

บทที่ 2

แนวคิด ทฤษฎีและเอกสารงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ค้นคว้าเอกสารและทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์ข้อสอบจากรายงานการวิเคราะห์ข้อสอบ และเว็บไซต์ต่างๆ โดยนำเสนอข้อมูลดังรายละเอียดต่อไปนี้

ความสำคัญของการวิเคราะห์ข้อสอบ

การวิเคราะห์ข้อสอบเป็นการตรวจสอบคุณภาพของข้อสอบเป็นรายข้อเพื่อพิจารณาว่าข้อสอบแต่ละข้อนั้น มีระดับค่าความยากหรือค่าความง่าย (Difficulty index or Easiness) และค่าอำนาจจำแนกของข้อสอบ (Discriminant index) เพียงใด รวมทั้งพิจารณาถึงประสิทธิภาพของตัวลงในข้อเลือกตอบของข้อสอบข้อนั้นด้วย ผลการวิเคราะห์จะทำให้ทราบว่าข้อสอบแต่ละข้อมีความเหมาะสมมากน้อยเพียงใด ข้อสอบที่มีคุณภาพจะสามารถนำไปวัดและประเมินผลได้อย่างเที่ยงตรงและเชื่อมั่นได้ (ปวรส บุตะเขียว, 2555)

ข้อสอบที่วิเคราะห์คุณภาพรายข้อ ลักษณะข้อสอบจะเป็นแบบเลือกตอบแบบถูก-ผิดหรือแบบเติมคำสั้นๆ แต่ต้องมีรูปแบบเดียวกันทั้งฉบับซึ่งมีวิธีวิเคราะห์ 2 แบบ คือ

1. การวิเคราะห์ข้อสอบแบบเลือกตอบแบบอิงกลุ่ม

การวิเคราะห์ข้อสอบแบบเลือกตอบแบบอิงกลุ่มเป็นการวิเคราะห์ข้อสอบที่ใช้ผลการสอบไปประเมินแบบอิงกลุ่ม โดยนำความแตกต่างระหว่างบุคคลมาเป็นแนวคิดในการประเมินผล การวิเคราะห์ข้อสอบแบบอิงกลุ่มนี้ จึงเป็นการวิเคราะห์โดยนำผลคะแนนของผู้เรียนแต่ละคนไปเปรียบเทียบกับผลคะแนนของผู้เรียนคนอื่นๆ ในกลุ่มเดียวกันเพื่อจำแนกคนที่เรียนเก่งออกจากคนที่เรียนอ่อน ดัชนีบ่งชี้คุณภาพของข้อสอบสำหรับแบบสอบอิงกลุ่ม ได้แก่ ระดับความยากง่าย (p) และอำนาจจำแนก (r) ซึ่งข้อสอบที่ดีจะต้องมีระดับความยากง่ายพอเหมาะ และต้องมีอำนาจการจำแนกที่ดี

1.1 ระดับความยากง่าย (Difficulty) หมายถึง สัดส่วนของผู้ตอบถูกจากคนทั้งหมดที่ตอบในข้อนั้น เช่น ข้อสอบข้อหนึ่ง มีผู้ทำข้อสอบ 50 คน ผู้ที่ตอบถูกมี 40 คน ข้อสอบข้อนี้จะมีค่าความยากง่าย = 0.80 ค่าสัดส่วนนี้เรียกว่า ดัชนีความยากง่าย (Index of difficulty) นิยมใช้อักษร P แทน ค่า P ที่ได้จากการคำนวณจะอยู่ระหว่าง 0 ถึง 1 สุมาลี จันทร์ชโล (2542: 136) ได้ให้ข้อเสนอแนะในการเลือกข้อสอบจากการพิจารณาค่า P ดังนี้

ตารางที่ 2.1 ข้อเสนอแนะในการเลือกข้อสอบจากค่าความยากง่าย

ค่าความยากง่าย (P)	ความหมาย	ข้อเสนอแนะ
.81 - 1.00	ง่ายมาก	ควรตัดทิ้ง
.61 - .80	ค่อนข้างง่าย	ดีพอใช้ ควรเก็บไว้ใช้
.41 - .60	ปานกลาง	ดีมากเก็บไว้ใช้
.20 - .40	ค่อนข้างยาก	ดีพอใช้ ควรเก็บไว้ใช้
.00 - .19	ยากมาก	ควรตัดทิ้ง

สูตรในการคำนวณค่าความยากง่าย (พวงรัตน์ ทวีรัตน์ 2540: 129) คือ

$$P = R / N$$

P = ค่าความยากง่ายของคำถามข้อนั้น

R = จำนวนผู้ตอบถูกในข้อนั้น

N = จำนวนผู้ตอบทั้งหมด

1.2 อำนาจจำแนก (Discrimination power) หมายถึง ความสามารถของข้อสอบในการจำแนกหรือแยกให้เห็นความแตกต่างระหว่างผู้สอบที่มีผลสัมฤทธิ์ต่างกัน เพื่อใช้พยากรณ์หรือบ่งชี้ความแตกต่างที่เห็นชัดในด้านความสามารถ เช่น จำแนกคนเก่งกับคนอ่อนจากกันได้ โดยถือว่าคนเก่งควรทำข้อสอบข้อนั้นได้ ส่วนผู้ที่อ่อนไม่ควรทำข้อสอบข้อนั้นได้ ดัชนีอำนาจจำแนกหรือค่าอำนาจจำแนก (Index of Discrimination) คัดจากสัดส่วนของผลต่างระหว่างจำนวนผู้ตอบถูกในกลุ่มที่ได้คะแนนสูงกับกลุ่มที่ได้คะแนนน้อย ใช้แทนด้วยอักษร r และมีค่าระหว่าง - 1 ถึง + 1 ข้อคำถามที่ดีจะต้องมีค่าอำนาจจำแนกเป็นบวก กรณีที่ค่า r ติดลบแสดงว่าข้อสอบข้อนั้นจำแนกกลับ คนเก่งทำไม่ได้แต่คนอ่อนทำได้ ถือว่าเป็นข้อสอบที่ไม่ดีควรตัดทิ้ง ค่าอำนาจจำแนกที่ดีควรมีค่าตั้งแต่ 0.2 ขึ้นไป (ปวรส บุตะเขียว, 2555) และยังมีค่าเข้าใกล้ +1 ยิ่งดี (Chase 1978: 140) นอกจากนี้ อาจารย์ผู้สอนควรตรวจสอบการจัดการเรียนสอนของตนว่าเพราะเหตุใดผู้ที่เรียนเก่งจึงไม่เข้าใจในเรื่องที่สอน

หากพิจารณาถึงความสัมพันธ์ระหว่างความยากง่ายและอำนาจจำแนกจะพบว่า ถ้าข้อสอบมีความยากง่าย 0.50 แสดงว่า ผู้เรียนจำนวนครึ่งหนึ่งทำข้อสอบนั้นถูก อำนาจจำแนกมีค่าสูงสุด แต่หากข้อสอบยากมากหรือง่ายมาก ค่าอำนาจจำแนกจะลดลง (สุมาลี จันทร์ชลอ 2542: 138)

สูตรในการคำนวณค่าอำนาจจำแนก (พวงรัตน์ ทวีรัตน์ 2540: 130) คือ

$$r = \frac{R_U - R_e}{N / 2}$$

r = ค่าอำนาจจำแนกของคำถามข้อนั้น

R_U = จำนวนผู้ตอบถูกในข้อนั้น ในกลุ่มเก่ง

R_e = จำนวนผู้ตอบถูกในข้อนั้น ในกลุ่มอ่อน

N = จำนวนคนเข้าสอบ

ตารางที่ 2.2 เกณฑ์การแปลความหมายค่าอำนาจจำแนก (r) ของข้อสอบ

ค่าอำนาจจำแนก (r)	ความหมาย
0.80 - 1.00	จำแนกได้ดีมาก
0.60 - 0.79	จำแนกได้ดี
0.40 - 0.59	จำแนกได้ปานกลาง
0.20 - 0.39	จำแนกได้พอใช้
0.00 - 0.19	จำแนกได้น้อยมาก
-1.00 - 0.09	ใช้ไม่ได้

(ล้วน สายยศและ อังคณา สายยศ, 2543, หน้า 185)

ประสิทธิภาพของตัวลวง (สุระ ชั้นที่แก้ว, 2555, หน้า 15)

1. สัดส่วนของผู้ที่เลือกตัวลวง (P_w) หมายถึง สัดส่วนของจำนวนคนที่เลือกตัวลวงนั้นๆ เช่น ข้อสอบข้อหนึ่งมีคนสอบ 100 คน มีคนเลือกตัวลวง ก. 25 คน แสดงว่าสัดส่วนผู้เลือกตัวลวง ก. 0.25 หรือ 25%

P_w จะมีค่า 0 ถึง 1 ตัวลวงที่ดีควรมีค่า P_w ตั้งแต่ 0.05 ขึ้นไป

2. อำนาจจำแนกของตัวลวง (r_w) หมายถึง ผลต่างระหว่างสัดส่วนของคนในกลุ่มอ่อนที่เลือกตัวลวงกับสัดส่วนของคนในกลุ่มเก่งที่เลือกตัวลวงนั้นๆ

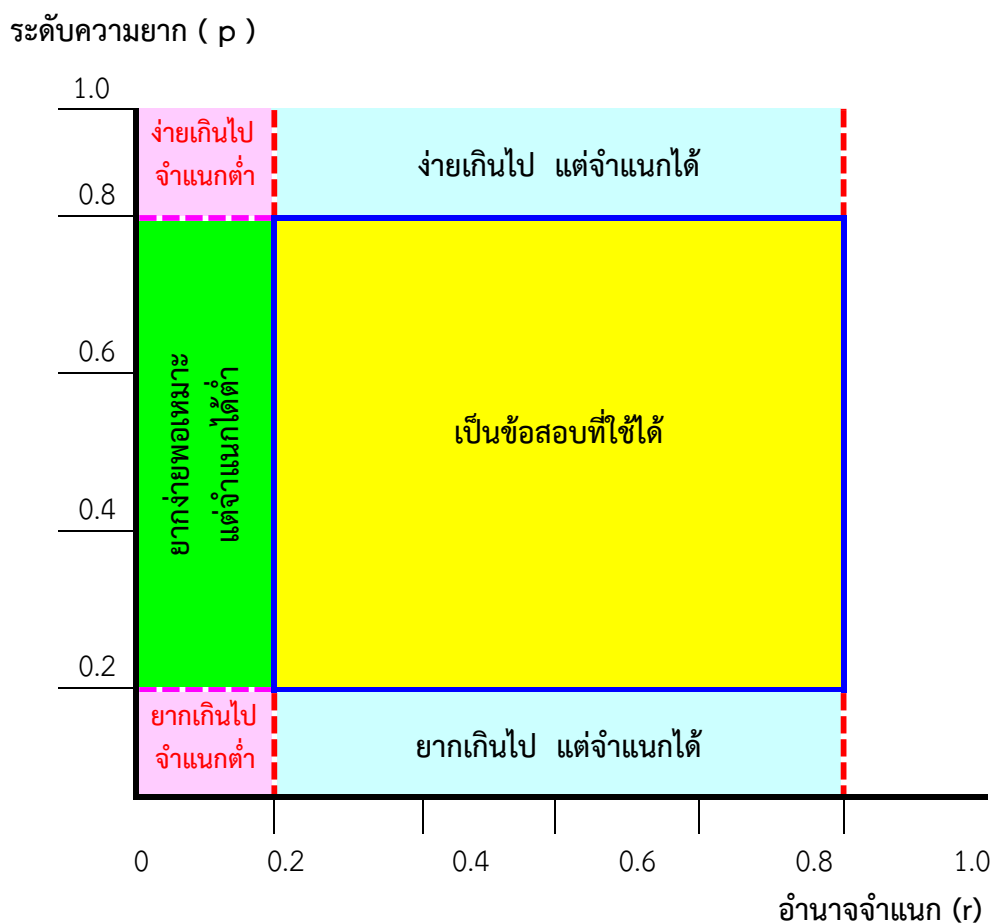
r_w มีค่าตั้งแต่ -1 ถึง 1 ตัวลวงที่ดีควรมีค่า r_w ตั้งแต่ 0.05 ขึ้นไป

การเลือกข้อสอบจากผลการวิเคราะห์ข้อสอบ

เมื่อวิเคราะห์ข้อสอบได้ค่าความยากง่ายและอำนาจการจำแนกแล้วควรนำมาจัดทำกราฟซึ่งจะแสดงการกระจายของข้อสอบและชี้ให้เห็นว่าข้อสอบข้อใดบ้างที่อยู่ในเกณฑ์ใช้ได้หรือใช้ไม่ได้

(ปวรส บุตะเขี้ยว, 2555)

ภาพที่ 2.1 แสดงการกระจายของข้อสอบ



2. การวิเคราะห์ข้อสอบแบบเลือกตอบแบบอิงเกณฑ์

การวิเคราะห์ข้อสอบแบบอิงเกณฑ์มีจุดมุ่งหมายเพื่อดูว่าผู้เรียนมีความรู้ ความสามารถ อะไรบ้างหรือผลที่เกิดจากการเรียนการสอนตามเกณฑ์หรือมาตรฐานที่ตั้งไว้เพียงใด มากกว่าที่จะ เปรียบเทียบกับกลุ่มว่าใครเก่งกว่าใคร เนื่องจากการวิเคราะห์ข้อสอบแบบอิงเกณฑ์เกี่ยวข้องกับ ประสิทธิภาพในการวัดผลของการจัดการเรียนการสอนโดยตรง ถ้าหากอาจารย์ใช้ข้อสอบชุดเดียวกัน ทดสอบก่อนเรียน (Pre – Test) และหลังเรียน (Post – Test) ข้อสอบอิงเกณฑ์ที่มีประสิทธิภาพควร ตั้งอยู่บนสมมติฐานที่ว่ากระบวนการเรียนการสอนจะส่งผลโดยตรงต่อการเรียนรู้ของผู้เรียน เมื่อการ เรียนการสอนสิ้นสุดลงควรมีผู้ตอบถูกมากกว่าก่อนเรียน จึงเกิดแนวคิดของดัชนีความไวของผลการ เรียนการสอนเกิดขึ้น

ดัชนีบ่งชี้คุณภาพของข้อสอบสำหรับแบบสอบอิงเกณฑ์ ได้แก่ ความสอดคล้องระหว่าง ข้อสอบกับจุดมุ่งหมาย (Item – Objective Congruence: IOC) ความยากของข้อสอบ (Item Difficulty: P) และอำนาจจำแนกหรือดัชนีความไว (Sensitivity Index: S_i)

2.1 ความสอดคล้องระหว่างข้อสอบกับจุดมุ่งหมาย (IOC) หมายถึง ลักษณะของข้อสอบที่ เป็นตัวแทนของจุดมุ่งหมายการเรียนรู้ที่กำหนดไว้ วิธีการวัดความสอดคล้องสามารถทำได้โดยอาศัย การตัดสินของผู้ทรงคุณวุฒิไม่ต่ำกว่า 3 คนเป็นผู้พิจารณา เช่น ให้ผู้ทรงคุณวุฒิประเมินค่าความ

สอดคล้องของข้อสอบแต่ละข้อกับจุดมุ่งหมายการเรียนรู้ที่มุ่งวัดโดยใช้มาตราส่วนแบบประเมินค่า 5 ระดับ (Likert Scale) แล้วนำมาคำนวณหาค่าเฉลี่ย ถ้าค่าเฉลี่ย (\bar{X}) ≥ 4.00 ถือว่าสอดคล้องหรือถ้าใช้แบบประเมินค่า 3 ระดับถือว่าข้อสอบนั้นสอดคล้องกับจุดมุ่งหมายเมื่อ $\bar{X} \geq 80\%$

2.2 ความยากของข้อสอบ (Item Difficulty: P_i) หมายถึง สัดส่วนของผู้ตอบข้อสอบข้อนั้นๆ ได้ถูกต้อง โดยพิจารณาจากกลุ่มที่ได้เรียนเรื่องนั้นมาแล้วกับกลุ่มที่ไม่ได้เรียนเรื่องนั้น ข้อสอบที่ง่ายสำหรับกลุ่มที่เคยเรียนสามารถเป็นข้อสอบอิงเกณฑ์ที่ดีที่สุดเพราะอาจารย์ผู้สอนสามารถจัดการเรียนการสอนให้ผู้เรียนส่วนใหญ่เกิดการเรียนรู้และทำข้อสอบได้หรืออาจเป็นเพราะข้อสอบง่ายเกินไปจึงทำให้ผู้เรียนตอบถูกจากความรู้นี้โดยไม่ได้เรียนรู้เพิ่มเติม ค่า P_i จะอยู่ระหว่าง 0 ถึง 1

ดังนั้น การวิเคราะห์ข้อสอบอิงเกณฑ์จึงจำเป็นต้องตรวจสอบว่าก่อนเรียนมีสัดส่วนของคำตอบถูกมากน้อยเพียงใดเพื่อนำผลมาเปรียบเทียบกันตอนหลังเรียน

การวิเคราะห์ข้อสอบแบบอิงเกณฑ์ จะพิจารณาค่าความยากของข้อสอบแตกต่างจากแบบอิงกลุ่ม เนื่องจากไม่ได้เน้นที่จะนำค่าความยากง่ายมา เพื่อเลือกข้อสอบแต่จะเน้นที่คุณภาพในการสอน กล่าวคือ ถ้าอาจารย์ยังไม่ได้สอนเนื้อหาข้อสอบควรจะยาก คือ มีค่า P_i ต่ำ แต่ถ้าได้สอนแล้วและสอนอย่างมีประสิทธิภาพส่งผลให้ผู้เรียนเกิดการเรียนรู้ ข้อสอบควรจะง่าย คือ มีค่า P_i สูง

ลักษณะข้อสอบอิงเกณฑ์ที่ดีควรมีค่าความยากของข้อสอบ ดังนี้ (ศิริชัย กาญจนวาสี, 2552 หน้า 243)

ตารางที่ 2.3 ลักษณะข้อสอบอิงเกณฑ์ที่ดี

กลุ่มผู้เรียน	ความยากของข้อสอบ (P_i)
1. กลุ่มที่เคยเรียนมาแล้ว	<ul style="list-style-type: none"> • ควรมีค่า P_i สูง (0.70 - 1.00) • ถ้า $P_i < 0.70$ ควรปรับปรุงหรือตัดทิ้ง
2. กลุ่มที่ยังไม่เคยเรียน	<ul style="list-style-type: none"> • ควรมีค่า P_i ต่ำ (0 - 0.5) • ถ้า $P_i > 0.5$ แสดงว่าข้อสอบง่ายเกินไป (ควรปรับปรุงหรือตัดทิ้ง)

2.3 อำนาจจำแนกหรือดัชนีความไว (Sensitivity Index: S_p) หมายถึง ผลต่างระหว่างสัดส่วนของผู้ที่ตอบถูกหลังเรียนกับสัดส่วนผู้ที่ตอบถูกก่อนเรียน

S_p มีค่าระหว่าง - 1 ถึง 1 ค่า S_p ควรมีค่าเป็นบวก ยิ่งเข้าใกล้ 1 ยิ่งมีความไว แสดงว่าข้อสอบข้อนั้นสามารถจำแนกผู้รอบรู้จากผู้ไม่รอบรู้ได้ดี และสามารถจำแนกการสอนได้อย่างมีประสิทธิภาพ ดังนั้น อำนาจจำแนกหรือดัชนีความไวของข้อสอบ (S_p) ควรมีค่า > 0

ลักษณะข้อสอบอิงเกณฑ์ที่สามารถจำแนกการสอนได้อย่างมีประสิทธิภาพควรมีค่า S_p ดังนี้ (พิชิต ฤทธิ์จรูญ, 2552, หน้า 145)

ตารางที่ 2.4 ลักษณะข้อสอบอิงเกณฑ์ที่สามารถจำแนกการสอนได้อย่างมีประสิทธิภาพ

อำนาจจำแนกหรือดัชนีความไวของข้อสอบ (S_r)
ตั้งแต่ 0.40 ขึ้นไป

ซึ่งค่า S_r นี้ จะแตกต่างจากค่าอำนาจจำแนก (r) ของข้อสอบแบบอิงกลุ่มคือ ค่าอำนาจจำแนก (r) ของข้อสอบแบบอิงกลุ่มเป็นดัชนีที่ชี้ให้เห็นว่าสามารถจำแนกคนที่มีผลสัมฤทธิ์ทางการเรียนสูงและต่ำได้มากน้อยเพียงใด

ข้อจำกัดของดัชนีความไว

1. ต้องสอบ 2 ครั้ง คือ ก่อนเรียน และหลังเรียน จึงจะคำนวณค่าได้
2. S_r ที่มีค่าต่ำหรือมีค่าเป็นลบ อาจเป็นผลจากข้อสอบหรือการจัดการเรียนการสอนไม่มีประสิทธิภาพ
3. กรณีที่มีระยะห่างระหว่างการสอบครั้งแรกและครั้งที่ 2 ใกล้เคียง คำตอบหรือคะแนนครั้งหลังอาจได้รับอิทธิพลจากการสอบครั้งแรกได้

2.4 ประสิทธิภาพของตัวลวง เมื่อสร้างข้อสอบอิงเกณฑ์แบบหลายตัวเลือก (Multiple Choices) ผู้ที่เลือกตัวลวง ถือว่า ตอบผิด ซึ่งแสดงให้เห็นว่าผู้นั้นยังไม่สัมฤทธิ์ผลตามเป้าหมายของการวัดในการวิเคราะห์ประสิทธิภาพตัวลวงทำโดยการตรวจสอบความถี่ของผู้ที่เลือกตัวลวงนั้นๆ และควรดูจากผลการสอบก่อนเรียน (Pre - Test) ถ้าตัวลวงใดมีผู้เลือกในสัดส่วนที่สูงถือว่าเป็นตัวลวงที่ใช้ได้ แต่ถ้าตัวลวงใดมีผู้เลือกน้อย แสดงว่าตัวลวงนั้นขาดประสิทธิภาพ สมควรที่จะต้องมีการปรับปรุงแก้ไขให้ดีขึ้น (ศิริชัย กาญจนวาสี, 2552 หน้า 241) และตัวลวงที่จัดว่าเป็นตัวลวงที่ดีนั้น ผู้ที่เรียนอ่อนควรจะเลือกตอบมากกว่าผู้ที่มีผลการเรียนดี

ประโยชน์ของการวิเคราะห์ข้อสอบ

1. ทำให้ทราบข้อมูลพื้นฐานของตัวข้อสอบและคำตอบ รวมถึงข้อสอบแต่ละข้อได้ทำหน้าที่วัดผลสัมฤทธิ์อย่างมีประสิทธิภาพเพียงใด
2. ชี้ให้เห็นถึงจุดอ่อนที่ผู้สอนหรือผู้เรียนต้องปรับปรุงแก้ไขเพื่อพัฒนาความสามารถและทักษะของผู้เรียนให้เป็นตามที่คาดหวัง
3. เป็นพื้นฐานสำคัญในการปรับปรุงพัฒนาการเรียนการสอน โดยเฉพาะอย่างยิ่งด้านพฤติกรรมการณ์เรียนรู้ที่เหมาะสมสำหรับเนื้อหาวิชานั้นๆ
4. ช่วยเพิ่มทักษะในการสร้างข้อสอบ ทำให้ทราบถึงอำนาจการจำแนก ระดับความยากง่าย ประสิทธิภาพของตัวลวง ตลอดจนการเขียนข้อสอบในลักษณะใดจึงจะได้ข้อสอบที่ดี
5. ทำให้สามารถคัดเลือกข้อสอบที่มีคุณภาพมารวมเป็นฉบับข้อสอบที่ดี สามารถจัดทำข้อสอบคู่ขนานและเป็นรากฐานสำคัญในการพัฒนาแบบสอบมาตรฐาน

ข้อสอบที่ดีควรมีค่าความยากง่ายที่เหมาะสมและมีอำนาจจำแนกสูง ค่าเฉลี่ยความยากง่ายของข้อสอบทั้งฉบับ ควรมีค่าประมาณ 0.50 แต่อย่างไรก็ตามในการคัดเลือกข้อสอบจะต้องคำนึงถึงความสมดุลระหว่างข้อสอบที่มีสถิติเหมาะสมกับข้อสอบที่วัดครอบคลุมจุดประสงค์ และเนื้อหาที่ต้องการด้วย บางครั้งอาจมีความจำเป็นที่จะต้องใช้ข้อสอบที่มีอำนาจการจำแนกไม่สูงนัก เพื่อให้มีข้อสอบที่วัดครอบคลุมเนื้อหาที่ต้องการ

1. การวิเคราะห์ข้อสอบแบบเลือกตอบแบบอิงกลุ่ม

ระดับความยากง่าย (p) ถ้าค่า p ยิ่งมากข้อสอบข้อนั้นจะง่ายและค่า p ยิ่งน้อย ข้อสอบนั้นจะยาก ทั้งนี้ข้อสอบที่ดีควรมีค่า p ระหว่าง 0.20 - 0.80

อำนาจการจำแนก (r) เป็นการจำแนกผู้เรียนออกจากกันตามระดับความรู้ความสามารถ ค่า r ควรมีค่าตั้งแต่ 0.20 ขึ้นไป

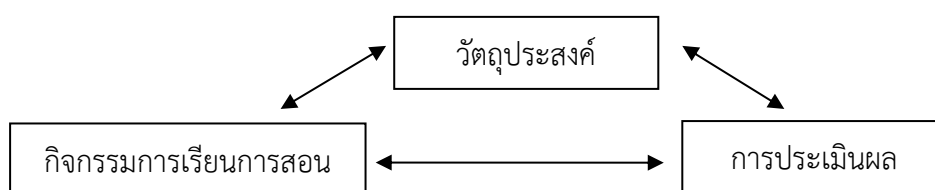
2. การวิเคราะห์ข้อสอบแบบเลือกตอบแบบอิงเกณฑ์

โดยการตรวจสอบค่าดัชนีความไว (S_r) ของผลการเรียน จะช่วยสะท้อนให้เห็นถึงความสามารถก่อนเรียนกับผลการเรียนรู้หลังการเรียนการสอนแล้ว ข้อสอบที่ดีควรมี S_r ตั้งแต่ 0.40 ขึ้นไป

ผลการวิเคราะห์ข้อสอบ ถ้าพบว่าข้อสอบไม่มีคุณภาพควรกลับมาพิจารณาว่าเกิดจากสาเหตุใด ข้อคำถามไม่มีความเป็นปรนัยหรือไม่ รวมทั้งต้องพิจารณาที่ตัวเลือกด้วยว่าเป็นไปตามหลักการเขียนข้อสอบที่ดีหรือไม่ และดำเนินการปรับปรุงข้อสอบรวมทั้งนำไป Try out อีกครั้ง เพื่อวิเคราะห์ประสิทธิภาพของข้อสอบต่อไป

ทฤษฎีการวิเคราะห์ข้อสอบ

วรรณดี แสงประทีปทอง (2556) การจัดการศึกษาประกอบด้วยองค์ประกอบที่สำคัญ 3 ส่วน ได้แก่ วัตถุประสงค์ กิจกรรมการเรียนการสอนและการประเมินผล องค์ประกอบทั้ง 3 มีความสัมพันธ์กันดังนี้



วัตถุประสงค์เป็นเป้าหมายของการจัดการศึกษา และเป็นตัวบ่งชี้การจัดการเรียนการสอนและการประเมินผล

กิจกรรมการเรียนการสอนเป็นกระบวนการที่จะให้ผู้เรียนได้พัฒนาไปตามวัตถุประสงค์และเป็นแนวทางในการประเมินผล

ส่วนการประเมินผลเป็นเครื่องชี้วัดผลที่เกิดขึ้นในตัวผู้เรียนซึ่งเกิดจากการจัดกิจกรรมการเรียนการสอนตามวัตถุประสงค์และเป็นเครื่องชี้วัดความเหมาะสมของการจัดกิจกรรม ดังนั้น การประเมินผลจึงมีความสำคัญอย่างมากต่อการจัดการศึกษา

ในการประเมินผลอาจารย์ผู้สอนจะใช้เครื่องมือชนิดต่างๆ เพื่อวัดความรู้ความสามารถที่เกิดขึ้นในตัวผู้เรียน เครื่องมือที่สร้างขึ้นนี้ ถ้ามีคุณสมบัติเป็นเครื่องมือวัดที่ดีจะทำให้การวัดความรู้ความสามารถของผู้เรียนทำได้อย่างถูกต้องตรงกับความสามารถที่แท้จริง เครื่องมือวัดที่นิยมใช้มากที่สุด คือ แบบทดสอบ โดยทั่วไปเป็นแบบทดสอบที่ผู้สอนสร้างขึ้นเอง ดังนั้นจึงจำเป็นต้องมีการตรวจสอบคุณภาพของเครื่องมือวัด ซึ่งเรียกว่าการวิเคราะห์ข้อสอบเพื่อนำผลการวิเคราะห์มาใช้เป็นข้อมูลในการปรับปรุงข้อสอบให้มีคุณภาพ และสามารถวัดความรู้ความสามารถที่แท้จริงของผู้เรียนได้

การวิเคราะห์ข้อสอบ วิเคราะห์ได้ 2 ลักษณะ คือ วิเคราะห์ข้อสอบเป็นรายข้อ และวิเคราะห์ข้อสอบทั้งฉบับ วิธีการวิเคราะห์สามารถทำได้หลายแนวทางขึ้นกับทฤษฎีการวัดผลที่นำมาใช้ ได้แก่ ทฤษฎีการทดสอบแบบคลาสสิกหรือเรียกว่าทฤษฎีการทดสอบแบบดั้งเดิม และทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT)

ความหมายของคำศัพท์ที่ใช้ในการวิเคราะห์ข้อสอบ

1. ความยากง่าย (Difficulty) คือ เปอร์เซนต์ของจำนวนผู้สอบที่สามารถทำข้อสอบนั้นได้ ถูกต้องถ้าข้อสอบข้อใดมีเปอร์เซนต์ของจำนวนผู้ตอบข้อสอบถูกมาก แสดงว่าข้อสอบนั้นง่าย ถ้าข้อสอบข้อใดมีเปอร์เซนต์ของจำนวนผู้ตอบข้อสอบถูกน้อย แสดงว่าข้อสอบข้อนั้นยาก ผลการคำนวณที่ได้จะใช้สัญลักษณ์ p เรียกว่า ความยากง่ายของข้อสอบ ซึ่งจะมีค่าอยู่ระหว่าง 0 ถึง 1

2. อำนาจจำแนก (Discrimination) คือ ข้อสอบข้อเดียวกันนำไปใช้สอบกับเด็กกลุ่มเก่งและกลุ่มอ่อน เด็กกลุ่มเก่งควรตอบข้อสอบข้อนั้นถูกต้องมากกว่าเด็กกลุ่มอ่อน ผลการคำนวณที่ได้จะใช้สัญลักษณ์ r เรียกว่า อำนาจจำแนก ซึ่งจะมีค่าอยู่ระหว่าง - 1 ถึง 1

3. ประสิทธิภาพของตัวถูก ตัวถูกที่เราต้องการให้เด็กเก่งเลือกตอบถูกมากกว่า เด็กอ่อน ถ้าเป็นเช่นนี้ค่า r จะเป็นบวก ข้อสอบที่มีค่า r เป็นบวกมากยิ่งดี ข้อสอบวัดผลสัมฤทธิ์ทางการเรียนควรมีค่า r ตั้งแต่ .20 ขึ้นไป ในทางกลับกันถ้าข้อสอบข้อใดตัวถูกถูกมีจำนวนเด็กอ่อนเลือกตอบมากกว่าเด็กเก่ง ค่า r จะเป็นลบหรือจำนวนเด็กเก่งและเด็กอ่อนไม่เลือกเลย ค่า r จะเป็น 0 ข้อสอบประเภทนี้ไม่มีประสิทธิภาพ ควรจะปรับปรุงเสียใหม่หรือไม่ก็ตัดทิ้งไป

4. ประสิทธิภาพของตัวลวง ตัวลวง เราต้องการลวงเด็กอ่อนมากกว่าลวงเด็กเก่ง ถ้าเป็นเช่นนี้ค่า r จะเป็นบวก ตัวลวงที่มีประสิทธิภาพควรมีค่า r เป็นบวก หรือมีค่า r มากกว่า 0 ในทางกลับกันตัวลวงใดมีจำนวนเด็กเก่งไปหลงตอบมากกว่าเด็กอ่อน ค่า r จะเป็นลบ หรือจำนวนเด็กเก่งและเด็กอ่อนเลือกตอบเท่ากันหรือทั้งเด็กเก่งและเด็กอ่อนไม่เลือกเลย ค่า r จะเป็น 0 ข้อสอบประเภทนี้ไม่มีประสิทธิภาพ ควรปรับปรุงเสียใหม่หรือไม่ก็ตัดทิ้งไป

5. ค่าความเชื่อมั่นของข้อสอบ (Reliability) คือ ความคงที่ของคะแนน กล่าวคือถ้านำข้อสอบฉบับหนึ่งไปสอบกับเด็กคนเดิมกี่ครั้งก็ได้คะแนนคงเดิมหรือใกล้เคียงกับคะแนนเดิมแสดงว่าข้อสอบฉบับนั้นมีความเชื่อมั่นสูง ความเชื่อมั่นของข้อสอบจะมีค่าอยู่ระหว่าง - 1.00 ถึง 1.00 ข้อสอบที่ดีควรมีค่าความเชื่อมั่นสูง ยิ่งสูงยิ่งดี ไม่ควรติดลบ ส่วนจะมีค่าเท่าไรถึงจะใช้ได้นั้นยังไม่มีเกณฑ์ที่แน่นอนทั้งนี้ เพราะค่าความเชื่อมั่นจะสูงหรือต่ำนั้น นอกจากจะขึ้นอยู่กับตัวข้อสอบแล้วยังขึ้นอยู่กับองค์ประกอบอื่นๆ อีก เช่น จำนวนข้อของข้อสอบ ยิ่งมีมากข้อความเชื่อมั่นก็จะยิ่งสูง ความแตกต่าง

ของกลุ่มเด็กที่ไปสอบถ้าเด็กมีความสามารถแตกต่างกันมาก การกระจายของคะแนนก็มีมากก็จะให้ค่าความเชื่อมั่นสูง

ทฤษฎีการทดสอบแบบคลาสสิก

ทฤษฎีการทดสอบแบบคลาสสิกหรือทฤษฎีการทดสอบแบบดั้งเดิม ใช้วิธีวิเคราะห์ข้อสอบเป็นรายข้อเพื่อวิเคราะห์ค่าความยาก (Item difficulty) ค่าอำนาจจำแนก (Discrimination power) และประสิทธิภาพของตัวลวง (Effectiveness of distracters) ส่วนการวิเคราะห์ข้อสอบทั้งฉบับจะวิเคราะห์ค่าความตรงหรือความเที่ยงตรง (Validity) ค่าความเที่ยงหรือความเชื่อมั่น (Reliability) ของแบบทดสอบ

ทฤษฎีการตอบสนองข้อสอบ

การวิเคราะห์ข้อสอบตามทฤษฎีนี้ เป็นการวิเคราะห์แบบรายข้อ เพื่อหาค่าความยาก ค่าอำนาจจำแนก ค่าความน่าจะเป็นของการเดาถูก ค่าฟังก์ชันสารสนเทศของข้อสอบ (Item information function) ส่วนการวิเคราะห์ข้อสอบทั้งฉบับจะวิเคราะห์ค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test information function) และความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถ ซึ่งถ้าความคลาดเคลื่อนมาตรฐานของการประมาณค่ามีค่าต่ำ แสดงว่าการประมาณค่าความสามารถของผู้เรียนมีความแม่นยำสูง

สำหรับข้อสอบที่ใช้ในการวัดผลระหว่างเรียนหรือใช้วัดผลปลายภาคเรียนนั้น ส่วนใหญ่เป็นข้อสอบที่อาจารย์ผู้สอนสร้างขึ้น และยังไม่ได้พัฒนาเป็นแบบทดสอบมาตรฐาน ดังนั้น ในการวิเคราะห์ข้อสอบจึงควรวิเคราะห์ตามแนวทฤษฎีการทดสอบแบบคลาสสิก ซึ่งจากการวิเคราะห์ จะได้ทราบประสิทธิภาพของตัวลวงด้วย ทำให้สามารถปรับปรุงข้อสอบได้ทั้งตัวคำถาม คำตอบถูกและตัวลวง อันจะนำไปสู่การพัฒนาข้อสอบให้เป็นมาตรฐาน ขณะที่การวิเคราะห์ข้อสอบตามแนวทฤษฎีการตอบสนองข้อสอบเหมาะที่จะใช้วิเคราะห์แบบทดสอบที่ได้พัฒนาแล้ว ดังนั้น ในการกล่าวถึงการวิเคราะห์ข้อสอบต่อไปนี้จะกล่าวเฉพาะการวิเคราะห์ข้อสอบตามแนวทฤษฎีการทดสอบแบบคลาสสิก

ทฤษฎีการทดสอบแบบคลาสสิกนั้นเชื่อว่าคะแนนที่ได้จากการทดสอบแต่ละครั้ง (X) ประกอบด้วยคะแนนความสามารถที่แท้จริง (T) ของผู้สอบและความคลาดเคลื่อนในการวัด (E) ซึ่งสามารถเขียนเป็นสมการทางคณิตศาสตร์ได้ดังนี้

$$X = T + E$$

เนื่องจากการทดสอบแต่ละครั้งจะเกิดความคลาดเคลื่อนขึ้นได้ไม่มากก็น้อยแต่เป้าหมายของการวัดต้องการให้คะแนนที่วัดหรือทดสอบได้ มีค่าใกล้เคียงคะแนนความสามารถที่แท้จริงของผู้สอบมากที่สุด ดังนั้น จึงต้องพยายามหาวิธีการต่างๆ เพื่อให้เกิดความคลาดเคลื่อนน้อยที่สุดเท่าที่จะทำได้

การวิเคราะห์ข้อสอบ

1. การวิเคราะห์ข้อสอบรายข้อ เป็นการตรวจสอบคุณภาพของข้อสอบแต่ละข้อ โดยพิจารณาส่วนสำคัญ 3 ประการ ได้แก่ ค่าความยาก ค่าอำนาจจำแนก และประสิทธิภาพของตัวลวง

1.1 ค่าความยากของข้อสอบ

1.1.1 ความยากของข้อสอบเปรียบเทียบได้กับความชันของเนิน ข้อสอบใดยากมากก็เหมือนเนินที่ชันมาก เด็กป็นไม่ค่อยไหวทำผิดมากจัดเป็นข้อสอบที่ยากมาก และในทำนองเดียวกัน ถ้าข้อสอบข้อใดมีผู้ทำถูกมากแสดงว่าเป็นข้อสอบที่ง่ายหรือมีระดับความยากต่ำ ความหมายของความยากและการคำนวณค่าความยาก ความยากของข้อสอบ คือ สัดส่วนหรือร้อยละของผู้สอบที่ตอบข้อสอบนั้นถูกค่าความยากคำนวณได้จากสูตร

$$\text{ความยาก (p)} = \frac{\text{จำนวนผู้สอบที่ตอบตัวเลือกนั้น}}{\text{จำนวนผู้สอบทั้งหมด}}$$

$$\text{หรือความยาก (p)} = \frac{\text{จำนวนผู้สอบที่ตอบตัวเลือกนั้น} \times 100}{\text{จำนวนผู้สอบทั้งหมด}}$$

$$\text{หรือ } (p) = \frac{H+L}{N_H + N_L}$$

เมื่อ H คือ จำนวนผู้สอบในกลุ่มที่ได้คะแนนรวมสูงที่ตอบตัวเลือกนั้น

N คือ จำนวนผู้สอบในกลุ่มที่ได้คะแนนรวมต่ำที่ตอบตัวเลือกนั้น

N_H คือ จำนวนคนในกลุ่มที่ได้คะแนนรวมสูง

N_L คือ จำนวนคนในกลุ่มที่ได้คะแนนรวมต่ำ

1.1.2 ลักษณะของความยาก มีลักษณะดังนี้

1) ค่าความยากของข้อสอบในรูปสัดส่วน (p) มีค่าอยู่ระหว่าง 0 ถึง 1 และค่าความยากของข้อสอบในรูปร้อยละ (p) มีค่าอยู่ระหว่าง 0 ถึง 100

2) ข้อสอบข้อใดที่ผู้สอบทำถูกหมดทุกคน แสดงว่าข้อสอบง่ายมากมีค่าความยาก (p) เท่ากับ 1.0 หรือ P เท่ากับ 100

1.1.3 ข้อสอบข้อใดที่ผู้สอบทำผิดหมดทุกคน แสดงว่าข้อสอบยากมาก มีค่าความยาก (p) เท่ากับ 0 หรือ P เท่ากับ 0

1.1.4 ข้อสอบข้อใดที่ผู้สอบครึ่งหนึ่งทำถูกและผู้สอบอีกครึ่งหนึ่งทำผิด แสดงว่าเป็นข้อสอบที่ยากปานกลาง หรือยากพอเหมาะ

1.1.5 ข้อสอบข้อใดที่มีค่าความยากสูง แสดงว่าข้อสอบข้อนั้น มีผู้ตอบถูกจำนวนมากจึงเป็นข้อสอบที่ง่าย

1.1.6 ข้อสอบข้อใดที่มีค่าความยากต่ำ แสดงว่าข้อสอบข้อนั้น มีผู้ตอบถูกจำนวนน้อยจึงเป็นข้อสอบที่ยาก

1.2 ค่าอำนาจจำแนกข้อสอบ

1.2.1 ค่าอำนาจจำแนกข้อสอบ หมายถึง ความสามารถของข้อสอบที่จะจำแนกความแตกต่างของสิ่งที่ต้องการวัดโดยสามารถจำแนกกลุ่มผู้สอบที่ได้คะแนนรวมสูงออกจากกลุ่มผู้ที่ได้คะแนนรวมต่ำ การคำนวณค่าอำนาจจำแนกอาจคำนวณได้โดยใช้สูตรอย่างง่าย (r) และสูตรสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล (Point biserial correlation coefficient; r_{pb})

1) การคำนวณค่าอำนาจจำแนก (r) โดยใช้สูตรอย่างง่ายเป็นการเปรียบเทียบจำนวนผู้ตอบในกลุ่มที่ได้คะแนนรวมสูงกับกลุ่มที่ได้คะแนนรวมต่ำที่ตอบข้อสอบข้อนั้น ถูก การคำนวณใช้สูตร

$$r = \frac{H - L}{N_H} \quad \text{หรือ} \quad r = \frac{H - L}{N_L}$$

โดยที่สัญลักษณ์แต่ละตัว มีความหมายเหมือนในการคำนวณค่าความยาก

2) การคำนวณค่าอำนาจจำแนกโดยใช้สูตรสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลวิธีนี้ใช้ในกรณีที่การกระจายของคะแนนรวมหรือกระจายคะแนนของผู้สอบที่ตอบข้อสอบผิดไม่เป็นโค้งปกติ การคำนวณค่า r_{pb} ใช้สูตร

$$r_{pb} = \frac{\bar{X}_p - \bar{X}_q}{S_x} = \sqrt{pq}$$

เมื่อ \bar{X}_p = คือค่าเฉลี่ยของคะแนนของผู้สอบที่ตอบข้อสอบข้อนั้นถูก
 \bar{X}_q = คือค่าเฉลี่ยของคะแนนของผู้สอบที่ตอบข้อสอบข้อนั้นผิด
 S_x = คือค่าเบี่ยงเบนมาตรฐานของคะแนนรวมของผู้สอบทั้งหมด
 P = คือสัดส่วนของผู้สอบที่ตอบข้อสอบนั้นถูก
 q = คือสัดส่วนของผู้สอบที่ตอบข้อสอบนั้นผิด

1.2.2 ลักษณะของอำนาจจำแนก

- 1) อำนาจจำแนกมีค่าตั้งแต่ - 1.00 ถึง 1.00
- 2) ข้อสอบข้อใดที่ผู้สอบในกลุ่มได้คะแนนรวมสูงทำถูกทุกคน แต่ผู้สอบในกลุ่มที่ได้คะแนนรวมต่ำทำผิดทุกคน อำนาจจำแนกมีค่าเท่ากับ 1 ข้อสอบข้อนั้นเป็นข้อสอบที่มีอำนาจจำแนกเป็นเลิศ
- 3) ข้อสอบข้อใดที่ผู้สอบในกลุ่มได้คะแนนรวมสูงทำผิดทุกคน แต่ผู้สอบในกลุ่มที่ได้คะแนนรวมต่ำทำถูกทุกคน อำนาจจำแนกมีค่าเท่ากับ - 1 ข้อสอบข้อนั้นเป็นข้อสอบที่ไม่ดี
- 4) ข้อสอบข้อใดที่ผู้สอบในกลุ่มได้คะแนนรวมสูงตอบถูกเท่าๆ กับผู้สอบในกลุ่มที่ได้คะแนนรวมต่ำ อำนาจจำแนกจะเท่ากับ 0 หรือมีค่าใกล้ศูนย์ ข้อสอบข้อนั้นมีอำนาจจำแนกต่ำ
- 5) โดยทั่วไปการสร้างข้อสอบต้องการข้อสอบที่มีอำนาจจำแนกเป็นบวก ถ้าอำนาจจำแนกที่เป็นบวกยิ่งมีค่ามากก็ยิ่งดี

1.3 ประสิทธิภาพของตัวลวง

ตัวเลือกที่เป็นลวงเป็นส่วนประกอบที่สำคัญของข้อสอบ ตัวลวงที่มีประสิทธิภาพต้องสามารถลวงผู้ตอบที่ไม่มีความสามารถได้ การวิเคราะห์ประสิทธิภาพของตัวลวงทำได้โดยวิเคราะห์ความยากและจำแนกอำนาจของตัวลวงแต่ละตัว

การคำนวณความยากของตัวลวงทำโดยวิธีเดียวกับการคำนวณความยากของตัวถูก โดยตัวลวงที่ดีจะต้องเป็นตัวลวงที่มีผู้เลือกตอบบ้าง ในทางปฏิบัติใช้ตัวลวงที่มีผู้เลือกตอบอย่างน้อยร้อยละ 5 และมีผู้เลือกตอบคำตอบนั้นน้อยกว่าตัวถูก ($.05 \leq p \text{ ตัวลวง} < p \text{ ตัวถูก}$) คือ ค่าความยากของตัวลวงจะมีค่าน้อยกว่าหรือเท่ากับ 0.05 และค่าความยากของตัวลวงจะมีค่าน้อยกว่าความยากของตัวถูก ซึ่งค่าอำนาจจำแนกตัวลวงของตัวลวง คำนวณได้จากสูตร

$$r = \frac{L - H}{N_H} \quad \text{หรือ} \quad r = \frac{L - H}{N_L}$$

$$\text{หรือ} \quad r = P_L - P_H$$

เมื่อ P_L คือสัดส่วนของผู้สอบที่ได้คะแนนรวมต่ำที่ตอบตัวเลือกนั้น

P_H คือสัดส่วนของผู้สอบในกลุ่มที่ได้คะแนนรวมสูงที่ตอบตัวเลือกนั้น

การวิเคราะห์ข้อสอบแบบอิงกลุ่มนอกจากวิธีการคำนวณสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล และวิธีใช้สูตรอย่างง่ายแล้วยังสามารถวิเคราะห์โดยใช้ตารางวิเคราะห์ข้อทดสอบของ จุง เต ฟาน (Chung - The Fan) ตารางนี้จะช่วยประเมินค่าความยากและอำนาจจำแนกของข้อสอบแต่ละข้อ (ทั้งตัวถูกและตัวลวง) ถ้าทราบร้อยละของผู้สอบในกลุ่มที่มีคะแนนรวมสูง และผู้สอบในกลุ่มที่มีคะแนนรวมต่ำที่ตอบข้อสอบแต่ละข้อ การใช้ตารางวิเคราะห์ข้อทดสอบของ จุง เต ฟาน นี้ มีข้อตกลงว่าจะแจกแจงคะแนนของกลุ่มผู้สอบที่นำมาวิเคราะห์ ต้องมีลักษณะเป็นโค้งปกติ และใช้เทคนิคร้อยละ 27 คือนำข้อมูลผลการสอบของกลุ่มผู้สอบที่มีคะแนนรวมสูงร้อยละ 27 และกลุ่มผู้สอบที่มีคะแนนรวมต่ำร้อยละ 27 มาวิเคราะห์ในทางปฏิบัติเทคนิค 27% นี้ ใช้สำหรับการสอบที่มีผู้สอบจำนวนมากเพื่อให้ได้การแจกแจงคะแนนของผู้สอบเป็นไปตามข้อตกลงเบื้องต้น ซึ่งมักถือเกณฑ์ว่าผู้สอบในแต่ละกลุ่มที่นำมาวิเคราะห์ควรมีก่อนละประมาณ 100 คน นั่นคือจำนวนผู้สอบทั้งหมดจะต้องมีตั้งแต่ 370 คนขึ้นไป และในกรณีที่มีผู้สอบจำนวนน้อยก็ไม่นิยมใช้เทคนิค 27% เนื่องจากอาจให้ผลการวิเคราะห์ที่คลาดเคลื่อนได้

โดยทั่วไปแบบทดสอบวัดผลสัมฤทธิ์ที่อาจารย์ผู้สอนสร้างขึ้นมักใช้กับผู้สอบจำนวนไม่มากนัก การวิเคราะห์ข้อสอบวิเคราะห์โดยใช้สูตรอย่างง่าย การพิจารณาตัดสินข้อสอบว่ามีคุณภาพนิยมใช้เกณฑ์ว่าถ้าข้อสอบมีค่าของความยากของตัวถูกระหว่าง 0.20 ถึง 0.80 และมีค่าอำนาจจำแนกของตัวถูกตั้งแต่ .20 ขึ้นไป จะถือว่าเป็นข้อสอบที่มีคุณภาพ ($.20 \leq p \text{ .80}$ และ $r \geq .20$)

2. การวิเคราะห์ข้อสอบทั้งฉบับ การวิเคราะห์ข้อสอบทั้งฉบับเป็นการตรวจสอบคุณภาพของเครื่องมือวัด คุณสมบัติที่สำคัญของเครื่องมือวัดมี 2 ประการ คือ ความตรงหรือความเที่ยงตรง (Validity) และความเที่ยงหรือความเชื่อมั่น (Reliability) มีรายละเอียดดังนี้

2.1 ความตรงหรือความเที่ยงตรง (Validity)

ความตรง หมายถึง ความสามารถของเครื่องมือที่สามารถวัดได้ในสิ่งที่ต้องการวัดได้เป็นความสอดคล้องระหว่างผลการวัดกับสิ่งที่ต้องการวัด ความตรงที่ใช้ในการทดสอบ จำแนกเป็น 3 ชนิด ได้แก่

2.1.1 ความตรงตามเนื้อหา (Content Validity) หมายถึง ความสอดคล้องของเนื้อหา สารของข้อสอบกับเนื้อหาวิชาที่สอน หรือข้อสอบสอดคล้องและครอบคลุมเนื้อหา และวัตถุประสงค์ของวิชาที่สอน การตรวจสอบความตรงตามเนื้อหา ทำได้โดยการวิเคราะห์เนื้อหาของแบบทดสอบทั้งฉบับอาศัยผู้เชี่ยวชาญเป็นผู้พิจารณาว่าเนื้อหาสาระของแบบทดสอบสอดคล้องกับแบบเรียน รายละเอียดของวิชาและหลักสูตรหรือไม่ ซึ่งในทางปฏิบัติจะใช้ตารางวิเคราะห์หลักสูตรเป็นตัวเทียบ

2.1.2 ความตรงตามโครงสร้าง (Construct Validity) หมายถึง ความสามารถของแบบทดสอบ ที่จะวัดคุณลักษณะหรือพฤติกรรมตามโครงสร้างทฤษฎีได้ การตรวจสอบความตรงตามโครงสร้างมีหลายวิธี ได้แก่

1) การเทียบกลุ่มรู้ชุด (Known Group) วิธีการนี้ จะนำแบบทดสอบที่สร้างขึ้น ไปใช้กับกลุ่มที่ทราบคุณลักษณะทางจิตวิทยาตามที่ต้องการวัด โดยใช้ 2 กลุ่ม ที่มีลักษณะตรงข้ามกันแล้วนำมาเทียบกัน ถ้าความแตกต่างของค่าที่วัดได้จากกลุ่มทั้งสอง ด้วยการทดสอบที (t - test) ถ้าความแตกต่างมีนัยสำคัญเชิงสถิติ แสดงว่าแบบทดสอบนั้นมีความตรงตามโครงสร้าง

2) การวิเคราะห์องค์ประกอบ (Factor Analysis) วิธีการนี้ อาศัยวิธีการทางสถิติสำหรับตรวจหาคุณสมบัติทางจิตวิทยา ด้วยการวิเคราะห์ความสัมพันธ์ระหว่างข้อสอบแต่ละข้อว่าข้อสอบทั้งหมดนั้น วัดองค์ประกอบอะไรบ้าง ถ้าองค์ประกอบที่วัดตรงตามทฤษฎีหรือสมมติฐานที่ตั้งไว้ แสดงว่าแบบทดสอบมีความตรงตามโครงสร้าง

3) การหาความสัมพันธ์กับเกณฑ์ที่มีโครงสร้างเหมือนกัน วิธีนี้ทำโดยหาสัมประสิทธิ์สหสัมพันธ์ ระหว่างคะแนนจากแบบทดสอบที่เราสร้างขึ้น กับแบบทดสอบอื่นที่วัดในโครงสร้าง หรือทฤษฎีด้วยกัน ซึ่งพิสูจน์ไว้แล้วว่ามีความตรงตามโครงสร้าง ถ้าแบบทดสอบที่สร้างขึ้นใหม่ มีสหสัมพันธ์กับแบบทดสอบที่เป็นเกณฑ์สูง แสดงว่าแบบทดสอบที่สร้างขึ้น มีความตรงตามโครงสร้าง

2.1.3 ความตรงตามเกณฑ์ (Criterion - Related Validity) เป็นการพิจารณาความสัมพันธ์ระหว่างแบบทดสอบที่สร้างขึ้นกับเกณฑ์ภายนอกบางอย่าง ซึ่งเป็นสภาพความเป็นจริงที่ได้จากการปฏิบัติงาน ความตรงตามเกณฑ์ที่เกี่ยวข้อง แบ่งเป็น 2 ประเภท ดังนี้

1) ความตรงตามสภาพ (Concurrent Validity) เป็นความสามารถของแบบทดสอบที่วัดได้ตรงกับสภาพความเป็นจริงของบุคคลในขณะนั้น เช่น แบบทดสอบวัดผลสัมฤทธิ์วิชาภาษาอังกฤษนำไปให้ผู้เรียนคนหนึ่งสอบ ปรากฏว่าได้คะแนนสูง ซึ่งในสภาพความเป็นความเป็นจริง ผู้เรียนมีความสามารถทางภาษาอังกฤษสูงจริง แสดงว่าแบบทดสอบวัดได้ตรงตามสภาพที่เป็นอยู่

2) ความตรงตามการพยากรณ์ (Predictive Validity) เป็นความสามารถของแบบทดสอบที่วัดผลได้ตรงกับสภาพความเป็นจริงที่เกิดขึ้นในอนาคต เช่น แบบทดสอบคัดเลือกบุคคลเข้าศึกษาต่อ เมื่อไปใช้สอบคัดเลือกบุคคลเข้าศึกษาต่อ ปรากฏว่าผู้ที่สอบคัดเลือกได้คะแนนดี เมื่อเข้าศึกษาต่อมีผลการเรียนอยู่ในเกณฑ์ที่ดี แสดงว่าแบบทดสอบมีความตรงตามการพยากรณ์

2.2 ความเที่ยงหรือความเชื่อมั่น (Reliability)

ความเที่ยง หมายถึง ความคงที่ของคะแนนที่วัดได้แต่ละครั้ง วิธีการหาค่าความเที่ยงของแบบทดสอบทำได้หลายวิธี คือ

2.2.1 วิธีสอบซ้ำ การหาความเชื่อมั่นโดยวิธีสอบซ้ำ เป็นการหาความสัมพันธ์ของคะแนนจากการทำแบบทดสอบฉบับเดียวกันสองครั้ง โดยทิ้งช่วงห่างให้เหมาะสม (ประมาณ 2 สัปดาห์) การหาความเชื่อมั่น โดยวิธีนี้เป็น การตรวจสอบความคงที่ของการแสดงออกของผู้สอบสองครั้งว่าจะมีความคงที่หรือไม่ วิธีการนี้มีจุดอ่อนที่ความแปรเปลี่ยนภายในตัวผู้สอบ ในระหว่างทิ้งช่วงการสอบ ดังนั้น การหาความเชื่อมั่นโดยวิธีนี้ ควรนำไปใช้กับแบบทดสอบวัดคุณลักษณะที่ค่อนข้างจะคงที่ไม่แปรเปลี่ยนโดยง่าย

2.2.2 วิธีแบบทดสอบคู่ขนาน การหาความเชื่อมั่นโดยใช้วิธีแบบทดสอบคู่ขนาน เป็นการหาความสัมพันธ์ของคะแนน จากการนำแบบทดสอบ 2 ฉบับที่เทียบเท่ากัน ไปสอบกับกลุ่มบุคคลเดียวกัน วิธีการนี้มีจุดอ่อนที่ความเป็นคู่ขนานกันของแบบทดสอบ 2 ฉบับ ซึ่งสร้างได้ยาก

2.2.3 วิธีหาความสอดคล้องภายใน แบ่งเป็น 2 วิธี คือ

1) วิธีแบ่งครึ่งแบบทดสอบ การหาความเที่ยงโดยวิธีนี้ เป็นการหาความสัมพันธ์ของคะแนน จากการใช้แบบทดสอบฉบับเดียวและสอบเพียงครั้งเดียว โดยนำผลการสอบมาแบ่งเป็นข้อมูล 2 ชุด โดยอาจแบ่งเป็น ข้อคู่-ข้อคี่ แบ่งเป็นครึ่งฉบับแรก ครึ่งฉบับหลัง จากการหาค่าสัมประสิทธิ์สหสัมพันธ์ จะได้สัมประสิทธิ์ความเชื่อมั่นของแบบทดสอบครึ่งฉบับ แล้วจึงนำไปปรับขยายเป็นสัมประสิทธิ์สหสัมพันธ์ของแบบทดสอบทั้งฉบับ การคำนวณโดยใช้สูตร

$$r_{tt} = \frac{2r_{xy}}{1+r_{xy}}$$

เมื่อ r_{tt} คือค่าความเที่ยงของเครื่องมือทั้งฉบับ

r_{xy} คือค่าความเที่ยงของเครื่องมือครึ่งฉบับ

ถ้าค่าความเที่ยงของเครื่องมือทั้งฉบับมีค่าสูง (มีค่าใกล้ 1.0) แสดงว่าเครื่องมือทั้ง 2 ส่วน มีความเป็นเอกพันธ์ หรือแสดงว่าเครื่องมือมีความสอดคล้องภายใน

การแบ่งเครื่องมือออกเป็น 2 ส่วน ต้องแบ่งอย่างมีเหตุผล กล่าวคือ ถ้าจัดข้อคำถามที่มีลักษณะเท่าเทียมกันออกเป็น 2 ส่วน ในลักษณะ ข้อคู่-ข้อคี่ การแบ่งผลการวัดเป็น 2 ส่วน เพื่อคำนวณหาค่าความเที่ยงของเครื่องมือครึ่งฉบับ ก็จะได้แบ่งในลักษณะผลการวัดจากข้อคู่ส่วนหนึ่ง และผลจากการวัดจากข้อคี่ส่วนหนึ่ง แต่ถ้าจัดข้อคำถามเป็น 2 ส่วน ในลักษณะครึ่งแรกและครึ่งหลัง การนำผลการวัดไปวิเคราะห์ก็ต้องแบ่งในลักษณะครึ่งแรกและครึ่งหลังให้สอดคล้องกัน การคำนวณค่าความเที่ยงโดยวิธีแบ่งครึ่งนี้จะมีลักษณะผสมผสานของวิธีการสอบซ้ำ และวิธีใช้เครื่องมือที่เป็นคู่ขนาน ซึ่งแก้ปัญหาการเก็บข้อมูล 2 ครั้ง มาเป็นการเก็บข้อมูลจากกลุ่มตัวอย่างเพียงครั้งเดียว และนำผลการวัดมาแบ่งเป็น 2 ชุด เพื่อคำนวณค่าความเที่ยง ดังนั้น ค่าความเที่ยงที่คำนวณได้ในตอนแรก (r_{xy}) จึง

เป็นค่าความเที่ยงของเครื่องมือเพียงครั้งฉบับต้องนำมาหาค่าความเที่ยงของเครื่องมือทั้งฉบับ (r_{tt}) อีกครั้งหนึ่ง

2) วิธีหาจากสูตรคูเดอร์และริชาร์ดสัน การหาความเที่ยงโดยวิธีนี้ เป็นการหาความสัมพันธ์ของคะแนนจากการใช้แบบทดสอบฉบับเดียวและการสอบเพียงครั้งเดียว โดยการนำผลการสอบมาคำนวณค่าสัมประสิทธิ์ โดยใช้สูตรของคูเดอร์และริชาร์ดสัน สูตรที่ใช้ มี 2 สูตร คือ สูตร KR - 20 กับสูตร KR - 21

สูตร KR - 20

$$KR20 = \frac{k}{k-1} \left[1 - \frac{\sum pq}{S_x^2} \right]$$

เมื่อ	k	คือจำนวนข้อของแบบทดสอบ
	S_x^2	คือความแปรปรวนของคะแนนรวม
	P	คือสัดส่วนของผู้ตอบข้อสอบถูกในแต่ละข้อ
	q	คือสัดส่วนของผู้ตอบข้อสอบผิดในแต่ละข้อ (1-p)

สูตร KR - 21

$$KR-21 = \frac{k}{k-1} \left(1 - \frac{\bar{x}(k-\bar{x})}{k\sigma_r^2} \right)$$

เมื่อ	k	คือจำนวนของแบบทดสอบ
	\bar{X}	คือค่าเฉลี่ยของคะแนนรวม
	S_x^2	คือความแปรปรวนของคะแนนรวม

สูตร KR - 20 และ KR - 21 นี้ใช้ได้เฉพาะการหาความเที่ยงของแบบทดสอบที่ให้คะแนนแต่ละข้อเป็นแบบ 0 กับ 1 เท่านั้น สูตร KR - 21 ใช้กรณีข้อสอบทุกข้อมีความยากเท่ากัน ซึ่งในทางปฏิบัติต้องพิจารณาเงื่อนไขที่เป็นจริงด้วย

ความเชื่อมั่นของข้อสอบแสดงถึงความคงเส้นคงวาของคะแนนที่ได้จากการใช้แบบทดสอบชุดนั้น ซึ่งจะมีผลมาจาก 3 องค์ประกอบ คือ

1. ความสัมพันธ์ระหว่างข้อสอบแต่ละข้อ หากข้อสอบแต่ละข้อมีความสัมพันธ์กันในเชิงบวก ค่าสัมประสิทธิ์ความเชื่อมั่นก็จะสูง

2. ความยาวหรือจำนวนข้อของข้อสอบ ข้อสอบที่มีจำนวนข้อสอบมากกว่าจะมีค่าความเชื่อมั่นสูงกว่าข้อสอบที่จำนวนข้อน้อยกว่า ในกรณีที่คุณสมบัติด้านอื่นๆ ทัดเทียมกัน

3. เนื้อหาของแบบทดสอบ ข้อสอบที่วัดเนื้อหาแตกต่างกันมาก จะมีค่าความเชื่อมั่นต่ำ ค่าสัมประสิทธิ์ความเชื่อมั่นมีค่าระหว่าง 0 - 1 สามารถคำนวณโดยใช้ค่าสัมประสิทธิ์ครอนบาค (Cronbach's Alpha) ใช้กับแบบทดสอบแบบเลือกตอบ ที่มีข้อที่ถูกเพียงข้อเดียว (ตอบถูกได้คะแนน 1 ตอบผิดได้คะแนน 0) และจะได้ค่าเท่ากับการใช้สูตร KR - 20

การแปลความหมายของค่าสัมประสิทธิ์ความเชื่อมั่น มีแนวทางดังนี้ (Office of Educational Assessment, University of Washington, 2005)

ค่าสัมประสิทธิ์ความเชื่อมั่น	การแปลความหมาย
$\geq .90$	ยอดเยี่ยม เทียบได้กับข้อสอบมาตรฐาน
.80 - .90	ดีมาก สำหรับใช้เป็นแบบทดสอบรายวิชาในชั้นเรียน
.70 - .80	ดี สำหรับใช้เป็นแบบทดสอบรายวิชาในชั้นเรียน อาจมีข้อสอบบางข้อที่จำเป็นต้องปรับปรุง
.60 - .70	ค่อนข้างต่ำ แบบทดสอบชุดนี้ ควรมีการวัดด้วยเครื่องมืออื่นเสริม ก่อนจะตัดสินใจตัดสินเกรด มีข้อสอบบางข้อที่ต้องปรับปรุงคุณภาพ
.50 - .60	ควรปรับปรุง จำเป็นต้องมีการวัดด้วยเครื่องมืออื่นเสริม ก่อนจะตัดสินใจตัดสินเกรด
$\leq .50$	ไม่น่าเชื่อถือ จำเป็นต้องปรับปรุง และไม่ควรถูกใช้เป็นหลักในการตัดสินใจตัดสินเกรด

3) วิธีหาจากสูตรประสิทธิแอลฟา การหาค่าความเที่ยงโดยใช้สูตรของครอนบาค (Cronbach) นี้ปรับมาจากสูตร $KR - 20$ ใช้หาความเที่ยงของเครื่องมือวัด ที่ให้คะแนนแตกต่างกันไปในแต่ละข้อ โดยไม่จำเป็นต้องเป็นระบบ การให้คะแนน แบบ 1 กับ 0 สูตรการคำนวณเป็นดังนี้

$$\alpha = \frac{k}{k-1} \frac{(1 - \sum S_i^2)}{S_x^2}$$

เมื่อ	α	คือค่าความเที่ยง
	K	คือจำนวนข้อของแบบวัด
	S_i^2	คือผลรวมความแปรปรวนของคะแนนรายข้อ
	S_x^2	คือความแปรปรวนของคะแนนรวม

การหาค่าความเที่ยงโดยใช้สูตรสัมประสิทธิ์แอลฟา สามารถหาได้โดยใช้ผลการสอบจากแบบทดสอบฉบับเดียว นำไปสอบกับบุคคลกลุ่มเดียว และนำไปใช้กันได้อย่างกว้างขวาง โดยไม่จำกัดเฉพาะแบบทดสอบที่ให้คะแนนแบบ 1 กับ 0

จุดมุ่งหมายของการวิเคราะห์ข้อสอบมีสองประการ คือ ประการแรกเพื่อคัดเลือกและปรับปรุงข้อสอบที่มีคุณภาพไว้ในโอกาสต่อไป ข้อสอบที่จะนำมาวิเคราะห์ต้องมีการตรวจให้คะแนนแบบตอบถูกให้ 1 ตอบผิดให้ 0 ถ้าเป็นข้อสอบแบบเลือกตอบจะเหมาะสมที่สุด เพราะนอกจากจะวิเคราะห์ข้อสอบรายข้อแล้วยังวิเคราะห์ถึงรายตัวเลือกได้ด้วย และนำผลการวิเคราะห์ไปปรับปรุงตัวคำถาม และตัวเลือกให้ดีขึ้นมีประสิทธิภาพ วิธีการวิเคราะห์ข้อสอบนี้วิเคราะห์ได้ทั้งข้อสอบวัดผลสัมฤทธิ์ทางการเรียน ข้อสอบสอบคัดเลือกเข้าศึกษาต่อและข้อสอบวัดความถนัดทางการเรียน ประการที่สองของการวิเคราะห์ข้อสอบ คือ เพื่อทราบคุณภาพของการเรียนการสอน อันจะเป็นแนวทางในการปรับปรุงการเรียนการสอนต่อไป บางครั้งผู้สอนจะพบว่าทำไมผู้เรียนส่วนใหญ่ทำข้อสอบไม่ได้ทั้งที่

ตั้งใจสอนอย่างเต็มที่หรือข้อสอบบางข้อที่ผู้เรียนอ่อนตอบข้อสอบถูก แต่ผู้เรียนเก่งกลับตอบข้อสอบผิดหรือคะแนนสอบของผู้เรียนกระจุกไม่กระจายตามที่ควรจะเป็นเหล่านี้เป็นต้น น่าจะมีการวิเคราะห์หาสาเหตุเพื่อหาวิธีการปรับปรุงแก้ไขหรือจะหาวิธีสอนอย่างไรจึงจะบรรลุผลอย่างไรก็ตาม การวิเคราะห์ข้อสอบ ยังถือได้ว่าเป็นการให้ความยุติธรรมแก่ผู้เรียนอีกด้วยถ้าข้อสอบดี มีคุณภาพ มีความเที่ยงตรง (Validity) มีความเชื่อมั่น (Reliability) ก็จะทำให้การประเมินผลการเรียนการสอนถูกต้องตามความเป็นจริง ทำให้มั่นใจว่าระดับคะแนนหรือเกรดที่ให้ไม่คลาดเคลื่อน

งานวิจัยที่เกี่ยวข้อง

นันทพร หาญวิทย์สกุล (2546) ได้จัดทำรายงานการวิเคราะห์ข้อสอบวิชาหลักการประชาสัมพันธ์ โดยมีวัตถุประสงค์เพื่อจัดทำข้อสอบมาตรฐานรายวิชาหลักการประชาสัมพันธ์ รหัส 2201 - 2701 หลักสูตรประกาศนียบัตรวิชาชีพเพื่อให้ได้ข้อสอบที่มีคุณภาพสำหรับนำไปจัดทำเป็นคลังข้อสอบ และนำผลการวิเคราะห์ข้อสอบรายข้อแจ้งผู้ออกข้อสอบทราบเพื่อปรับปรุงแก้ไขข้อสอบเดิม และปรับปรุงการออกข้อสอบในครั้งต่อไปให้ได้มาตรฐานมากขึ้น ผู้วิจัยได้นำแบบทดสอบรายวิชาหลักการประชาสัมพันธ์ จำนวน 4 ชุด รวม 195 ข้อทำการ ทดสอบกับนักเรียนห้อง ชกข.2/1 แผนกวิชาการตลาด วิทยาลัยบริหารธุรกิจและการท่องเที่ยวกรุงเทพในภาคเรียนที่ 1/2552 แล้วพบว่าข้อสอบที่ดีควรเก็บไว้มีจำนวน 87 ข้อ คิดเป็นร้อยละ 44.61 ข้อสอบที่ควรปรับปรุงมีจำนวน 49 ข้อ คิดเป็นร้อยละ 25.12 ส่วนข้อสอบที่ควรตัดทิ้งมีจำนวน 51 ข้อ คิดเป็นร้อยละ 26.15 เป็นไปตามวัตถุประสงค์ของผู้ทำการศึกษาที่ต้องการวัดคุณภาพของข้อสอบให้เป็นไปตามมาตรฐาน และนำผลการวิเคราะห์ไปปรับปรุงคุณภาพของข้อสอบให้ดียิ่งขึ้น

รังสรรค์ ไกรศรานนท์ อธิบายเรื่องการวิเคราะห์ข้อสอบด้วยเทคนิค 25% ว่าการวิเคราะห์ข้อสอบมีจุดมุ่งหมายเพื่อคัดเลือกและปรับปรุงข้อสอบที่มีคุณภาพไว้ในโอกาสต่อไป ข้อสอบที่จะนำมาวิเคราะห์ต้องมีการตรวจให้คะแนนแบบตอบถูกให้ 1 ตอบผิดให้ 0 ถ้าเป็นข้อสอบแบบเลือกตอบจะเหมาะสมที่สุด เพราะนอกจากจะวิเคราะห์ข้อสอบรายข้อแล้วยังวิเคราะห์ถึงรายตัวเลือกได้ด้วย และนำผลการวิเคราะห์ไปปรับปรุงตัวคำถาม และตัวเลือกให้ดีมีประสิทธิภาพ วิธีการวิเคราะห์ข้อสอบวิเคราะห์ได้ทั้งข้อสอบวัดผลสัมฤทธิ์ทางการเรียน ข้อสอบสอบคัดเลือกเข้าศึกษาต่อและข้อสอบวัดความถนัดทางการเรียน นอกจากนั้นการวิเคราะห์ข้อสอบก็เพื่อทราบคุณภาพของการเรียนการสอนเป็นแนวทางในการปรับปรุงการเรียนการสอน ดังนั้น การวิเคราะห์ข้อสอบยังถือได้ว่าเป็นการให้ความยุติธรรมแก่ผู้เรียน ถ้าข้อสอบดี มีคุณภาพ มีความเที่ยงตรง (Validity) มีความเชื่อมั่น (Reliability) ก็จะทำให้การประเมินผลการเรียนการสอนถูกต้องตามความเป็นจริง ทำให้มั่นใจว่าระดับคะแนนหรือเกรดที่ให้ไม่คลาดเคลื่อน

เชิดศักดิ์ ไธรมณีนรัตน์ (2553) อธิบายว่าการวิเคราะห์ข้อสอบปรนัย (Item analysis) เป็นการใช้วิธีการทางสถิติเพื่อวิเคราะห์คำตอบที่ผู้สอบตอบข้อสอบปรนัยในการสอบครั้งหนึ่งเพื่อประเมินว่าข้อสอบที่นำมาใช้ในการสอบครั้งนั้นมีคุณสมบัติอย่างไร ทำงานได้ตามที่ต้องการหรือไม่ มีระดับความยากง่ายของข้อสอบเหมาะสมหรือไม่ มีข้อบกพร่องหรือไม่ และควรได้รับการปรับปรุงแก้ไขอย่างไร การวิเคราะห์ข้อสอบเป็นศาสตร์ที่ได้รับการพัฒนาอย่างต่อเนื่องมาเป็นเวลานาน มีเทคนิค

และวิธีการต่างๆ มากมายที่ผู้วิเคราะห์สามารถใช้เพื่อบอกคุณสมบัติของข้อสอบแต่ละข้อ ตั้งแต่วิธีการง่ายๆ ไปจนถึงวิธีการที่มีความซับซ้อนมาก โดยแต่ละเทคนิคการวิเคราะห์ก็มีจุดประสงค์แตกต่างกันไป ตั้งแต่การบอกระดับความยากง่าย การบอกถึงความสามารถในการแยกผู้สอบที่เก่งออกจากผู้สอบที่ไม่เก่ง ไปจนถึงเทคนิคขั้นสูงที่สามารถบอกได้ว่าข้อสอบมีความลำเอียงต่อผู้สอบเพศใดเพศหนึ่ง หรือผู้สอบจากสถาบันใดสถาบันหนึ่งเป็นพิเศษหรือไม่ มีการเดาข้อสอบมากน้อยเพียงใด ผู้สอบรู้ข้อสอบมาก่อนเข้าสอบหรือไม่หรือมีความน่าจะเป็นมากน้อยเพียงใด ที่ผู้สอบลอกคำตอบ ในบทความนี้ผู้นิพนธ์ไม่ได้ตั้งเป้าประสงค์ที่จะรวบรวมและอภิปรายเทคนิคการวิเคราะห์ข้อสอบทุกวิธีที่มีใช้อยู่ในปัจจุบัน แต่ต้องการเพียงนำเสนอความรู้พื้นฐานที่เกี่ยวกับการวิเคราะห์ข้อสอบและอธิบายถึงวิธีการวิเคราะห์ข้อสอบที่นิยมใช้กันในทางแพทยศาสตรศึกษา โดยเฉพาะในประเทศไทย โดยประสงค์ให้อาจารย์ผู้อ่านสามารถนำเอาความรู้ที่ได้จากบทความนี้ไปใช้แปลผลการวิเคราะห์ข้อสอบที่ตนเองเกี่ยวข้อง และดำเนินการปรับปรุงคุณภาพของข้อสอบได้อย่างเหมาะสม

Haladyna Downing และ Ro driguez (2002) การวิจัยครั้งนี้พบว่าข้อสอบ multiple choice ที่มีตัวลวงคล้ายคลึงกันทั้งชุดกับคำตอบที่ถูกต้อง จะทำให้ข้อสอบข้อนั้นเป็นข้อสอบที่ยากมากขึ้น และ จากการวิจัยของ Smith and Smith (1988) ที่พบว่า การใช้ตัวลวงที่มีคำตอบที่ถูกต้องแตกต่างกันอย่างสิ้นเชิง หรือตัวลวงที่มีความคล้ายคลึงกันกับคำตอบที่ถูกต้อง นั้นความยากง่ายของข้อสอบจะขึ้นอยู่กับภาระเขียนโจทย์คำถามมากกว่า โดยบอกว่า โจทย์คำถามที่มีลักษณะเป็นคำถามจะง่ายกว่า โจทย์ที่มีลักษณะข้อคำถามปลายเปิดโดยเว้นไว้เติมคำตอบ การวิจัยครั้งนี้ถูกออกแบบให้มี 5 ขั้นตอน โดยการสร้างและวิเคราะห์ข้อสอบในการสอบขึ้นทะเบียนใบขับขี่ และนำข้อสอบแบบต่างๆ ไปใช้กับนักเรียนโรงเรียนต่างๆ ลักษณะข้อสอบมีหลายแบบ แต่มีลักษณะข้อสอบคู่ขนานกัน แต่ละขั้นตอนมีผู้เชี่ยวชาญวิเคราะห์และตรวจสอบข้อสอบแต่ละชุดอย่างละเอียด การวิจัยครั้งนี้มีการตั้งสมมุติฐาน 2 ข้อคือ 1. ลักษณะของตัวลวงที่มีความคล้ายคลึงกันทั้งชุดกับคำตอบที่ถูกต้องจะทำให้ข้อสอบข้อนั้นยากมากกว่าตัวลวงที่มีคำตอบที่ถูกต้องแตกต่างกันอย่างสิ้นเชิง และ 2. ลักษณะโจทย์คำถามที่มีลักษณะคำถามง่ายกว่า ลักษณะโจทย์คำถามที่มีลักษณะปลายเปิด เว้นไว้ให้ตอบ ผลการวิจัยครั้งนี้ พบว่า ตัวลวงของข้อคำถามที่มีความสั้นยาวของตัวอักษรแต่ละข้อตัวลวง ใกล้เคียงกัน เนื้อหาของคำตอบคล้ายคลึงกัน กับข้อคำถามที่ถูกต้อง จะทำให้ข้อสอบข้อนั้นเป็นข้อสอบที่ยากมากกว่าข้อสอบที่มีตัวลวงที่มีเนื้อหาของคำตอบแตกต่างกันกับคำตอบที่ถูกต้อง อย่างมีนัยสำคัญทางสถิติ ที่ระดับ 0.01 ($p < 0.01$) เพราะตัวลวงที่มีเนื้อหาแตกต่างกัน จะง่ายต่อการคาดเดา เป็นไปตามสมมุติฐานที่ตั้งไว้ ข้อ 1 ส่วนสมมุติฐานที่ 2 ผลการวิจัย ไม่พบว่า ลักษณะการตั้งโจทย์คำถามที่เป็นข้อคำถามหรือโจทย์คำถามที่เป็นข้อความปลายเปิด ไม่พบความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ในเรื่องความยากง่ายของข้อสอบ

Caldwell, D. J. & Pate, A. M. (1990). การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบการใช้รูปแบบคำถามตามแนวปฏิบัติที่เป็นมาตรฐานและไม่มาตรฐาน ข้อสอบที่มีแนวปฏิบัติที่เป็นมาตรฐาน ภายในข้อคำถามจะมีหลักการในการสร้างคำถาม ดังนี้ 1. การใช้คำถามเชิงบวก หลีกเลี่ยงคำถามเชิงลบ เช่น ยกเว้น ไม่ใช่ 2. ในแต่ละคำถามควรใช้ตัวเลือกเพียง 3 ตัวเลือก 3. การใช้ตัวเลือก “ไม่มีข้อใดถูกต้อง” ไม่แนะนำให้ใช้บ่อย รูปแบบการวิจัยเป็นข้อสอบปรนัย 15 ข้อ คำถาม ทดสอบกับนักศึกษาเภสัชศาสตร์ โดยนักศึกษาจำนวน 55 คน ตอบแบบสอบถามที่สร้างตามแนวปฏิบัติที่เป็น

มาตรฐาน เรียกว่า standard scale และนักศึกษา 54 คน ตอบแบบสอบถามที่ไม่ได้สร้างจากแนวปฏิบัติ เรียกว่า nonstandard scale สัดส่วนของคำถามเป็นการวัดด้านความรู้-ความจำ ร้อยละ 60 เข้าใจ ร้อยละ 20 และ นำไปใช้ ร้อยละ 20 ผลการวิจัย พบว่า ข้อสอบที่ถูกสร้างตามแนวปฏิบัติที่เป็นมาตรฐาน นักศึกษาจะได้คะแนนสูงกว่ากลุ่มที่ทำข้อสอบที่ไม่ได้มาตรฐาน โดยค่าคะแนนของผู้สอบข้อสอบ nonstandard scale มีความยากกว่า ข้อสอบ standard scale ถึง 12.7 คะแนน แนวปฏิบัติที่แนะนำให้หลีกเลี่ยงการใช้ตัวลวง ไม่มีข้อใดถูกต้อง แสดงให้เห็นความแตกต่างอย่างมีนัยสำคัญทางสถิติ ของคะแนนเฉลี่ย ระหว่างกลุ่ม ที่ทำ ข้อสอบ standard scale และกลุ่ม ที่ทำข้อสอบ nonstandard scale ร้อยละ 53.6 และ 41.3 ตามลำดับ ($p < 0.001$) จากผลการวิจัยสรุปได้ว่า ข้อสอบที่ไม่ได้มาตรฐานจะมีความยาก ทำให้นักศึกษาได้คะแนนน้อยกว่าข้อสอบที่เป็นมาตรฐาน แต่ไม่สามารถแยกแยะกลุ่ม ถูก/ผิด (เก่ง/อ่อน) ได้ และนักศึกษาที่เรียนอ่อนมักจะทำผิด ดังนั้นการใช้รูปแบบการสร้างข้อคำถามและตัวลวงควรมีการทดสอบเชิงโครงสร้างของคำถาม และความสอดคล้องระหว่างข้อถูกและผิด

บทความวิจัยเรื่อง การตรวจสอบการเคลื่อนของค่าความยากง่ายและอำนาจจำแนกของข้อสอบที่ใช้ซ้ำหลายครั้ง เมื่อข้อสอบถูกใช้เป็นเวลาหลายปี ค่าความยากง่ายและอำนาจจำแนกอาจเปลี่ยนแปลงไป การเปลี่ยนแปลงอาจเนื่องจากการเปลี่ยนด้านวิธีการสอน ข้อสอบถูกเปิดเผยโดยผู้สอบครั้งก่อนๆ หรือการเปลี่ยนแปลงจุดเน้นของหัวข้อและเนื้อหาวิชาเมื่อมีการพัฒนาหลักสูตร โดยเฉพาะเนื้อหาทางด้านเทคโนโลยี ปัจจัยเหล่านี้เป็นเหตุทำให้เกิดการเปลี่ยนแปลงของค่าความยากง่าย (P) และค่าอำนาจจำแนกของข้อสอบ (R) ซึ่งเรียกว่า การเคลื่อนของค่า P และค่า R (Item parameter drift) ผู้วิจัยศึกษาเปรียบเทียบเครื่องมือวัดการเคลื่อนที่ของค่าความยากง่ายและอำนาจจำแนกของข้อสอบที่ใช้หลายครั้ง 3 วิธี คือ 1. BILOG - MG (Zimowski, Muraki, Mislavy, & Bock, 2003) ที่ใช้สำหรับตรวจสอบการเคลื่อนของค่าความยากง่าย 2. CUSUM (Veerkamp & Glas, 2000) ที่ใช้ตรวจสอบการเคลื่อนของค่าความยากง่ายและอำนาจจำแนก และ 3. Kim, Cohen, & Park's χ^2 test (1995) ที่ใช้ตรวจสอบการเคลื่อนของค่าความยากง่ายและอำนาจจำแนกของข้อสอบที่ใช้กลุ่มผู้สอบต่างกัน การเก็บรวบรวมข้อมูลใช้วิธีการจัดสอบขึ้นใหม่และใช้ชุดคะแนนจากการสอบจริง โดยใช้คะแนนจากชุดข้อสอบวิชาประวัติศาสตร์และการเมืองสหรัฐอเมริกา จำนวน 100 ข้อ (10 ข้อเป็นข้อสอบที่มีข้อมูลการเคลื่อนของ P, R 90 ข้อเป็นข้อสอบที่ไม่เคลื่อน) แล้วศึกษาเปรียบเทียบการเคลื่อนของค่า P, R ของการสอบครั้งที่ 3, 4, 5 โดยใช้วิธีทั้ง 3 การวิจัยพบว่า วิธี BILOG - MG และ KCP's χ^2 test เป็นวิธีการตรวจสอบที่ใช้ง่ายและเที่ยงตรงได้ผลดีกว่า CUSUM โดยจะพบการเคลื่อนของค่า P, R เสมอ ในข้อสอบที่มีประวัติการเคลื่อน ทั้ง 10 ข้อ (ตั้งแต่การวัดครั้งที่ 3, 4, 5) และวิธีการทั้ง 3 ยังไม่สามารถตรวจสอบการเคลื่อนของค่า P, R ในข้อสอบ 90 ข้อที่ไม่มีประวัติการเคลื่อนได้

Tasdemir, M. (2010) การศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบความยากง่าย อำนาจการจำแนกในการประเมินผลสัมฤทธิ์ทางการเรียนระหว่างข้อสอบแบบหลายตัวเลือกกับแบบถูกผิด รูปแบบการวิจัยเป็นแบบเชิงพรรณนา กลุ่มตัวอย่าง คือ นักศึกษาระดับปริญญาตรีชั้นปีที่สี่จำนวน 252 คน ในปีการศึกษา 2007 - 2008 เครื่องมือที่ใช้เก็บข้อมูลเป็นข้อสอบปลายภาครวม 100 ข้อ โดยแบ่งเป็นข้อสอบแบบหลายตัวเลือก 50 ข้อ และแบบถูกผิดจำนวนเท่ากันคือ 50 ข้อ ซึ่งมีโครงสร้างเนื้อหาแบบ

คู่ขนานกัน และผ่านการตรวจสอบจากผู้ทรงคุณวุฒิในด้านความตรงตามเนื้อหา หาค่าความเที่ยงจาก ความสอดคล้องภายใน โดยใช้สูตร Kuder - Richardson (KR 20) ได้ .779 และ Spearman -Brown ได้ .782 ซึ่งถือว่าเป็นแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีค่าความน่าเชื่อถือสูง วิเคราะห์ข้อมูลด้วย โปรแกรมสำเร็จรูป SPSS สถิติที่ใช้คือ Paired - t test เพื่อเปรียบเทียบความแตกต่างระหว่างข้อสอบ ทั้งสองประเภท โดยถือว่าข้อสอบที่ดีควรมีค่าอำนาจการจำแนกมากกว่า 0.19 ขึ้นไป ส่วนข้อสอบที่มีความยากง่ายพอเหมาะมีค่าความยากง่ายประมาณ 0.5 พบว่า คะแนนเฉลี่ยที่ได้จากการทดสอบด้วย ข้อสอบแบบหลายตัวเลือกและแบบถูกผิดในเพศชายและหญิงไม่มีความแตกต่างกันอย่างมีนัยสำคัญทาง สถิติ คะแนนเฉลี่ยที่ได้จากการทดสอบแบบหลายตัวเลือกและแบบถูกผิด 28 หัวข้อที่มีโครงสร้างเนื้อหา แบบคู่ขนานกัน มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติกล่าวคือมีข้อสอบ 13 ข้อที่ผู้เข้าสอบทำ คะแนนจากแบบทดสอบแบบหลายตัวเลือกได้สูงกว่าแบบถูกผิด ส่วนอีก 15 ข้อนั้นผู้เข้าสอบทำคะแนน จากแบบทดสอบแบบถูกผิดได้สูงกว่า ในขณะที่อีก 22 ข้อไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ความยากง่ายของแบบทดสอบแบบหลายตัวเลือกและแบบถูกผิดไม่มีความแตกต่างกันอย่างมีนัยสำคัญทาง สถิติ (ข้อสอบที่มีความยากคือ $1 - P \geq 0.5$) โดยพบว่าข้อสอบแบบหลายตัวเลือกมีข้อสอบที่ค่อนข้างยาก 10 ข้อ ส่วนอีก 40 ข้อค่อนข้างง่าย ส่วนข้อสอบแบบถูกผิดมี 5 ข้อที่ค่อนข้างยาก 45 ข้อค่อนข้างง่าย อำนาจการจำแนกของแบบทดสอบแบบหลายตัวเลือกและแบบถูกผิด ไม่มีความแตกต่างกันอย่างมี นัยสำคัญทางสถิติ โดยข้อสอบแบบหลายตัวเลือกมีค่าอำนาจการจำแนกต่ำ มี 10 ข้อที่ มีค่าอำนาจการ จำแนกอยู่ในเกณฑ์ปกติหรือใช้ได้ 40 ข้อ ส่วนข้อสอบแบบถูกผิดมีค่าอำนาจการจำแนกต่ำ 14 ข้อ ที่มี ค่าอำนาจการจำแนกอยู่ในเกณฑ์ปกติใช้ได้ 36 ข้อ