่ 2 ผลจากการสอบแบบทดสอบนำร่องในครั้งที่สองในแต่ละฉบับจะแยกเป็น 2 พวก ซึ่งเป็นการจัดจำแนกในขั้นสุดท้าย ดังในภาพ 3 โดยใช้แบบทดสอบวัดผลชุดละ 20 ข้อ ดังนั้น ผู้สอบแต่ละคนจะตอบข้อสอบ 40 ข้อ

ภาพ 2 แสดงการทดสอบแบบเลอว์ริงสองชั้นข้อสอบในแบบทดสอบนำร่องกระจายไปตามระดับความยาก

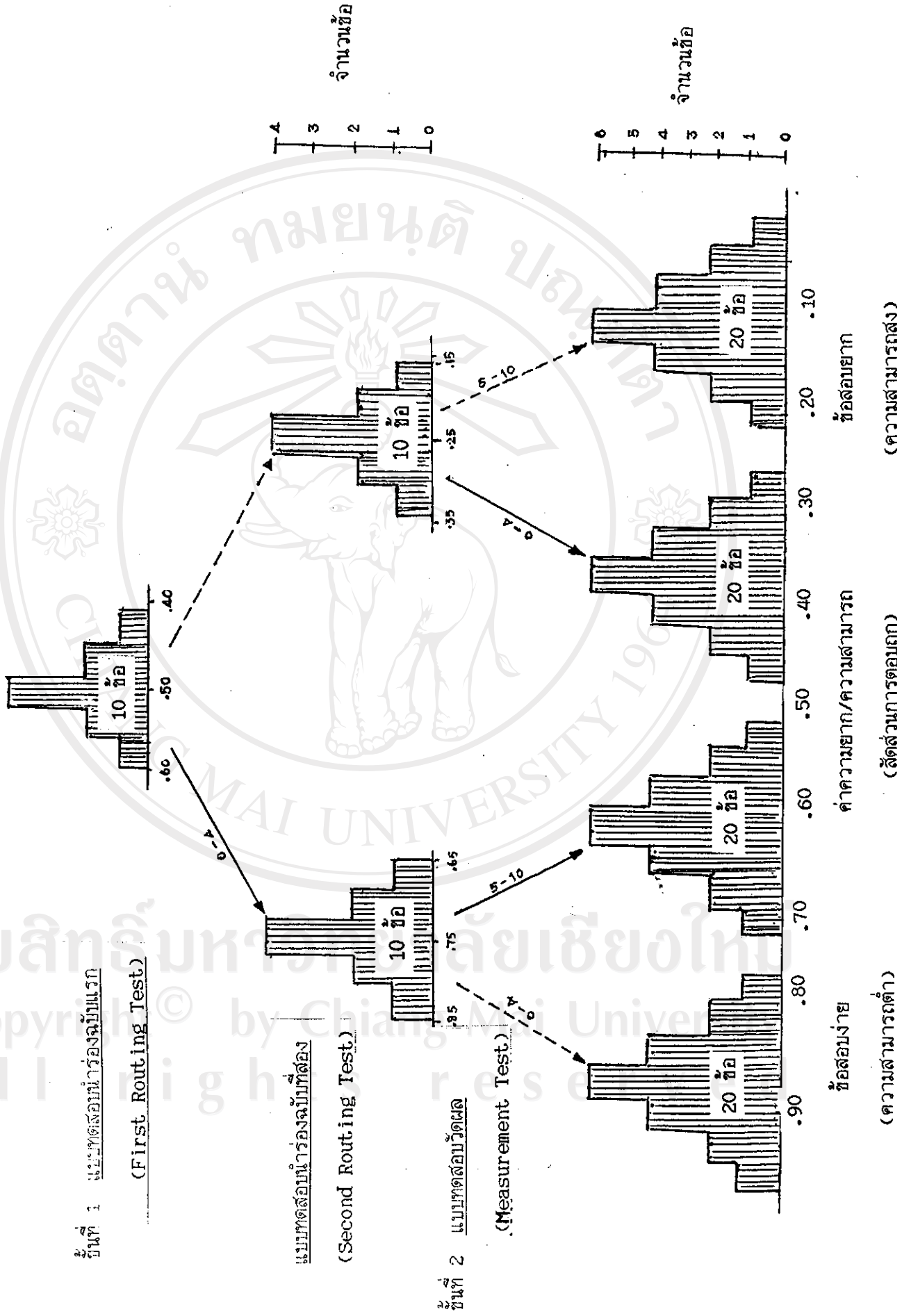


ข้อสอบง่าย (ความสามารถต่ำ)

ค่าความยาก/ความสามารถ (สัดส่วนการตอบถูก)

ข้อสอบยาก (ความสามารถสูง)

ภาพ 3 แสดงการสอบเทียบเลขวิธีสองขั้นที่ใช้การนำร่อง 2 ครั้ง



ลอร์ด (Lord, 1977:126) กล่าวว่า การตัดสินใจที่จะใช้การทดสอบเทเลอร์วิธีสองชั้นนั้นจะต้องคำนึงถึงสิ่งต่อไปนี้

1. ความยาวของแบบทดสอบนำร่อง
2. ความยาวของข้อสอบแต่ละชุดในชั้นที่สอง
3. จำนวนของข้อสอบแต่ละฉบับในชั้นที่สอง
4. ระดับของความยากของแบบทดสอบนำร่อง
5. ระดับความยากแต่ละฉบับในชั้นที่สอง
6. วิธีการให้คะแนนของแบบทดสอบนำร่อง
7. คะแนนจุดตัดในแบบทดสอบนำร่อง เพื่อแยกไปทำข้อสอบในชั้นที่สอง
8. วิธีการให้คะแนนในแบบทดสอบชั้นที่สอง
9. วิธีการรวมคะแนนจากแบบทดสอบในชั้นที่ 1 และ 2

ถ้าแบบทดสอบนำร่องสั้นมาก การนำร่องก็จะมีประสิทธิภาพ ถ้ามันยาวเกินไป ข้อสอบในชั้นที่สองก็จะสั้นกว่าที่ควรจะเป็น ถ้าจำนวนของแบบทดสอบในชั้นที่สองมีน้อยก็จะทำให้การทดสอบไม่ครอบคลุม ถ้ามีแบบทดสอบในชั้นที่สองมากการวัดผลก็จะเป็นที่ยอมรับค่าใช้จ่ายในการพัฒนาข้อสอบก็จะมากเกินไป (Lord, 1977:126) การเลือกแบบทดสอบแต่ละฉบับในชั้นที่สอง และความเหมาะสมในการกำหนดจุดตัดในแบบทดสอบนำร่องเป็นปัจจัยที่สำคัญมากที่สุด (Lord, 1977:127) การกำหนดจุดแยกที่ไม่ดีอาจจะทำให้ผู้สอบจำนวนมาก ๆ ไปสอบแบบทดสอบในชั้นที่สองชุดเดียวกัน (Betz and Weiss, 1973 cited in Lord, 1977:127)

การให้คะแนน (Scoring)

ความแตกต่างกันของแบบทดสอบวัดผล (Measurement Test) ซึ่งประกอบด้วยข้อสอบที่แตกต่างกันตามระดับความยาก วิธีการให้คะแนนของการตอบถูกที่ใช้ในแบบทดสอบดั้งเดิมไม่เหมาะสมกับการทดสอบแบบสองชั้น ดังนั้นจึงต้องมีการพัฒนาวิธีการให้คะแนนแบบใหม่

การให้คะแนนตามค่าความยากเฉลี่ย เป็นวิธีการให้คะแนนแบบหนึ่งของการทดสอบแบบสองชั้น วิธีการนี้คือการคิดค่าความยากเฉลี่ยจากข้อที่ทำถูกทั้งหมดของผู้สอบแต่ละคน คะแนนค่าความยากเฉลี่ยอาจคิดเพียงคำตอบที่ตอบในแบบทดสอบวัดผล (Measurement Test) หรืออาจจะรวมทั้งจำนวนข้อที่ทำในแบบทดสอบนำร่องด้วยก็ได้

ลอร์ด (Lord, 1971 cited in Weiss, 1974:9) ได้พัฒนาวิธีความเป็นไปได้สูงสุด (Maximum Likelihood) เพื่อประมาณความสามารถจากผลการตอบจากแบบทดสอบสองชั้น โดยมีเงื่อนไขว่าการประมาณค่าความสามารถกระจายเป็นปกติ (ค่าเฉลี่ย = 0, ส่วนเบี่ยงเบนมาตรฐาน = 1) รูปแบบที่ลอร์ดเสนอขึ้นอยู่กับข้อตกลงที่ว่าแบบทดสอบนำร่อง (Routing Test) และแบบทดสอบวัดผล (Measurement Test) อยู่ในลักษณะเป็นกลุ่มความยาก (peaked) คือประมาณ .50 การให้คะแนนใช้จำนวนข้อตอบถูก ในแบบทดสอบนำร่องและแบบทดสอบวัดผลของแต่ละคนเทียบกับจำนวนข้อในแบบทดสอบ แปลงค่าคะแนน ค่าความยากและค่าอำนาจจำแนกของข้อสอบ (ซึ่งมีข้อตกลงว่าเท่ากันหมดทุกข้อ) ค่าเหล่านี้จะใช้เป็นตัวกำหนดในการแยกประมาณความสามารถของแบบทดสอบนำร่อง และแบบทดสอบวัดผล การประมาณค่าความสามารถตัวสุดท้ายจะถูกกำหนดโดยผลรวม ส่วนกลับของน้ำหนักการประมาณค่าความสามารถ 2 ตัว โดยการประมาณความแปรปรวนจากการประยุกต์ใช้วิธีนี้ เบตซ์และไวส์ (Betz and Weiss, 1973 cited in Weiss, 1974:9) ชี้ให้เห็นว่าการประมาณค่าความสามารถได้มาโดยวิธีการที่ไม่เที่ยงตรงตามหลักตรรกวิทยา เมื่อใช้การทดสอบจริง เบตซ์และไวส์ได้ศึกษาโดยใช้แบบทดสอบนำร่อง 10 ข้อ และแบบทดสอบวัดผล 30 ข้อ จำนวน 4 ฉบับ ใช้วิธีการกำหนดน้ำหนักของลอร์ดปรากฏว่าให้ข้อสรุปที่ขาดเหตุผล ดังตัวอย่าง ผู้สอบคนหนึ่งตอบถูกข้อสอบ 4 ข้อ ในแบบทดสอบนำร่อง และทำข้อสอบในแบบทดสอบวัดผล 6 ข้อ จะได้ค่าประมาณความสามารถสูงกว่า ผู้ทดสอบที่ได้คะแนนเดียวกันในแบบทดสอบนำร่องแต่ตอบถูกมากกว่า (7-12 ข้อ) ในแบบทดสอบวัดผลชุดเดียวกัน

เคลียร์ และ ลินน์ (Clearey, 1968 and Linn, 1969 cited in Weiss, 1974:9) ได้พัฒนาวิธีการให้คะแนนแบบอื่น ๆ ในการทดสอบเทเลอร์วิธีสองชั้น แต่วิธีเหล่านี้ก็

ไม่ได้ใช้กันแพร่หลาย ซึ่งวิธีการของเขายูบย่นพื้นฐานวิธีการถดถอย (Regression) ที่จะประมาณคะแนน

ข้อดีและข้อจำกัดของการทดสอบเทเลอร์วิธีสองขั้น

ข้อดีของการทดสอบเทเลอร์วิธีสองขั้นเมื่อเปรียบเทียบกับแบบทดสอบดั้งเดิม ถึงแม้ว่าวิธีสองขั้นจะมีลักษณะเช่นเดียวกับแบบทดสอบดั้งเดิม 2 ฉบับ โดยให้ทำแบบทดสอบนำร่องก่อน เมื่อทราบคะแนนแล้วจึงให้แบบทดสอบวัดผล ดังนั้นจึงเป็นสิ่งบอกให้รู้ถึงระดับความสามารถของผู้สอบเพื่อจะจัดข้อสอบให้เหมาะสมกับความสามารถ โดยปกติแล้วแบบทดสอบนำร่องจะสั้นเมื่อเปรียบเทียบกับแบบทดสอบวัดผล โดยที่แบบทดสอบวัดผลนั้นจะให้ค่าสาระประโยชน์ (Information) ต่อข้อมากกว่า และด้วยเหตุนี้จะเป็นประโยชน์ช่วยลดผลกระทบอันอาจเกิดจากแบบทดสอบที่ยากเกินไปหรือง่ายเกินไปกับผู้สอบในแบบทดสอบแบบดั้งเดิม

ข้อดีของการทดสอบวิธีสองขั้นมากกว่าการทดสอบเทเลอร์วิธีอื่นอีกประการหนึ่งคือ การทดสอบคล้ายกับการทดสอบแบบดั้งเดิมที่ใช้กันอยู่ ซึ่งตอนในกระดาษคำตอบ สามารถตรวจให้คะแนนได้ง่าย ๆ

โดยหลักของเหตุผลแล้วการทดสอบเทเลอร์สองขั้นมีข้อจำกัด 2 ประการ คือ

ประการแรก ความคลาดเคลื่อนในการนำร่อง ความคลาดเคลื่อนนี้คือความคลาดเคลื่อนในการกำหนดแบบทดสอบวัดผลให้กับผู้สอบ เพราะความคลาดเคลื่อนจากการวัดในแบบทดสอบนำร่อง ความคลาดเคลื่อนที่ปรากฏประการแรกสำหรับผู้สอบที่ได้คะแนนตกอยู่ใกล้คะแนนจุดแยกที่กำหนดให้ได้รับแบบทดสอบที่ต่างกันไป ซึ่งจากผลการศึกษาของทั้งลินน์ (1969) แองกอฟฟ์ และฮัดเดิลสตัน (Linn, 1969 Angoff and Huddleston, 1958 cited in Weiss, 1974:10) ได้แสดงให้เห็นถึงความคลาดเคลื่อนของการนำร่องมากถึงประมาณ 20 เปอร์เซ็นต์ อย่างไรก็ตาม เบตซ์และไวส์ (Betz and Weiss, 1973 cited in Weiss, 1974:10) แสดงให้เห็นถึงความคลาดเคลื่อนในการนำร่องที่มีประมาณ 4 - 5 เปอร์เซ็นต์ ในการทดสอบแบบทดสอบสองขั้นด้วยคอมพิวเตอร์ และวิธีการนำร่อง 2 ครั้งกับผู้สอบ จะทำให้เกิดความเหมาะสมในการใช้แบบทดสอบวัดผลมากขึ้น

ประการที่สองคือจำนวนข้อคำถามที่ใช้สอบกับผู้สอบ ผลการวิจัยด้วยแบบจำลองแปรผัน (Variable Branching Models) ซึ่งเป็นวิธีหนึ่งของการทดสอบเทเลอร์ เช่น วิธีของเบลล์ (Bayesian Strategies) (Weiss, 1974 : 54 - 61) และแบบทดสอบปรับระดับ (Stradative Test) (Weiss, 1974:50-54) ดูจะชี้ให้เห็นว่าความแตกต่างของจำนวนข้อที่กำหนดให้สอบนั้นจะบอกถึงระดับของความเที่ยงตรงในการวัด ซึ่งรูปแบบของวิธีสองชั้นนั้นกำหนดเกณฑ์ได้แน่นอนว่าผู้สอบแต่ละคนต้องตอบข้อสอบทั้งหมดในแบบทดสอบแบบดั้งเดิม 2 ฉบับ คือแบบทดสอบนำร่องและแบบทดสอบวัดผล ดังนั้นวิธีการนี้จึงจัดจำนวนข้อสอบไม่ค่อยจะเหมาะสมกับความสามารถที่แตกต่างกันของผู้สอบ

งานวิจัยที่เกี่ยวข้องกับการทดสอบเทเลอร์วิธีสองชั้น

ลาร์กิน และไวส์ (Larkin and Weiss, 1975:1-27) ได้ศึกษาเปรียบเทียบแบบทดสอบวิธีสองชั้นกับแบบทดสอบรูปปิรามิด โดยแบบทดสอบทั้งสองสร้างจากกลุ่มข้อสอบจำนวน 369 ข้อ ซึ่งข้อสอบเป็นแบบเลือกตอบชนิด 5 ตัวเลือก ถามเกี่ยวกับคำศัพท์ต่ำกว่าระดับวิทยาลัย กลุ่มข้อสอบพิจารณาจากการประมาณค่าพารามิเตอร์ ตามทฤษฎีการตอบข้อคำถามแต่ละข้อจะมีค่าสัมประสิทธิ์สหสัมพันธ์แบบไบซีเรียลตั้งแต่ .30 ขึ้นไป แบบทดสอบวิธีสองชั้นประกอบด้วยแบบทดสอบนำร่อง 10 ข้อ ผู้สอบที่ทำแบบทดสอบนำร่องได้ 0-4, 5-6, 7-8 และ 9-10 คะแนน จะถูกกำหนดให้แยกไปตอบแบบทดสอบวัดผลในชุดที่ง่ายสุด หรือชุดที่ยากขึ้นชุดใดชุดหนึ่งตามลำดับของคะแนน แบบทดสอบวัดผลมี 4 ชุด ๆ ละ 30 ข้อ การให้คะแนนแบบทดสอบวิธีสองชั้นใช้วิธีประมาณค่าความสามารถตามวิธีของลอร์ด

แบบทดสอบรูปปิรามิดที่ใช้ศึกษามี 3 ฉบับ แต่ละฉบับมี 15 ชั้น มีค่าความแตกต่างระหว่างค่าความยากของข้อสอบในชั้นที่อยู่ติดกันคงที่ ใช้กฎการแยกทางแบบเพิ่ม 1/ลด 1 ใช้วิธีการให้คะแนน 4 วิธี คือ 1) ให้ตามจำนวนข้อที่ตอบถูก 2) ให้ตามค่าความยากเฉลี่ยของข้อสอบที่ทำ 3) ให้ตามค่าความยากเฉลี่ยของข้อสอบที่ตอบถูก และ 4) ให้ตามค่าความยากของข้อสอบข้อสุดท้ายที่ทำ ดำเนินการสอบโดยใช้เครื่องคอมพิวเตอร์ ทดสอบกับนักศึกษา 111 คน

ผลการศึกษาพบว่า ในการทดสอบคะแนนเฉลี่ยที่ได้รับจากการวัดลำดับการทดสอบก่อน หลังระหว่างแบบทดสอบรูปปริมาตร กับแบบทดสอบวิธีสองชั้น โดยใช้การวิเคราะห์ความแปรปรวน ปรากฏว่า ไม่มีความแตกต่างอย่างมีนัยสำคัญทางสถิติ จึงถือได้ว่าการจัดให้ผู้สอบได้สอบแบบ ทดสอบชนิดใดก่อนจะ ไม่มีผลกระทบต่อคะแนนที่ได้รับ ในด้านการกระจายของคะแนนนั้นคะแนนของ แบบทดสอบรูปปริมาตรที่ได้จากการให้คะแนนทั้ง 4 วิธี มีการกระจายแบบรูปโค้งเบ้ทางบวกเล็กน้อย คะแนนที่ให้ตามจำนวนข้อที่ตอบถูกจะมีความเบ้มากที่สุด และแตกต่างจากโค้งปกติอย่างมีนัย สำคัญทางสถิติ ส่วนคะแนนที่ให้ตามค่าความยากเฉลี่ยของข้อสอบที่ตอบถูก ก็จะไม่มีความเบ้เลย คะแนนที่ให้ตามค่าความยากเฉลี่ยของข้อสอบที่ทำและคะแนนที่ให้ตามค่าความยากของข้อสอบข้อสุดท้ายที่ทำคะแนน จะมีการกระจายแบบ โค้งแบนแต่ก็ไม่แตกต่างจากโค้งปกติอย่างมีนัยสำคัญทางสถิติ คะแนนที่ให้ตามจำนวนของข้อที่ตอบถูก และคะแนนที่ให้ตามค่าความยากเฉลี่ยของข้อสอบที่ตอบถูก จะมีการกระจายแบบ โค้ง โค้งขึ้นเล็กน้อย คะแนนที่ให้ตามค่าความยากเฉลี่ยของข้อที่ทำจะมีการ กระจายแบบ โค้งแบนมากที่สุด เมื่อพิจารณาทั้งความเบ้และความโด่งของการกระจายของคะแนน แล้วคะแนน ที่ให้ตามค่าความยากเฉลี่ยของข้อสอบที่ตอบถูกจะแตกต่าง ไปจากโค้งการกระจายปกติ น้อยที่สุด การกระจายของคะแนนที่ได้จากแบบทดสอบชนิด 2 ชั้น จะอยู่ในรูปโค้งเบ้ทางบวกเล็กน้อยและ โค้งเล็กน้อยแต่ทั้ง 2 กรณี ก็มีความแตกต่างอย่างมีนัยสำคัญทางสถิติกับโค้งการกระจาย ปกติ ความเบ้ของโค้งการแจกแจงคะแนนยังมากกว่าโค้งความเบ้ของคะแนนที่ให้โดยวิธีต่าง ๆ ของแบบทดสอบรูปปริมาตรทุกวิธี ความโด่งของ โค้งการกระจายของคะแนนจากแบบทดสอบ 2 ชั้น ก็ยังมียอดแหลมกว่าโค้งความโด่งที่ได้จากแบบทดสอบรูปปริมาตรด้วย เมื่อพิจารณาถึงแบบทดสอบ ย่อยต่าง ๆ ของแบบทดสอบชนิด 2 ชั้น พบว่า แบบทดสอบนำร่องมีความยากของข้อสอบที่ เหมาะสม คือ คะแนนเฉลี่ยข้อถูกเท่ากับ 5.58 จาก 10 คะแนน เมื่อคิดจาก 111 คน ส่วน เบี่ยงเบนมาตรฐานของคะแนนข้อถูกมีค่ามาก (2.61) แสดงว่าแบบทดสอบนำร่องสามารถจำแนก ผู้สอบออกตามความสามารถได้ดี คะแนนเฉลี่ยข้อถูกของแบบทดสอบวัดผลรวมทั้ง 4 ชุด เท่ากับ 18.56 จาก 30 คะแนนมีค่ามากกว่าครึ่งหนึ่งของจำนวนข้อในแบบทดสอบวัดผล เมื่อพิจารณาแยก คะแนนในแต่ละฉบับ คะแนนเฉลี่ยข้อถูกจะอยู่ระหว่าง 17.00 ถึง 19.93 หรือประมาณ

57 เปอร์เซนต์ ถึง 66 เปอร์เซนต์ ของจำนวนข้อสอบทั้ง 30 ข้อ แสดงว่าความยากของแบบทดสอบวัดผลเหมาะกับกลุ่มผู้สอบในแต่ละกลุ่มของผู้ที่แยกมาสอบในแบบทดสอบชุดนี้ อย่างไรก็ตาม จะมีการเกิดขึ้นในการทำแบบทดสอบวิธีสองชั้นมากกว่าการทำแบบทดสอบรูปปิรามิด การให้คะแนนในแบบทดสอบวัดผลทั้ง 4 ฉบับ พบว่าให้ผลเกือบคงที่ในแบบทดสอบวัดผล 3 ฉบับแรก อย่างไรก็ตามแบบทดสอบวัดผลฉบับที่ 1 ให้คะแนนกระจายกว้างกว่าคะแนนจากแบบทดสอบวัดผลฉบับอื่น ๆ และแบบทดสอบนำร่องเพื่อจัดกลุ่มผู้สอบให้แบบทดสอบวัดผลฉบับที่มีความยากน้อยที่สุดหรือฉบับที่ 4 จะมีส่วนเบี่ยงเบนมาตรฐานมากที่สุดถึง 1.15

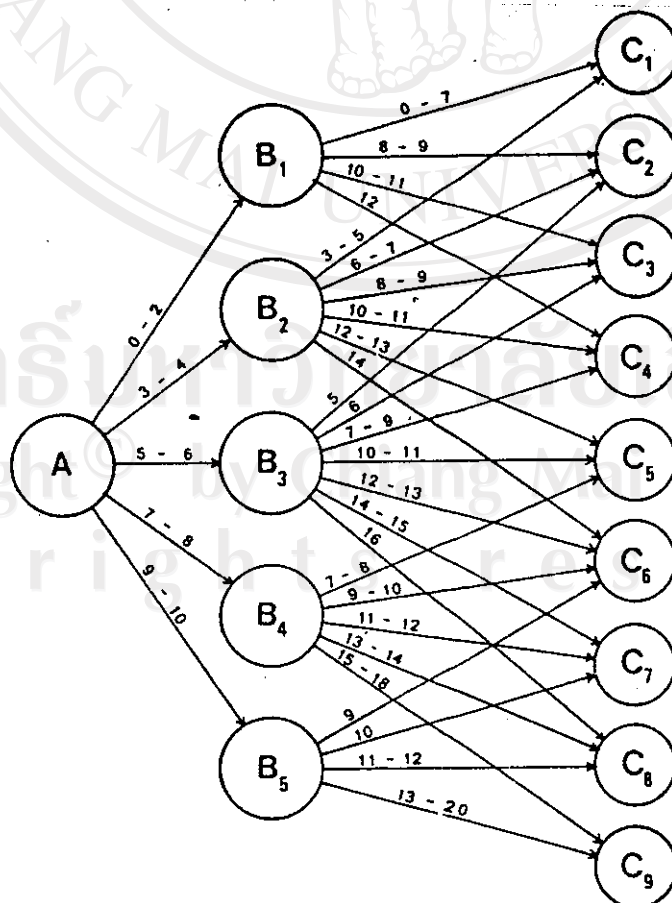
ความสัมพันธ์ของคะแนนที่ได้จากแบบทดสอบรูปปิรามิดกับแบบทดสอบวิธีสองชั้น พบว่าค่าสหสัมพันธ์แบบเพียร์สัน-โปรดักโมเมนต์ เท่ากับ .84, .81, .79 และ .83 ซึ่งเป็นค่าสหสัมพันธ์ของคะแนนที่ได้จากแบบทดสอบรูปปิรามิดที่ให้คะแนนตามวิธีที่ 1 ถึง 4 กับคะแนนที่ได้จากแบบทดสอบวิธีสองชั้น ตามลำดับ เกี่ยวกับค่าความเชื่อมั่นคงที่ภายในใช้ค่าสัมประสิทธิ์ความเชื่อมั่นของ ฮอยยี่ พบว่าแบบทดสอบวิธีสองชั้นนั้น แบบทดสอบนำร่องมีค่า .72 (10 ข้อ กลุ่มตัวอย่าง 111 คน) และแบบทดสอบวัดผลฉบับที่ 1 ถึง 4 มีค่า .84, .66, .75 และ .72 ตามลำดับ (ทุกฉบับมี 30 ข้อ และมีจำนวนกลุ่มตัวอย่าง 21, 20, 27 และ 43 ตามลำดับ) แสดงให้เห็นว่า แบบทดสอบวัดผลที่มีค่าอำนาจจำแนกเฉลี่ยมากที่สุดจะมีค่าความเชื่อมั่นต่ำสุด (แบบทดสอบวัดผลฉบับที่ 2) แบบทดสอบที่มีค่าความยากเฉลี่ยมากที่สุดจะมีค่าความเชื่อมั่นสูงสุด (แบบทดสอบวัดผลฉบับที่ 1) ส่วนแบบทดสอบวัดผลฉบับที่ 4 ที่มีค่าความยากเฉลี่ยน้อยที่สุด จะมีค่าความเชื่อมั่นเท่ากับแบบทดสอบนำร่อง

รูปแบบของการทดสอบเทเลอร์ที่จะเสนอต่อไปนี้เป็นรูปแบบที่จะใช้ในการวิจัย ซึ่งอยู่ในลักษณะของวิธีสองชั้น ซึ่งจะเสนอเป็นลำดับดังต่อไปนี้

5.3.2 แบบทดสอบเทเลอร์แบบแตกแขนง (Branched Test)

แบบทดสอบเทเลอร์ชนิดนี้เป็นรูปแบบ ที่เสนอโดยฟิชเชอร์และเพนเดิล

(Fischer and Pendl, 1980 : 178) ซึ่งมีลักษณะใกล้เคียงกับแบบทดสอบวิธีสองขั้นมากรั้งได้อะแกรมภาพ 4 แบบทดสอบ A เป็นแบบทดสอบชุดแรก ซึ่งเรียกว่า แบบทดสอบกรอง (Screening Test) หรือแบบทดสอบนำร่อง (Pilot Test) เพื่อจะใช้สอบผู้สอบทุกคนประกอบด้วย ข้อสอบจำนวน 10 ข้อ มีค่าพารามิเตอร์กระจายอยู่ในช่วง -1.53 ถึง 1.39 (ดูตาราง 1) ซึ่งแบบทดสอบฉบับนี้จะให้การประมาณค่าความสามารถของผู้สอบอย่างคร่าว ๆ (หยยาบ) ผลคะแนนจากการสอบในแบบทดสอบ A เป็นเครื่องชี้เพื่อให้ผู้สอบไปทำแบบทดสอบชุด B คือ $B_1, B_2, \dots, B_5$ ซึ่งแบบทดสอบในชุดนี้สร้างขึ้นเพื่อจำแนกความสามารถให้แคบมากยิ่งขึ้น ผลรวมของการตอบในแบบทดสอบ A และ B ก็จะเป็นเครื่องชี้ให้ไปทำแบบทดสอบในชุด C ซึ่งแบบทดสอบแต่ละฉบับในชุด C จะถูกคัดเลือกเข้ามาพ่วงต่อเพื่อจะวัดระดับความสามารถที่ระดับสูงสุด ค่าพารามิเตอร์ของข้อสอบในแบบทดสอบชุด B ตั้งแต่ $B_1, \dots, B_5$ ค่าความยากของข้อสอบจะไม่คาบเกี่ยวกัน แต่ในแบบทดสอบชุด C จะมีค่าความยากของข้อสอบส่วนหนึ่งคาบเกี่ยวกับแบบทดสอบอีกฉบับหนึ่งที่อยู่ติดกัน (ดูตาราง 1)



ภาพ 4 ไดอะแกรมแสดงแบบทดสอบเทเลออร์แบบแตกแขนง (Brachted Test)

จากไดอะแกรมในภาพ 4 ผู้สอบที่ทำคะแนนในแบบทดสอบ A ได้ 4 คะแนน ถ้าจะเลือกแบบทดสอบ B_2 ให้ทำผลคะแนนจากการสอบในแบบทดสอบ A และ B_2 จะเป็นดัชนีชี้สำหรับเลือกแบบทดสอบในชุด C ต่อไป

ตาราง 1 แสดงการจัดค่าความยากของข้อสอบเข้าฉบับของแบบทดสอบเทเลอว์แบบแตกแขนง (Branched Test)

A	B_1	B_2	B_3	B_4	B_5
-1.53	-2.43	-1.35	-.24	.79	1.61
-1.26	-2.37	-1.31	-.16	.95	1.71
-.95	-2.16	-1.16	-.07	1.02	1.72
-.61	-2.08	-1.06	.00	1.14	1.92
-.29	-1.98	-.97	.13	1.20	2.13
.02	-1.87	-.91	.23	1.31	2.21
.39	-1.69	-.77	.29	1.39	2.24
.75	-1.69	-.58	.40	1.51	2.32
1.12	-1.61	-.53	.53	1.61	2.61
1.39	-1.51	-.42	.61	1.71	2.78

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
-2.43	-1.90	-1.37	-.82	-.36	.18	.65	1.16	1.68
-2.37	-1.84	-1.26	-.78	-.28	.24	.75	1.23	1.72
-2.16	-1.69	-1.16	-.66	-.16	.34	.84	1.39	1.77
-2.08	-1.61	-1.06	-.52	-.08	.42	.95	1.47	1.92
-1.98	-1.42	-.95	-.45	.05	.56	1.06	1.53	2.03
-1.90	-1.37	-.82	-.36	.18	.65	1.16	1.68	2.17
-1.84	-1.26	-.78	-.28	.24	.75	1.23	1.72	2.20
-1.69	-1.16	-.66	-.16	.34	.84	1.39	1.77	2.30
-1.61	-2.06	-.52	-.08	.42	.95	1.47	1.92	2.58
-1.42	-.95	-.45	.05	.56	1.06	1.53	2.03	2.72

5.3.3 การทดสอบเทเลอร์แบบทดสอบตามสภาพ (Status Tailoring)

การทดสอบเทเลอร์แบบนี้ ต้องจัดแบบทดสอบไว้เลือก เพื่อให้เหมาะสมกับความสามารถของนักเรียนแต่ละคน คือ มีฉบับง่าย ปานกลาง ยาก หรือยากมากขึ้น การทดสอบ ครูผู้สอนในวิชาที่ทดสอบจะรู้ข้อมูลเบื้องต้นของนักเรียนแต่ละคนได้ดี จากเกรดเฉลี่ยผลการเรียน ดังนั้นข้อมูลเหล่านี้จะช่วยให้ครูเลือกข้อสอบฉบับใดฉบับหนึ่งได้ โดยที่แบบทดสอบนั้นเหมาะสมกับความสามารถของนักเรียน (Wright and Stone, 1978:153-156)

5.3.4 การทดสอบเทเลอร์แบบทดลองทำ (Performance Tailoring)

การทดสอบชนิดนี้ต้องอาศัยแบบทดสอบนำร่อง (Pilot Test) ซึ่งมีข้อสอบ

ประมาณ 5 - 10 ข้อ ซึ่งมีค่าความยากข้อสอบกระจายให้เต็มมาตราวัดความสามารถ โดยที่ค่าความยากเฉลี่ยควรอยู่ประมาณ 0.00 ผลจากการทำในแบบทดสอบนำร่อง จะได้ค่าประมาณความสามารถ และความคลาดเคลื่อนมาตรฐานของค่าความสามารถของผู้สอบแต่ละคน ซึ่งค่าประมาณทั้งสองค่านี้จะนำไปใช้ประโยชน์ในการจัดข้อสอบชุดใหม่ เพื่อให้เหมาะสมกับระดับความสามารถที่ประมาณได้ โดยที่ค่าความยากของข้อสอบจะกระจายอยู่ในช่วง $\pm$ ความคลาดเคลื่อนมาตรฐานของค่าประมาณความสามารถ นั่นคือ ค่าความยากของข้อสอบต่ำสุดจะเท่ากับค่าความสามารถ ลบด้วยค่าความคลาดเคลื่อนมาตรฐานของค่าความสามารถ และค่าความยากของข้อสอบมากที่สุดจะเท่ากับค่าความสามารถบวกด้วยค่าความคลาดเคลื่อนมาตรฐานของค่าความสามารถ ผลจากการสอบในแบบทดสอบชุดใหม่นี้จะนำไปประมาณความสามารถ และค่าความคลาดเคลื่อนมาตรฐานของค่าความสามารถใหม่ (Wright and Stone, 1979:156-160)

5.3.5 การทดสอบเทเลอร์แบบให้ตัดสินใจ (Self-Tailoring)

ไรท์และสโตน (Wright and Stone, 1979:161-164) ได้เสนอวิธีการทดสอบแบบเทเลอร์วิธีนี้ โดยจัดค่าความยากของข้อสอบเรียงลำดับจากง่ายสุดไปจนถึงยากที่สุด ผู้สอบแต่ละคนจะมีจุดเริ่มต้นในการทดสอบต่างกัน โดยที่ข้อสอบที่เป็นจุดเริ่มต้นนี้ผู้สอบคิดว่าจะเหมาะสมกับระดับความสามารถของตนเอง กล่าวคือไม่ง่ายจนเกินไปที่จะตอบได้ถูก การสอบก็จะตอบข้อสอบข้อต่อไป ซึ่งมีระดับความยากสูง ๆ และจะหยุดสอบเมื่อผู้สอบตอบผิดติดต่อกัน 3 ข้อ ผลจากการตอบของผู้สอบแต่ละคน จะนำไปประมาณค่าความสามารถต่อไป

จากที่กล่าวถึงมาแล้ว จะเห็นว่า การทดสอบเทเลอร์แบบทดสอบตามสภาพ (Status Tailoring) การทดสอบเทเลอร์แบบให้ทดลองทำ (Performance Tailoring) และการทดสอบเทเลอร์แบบให้ตัดสินใจเลือกเอง (Self-Tailoring) คือการทดสอบเทเลอร์วิธีสองขั้น (Two - Stage Strategies) นั่นเอง โดยที่การทดสอบเทเลอร์แบบทดสอบตามสภาพและการทดสอบเทเลอร์แบบให้ตัดสินใจเลือกเองนั้นใช้ครูผู้สอนและตัวผู้สอบเอง เป็นตัวนำร่องแทนการใช้แบบทดสอบนำร่อง สำหรับการทดสอบเทเลอร์แบบให้ทดลองทำ (Performance

Tailoring) ก็ใช้แบบทดสอบนำร่อง (Pilot Test) เพื่อใช้เลือกแบบทดสอบที่เหมาะสม เช่นเดียวกับการทดสอบวิธีสอบขึ้นที่กล่าวมาแล้ว

ในการทดสอบเทเลอร์แบบให้ตัดสินใจเลือกเอง (Self-Tailoring) จะมีการทดสอบความเหมาะสมของผู้สอบและการควบคุมคุณภาพ (Person Fit and Quality Control) เมื่อผู้สอบจะต้องตอบข้อสอบที่มีความยากสูงขึ้น ๆ ในการประมาณความสามารถของผู้สอบ จะต้องพิจารณาเหตุการณ์ที่ผู้สอบแต่ละคนตอบนั้น เป็นไปตามความคาดหวังหรือไม่ กล่าวคือ ในการสอบเป็นไปได้ที่จะแนบรวมเท่ากัน แต่เหตุการณ์ในการทดสอบนั้นแตกต่างกัน เช่น ในลักษณะที่ 1 ตอบข้อง่าย ๆ ถูกและเมื่อข้อสอบยากขึ้นตอบผิด ซึ่งลักษณะเช่นนี้ไปตามที่คาด ในลักษณะที่ 2 ตอบข้อง่าย ๆ ผิดแต่ในข้อสอบที่มีระดับความยากสูงกลับตอบถูก ซึ่งลักษณะเช่นนี้ไม่เป็นไปตามที่คาดหวัง

ดังนั้นโมเดลราชด์จึงได้ประยุกต์วิธีการเพื่อจะประเมินความเที่ยงตรง จากผลการตอบของผู้สอบแต่ละคน โดยนำผลการตอบของแต่ละคนไปเปรียบเทียบกับผลการตอบที่คาดว่าจะ เป็น โดยใช้วิธีการเปรียบเทียบค่าความเหมาะสม (fit) ซึ่งจะเป็นดัชนีชี้ว่าผลการตอบของผู้สอบเป็นไปตามที่โมเดลคาดหวังหรือไม่ ซึ่งไรท์และสโตน ได้เสนอวิธีการสรุปได้ดังนี้ (Wright, and Stone. 1979:165-169)

$$1. \text{หาค่า } Z_{v_1}^2 = \exp[2x_{v_1} - 1](b_1 - \theta_{v_1})$$

เมื่อ $Z_{v_1}^2$ คือ ความคลาดเคลื่อนมาตรฐานกำลังสองของเหตุการณ์
ที่ระดับความสามารถ θ_{v_1} ทำข้อสอบ b_1

X คือ ผลการตอบ (ผิด 0 ถูก 1)

$$2. \text{หา } V_{v_1} \text{ จากสูตร}$$

$$V_{v_1} = Z_{v_1}^2 / (L-1)$$

เมื่อ V_{v_1} คือ mean Square

$$3. \text{หาค่า } t_{v_1} \text{ จากสูตร}$$

$$t_{v_1} = \sqrt{[\ln(V) + (V-1)] [(L-1)/8]} \quad N(0,1)$$

4. เปรียบเทียบค่า t_u กับค่า t จากตารางทางเดียว

ไรท์และสโตน (Wright and Stone, 1979:169) ได้ให้แนวทางในการแปลความหมายไว้ดังนี้

- ถ้า $t < 3$ จะยอมรับผลการวัดของผู้สอบว่ามีความเที่ยงตรง
- ถ้า $3 < t < 5$ สามารถที่จะหาจุดที่ทำให้ผลการตอบไม่เหมาะสม (Misfit)
- ถ้า $t > 5$ จะปฏิเสธผลการวัด

5.4 การประมาณค่าความสามารถและค่าความคลาดเคลื่อนที่ใช้ในการทดสอบเทเลอร์

เนื่องจากการวิจัยครั้งนี้ใช้ราชคโมเดล ดังนั้นจึงขอกล่าวเฉพาะวิธีการประมาณค่าความสามารถและความคลาดเคลื่อนของราชคโมเดลเท่านั้น คือโดยวิธี UCON, PROX และ UFORM (Wright and Stone, 1979:142-144)

เมื่อนักเรียนทำแบบทดสอบผลการตอบคือ คะแนนจากการทดสอบ สิ่งที่ต้องการคือ ใช้คะแนนดิบจากการทดสอบ (r) เพื่อที่จะประมาณค่าความสามารถ (θ) ซึ่งจะต้องอ้างอิงถึงโมเดลที่ใช้วัด

$$P(X_1 = 1) = \pi_1 = \exp(\theta - b_1) / [1 + \exp(\theta - b_1)] \quad (5.1)$$

ค่าพารามิเตอร์ความยากข้อสอบคือ b และ θ คือ ค่าพารามิเตอร์ความสามารถของผู้สอบ โดยที่ผลการตอบของผู้สอบที่สังเกตได้คือ X_1 ตามโมเดลผลการตอบข้อสอบแต่ละข้อ เราสามารถที่จะคาดหวัง (expect) ผลการตอบอยู่ในรูปของโอกาสในการตอบถูก ซึ่งโอกาสในการตอบถูกแต่ละข้อนั้นจะแกว่งอยู่ในช่วง 0 ถึง 1 ดังนั้นค่าคาดหวัง (expect) ในการตอบข้อสอบข้อที่ X_1 คือ

$$E(X_1) = \pi_1 \quad (5.2)$$

เมื่อคะแนนจากแบบทดสอบ $r = \sum X_1$ คือผลรวมของการตอบข้อสอบ ค่าความคาดหวังจะได้

$$E(r) = E(\sum X_1) = \sum E(X_1) = \sum \pi_1 \quad (5.3)$$

การแทนค่าโอกาสในการตอบถูก (π) ในการวัดความสามารถ (θ_r) เพื่อที่จะประมาณค่าพารามิเตอร์ความสามารถ (θ) บนพื้นฐานของคะแนนดิบ (r) และค่าความยากข้อสอบ (b_1) แทนค่าพารามิเตอร์ความยาก จะได้สมการความสัมพันธ์ของคะแนน (r) และความสามารถ (θ_r)

$$r = \exp(\theta - b_1) / [1 + \exp(\theta - b_1)] \quad (5.4)$$

จากสมการคะแนนจากผลการสอบ (r) และค่าความยาก (b_1) ของข้อสอบที่สอบก็จะสามารถวัดค่า θ_r ได้

การประมาณโดยวิธี UCON

จากสมการ 6.4 การประมาณค่าการวัดโดยวิธี UCON นั้นจะให้การประมาณค่ากลับไปกลับมาหลายครั้งในรูปสมการ

$$b_{r,1}^{j+1} = b_{r,1}^j + (r - P_{r,1}^j) / [P_{r,1}^j(1 - P_{r,1}^j)]$$

$$\text{เมื่อ } P_{r,1}^j = \exp(\theta_r^j - b_{r,1}^j) / [1 + \exp(\theta_r^j - b_{r,1}^j)]$$

ค่าประมาณความสามารถครั้งแรก (θ_r^j) จะเป็น

$$\theta_r^0 = \ln[r / (L - r)]$$

การประมาณค่าโดยวิธี UCON ทำกลับไปกลับมา 3 หรือ 4 รอบ ซึ่งจะหยุดก็ต่อเมื่อค่าประมาณความสามารถก่อนครั้งสุดท้าย และครั้งสุดท้ายแตกต่างกันน้อยกว่า .01 logits

$$|\theta_r^{j+1} - \theta_r^j| < .01 \text{ logits}$$

การประมาณค่าโดยวิธี PROX

โดยวิธีนี้ไม่ต้องประมาณค่าหลายรอบหรือกลับไปกลับมา

1. หาค่าความสามารถจากสูตร

$$\theta = h + X \ln [r / (L - r)]$$

เมื่อ θ = ค่าประมาณความสามารถ

$$h = b_1 / L$$

$$X = 1 + S_b^2 / (2.89)$$

L = จำนวนข้อของแบบทดสอบ

r = คะแนนดิบที่สอบได้

S_p^2 = ความแปรปรวนของค่าความยากข้อสอบ

2. หาค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถ

$$S = X \sqrt{L/r (L - r)}$$

การประมาณค่าโดยวิธี UFORM

1. ประมาณค่าความสามารถ

$$\theta = h + X_{fw}$$

เมื่อ θ = ค่าประมาณความสามารถ

$$h = b_1/L$$

X_{fw} = ค่าที่ได้จากการนำค่า f และค่า w ไปเปิดตารางอยู่ในหนังสือ

Best Test Design ของไรท์และสโตน (Wright and Stone, 1979)

$$r = r/L$$

$$w = [(b_L + b_{L-1} - b_2 - b_1)/2][L/(L-2)]$$

2. ค่าความคลาดเคลื่อนมาตรฐาน

$$S = \sqrt{C_{fw}} / \sqrt{L}$$

เมื่อ $\sqrt{C_{fw}}$ คือ ค่าที่ได้จากการนำค่า f และค่า w ไปเปิดตาราง

5.5 การทดสอบทฤษฎีเบย์ส (Bayesian Strategies)

วิธีของเบย์สเป็นการทดสอบทฤษฎีเบย์ส

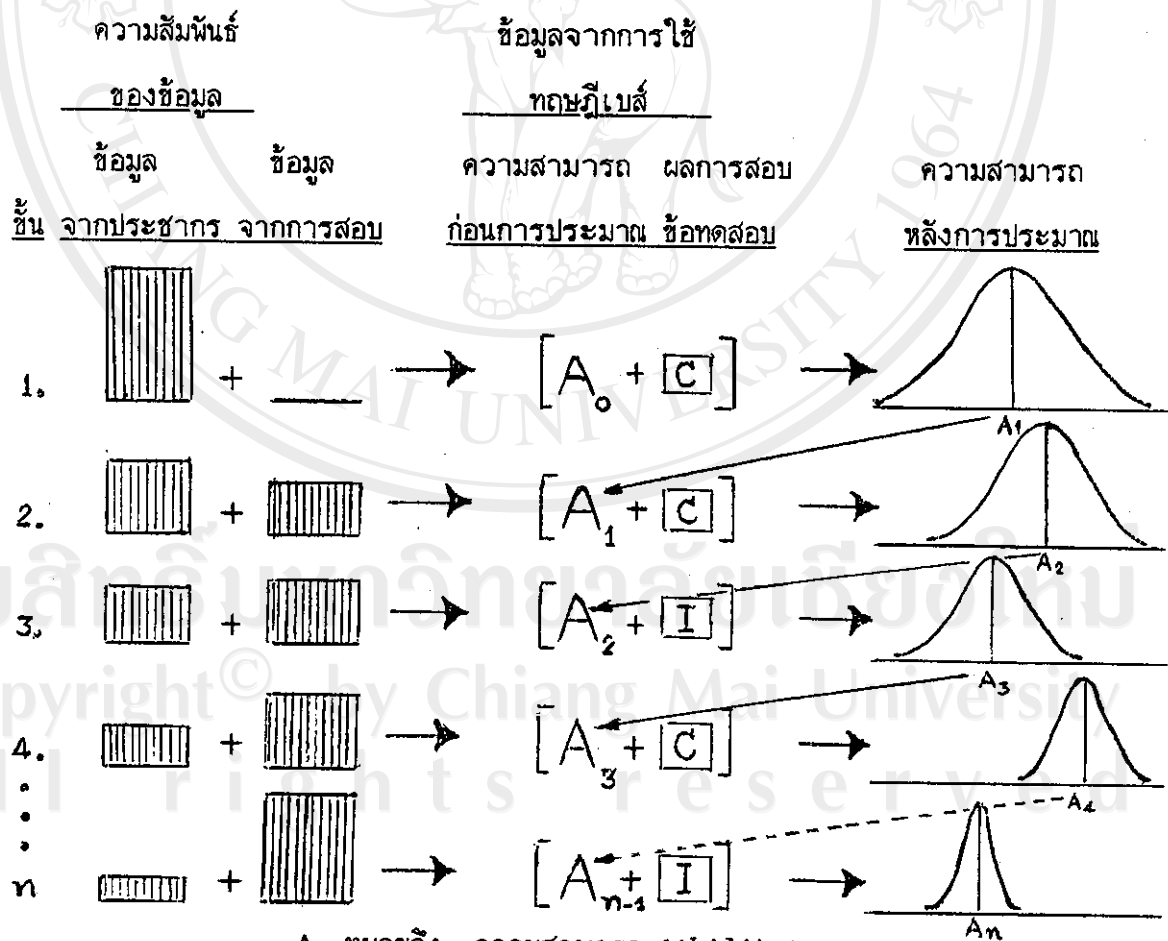
ซึ่งประยุกต์มาจากทฤษฎีเบย์ส (Bayes' Theorem)

กระบวนการทดสอบวิธีเบย์ส มีดังนี้

1. ประเมินความสามารถของผู้สอบ และความคลาดเคลื่อนมาตรฐานของการประมาณความสามารถ ซึ่งถือว่าเป็นข้อมูลที่ได้มาก่อน

2. คัดเลือกข้อสอบจากกลุ่มข้อสอบ ที่มีค่าความยากใกล้เคียงกับระดับความสามารถของซึ่งเป็นข้อมูลอยู่ก่อนแล้วให้มากที่สุด ผลจากการตอบถูกหรือผิดจะถูกประมาณค่าความสามารถใหม่ เพื่อเป็นข้อมูลเบื้องต้นเพื่อจะคัดเลือกข้อสอบที่มีค่าความยากใกล้เคียงกับระดับความสามารถของผู้สอบต่อไป โดยวิธีเบสส์ ในการสอบข้อสอบแต่ละข้อ ผลจากการสอบแต่ละข้อนั้นความคลาดเคลื่อนมาตรฐานของค่าประมาณความสามารถจะลดลงต่ำลง ไปตามลำดับ ดังภาพ 5

ภาพ 5 แสดงวิธีการทดสอบโดยวิธีเบสส์



A หมายถึง ความสามารถ (Ability)

C หมายถึง ผลการตอบถูก (Correct)

I หมายถึง ผลการตอบผิด (Incorrect)

จากภาพ 5 แสดงให้เห็นลักษณะของการทดสอบวิธีเบส์ได้ 2 ประการ คือ

1. การประเมินผลการตอบตามลำดับผลการตอบในแต่ละข้อ คือ ถูก หรือ ผิด จะมีผลต่อการประมาณความสามารถ ถ้าตอบถูกค่าประมาณความสามารถก็จะสูงขึ้น ในทางตรงกันข้าม ถ้าตอบผิดก็จะทำให้ค่าประมาณความสามารถต่ำลง

2. ค่าความคลาดเคลื่อนของการประมาณค่าความสามารถ ในแต่ละขั้นจะลดลงตั้งนั้น การกระจายของโค้งปกติที่แสดงในภาพ 5 หลังจากการประมาณความสามารถจากผลการตอบจะแคบลง ๆ ตามขั้นการสอบจนถึงขั้นที่ n ความคลาดเคลื่อนของค่าความสามารถก็จะแสดงให้เห็นขอบเขตของความสามารถที่แท้จริง

สูตรสำหรับการประมาณความสามารถ และความคลาดเคลื่อนภายหลังการตอบข้อสอบแต่ละข้อ โอเวน (Owen, 1975 cited in Thorndike, 1982) เป็นดังนี้

ในการนี้ผลการตอบถูก

$$\theta_{m+1} = \theta_m + (1 - c_g) \left[\frac{\sigma_m^2}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}} \right] \left[\frac{\phi(D)}{c_g + (1 - c_g) \Phi(-D)} \right]$$

และ

$$\sigma_{m+1}^2 = \sigma_m^2 \left\{ 1 - \left[\frac{1 - c_g}{1 + \frac{1}{a_g^2 \sigma_m^2}} \right] \left[\frac{\phi(D)}{A} \right] \left[\frac{(1 - c_g)\phi(D)}{A} - D \right] \right\}$$

ถ้าผลการตอบผิด

$$\theta_{m+1} = \theta_m - \left(\frac{\sigma_m^2}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}} \right) \left(\frac{\phi(D)}{\Phi(D)} \right)$$

และ

$$\sigma_{m+1}^2 = \sigma_m^2 \left[1 - \left(\frac{\phi(D)}{1 + \frac{1}{a_g^2 \sigma_m^2}} \right) \left(\frac{\phi(D) + D}{\Phi(D)} \right) \right]$$

เมื่อ θ_{m+1}	คือ	ค่าประมาณความสามารถจากผลการตอบข้อที่ต้องการประมาณ (posterior estimate)
θ_m	คือ	ค่าประมาณความสามารถก่อนตอบข้อที่ต้องการประมาณ (prior estimate)
C_g	คือ	โอกาสในการเดาข้อสอบถูก ซึ่งกำหนดไว้ให้เท่ากันในทุกข้อที่ตอบถูกเท่ากัน .20
a_g^2	คือ	ค่าอำนาจจำแนกที่กำหนดไว้ให้เท่ากันทุกข้อเท่ากับ 1.00
σ_m^2	คือ	ค่าความแปรปรวนของการประมาณความสามารถก่อนข้อที่ต้องการประมาณ
$\phi(D)$	คือ	ค่า ordinate ของโค้งปกติ (Normal Curve) ที่จุด D (normal probability density function)
$\Phi(D)$	คือ	พื้นที่สะสมใต้โค้งปกติที่บนจุด D (cumulative normal distribution function)

$$D = \frac{b_g - \theta_m}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}}$$

$$A = c_g + (1 - c_g)\Phi(-D)$$

ตัวอย่างของการประมาณค่าความสามารถโดยวิธีเบย์ ถ้ามีข้อมูลค่าเริ่มแรกของการประมาณ (prior information) ของผู้สอบคนหนึ่งมีค่าความสามารถเท่ากับ .90 ความแปรปรวนของค่าประมาณความสามารถเท่ากับ .19 ข้อสอบทั้งหมดที่ใช้สอบมีค่าอำนาจจำแนก (a_g) = 1.00 และกำหนดค่าให้โอกาสในการเดาถูกเท่ากับ .20 (เพราะว่าถ้ากำหนดว่าไม่มีการเดา ($C_g = 0$) ผลการตอบถูกและตอบผิด ค่าประมาณที่ได้ก็จะเท่ากัน) ถ้าสามารถคัดเลือกข้อสอบที่มีความแตกต่าง ระหว่างค่าความสามารถและความยากของข้อสอบเท่ากับศูนย์ ($b_g - \theta_m = 0$) ในการสอบแต่ละขั้นได้และผลการตอบเป็น ถูก, ผิด, ถูก, ถูก, ผิด, ผิด,

จะสามารถประมาณค่าความสามารถ ค่าความแปรปรวน ค่าความคลาดเคลื่อนของค่าความสามารถได้ดังตาราง 2

แสดงการประมาณค่าการตอบโดยวิธีเบส์

1. เมื่อตอบข้อสอบข้อแรกถูกจะประมาณค่าความสามารถและค่าความแปรปรวนได้ดังนี้

$$\begin{aligned}\theta_{m+1} &= 0.90 + (1.0 - 0.20) \left(\frac{0.19}{\sqrt{1.0 + 0.19}} \right) \left(\frac{0.3989}{0.20 + (0.80)(0.50)} \right) \\ &= 0.90 + 0.092 = 0.992\end{aligned}$$

$$\begin{aligned}\sigma_{m+1}^2 &= 0.19 \left\{ 1.0 - \left(\frac{0.80}{1.0 + \frac{1.0}{0.19}} \right) \left(\frac{0.3989}{.20 + (0.80)(0.50)} \right) \right. \\ &\quad \left. \times \left(\frac{0.80 (0.3989)}{0.20 + (0.80)(0.50)} \right) \right\} \\ &= 0.181\end{aligned}$$

2. เมื่อตอบข้อที่สองผิด จะประมาณค่าความสามารถและค่าความแปรปรวนต่อจากข้อ

แรก ได้ดังนี้

$$\begin{aligned}\theta_{m+2} &= 0.992 - \left(\frac{0.180}{\sqrt{1.0 + 0.180}} \right) \left(\frac{0.3989}{0.50} \right) \\ &= 0.992 - 0.132 = 0.859\end{aligned}$$

$$\begin{aligned}\sigma_{m+2}^2 &= 0.181 \left\{ 1.0 - \left(\frac{0.3989}{1.0 + \frac{1.0}{0.18}} \right) \left(\frac{0.3989}{.25} \right) \right\} \\ &= 0.163\end{aligned}$$

ตาราง 2 แสดงผลการตอบและการประมาณค่าโดยวิธีเบส

ข้อสอบ	ผลการตอบ	ค่าประมาณ ความสามารถ	ความแปรปรวน (σ^2)	ความคลาดเคลื่อน (ϵ)
1	ถูก	.992	.181	.425
2	ผิด	.859	.163	.804
3	ถูก	.939	.157	.396
4	ถูก	1.017	.151	.389
5	ผิด	.905	.138	.371
6	ผิด	.802	.127	.356

จะเห็นว่าเมื่อจำนวนข้อเพิ่มขึ้นความคลาดเคลื่อนมาตรฐานจะลดลงด้วย มีข้อควรสังเกตคือค่าความคลาดเคลื่อนมาตรฐานสัมพันธ์โดยตรงกับค่าความเชื่อมั่น (Reliability) หรือค่ากำลังสองของค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าประมาณความสามารถ (θ) และค่าความสามารถจริง (θ) (Urry, 1977:183) ดังสมการ

$$\sigma_{\theta}^2 = 1 - \sigma^2 \quad (4.5.1)$$

ดังนั้นความเที่ยงตรงหรือค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าประมาณความสามารถ (θ) และความสามารถจริง (θ) จะได้

$$\rho_{\theta\theta} = \sqrt{1 - \sigma^2} \quad (4.5.2)$$

จากสมการ 4.5.1 และ 4.5.2 จะเห็นว่า เราสามารถเลือกระดับความเที่ยงตรงซึ่งในที่นี้ได้แก่ความเที่ยงตรงเชิงโครงสร้าง (Construct Validity : $\rho_{\theta\theta}$) โดยสังเกตค่าความคลาดเคลื่อนมาตรฐาน (ϵ) จากผลการสอบ ดังนั้นถ้าผู้สอบทำมาถึงข้อ 6

ซึ่งมีความคลาดเคลื่อน .356 จะมีความเที่ยงตรง (Construct Validity) เป็น .93 และความเชื่อมั่น .87

ดังนั้น การทดสอบในลักษณะนี้จะส่งผลดังนี้ คือ (ฉัตรนา นพพมา 2525 หน้า 10)

1. ควบคุมระดับความเที่ยงตรงได้
2. สามารถกำหนดระดับความเชื่อมั่นที่ต้องการวัดตามจำนวนข้อของการทดสอบได้

ข้อดีและข้อจำกัดของวิธีเบส

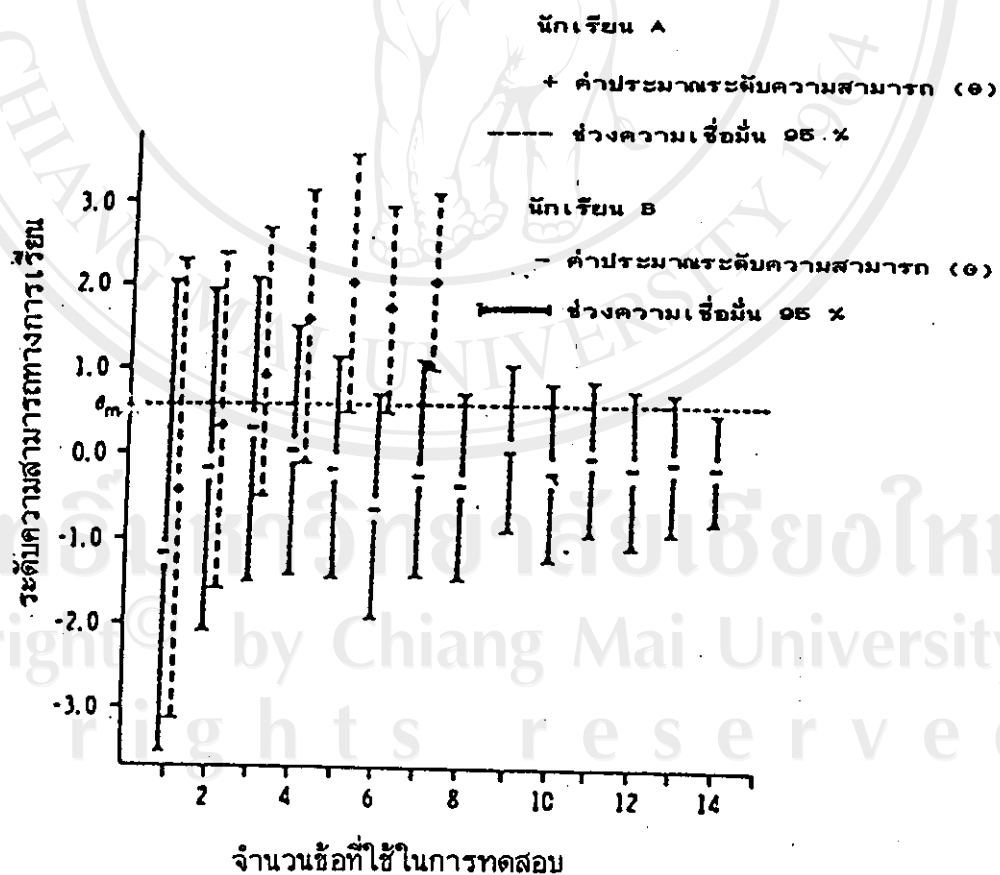
ข้อดีประการแรกของวิธีเบส คือการควบคุมขนาดของความคลาดเคลื่อนของการประมาณความสามารถของผู้สอบแต่ละคนซึ่งสามารถลดลงจนถึงระดับถึงที่พอใจได้ ซึ่งข้อดีอันนี้เป็นไปได้ที่จะประยุกต์ใช้การกำหนดจุดสิ้นสุดของการตอบกับเกณฑ์ที่ตั้งขึ้น (Weiss, 1974 : 59)

ข้อจำกัดวิธีเบส ต้องอาศัยกลุ่มข้อสอบที่มีขนาดใหญ่ ทั้งนี้เพื่อจะคัดเลือกข้อสอบให้เหมาะสมและใกล้เคียงกับความสามารถที่ถูกประมาณได้ให้มากที่สุด ซึ่งประสิทธิภาพของการทดสอบจะลดลงถ้าการจัดข้อสอบให้สอบแต่ละข้อไม่เหมาะสมกับระดับความสามารถ ดังนั้นจึงต้องใช้กลุ่มข้อสอบขนาดใหญ่ มีอำนาจจำแนกสูง และการกระจายของค่าพารามิเตอร์ของข้อสอบคือค่าความยากจะต้องกว้างเต็มพิสัยของค่าความสามารถ และเนื่องจากการทดสอบวิธีเบสจะต้องอาศัยการคำนวณที่รวดเร็วจากผลการสอบแต่ละข้อ ดังนั้นจึงมีความยุ่งยากพอสมควรในการดำเนินการสอบ (Weiss, 1974 : 61)

การนำวิธีเบสไปใช้และงานวิจัยที่เกี่ยวข้อง

การทดสอบเทเลอร์โดยวิธีเบสนี้ ได้นำไปใช้ในการทดสอบเทเลอร์ในคอมพิวเตอร์เพื่อใช้ในการทดสอบการรอบรู้ที่เรียกว่า Adaptive Mastery Testing (AMT) (Wiss and Kingbury, 1984:364) ซึ่งเป็นวิธีการที่จะกำหนดว่าระดับความสามารถ (θ) ของนักเรียนสูงกว่าหรือต่ำกว่าเกณฑ์ที่กำหนดเป็นกลุ่มรอบรู้หรือไม่รอบรู้ (θ_{cut}) กำหนดโดยผู้เชี่ยวชาญซึ่งแปลงจากเกณฑ์ตัดสินส่วนของการตอบถูกให้อยู่ในรูปของค่าความสามารถ(θ) การเริ่มต้นสอบนั้นมีข้อตกลง

ว่าระดับความสามารถของนักเรียน $\theta_i = \theta_m$ จากข้อตกลงดังกล่าวจะคัดเลือกข้อสอบที่ให้ค่าสาระประโยชน์สูงสุด (Maximum Information) ณ ที่ระดับความสามารถ θ_m ให้นักเรียนสอบผลจากการสอบถูกหรือผิด ก็จะได้ค่าประมาณความสามารถและช่วงความเชื่อมั่น ค่าความสามารถที่ประมาณได้ใหม่คือ θ_1 และช่วงความเชื่อมั่นของค่าความสามารถในขั้นต่อไปก็จะคัดเลือกข้อสอบ ที่ให้ค่าสาระประโยชน์สูงสุดกับระดับความสามารถ θ_1 ผลการสอบที่จะนำไปประมาณค่าความสามารถและช่วงความเชื่อมั่นใหม่ กระบวนการลักษณะนี้จะทำต่อเนื่องไปเรื่อย ๆ จนกระทั่ง ช่วงความเชื่อมั่นของค่าประมาณความสามารถไม่ครอบคลุม θ_m ผลก็จะได้ว่าผู้ที่อยู่สูงกว่า θ_m ก็จะเป็นผู้รอบรู้และต่ำกว่า θ_m ก็จะเป็นผู้ไม่รอบรู้



ภาพ 6 แสดงการทดสอบแบบเทเลอร์ที่ใช้กับการทดสอบโดยคอมพิวเตอร์เพื่อจำแนกความรอบรู้

คิงสเบอรีและไวส์ได้ใช้กระบวนการทดสอบแบบรอบรู้ชนิดนี้กับข้อมูลจริง (1979) การศึกษาประยุกต์ (1981) และสถานการณ์จำลองในคอมพิวเตอร์ Monte Carlo (Kingsbury and Weiss, 1983) โดยเปรียบเทียบกับ การทดสอบการรอบรู้แบบเดิมที่ให้คะแนนตามสัดส่วนของการตอบถูกและวิธีให้คะแนนวิธีเบส โดยใช้ข้อสอบแบบเดิมที่มีความยาว 10, 25 และ 50 ข้อ ศึกษาเกี่ยวกับกลุ่มข้อสอบที่เป็น 1) Unifrom 2) พารามิเตอร์ตัวเดียว (One-Parameter IRT, Rasch Model) 3) พารามิเตอร์ 2 ตัว (Two-Parameters IRT Model) 4) พารามิเตอร์ 3 ตัว (Three-Parameters IRT Model) พบว่า การทดสอบวิธีเบสเป็นการทดสอบที่มีแนวโน้มให้เห็นว่ามีความเที่ยงตรงในการจำแนกการรอบรู้ได้สูงและใช้จำนวนข้อสอบสั้นกว่าในทุกกลุ่มข้อสอบ

จากผลการวิจัยและข้อดีของการทดสอบโดยวิธีเบสดังกล่าวมาแล้วข้างต้น ผู้วิจัยจึงใช้การทดสอบวิธีเบสในการจำแนกความรอบรู้ที่แท้จริงของนักเรียน

6. การทดสอบแบบอิงเกณฑ์

แบบทดสอบอิงเกณฑ์เป็นแบบสอบที่นักทดสอบพึงให้ความสนใจอย่างจริงจัง เมื่อประมาณปี 1970 มา นี้ โดยกลาสเซอร์และนิทโก (Glaser and Nitko) ได้เป็นผู้เสนอแนวคิดไว้ว่า แบบสอบอิงเกณฑ์เป็นแบบสอบที่สร้างขึ้นอย่างพิถีพิถันเพื่อการวัดผล ซึ่งสามารถที่จะแปลความหมายได้โดยตรงตามมาตรฐานของการปฏิบัติ (performance) ที่กำหนดไว้ (Glaser and Nitko, 1971 อ้างถึงใน สุวัฒน์ สุขมลสันต์ 2528 หน้า 49) และอีกคำนิยามหนึ่งที่นักทดสอบมักอ้างอิงเสมอคือ แบบทดสอบอิงเกณฑ์ คือ แบบทดสอบซึ่งใช้เพื่อวัดสภาพที่แน่นอนของแต่ละบุคคลโดยอาศัยเกณฑ์บางอย่าง เช่น มาตรฐานของการปฏิบัติ หรือปริเขต (domain) ของพฤติกรรมที่ได้นิยามไว้อย่างดีแล้ว (Hambleton and others, 1978 : 1)

จากคำนิยามทั้งสองดังกล่าวแล้วจะเห็นได้ว่า แบบสอบอิงเกณฑ์มีลักษณะพิเศษ 2 ประการคือ 1) ส่วนประกอบของแบบสอบ ซึ่งประกอบด้วยข้อสอบที่สร้างจากวัตถุประสงค์เชิงพฤติกรรม และ 2) เกณฑ์ของพฤติกรรมหรือเกณฑ์ของการปฏิบัติที่ได้กำหนดไว้ก่อนแล้ว ซึ่งตาม

คำนิยามทั้งสอง หมายถึงกลุ่มของพฤติกรรม ซึ่งอาจจะมีหลายอย่าง หรือเพียงอย่างเดียวก็ได้ แต่ได้กำหนดขอบเขตเนื้อหาไว้อย่างชัดเจนแล้ว (Millman, 1974 อ้างถึงใน สุวัฒน์ สุขมลสันต์ 2528 หน้า 50) ดังนั้น แบบสอบอิงเกณฑ์อาจไม่จำเป็นต้องมีจุดตัดของคะแนนก็ได้

อนึ่ง พิลลินเนอร์ (Pilliner, 1979 อ้างถึงใน สุวัฒน์ สุขมลสันต์ 2528 หน้า 50) กล่าวว่า แบบสอบอิงเกณฑ์และแบบสอบอิงกลุ่มมีความแตกต่างกัน 5 อย่าง คือ

1. จุดมุ่งหมาย

แบบสอบอิงเกณฑ์ใช้เพื่อวัดระดับความสามารถของผู้สอบ โดยการเปรียบเทียบกับระดับความสามารถของผู้อื่นเป็นเกณฑ์ แต่แบบสอบอิงเกณฑ์ใช้เพื่อวัดระดับความสามารถของผู้สอบโดยการเปรียบเทียบกับระดับความสามารถที่กำหนดไว้เป็นเกณฑ์

2. การสร้าง

แบบสอบอิงกลุ่มเน้นที่การเลือกข้อสอบที่มีค่าอำนาจจำแนกสูงไว้ และตัดข้อสอบที่ไม่ดีทิ้ง แต่แบบสอบอิงเกณฑ์จะรวมข้อสอบทุกข้อที่วัดสิ่งที่ต้องการวัดไว้

3. การใช้

แบบสอบอิงกลุ่มใช้เพื่อการคัดเลือก และเพื่อการเปรียบเทียบระดับความรู้ความสามารถ แต่แบบสอบอิงเกณฑ์ใช้เพื่อการวินิจฉัยความรู้ความสามารถของผู้สอบ

4. สาระสนเทศที่ได้

แบบสอบอิงกลุ่มให้สาระสนเทศเกี่ยวกับผู้สอบกว้าง ๆ แต่แบบสอบอิงเกณฑ์ให้สาระสนเทศเกี่ยวกับผู้สอบได้อย่างเฉพาะ

5. การอ้างอิงสรุป

ผลการสอบของแบบสอบอิงกลุ่ม สามารถใช้อ้างอิงสรุปเกี่ยวกับความรู้ความสามารถของผู้สอบได้อย่างจำกัด และต้องใช้อ้างอิงกับค่าความตรงของแบบสอบ แต่ผลการสอบของแบบสอบอิงเกณฑ์สามารถใช้อ้างอิงสรุปได้กว้างกว่า สำหรับเนื้อหาที่เฉพาะเจาะจงว่า

แต่อย่างไรก็ตาม นักทดสอบยังมีความสับสนว่า แบบสอบอิงเกณฑ์คืออะไร และมีคุณสมบัติอย่างไรบ้าง ดังจะเห็นได้จากการศึกษาของ แยมเบิลตัน และคณะในปี 1978

(Hambleton and others, 1978 : 2) พบว่ามีคำนิยามเกี่ยวกับ แบบทดสอบอิงเกณฑ์ ประมาณ 600 คำนิยามที่ปรากฏในวรรณคดีต่าง ๆ ทางการศึกษา แต่คำนิยามเหล่านี้แตกต่างกันเท่า ๆ กับจำนวนผู้ที่ให้คำนิยาม แฮมเบิลตัน และคณะ (Hambleton and others, 1978:3) กล่าวว่า เมื่อจุดมุ่งหมายของสิ่งที่วัดของแบบสอบอิงกลุ่มสอดคล้องกับจุดมุ่งหมายของสิ่งที่วัดของแบบสอบอิงเกณฑ์ แบบสอบอิงกลุ่มสามารถใช้เป็นแบบสอบอิงเกณฑ์ได้และแบบสอบอิงเกณฑ์ก็สามารถใช้เป็นแบบสอบอิงกลุ่มได้ และ แบบทดสอบโดยมากสามารถใช้เป็นแบบสอบอิงกลุ่มและแบบสอบอิงเกณฑ์ได้ด้วย

ดังนั้น สิ่งที่แตกต่างกันอย่างเห็นได้ชัดระหว่างแบบสอบอิงกลุ่ม และแบบสอบอิงเกณฑ์ ก็คือ การแปลความหมายของคะแนน (interpretation) ส่วนลักษณะของเครื่องมือ (nature of the instruments) ของแบบสอบทั้ง 2 แบบ ยังไม่แตกต่างกันอย่างชัดเจน (Block and Dockrell, 1984 อ้างถึงใน สุวัฒน์ สุขมลสันต์ 2529 หน้า 51) ด้วยเหตุนี้ นักทดสอบบางท่านในปัจจุบัน เช่น เบรนนัน (Brennan, 1984 อ้างถึงใน สุวัฒน์ สุขมลสันต์ 2529 หน้า 51) ผู้ซึ่งมีชื่อเสียงมากทางทฤษฎีการอ้างอิงสรุป เชื่อว่า ความแตกต่างของแบบสอบอิงกลุ่ม และแบบสอบอิงเกณฑ์อยู่ที่การแปลความหมายของคะแนนเท่านั้น ส่วนประกอบของแบบสอบและการคัดเลือกข้อทดสอบนั้น เป็นเพียงสิ่งที่ทำให้แบบสอบสามารถใช้อ้างอิงสรุปผลสอบได้มากน้อยเพียงใดเท่านั้น ซึ่งแบบสอบทั้ง 2 แบบ อาจมีความแตกต่างกันหรือเหมือนกันก็ได้