การพัฒนาแบบทดสอบ

การประเมินลักษณะมนุษย์

สันทัด พรประเสริฐมานิต

โครงร่างการนำเสนอ

- แนวคิดของแบบทดสอบ
- การสร้างแบบทดสอบ
- การทดลองใช้แบบทดสอบ
- การวิเคราะห์ข้อคำถาม
- การทบทวนแบบทดสอบ

แนวคิดของแบบทดสอบ

- วัตถุประสงค์ของแบบทดสอบ (เช่น วิจัย, ประเมิน, ทำนาย, จำแนก)
- กลุ่มเป้าหมายของแบบทดสอบ (เช่น ผู้ใหญ่, เด็ก, ผู้ป่วย)
- มีกระบวนการดำเนินการแบบทดสอบอย่างไร (เช่น ใช้ผู้ประเมิน, วัดแบบกลุ่ม, ใช้คอมพิวเตอร์)

การสร้างแบบทดสอบ

- การสร้างมาตร (Scaling) คือ การสร้างกฎในการสร้างตัวเลขจากข้อคำถาม
- มีหลายรูปแบบ แต่วิธีที่ใช้มากที่สุดคือมาตรแบบผลรวมจากการประเมิน (Summated Rating Scale)
  - มาตรวัดแบบลิเคิร์ท
  - ผลรวมของข้อคำถามที่ถูก
  - วิธีนี้เป็นวิธีที่สถิติที่พัฒนาขึ้นมาส่วนใหญ่ ถูกออกแบบให้นำมาใช้ เช่น การวิเคราะห์องค์ประกอบ (Factor analysis)
- ดังนั้น ในวิชานี้จะพูดถึงเฉพาะการสร้างมาตรรูปแบบนี้

การสร้างแบบทดสอบ

- การเขียนข้อคำถาม ให้นึกถึงความตรงเชิงเนื้อหาเป็นหลัก
- รูปแบบของข้อคำถามมีจำนวนมาก ดูข้อเปรียบเทียบที่ตาราง 8.1
- เป็นส่วนที่ยากที่สุด เพราะต้องใช้จินตนาการเพื่อให้ข้อคำถามที่วัดภาวะสันนิษฐานให้ตรงจุด
- ให้นึกถึงกลุ่มตัวอย่างที่จะอ่านคำถามเสมอ ว่าอ่านรู้เรื่องหรือไม่ มีจุดที่ทำให้สับสนหรือมึนหรือไม่ มีเนื้อหาที่ไม่เกี่ยวข้องหรือไม่ ก่อให้เกิดอคติหรือไม่
- ตรวจสอบความถูกต้องเสมอ
- หากเป็นข้อคำถามปลายเปิด ต้องมีแนวทางในการให้คะแนนเสมอ

Score Points for 1-point Items
1 Point A one-point response is correct. The response indicates that the student has completed the task correctly.
0 Points A zero-point response is completely incorrect, irrelevant, or incoherent.
Score Points for 2-point Items
2 Points (Full credit) A two-point response is complete and correct. The response demonstrates a thorough understanding of the concepts and/or procedures embodied in the task. • Indicates that the student has completed all aspects of the task, showing the correct application of concepts and/or procedures • Contains clear, complete explanations, supporting work, or evidence when required
1 Point (Partial Credit) A one-point response is only partially correct. The response demonstrates only a partial understanding of the concepts and/or procedures embodied in the task. • Addresses some elements of the task correctly but may be incomplete • May contain a correct answer but with an incomplete explanation when required • May contain an incorrect answer but with an explanation or supporting work indicating a correct understanding of the concepts
0 Points A zero-point response is inaccurate or inadequate, irrelevant, or incoherent.

From TIMSS & PIRLS (2011)

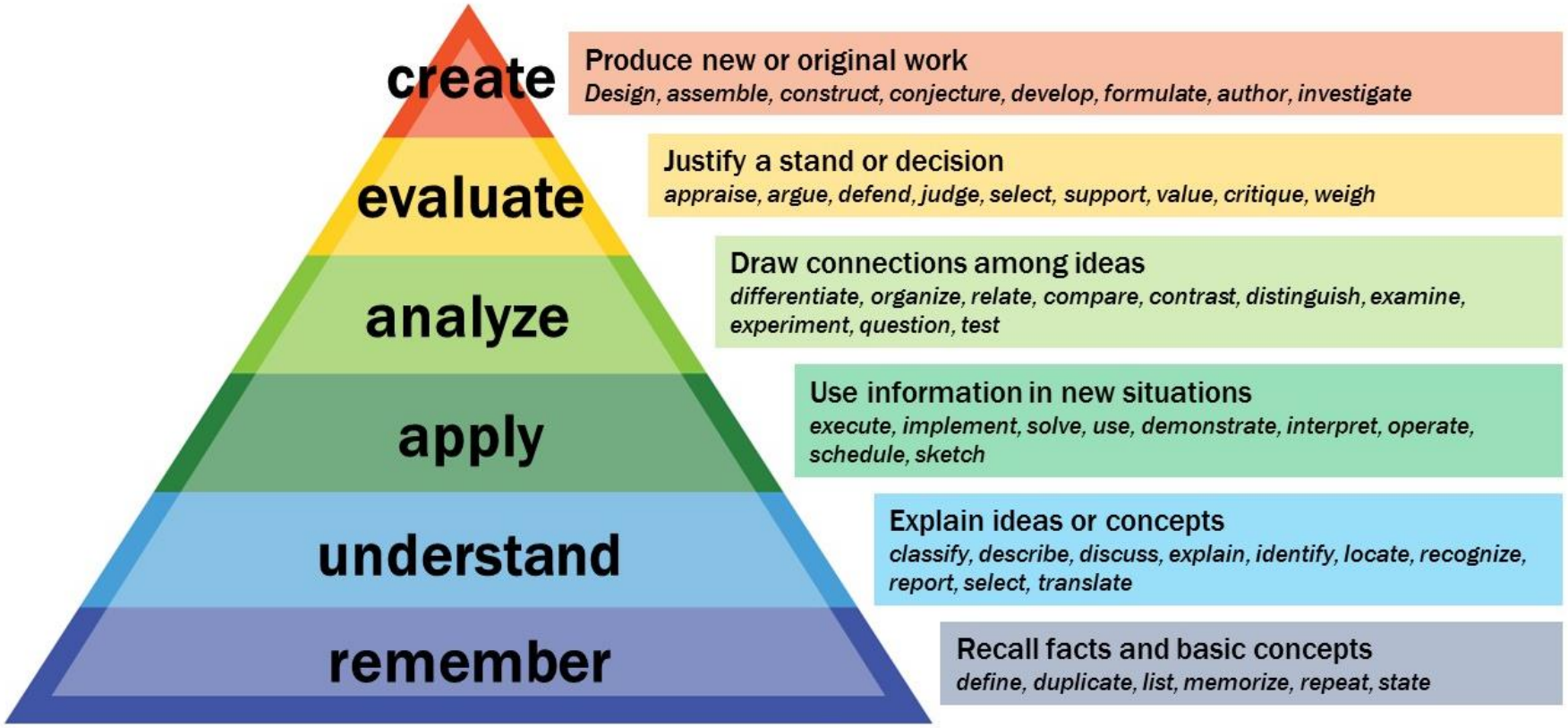
การสร้างแบบทดสอบ

- ข้อแนะนำในการเขียนข้อคำถาม
  - ไม่ใช่คำนิเสธ โดยเฉพาะอย่างยิ่งนิเสธซ้อนนิเสธ
  - ไม่ใช้ข้อคำถามที่มีสองเนื้อหา
  - ไม่ใช้ข้อคำถามที่มีเนื้อหาที่ไม่เกี่ยวข้อง
  - สั้น กระชับ

การสร้างแบบทดสอบ

- ข้อแนะนำในการเขียนข้อคำถาม (จำเพาะสำหรับข้อสอบ)
  - สำหรับข้อคำถามที่มีคำตอบถูกต้อง ข้อคำถามควรเป็นประโยคคำถามหรือการเติมคำในช่องว่าง
  - ตัวเลือกควรมีอำนาจในการลวง มีเนื้อหาใกล้เคียงกัน
  - ตัวลวงไม่ควรใช้ “ถูกต้องทั้งหมด” หรือ “ไม่ถูกต้องทั้งหมด”
  - จำนวนตัวลวงอาจเพิ่มหรือลดได้ ถ้าทุกข้อมีความเป็นไปได้ทั้งหมด
  - เนื้อหาจากข้อคำถามหนึ่งไม่ควรใช้อีกข้อคำถามหนึ่งได้

Bloom's Taxonomy



From Center for Teaching, Vanderbilt University

การสร้างแบบทดสอบ

- ข้อคำถามแบบจัดอันดับ หรือบังคับเลือก (Ipsative data) เป็นวิธีการลดการตอบสนองตามความคาดหวังของสังคมได้จริงเหรอ?

I am the sort of person who...

			M	L
1	A	I try out new activities.	●	○
	B	I consider other people's feelings.	○	○
	C	I like to understand the underlying theory.	○	●

Gordon Personal Profile-Inventory

Prog

Test Instructions

On the following screens are a number of descriptions of personal characteristics of people. These descriptions are grouped in sets of four. You are to examine each set and find the one description that is *most like you*. Then select the circle following the statement, in the column headed **Most**.

Next, examine the other three statements in the set and find the one description that is *least like you*; then select the circle following that statement, in the column headed **Least**. Leave the remaining two statements in the set unselected.

Here is an example set:

	Most	Least
has an excellent appetite	○	○
watches too much television	○	◉
follows a well-balanced diet	○	○
doesn't get enough exercise	◉	○

Suppose that you have read the four descriptive statements in the example and decided that, although several of the statements may apply to you to some degree, "doesn't get enough exercise" is *more like you* than any of the others. You would select the circle following the statement in the column headed **Most**, as shown in the example above.

You would then examine the other three statements to decide which one is *least like you*. Suppose that "watches too much television" is *less like you* than the other two. You would select the circle following that statement in the column headed **Least**, as shown in the example above.

For every set you should have *one* and *only one* selection in the **Most** column and *one* and *only one* selection in the **Least** column. There should be nothing selected for two of the statements.

In some cases it may be difficult to decide which statements you should select. Make the best decisions you can. There are no right or wrong answers. In each set you are to select two statements in the way in which they most *nearly apply to you*. Be sure to select one statement as being *most like you* and one as being *least like you*. Do this for every set. Do *not* skip any set.

Click **Next Page** to start the assessment.

Previous Page | Next Page

การทดลองใช้แบบทดสอบ

- เก็บข้อมูลกับกลุ่มตัวอย่างที่ใกล้เคียงกับกลุ่มตัวอย่างที่ต้องการจริง
- เก็บข้อมูลให้เหมาะสมกับการวิเคราะห์องค์ประกอบ (สอนบทถัดไป) ถ้าวิเคราะห์อย่างง่าย (ดังที่สอนในบทนี้) ให้ใช้เกณฑ์ 5 หรือ 10 คนต่อข้อ
- อาจมีการวิเคราะห์ข้อคำถามเชิงคุณภาพระหว่างทดลองใช้

การวิเคราะห์ข้อคำถาม

- การวิเคราะห์เชิงปริมาณ
- การวิเคราะห์เชิงคุณภาพ

การวิเคราะห์ข้อคำถามเชิงปริมาณ

- ความยากรายข้อ (Item difficulty) คือ สัดส่วนของกลุ่มตัวอย่างที่ตอบข้อดังกล่าวถูกต้อง ในมาตรวัดเชิงเจตคติ จะหมายถึงค่าเฉลี่ยของข้อคำถาม
- ค่าความยากรายข้อที่ดี ควรอยู่กึ่งกลางระหว่างสัดส่วนที่ได้จากการสุ่มและคะแนนเต็ม เช่น (.25 + 1) / 2 = .625 สำหรับข้อทางเลือก 4 ข้อ
- ควรมีการออกแบบส่วนผสมของความยาก (หรือค่าเฉลี่ย) ระหว่างข้อคำถาม
  - ถ้าใช้ในการแบ่งแยกความแตกต่างระหว่างบุคคล ความยากควรผสมกัน
  - ถ้าใช้ในการตัดสินว่าผ่านเกณฑ์หรือไม่ ควรเน้นระดับความยากที่เหมาะสมกับจุดตัด

การวิเคราะห์ข้อคำถามเชิงปริมาณ

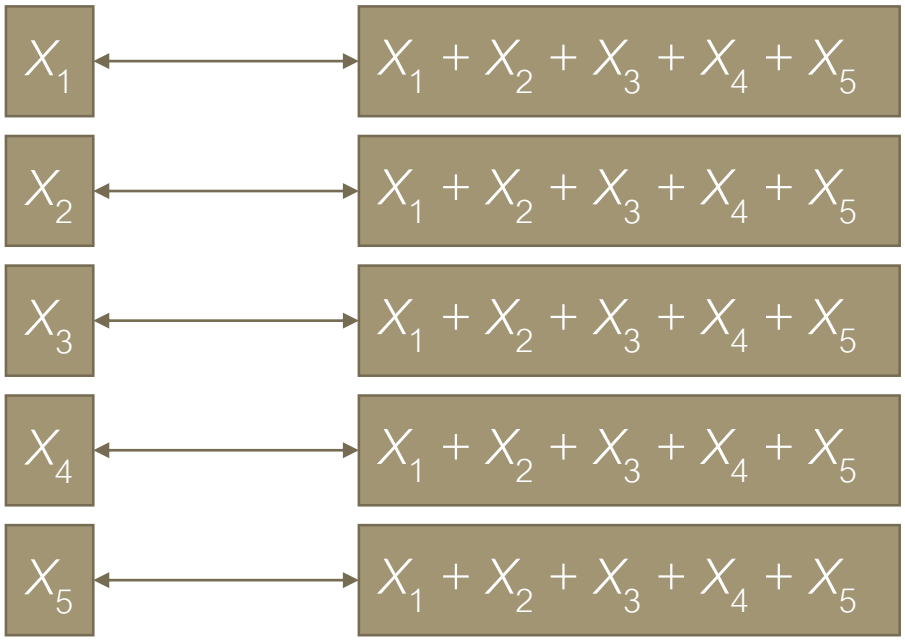
- ความเที่ยงของข้อคำถาม (Item-reliability index)
  - *SD* ของข้อคำถาม คูณกับ สหสัมพันธ์ระหว่างข้อคำถามและคะแนนรวม
- ความตรงของข้อคำถาม (Item-validity index)
  - *SD* ของข้อคำถาม คูณกับ สหสัมพันธ์ระหว่างข้อคำถามและเกณฑ์
- ค่าเหล่านี้ ไม่ได้มีค่าเป็นมาตรฐาน อาจมีค่ามากกว่า 1 ทำให้ค่าที่ได้แปลความหมายยาก
- นอกจากนี้ ความหมายในทางสถิติค่อนข้างแปลก เพราะสูตรไม่เหมือนทั้งสัมประสิทธ์ถดถอยหรือความแปรปรวนร่วม

การวิเคราะห์ข้อคำถามเชิงปริมาณ

- อำนาจจำแนก (Item discrimination) คือ ความสามารถของข้อคำถามในการแบ่งระหว่างคนมีภาวะสันนิษฐานสูงและต่ำ
- ในที่แท้จริง ควรจะเป็นน้ำหนักองค์ประกอบ (Factor loading) กล่าวคือ เมื่อภาวะสันนิษฐานเปลี่ยนแปลง 1 หน่วยแล้วคะแนนข้อคำถามเปลี่ยนแปลงไปเท่าไร
- ในอดีต การวิเคราะห์องค์ประกอบทำได้ยาก จึงใช้การประมาณค่าด้วย 3 ค่า คือ ความสัมพันธ์ระหว่างข้อคำถามและคะแนนรวม (Item total correlation), อำนาจจำแนกกลุ่มสูงต่ำ, ค่าอัลฟ่าเมื่อข้อคำถามถูกลบ

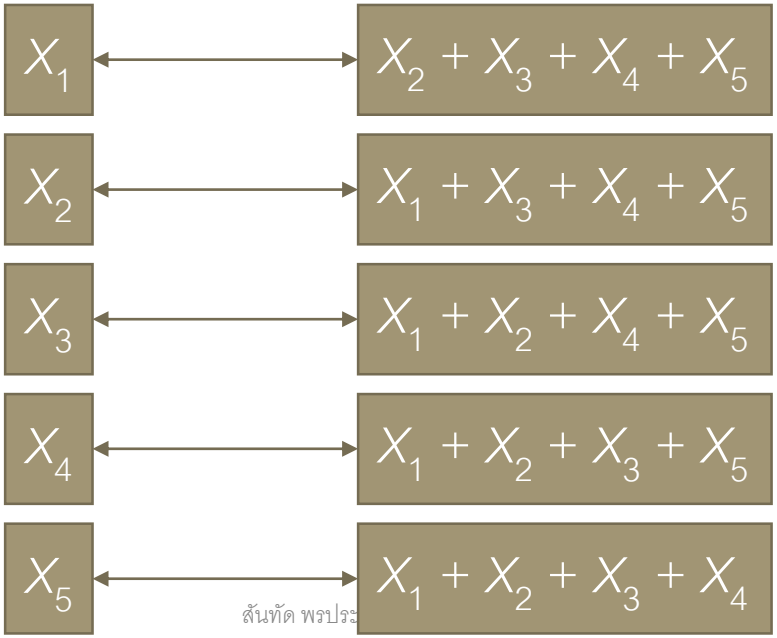
การวิเคราะห์ข้อคำถามเชิงปริมาณ

- ความสัมพันธ์ระหว่างข้อคำถามและคะแนนรวม (Item total correlation)
  - สมมติว่าข้อคำถามมี 5 ข้อ ค่านี้ของแต่ละข้อคือความสัมพันธ์ของแต่ละข้อกับคะแนนรวม



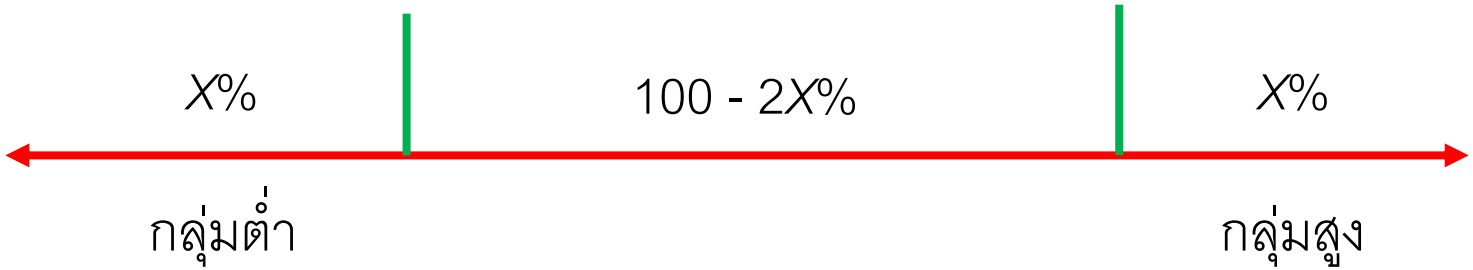
การวิเคราะห์ข้อคำถามเชิงปริมาณ

- ความสัมพันธ์ระหว่างข้อคำถามและคะแนนรวม (Item total correlation)
  - เนื่องจากคะแนนรวม มีคะแนนข้อคำถามที่ต้องการหาอำนาจจำแนกในการรวมคะแนนด้วย จึงตัดคะแนนข้อคำถามนั้นออกจากคะแนนรวมเพื่อลดการเฟ้อของค่าสหสัมพันธ์ เรียกว่า Corrected item total correlation (CITC)



การวิเคราะห์ข้อคำถามเชิงปริมาณ

- การเปรียบเทียบกลุ่มสูงกลุ่มต่ำ



$$d = \frac{M_{\text{กลุ่มสูง}} - M_{\text{กลุ่มต่ำ}}}{2}$$

- การเปรียบเทียบนี้ ให้ผลไม่แตกต่างจาก CITC (Preacher et al., 2005) ซ้ำยังทำให้กำลังในการวิเคราะห์ทางสถิติต่ำลงด้วย

การวิเคราะห์ข้อคำถามเชิงปริมาณ

- ค่าอัลฟ่า ถ้าข้อคำถามนั้นถูกลบออกไป (Alpha if item deleted)
  - เชื่อว่า ยิ่งค่านี้ยิ่งสูง ข้อคำถามดังกล่าวควรจะถูกลบออกจากมาตรวัด
  - แท้จริงแล้ว ค่านี้เป็นฟังก์ชันของ CITC

ให้ Y_i เป็นผลรวมของคะแนนยกเว้นข้อที่ *i* และ α_i เป็นค่าอัลฟ่าเมื่อลบข้อที่ *i*

$$r_{X_iY_i} = \frac{Cov(X_i, Y_i)}{\sqrt{Var(X_i)Var(Y_i)}} = \frac{\sum_{j=1,j\neq i}^{p} Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(Y_i)}}$$

$$\alpha_i = \left(\frac{p-1}{p-2}\right) \frac{\sum_{j=1,j\neq i}^{p} \sum_{k=1,k\neq i,j}^{p} Cov(X_j, X_k)}{Var(Y_i)}$$

จาก

$$\sum_{j=1,j\neq i}^{p} \sum_{k-1,k\neq i,j}^{p} Cov(X_j, X_k) + 2 \sum_{j=1,j\neq i}^{p} Cov(X_i, X_j) + \sum_{j=1}^{p} Var(X_j) = Var(Y)$$

ดังนั้น

$$\sum_{j=1,j\neq i}^{p}\sum_{k-1,k\neq i,j}^{p} Cov(X_j,X_k) + 2r_{X_iY_i}\sqrt{Var(X_i)Var(Y_i)} + \sum_{j=1}^{p} Var(X_j) = Var(Y)$$

$$\alpha_i = \left(\frac{p-1}{p-2}\right)\frac{Var(Y) - \sum_{j=1}^{p} Var(X_j) - 2r_{X_iY_i}\sqrt{Var(X_i)Var(Y_i)}}{Var(Y_i)}$$

$$\alpha_i = \left(\frac{p-1}{p-2}\right)\frac{Var(Y) - \sum_{j=1}^{p} Var(X_j)}{Var(Y_i)} - 2\left(\frac{p-1}{p-2}\right)r_{X_iY_i}\sqrt{\frac{Var(X_i)}{Var(Y_i)}}$$

แปรผกผันกัน

การวิเคราะห์ข้อคำถามเชิงปริมาณ

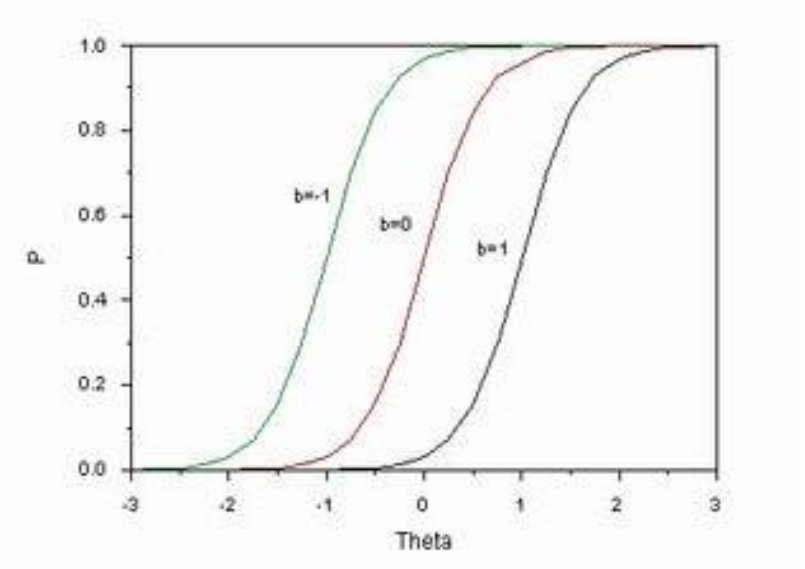
- แสดงให้เห็นว่าค่าอัลฟ่าเมื่อข้อคำถามถูกลบไป แปรผกผันกับ CITC
- เมื่อทราบค่าหนึ่งแล้วจะรู้อีกค่าหนึ่ง ดังนั้นไม่จำเป็นต้องพิจารณาจากทั้งสองค่า (แต่อาจรายงานทั้งสองค่าได้)
- ค่า CITC ดีกว่า เนื่องจากค่าสหสัมพันธ์สามารถอ่านค่าได้โดยตรง มีการทดสอบระดับนัยสำคัญชัดเจน

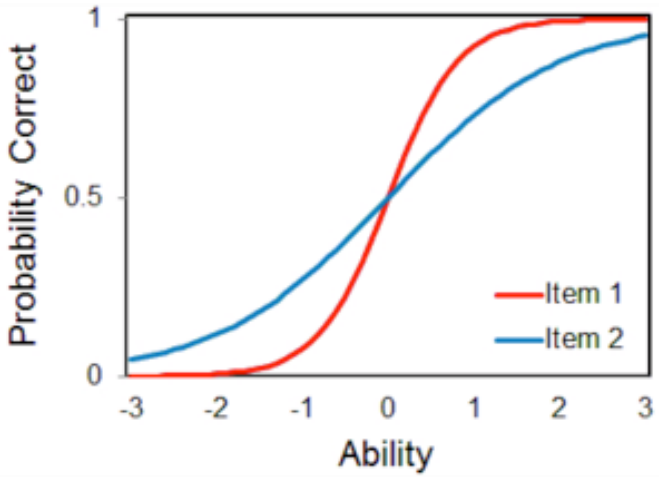
การวิเคราะห์ข้อคำถามเชิงปริมาณ

- การวิเคราะห์ตัวลวง (Analysis of item alternatives)
  - สัดส่วนที่ผู้ตอบเลือกตัวลวง สัดส่วนนี้จะมีค่าเท่าไรก็ได้ แต่ไม่ใช่ 0 มิเช่นนั้นตัวลวงจะไม่ได้ทำหน้าที่เป็นตัวลวงเลย
  - ข้อคำถามที่ดี ผู้ที่มีภาวะสันนิษฐานต่ำต้องมีแนวโน้มที่จะเลือกตัวลวงมากกว่าผู้ที่มีภาวะสันนิษฐานสูง
  - กล่าวคือ มีความสัมพันธ์ทางลบกับภาวะสันนิษฐาน ซึ่งอาจวิเคราะห์ผ่าน CITC หรือการวิเคราะห์กลุ่มสูงต่ำ

การวิเคราะห์ข้อคำถามเชิงปริมาณ

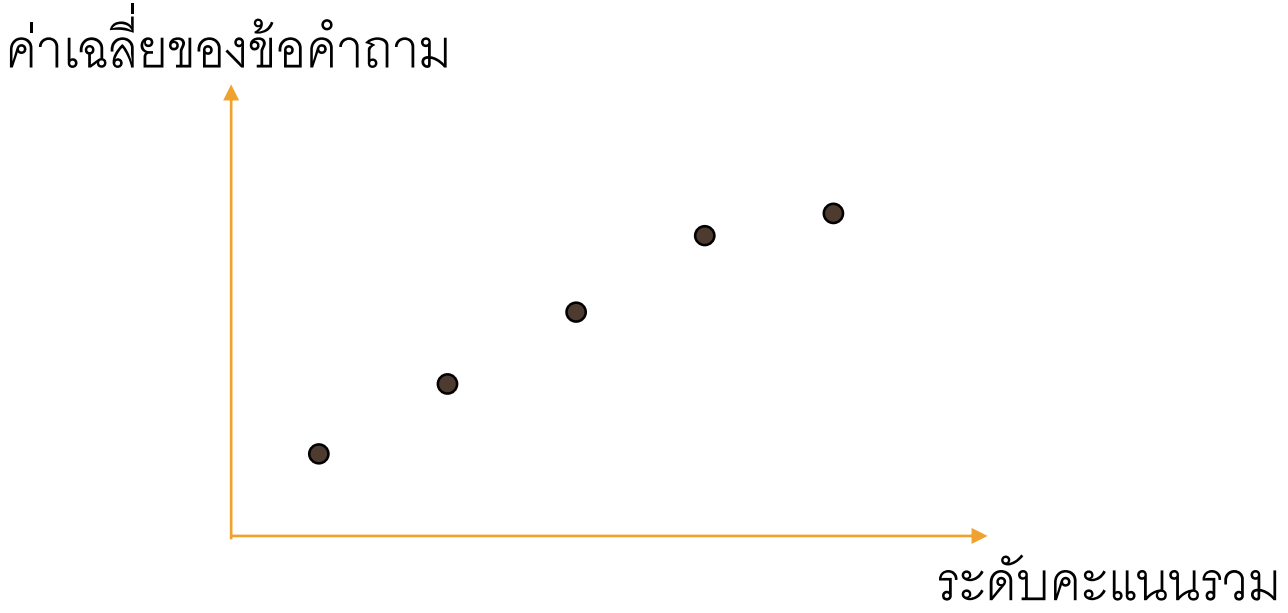
- วิธีทางหนึ่งในการวิเคราะห์ข้อคำถาม คือ นำความยากรายข้อ (Item difficulty) มาสร้างความสัมพันธ์กับระดับภาวะสันนิษฐาน
- เรียกว่า เส้นโค้งลักษณะของข้อคำถาม (Item Characteristic Curve)





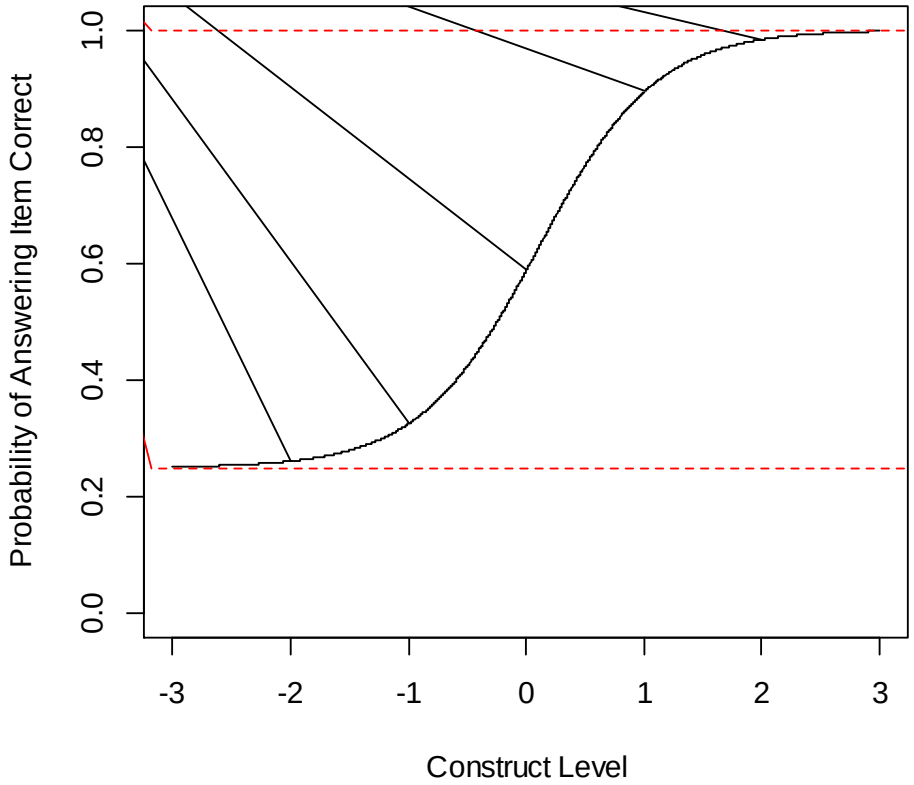
การวิเคราะห์ข้อคำถามเชิงปริมาณ

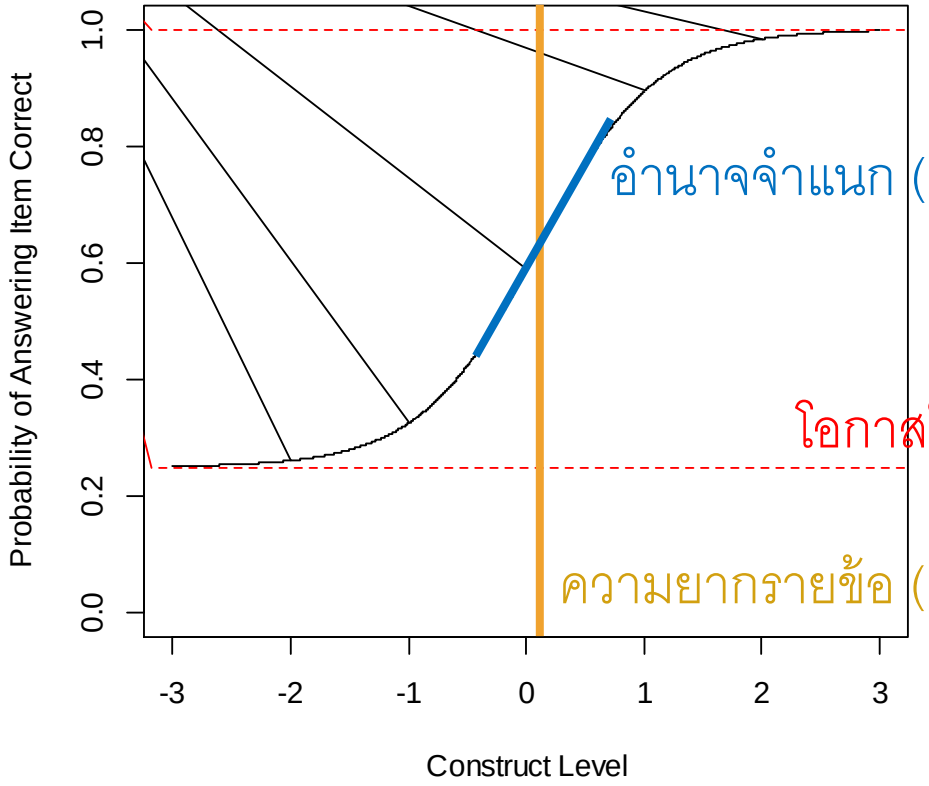
- วิธีอย่างง่ายที่สุด คือ ให้แบ่งคะแนนรวมออกเป็นกลุ่มๆ แล้วสังเกตค่าเฉลี่ยในแต่ละระดับของคะแนนรวม



การวิเคราะห์ข้อคำถามเชิงปริมาณ

- ทฤษฎีการตอบสนองของข้อคำถาม (Item response theory) สามารถสร้างโมเดลแสดงความสัมพันธ์ระหว่างข้อคำถามและคะแนนรวม





อำนาจจำแนก (Discrimination)

โอกาสในการเดาแฝง (Pseudo-guessing)

ความยากรายข้อ (Difficulty)

$$P(X = 1|\theta) = 0.25 + (1 - 0.25)\left(\frac{e^{2\theta - 0.2}}{1 + e^{2\theta - 0.2}}\right)$$

การวิเคราะห์ข้อคำถามเชิงคุณภาพ

- การวิเคราะห์ข้อคำถามเชิงคุณภาพ (Qualitative Item Analysis) เพื่อเข้าใจลักษณะของข้อคำถามมากขึ้น ในมุมมองของผู้ทดสอบ (ตาราง 8.3)
- การทำโดยคิดดังๆ (Think aloud) เป็นวิธีที่ทำให้รู้กระบวนการคิดในการตอบคำถามต่างๆ ทำให้ป้องกันการคิดบางรูปแบบล่วงหน้าในการทำแบบทดสอบ (เช่น ผ่านสูตรลัด ไม่ได้แสดงถึงลักษณะที่ต้องการ)

การทบทวนแบบวัด

- บางครั้ง แบบทดส

การทบทวนแบบวัด

- ทบทวนแบบวัดที่เสร็จแล้วว่ามีคุณสมบัติความเที่ยง ความตรง ตามที่พึงประสงค์หรือไม่
  - ถ้ามีคุณสมบัติที่ดี ก็เขียนคู่มือการใช้แบบวัด และสร้างรายงานคุณภาพของแบบวัด
  - ถ้าไม่มีคุณสมบัติที่ดี ต้องทบทวนแบบทดสอบใหม่ กลับไปขั้นตอนสร้างข้อคำถาม

การทบทวนแบบวัด

- บางครั้ง แบบวัดที่พัฒนามานานแล้ว ต้องมีการทบทวนคุณภาพของแบบทดสอบใหม่
- ข้อคำถามบางข้ออาจหลุดออกไปยังคนทั่วไปแล้ว หรือข้อคำถามบางข้อไม่ทันสมัยแล้ว
- หากเป็นมาตรวัดที่ใช้ในการประเมินรายบุคคล ต้องมีการเทียบคะแนนระหว่างแบบวัดชุดเก่าและชุดใหม่ เรียกว่า การทำให้แบบทดสอบเท่าเทียมกัน (Test Equating)

การปรับปรุงแบบทดสอบ

- ก่อนสร้างแบบวัดใหม่ ควรตรวจสอบแบบวัดเก่าก่อน
- หาแบบทดสอบเก่าได้อย่างไร
  - อินเทอร์เน็ต
  - งานวิจัย (เช่น งานวิจัยในวารสาร Assessment, วิทยานิพนธ์)
  - Tests in Print
  - Mental Measurements Yearbook
- แบบทดสอบเก่าเหมาะสมกับวัตถุประสงค์ของคุณหรือไม่
- ถ้าไม่เหมาะสม จะสามารถนำมาปรับปรุงได้หรือไม่ แล้วประเมินสิ่งที่ปรับปรุงแล้วได้อย่างไร

การปรับปรุงแบบทดสอบ

- แบบวัดอยู่ในภาษาอังกฤษ
  - การแปลไปกลับ (Translation and Back Translation) ให้แก้ไขจนกว่าผู้วิจัยเชื่อว่าข้อคำถามใกล้เคียงกันทั้งสองภาษา
  - ทดสอบความเข้าใจภาษาของแบบวัดที่แปลแล้ว
  - หากต้องเก็บข้อมูลจากทั้งสองกลุ่ม ต้องทดสอบความใกล้เคียงกันของแบบวัด (Measurement Invariance)
    - น้ำหนักองค์ประกอบ (Factor loadings) ต้องใกล้เคียงกัน
    - ค่าเฉลี่ยของข้อคำถามหลังจากควบคุมระดับคะแนนองค์ประกอบแล้ว หรือจุดตัดของข้อคำถาม (Item intercept) จะต้องใกล้เคียงกันระหว่างกลุ่ม

การปรับปรุงแบบทดสอบ

- แบบวัดอยู่ในภาษาอังกฤษ
  - หากเก็บข้อมูลจากแบบทดสอบที่แปลงแล้วเพียงอย่างเดียว อาจวิเคราะห์ความเที่ยงและความตรงเพิ่มเติม โดยเฉพาะอย่างยิ่ง การวิเคราะห์องค์ประกอบ แล้วดูว่าน้ำหนักองค์ประกอบใกล้เคียงกับแบบวัดดั้งเดิมหรือไม่
  - นำน้ำหนักองค์ประกอบมาหาความสัมพันธ์กันระหว่างสองกลุ่ม ควรมีค่า .8 ขึ้นไป
  - ไม่ควรตัดข้อคำถาม แต่ควรเก็บข้อมูลทั้งหมดด้วยข้อคำถามทั้งหมด แล้ววิเคราะห์แยกระหว่างมีข้อคำถามที่แย่และไม่มีข้อคำถามที่แย่

การปรับปรุงแบบทดสอบ

- กลุ่มตัวอย่างที่แตกต่างจากเดิม
  - ทดสอบความเข้าใจภาษาใหม่
  - วิเคราะห์น้ำหนักองค์ประกอบใหม่ แล้วหาสหสัมพันธ์ระหว่างน้ำหนักองค์ประกอบของกลุ่มตัวอย่างเก่าและกลุ่มตัวอย่างใหม่ ควรมีค่า .8 ขึ้นไป
  - เช่นเดิม ไม่ควรตัดข้อคำถาม แต่ควรวิเคราะห์แยกระหว่างข้อคำถามทั้งหมดและตัดข้อคำถามแล้ว

การปรับปรุงแบบทดสอบ

- แบบวัดใช้เวลานานเกินไป
  - มีผลการวิเคราะห์องค์ประกอบเดิม
    - ให้ใช้ผลการวิเคราะห์องค์ประกอบเดิม ตัดข้อคำถามที่มีน้ำหนักองค์ประกอบต่ำ โดยให้มีข้อคำถามในองค์ประกอบเก่าอย่างน้อย 3 ข้อ
    - เก็บข้อมูลใหม่ ใช้การวิเคราะห์องค์ประกอบแบบยืนยัน (Confirmatory factor analysis) เพื่อตรวจสอบว่าแบบทดสอบเก่ามีองค์ประกอบเหมือนเดิมหรือไม่
  - ไม่มีผลการวิเคราะห์องค์ประกอบเดิม
    - ตรวจสอบความครอบคลุมของเนื้อหาในแบบทดสอบเดิม
    - วิเคราะห์องค์ประกอบ แล้วตัดข้อคำถามผ่านการวิเคราะห์องค์ประกอบ

คาบต่อไป

- ใบงานที่ 4
  - หาความเที่ยงและความตรงจากข้อมูลจริง
  - เขียนรายงานสรุป
- ดูวีดีโอเรื่อง EFA