บทที่ 2
แนวคิด ทฤษฎี เอกสารและงานวิจัยที่เกี่ยวข้อง

วรรณกรรมที่เกี่ยวข้องกับการวิจัยเรื่อง ผลของการตัดข้อสอบที่มีอำนาจจำแนกติดลบออกต่อคะแนนการสอบและความเชื่อมั่นของข้อสอบทั้งฉบับ: รายวิชาวิทยาการระบาด ประกอบด้วย

1. ทฤษฎีการวิเคราะห์ข้อสอบ
2. รายวิชาวิทยาการระบาด
3. บทความและงานวิจัยที่เกี่ยวข้อง

ดังรายละเอียดต่อไปนี้

ทฤษฎีการวิเคราะห์ข้อสอบ

เชิดศักดิ์ ไอรมณีรัตน์ (2552) กล่าวถึงแนวคิดสำคัญในการวิเคราะห์ข้อสอบ โดยระบุว่าประกอบด้วย 2 ส่วนคือ การวิเคราะห์ข้อสอบรายข้อ (Item analysis) และการวิเคราะห์ข้อสอบโดยรวม (Test analysis) โดยมีรายละเอียดพอสังเขปดังนี้

1. การวิเคราะห์ข้อสอบรายข้อ (Item analysis) การวิเคราะห์ข้อสอบแต่ละข้อพิจารณา 3 ปัจจัย คือ

1.1 ความยากง่ายของข้อสอบ (Item difficulty, p) ความยากง่ายของข้อสอบวัดโดยใช้ค่า p ซึ่งย่อมาจาก Proportion of examinees answering item correctly (สัดส่วนของผู้ที่ตอบข้อสอบนั้นถูก) ซึ่งหาได้จากการนำจำนวนผู้สอบที่ตอบข้อนั้นถูก ต้องหารด้วยจำนวนผู้สอบที่ตอบข้อนั้นทั้งหมด หากข้อสอบข้อนั้นเป็นข้อสอบที่ง่ายผู้สอบทุกคนตอบถูก ค่า p ก็จะเป็น 1 หากไม่มีผู้สอบคนใดตอบถูกเลย ข้อสอบข้อนั้นก็จะมีค่า p เป็น 0 หากมีคนตอบถูก 70% ข้อสอบข้อนั้นก็มีค่า p เท่ากับ 0.7

สุมาลี จันทร์ชลอ (2542: 136) ได้ให้ข้อเสนอแนะในการเลือกข้อสอบจากการพิจารณาค่า P ดังตารางที่ 2.1

ตารางที่ 2.1 ข้อเสนอแนะในการเลือกข้อสอบจากค่าความยากง่าย

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

สูตรในการคำนวณค่าความยากง่าย (พวงรัตน์ ทวีรัตน์ 2540: 129)

$$P = R / N$$

เมื่อ P = ค่าความยากง่ายของคำถามข้อนั้น
R = จำนวนผู้ตอบถูกในข้อนั้น
N = จำนวนผู้ตอบทั้งหมด

1.2 ความสามารถในการจำแนกผู้สอบ หมายถึง ความสามารถของข้อสอบข้อหนึ่งๆ ในการแยกผู้สอบที่ทำคะแนนได้ดี ออกจากผู้สอบที่ทำคะแนนได้ไม่ดี ข้อสอบที่มีความสามารถในการแยกแยะได้ดีนั้น ผู้สอบที่ตอบข้อสอบข้อนั้นผิดมักจะได้คะแนนต่ำ ดัชนีที่ใช้วัดความสามารถในการจำแนกผู้สอบที่ใช้กันมากที่สุดในปัจจุบันคือค่า point-biserial correlation ซึ่งนิยมใช้อักษรย่อเป็น r ซึ่งสามารถคำนวณได้จากสูตรต่อไปนี้ (เชิดศักดิ์ ไอรมณีรัตน์, 2552)

$$r = \frac{M_p - M_q}{SD}\sqrt{pq}$$

เมื่อ M_p = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบถูก
M_q = คะแนนสอบเฉลี่ยของผู้สอบที่ตอบข้อสอบผิด

SD = ค่าเบี่ยงเบนมาตรฐานของคะแนนสอบ

P = สัดส่วนของผู้สอบที่ตอบข้อสอบถูกต่อผู้สอบทั้งหมด

q = สัดส่วนของผู้สอบที่ตอบข้อสอบผิดต่อผู้สอบทั้งหมด

ค่า point-biserial correlation ที่คำนวณได้มีค่าอยู่ในช่วง -1 ถึง 1 โดยค่าที่ติดลบหมายถึงข้อสอบข้อนั้นผู้ที่ตอบถูกมักสอบได้คะแนนรวมต่ำ แต่ผู้ที่ตอบผิดมักสอบได้คะแนนรวมสูง ในทางตรงกันข้าม หากค่า point-biserial ยิ่งสูงแสดงถึงข้อสอบที่มีความสามารถในการแยกแยะดี ผู้ที่ตอบข้อสอบข้อนั้นถูกมักทำคะแนนรวมได้สูง ข้อสอบที่ดีควรมีค่า point-biserial สูงกว่า 0.2 ข้อสอบที่ดีพอใช้ควรมีค่า point-biserial อยู่ในช่วง 0.1 – 0.19 ข้อสอบที่มีค่า point-biserial ต่ำกว่า 0.1 เป็นข้อสอบที่ไม่ดีนัก โดยเฉพาะข้อสอบที่มีค่า point-biserial ต่ำกว่า 0 ไม่ควรนำมาคิดคะแนน

1.3 ประสิทธิภาพของตัวลวง (distractor functionality) ตัวลวงที่มีประสิทธิภาพนั้นมีคุณสมบัติ 2 ประการคือ

1.3.1 มีผู้สอบเลือกตัวลวงนั้นไม่ต่ำกว่า ร้อยละ 5 ของจำนวนผู้สอบทั้งหมด

1.3.2 มีค่า point-biserial correlation ของตัวลวงนั้นเป็นลบ กล่าวคือตัวลวงที่ดีจะลวงให้ผู้สอบที่มีความรู้ไม่ดี (มีคะแนนต่ำ) มาเลือก หากตัวลวงใดมีค่า point-biserial correlation เป็นบวก ให้ทบทวนข้อสอบข้อนั้นดูว่าอาจเฉลยผิด หรือมีคำตอบที่ถูกต้องมากกว่า 1 ตัวเลือก ตัวลวงใดที่มีผู้สอบเลือกน้อย หรือลวงให้ผู้ที่มีความรู้ดีมาเลือกจัดเป็นตัวลวงที่ไม่ดี สมควรพิจารณาตัดทิ้งหรือปรับเปลี่ยน

2. การวิเคราะห์ข้อสอบโดยรวม (Test-analysis)

การวิเคราะห์ข้อสอบโดยรวมเป็นการพิจารณาว่าเมื่อข้อสอบทั้งชุดทำงานร่วมกันแล้วผลสอบที่ได้ออกมาเป็นอย่างไร มีระดับความยากง่ายเป็นอย่างไร มีการกระจายตัวของคะแนนเป็นอย่างไร มีความน่าเชื่อถือของคะแนนสอบมากน้อยเพียงใด ดัชนีต่างๆที่ต้องพิจารณาได้แก่

2.1 ความเที่ยงตรงของคะแนนสอบ (Internal consistency reliability) เป็นการตรวจสอบว่าคะแนนที่ได้ออกมานั้นมีความน่าเชื่อถือเพียงใด เป็นการตอบคำถามว่าหากนำผู้สอบมาสอบใหม่ในสภาวการณ์เดิม ด้วยข้อสอบที่มีระดับความยากง่ายเท่าเดิม และผู้สอบมีความรู้เท่าเดิมไม่ได้ไปศึกษาเพิ่มเติม จะได้คะแนนสอบเท่าเดิมหรือไม่

ดัชนีชี้วัดความเที่ยงตรงของคะแนนสอบที่นิยมใช้ในการรายงานผลสอบด้วยข้อสอบปรนัยคือค่าสัมประสิทธิ์อัลฟ่า (Coefficient Alpha) ซึ่งสามารถคำนวณได้จากสูตร

$$\propto = \frac{n}{\mathrm{n}-1}\left(1 - \frac{\sum Q_{xi}^2}{Q_x^2}\right)$$

เมื่อ $\propto$ = สัมประสิทธิ์ อัลฟ่า (Coefficient Alpha)

n = จำนวนชุดย่อยของข้อสอบที่ทำการแบ่งออกเพื่อหาความเที่ยง

Q_x^2 = การกระจายตัว (variance) ของคะแนนรวม

Q_{xi}^2 = การกระจายตัว (variance) ของคะแนนข้อสอบย่อยชุดที่ i

ค่าสัมประสิทธิ์อัลฟ่านี้มีค่าอยู่ในช่วง 0 - 1 ค่าต่ำแสดงว่าคะแนนที่ได้มีความเชื่อถือได้น้อย ไม่แตกต่างไปจากการเดาสุ่ม ค่าสูงแสดงว่าคะแนนที่ได้นั้นมีความน่าเชื่อถือมาก หากทำการทดสอบซ้ำคะแนนที่ได้ก็จะใกล้เคียงเดิม โดยทั่วไประดับของความเที่ยงตรงของคะแนนสอบที่ยอมรับได้นั้นขึ้นกับว่าต้องการนำคะแนนสอบไปใช้ทำอะไร หากการตัดสินผลสอบนั้นมีความสำคัญมากต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์อัลฟ่าไม่ต่ำกว่า 0.9 หากการตัดสินผลสอบนั้นมีความสำคัญปานกลางต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์อัลฟ่าในช่วง 0.8 - 0.89 หากการตัดสินผลสอบนั้นมีความสำคัญน้อย ต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์อัลฟ่าในช่วง 0.7 - 0.79

ประเด็นสำคัญที่ต้องพิจารณาคือเมื่อได้คะแนนสอบที่มีค่าสัมประสิทธิ์อัลฟ่าต่ำ จะต้องดำเนินการอย่างไรเพื่อพัฒนาให้การสอบครั้งต่อไปไม่ประสบปัญหาเรื่องความไม่น่าเชื่อถือของคะแนนสอบ ปัจจัยหลักที่จะช่วยความเที่ยงตรงของคะแนนสอบปรนัยมี 3 ปัจจัยคือ

1. เพิ่มจำนวนข้อสอบให้มากขึ้น ยิ่งมีข้อสอบมากข้อคะแนนที่ได้จะมีความเที่ยงตรงเพิ่มมากขึ้น

2. ปรับให้ข้อสอบมีการคละกันของข้อสอบที่ยากและง่ายอย่างเหมาะสม เพื่อปรับให้คะแนนมีการกระจายตัวมากขึ้น หากข้อสอบทั้งชุดประกอบไปด้วยข้อสอบที่ง่ายหมด ผู้สอบเกือบทั้งหมดได้คะแนนสูงมาก จะทำให้มีความแตกต่างของคะแนนน้อย โอกาสที่แยกแยะผู้สอบที่มีความรู้ดีออกจากผู้ที่มีความรู้ปานกลาง หรือไม่ดีได้อย่างมั่นใจเป็นไปได้น้อย ดังนั้นหากอาจารย์ปรับให้มีการคละกันของข้อสอบยากและง่ายอย่างเหมาะสม จะทำให้ผู้สอบมีระดับคะแนนแตกต่างกันมาก ค่าสัมประสิทธิ์อัลฟ่าก็จะสูงมากขึ้น

3. ปรับสภาวะแวดล้อมของการสอบให้เหมาะสม กำจัดสิ่งรบกวนสมาธิของผู้สอบให้เหมาะสม เช่น เสียงรบกวน แสงไฟที่ไม่เพียงพอ หรือไฟติดๆดับๆ เป็นต้น

2.2 การกระจายตัวของคะแนน และคะแนนเฉลี่ย (Standard deviation and mean Score) การตรวจดูลักษณะพื้นฐานของคะแนนสอบนี้ จะช่วยบอกได้คร่าวๆว่าการเรียนการสอนมีประสิทธิภาพเพียงใด หากนักเรียนทั้งชั้นเรียนเข้าใจเนื้อหาดี คะแนนสอบที่ได้ออกมาควรจะไม่มีการกระจายตัวมากนัก (คะแนนเกาะกลุ่มกัน) และคะแนนเฉลี่ยก็ควรจะค่อนข้างสูง หากคะแนนสอบของนักเรียนควรมีการกระจายตัวมากผิดปกติ แสดงว่าอาจมีปัญหาบางประการในการจัดการเรียนการสอน ทำให้นักเรียนบางคนมีความรู้ความเข้าใจดี แต่มีนักเรียนบางกลุ่มที่ไม่ค่อยเข้าใจ

2.3 ค่าความยากง่ายเฉลี่ยของข้อสอบ (Average difficult) จากการวิเคราะห์ข้อสอบราย ข้อ หลังจากได้ค่าความยากง่ายของข้อสอบแต่ละข้อ (p) เมื่อนำค่า p ของข้อสอบทุกข้อมาหาค่าเฉลี่ย จะได้ค่าความยากง่ายของข้อสอบทั้งชุด ค่าที่ได้จะใช้เป็นค่าดัชนีชี้วัดว่าข้อสอบทั้งชุดโดยรวมมีระดับความยากง่ายอย่างไร หากผู้สอบเป็นนักศึกษากลุ่มใหญ่พอที่จะตั้งสมมติฐานว่าระดับความสามารถมีการกระจายตัวอย่างเหมาะสม และไม่ต่างจากระดับความสามารถเฉลี่ยของกลุ่มผู้สอบปีก่อนๆ เราสามารถนำค่าความยากง่ายของข้อสอบทั้งชุดนี้มาเทียบได้ว่าข้อสอบที่ใช้มีความยากง่ายแตกต่างจากปีก่อนๆหรือไม่ ซึ่งอาจารย์อาจนำข้อมูลนี้มาใช้พิจารณาปรับเกณฑ์การตัดเกรดว่าต้องมีการปรับระดับคะแนนที่ได้เกรดต่างๆหรือไม่ อย่างไร

2.4 ค่าความสามารถในการแยกแยะผู้สอบเฉลี่ย (Average discrimination) การนำค่า point - biserial correlation ของข้อสอบทั้งชุดมาหาค่าเฉลี่ย เป็นการบอกคร่าวๆว่าโดยรวมแล้วข้อสอบชุดนี้มีความสามารถในการแยกแยะผู้สอบตามระดับความสามารถเพียงใด ยิ่งได้ค่าสูงก็ยิ่งดี แต่มีข้อควรระวังในการแปลผลในกรณีที่การเรียนการสอนเป็นไปได้ดี และผู้สอบทั้งหมด หรือเกือบทั้งหมดทำคะแนนได้สูง ค่า point - biserial correlation เฉลี่ยของข้อสอบทั้งชุดจะไม่สูงแต่ไม่ได้แปลว่าข้อสอบที่ใช้มีคุณภาพไม่ดี

รายวิชาวิทยาการระบาด

รายวิชานี้จัดการเรียนการสอนในชั้นปีที่ 2 เป็นวิชาภาคทฤษฎี หมวดวิชาพื้นฐานทางวิชาชีพและเป็นวิชาภาคบังคับ มีสาระเนื้อหาเกี่ยวกับแนวคิด หลักการที่เกี่ยวข้องกับวิทยาการระบาด รวมถึงธรรมชาติของการเกิดโรค การกระจายโรคในชุมชน การคำนวณและวิเคราะห์ สถิติชีพ ดัชนีชี้วัดภาวะสุขภาพของชุมชน บทบาทของพยาบาลในการคัดกรอง การเฝ้าระวัง การป้องกัน การควบคุมโรคและความพิการในชุมชน กิจกรรมการเรียนการสอนประกอบด้วย 1) การบรรยายประกอบสไลด์ PowerPoint

2) การมอบหมายให้นักศึกษาจัดทำรายงาน 2 ฉบับคือรายงานการวิเคราะห์สถานการณ์ด้านวิทยาการระบาดและรายงานการวิเคราะห์บทความด้านวิทยาการระบาด แผนการประเมินผลแสดงดังตารางที่ 2.2

ตารางที่ 2.2 แผนการประเมินผลลัพธ์การเรียนรู้รายวิชาวิทยาการระบาด

วิธีการประเมินผล/ผลการเรียนรู้ที่คาดหวัง	สัปดาห์ที่กำหนด	สัดส่วนของการประเมินผล
1. สอบ	สัปดาห์ที่ 9 และ 15	ร้อยละ 70
2. ประเมินพฤติกรรมด้านคุณธรรมและวินัย	ตลอดภาคการศึกษา	ร้อยละ 5
3. รายงาน 2 ฉบับ (ประกอบด้วย		
3.1 รายงานการวิเคราะห์สถานการณ์ด้านวิทยาการระบาด	สัปดาห์ที่ 1 - 3	ร้อยละ 10
3.2 รายงานการวิเคราะห์บทความด้านวิทยาการระบาด	สัปดาห์ที่ 3 - 6	ร้อยละ 10
4. การนำเสนอรายงาน อภิปราย และแสดงความคิดเห็นในชั้นเรียน	ตลอดภาคการศึกษา	ร้อยละ 5
	รวม	**ร้อยละ 100**

การประเมินผล

การประเมินผลการเรียนใช้แบบอิงกลุ่ม และอิงเกณฑ์ โดยใช้เกณฑ์ 60% ขึ้นไปสำหรับระดับคะแนน C กำหนดการกระจายตัวของข้อสอบตามระดับการวัดผลการเรียนรู้แสดงดังตารางที่ 2.3

ตารางที่ 2.3 การกระจายตัวของข้อสอบตามระดับการวัดผลการเรียนรู้

ขอบเขตเนื้อหา	จำนวนข้อสอบตามระดับการเรียนรู้				รวม
	รู้-จำ	เข้าใจ	นำไปใช้	วิเคราะห์	(ข้อ)
หน่วยที่ 1 ความรู้เบื้องต้นเกี่ยวกับวิทยาการระบาด (Introduction to Epidemiology)	1	5	1	3	10
หน่วยที่ 2 ปัจจัยสามทางระบาดวิทยา (Epidemiologic triad)	2	2	2	4	10
หน่วยที่ 3 ธรรมชาติของการเกิดโรค	2	2	2	4	10
หน่วยที่ 4 ปัจจัยด้านประชากร สถานที่และเวลา กับการเกิดและการกระจายโรคในชุมชน	2	3	3	2	10
หน่วยที่ 5 ดัชนีอนามัย	4	6	6	4	20
หน่วยที่ 6 การศึกษาทางระบาดวิทยา : นิยาม วัตถุประสงค์ ประโยชน์และการวัดอัตราเสี่ยงของการเกิดโรค	4	12	14	-	30
หน่วยที่ 7 กลวิธีระบาดวิทยา	-	10	10	-	20
หน่วยที่ 8 บทบาทของพยาบาลในการคัดกรองโรค	-	1	4	5	10
หน่วยที่ 9 ระบาดวิทยากับการบริการพยาบาลระดับปฐมภูมิ	2	1	4	3	10
หน่วยที่10 วิทยาการระบาดกับการบริการพยาบาลระดับทุติยภูมิและตติยภูมิ	2	1	4	3	10
รวม	**19**	**43**	**50**	**28**	**140**

การประเมินผลด้วยแบบสอบเป็นข้อสอบปรนัยชนิดหลายตัวเลือก (Multiple choice test) 4 ตัวเลือกและมีคำตอบที่ถูกที่สุดเพียงข้อเดียว เนื้อหาข้อสอบเป็นไปตามคำอธิบายรายวิชาที่กำหนดไว้ใน

หลักสูตร กำหนดวัตถุประสงค์การเรียนรู้เชิงพุทธิพิสัยไว้ 4 ระดับ คือ รู้-จำ เข้าใจ นำไปใช้และวิเคราะห์ การกระจายระดับของการวัดผลการเรียนรู้ขึ้นกับธรรมชาติของเนื้อหาและสัปดาห์ของการเรียน

บทความและงานวิจัยที่เกี่ยวข้อง

Burton (2004) ได้เขียนบทความเรื่องการหาดัชนีอำนาจจำแนกของข้อสอบช่วยปรับปรุงคุณภาพของข้อสอบได้จริงหรือไม่ สามารถสรุปบทความได้ว่าแบบทดสอบที่มีข้อคำตอบให้เลือกเป็นรายข้อและให้คะแนนเป็น 0 (ศูนย์) และ 1 จะมีคุณภาพหรือไม่สามารถบอกได้จากการหาดัชนีอำนาจจำแนกของข้อสอบแต่ละข้อ และเปรียบเทียบดัชนีกลุ่มสูงและกลุ่มต่ำได้ (U-L Index) เพราะว่าจำนวนผู้ที่ตอบแบบสอบถามถูกแต่ละข้อจะถูกแบ่งเป็น 2 กลุ่มคือกลุ่มที่มีคะแนนสูงสุดและกลุ่มที่มีคะแนนต่ำสุด U-L Index จะไม่มีความเที่ยงถ้าไม่ใช้กับกลุ่มทดสอบจำนวนมากและการกระจายของคะแนนรวม และไม่สามารถนำไปเปรียบเทียบกับแบบทดสอบแต่ละชุดได้ การหาค่าสหสัมพันธ์ของคะแนนแต่ละข้อและคะแนนรวมจะมีความตรงหรือน่าเชื่อถือมากกว่าและยอมรับในปัจจุบันแต่ก็มียังมีข้อบกพร่องที่คล้ายกัน เพราะว่าไม่สามารถแทนที่ในเรื่องของการระมัดระวังการใช้ถ้อยคำในแต่ละข้อคำถาม ข้อสอบจะมีความเที่ยงสูงเมื่อนำข้อที่มีค่าอำนาจจำแนกต่ำหรือไม่ดีออกไป นอกจากนี้ความน่าเชื่อถือของดัชนีอำนาจจำแนกของข้อสอบต้องคำนึงถึงผลกระทบของการเดาของผู้ตอบแบบทดสอบและความสมบูรณ์ของคำถาม ซึ่งสามารถค้นหาได้อย่างง่ายโดยใช้รูปแบบการประมวลผลทางคอมพิวเตอร์เป็นเครื่องมือในการสืบค้นและเรียนรู้

Phipps et al (2009) ได้ศึกษาความสัมพันธ์ระหว่างข้อคำถามชนิดกรณีศึกษา (Case based) กับ ไม่ใช่กรณีศึกษา (non-case based) ในแบบทดสอบชนิดเลือกตอบ (Multiple choice items) กับผลการวิเคราะห์ค่าความยากง่าย และดัชนีอำนาจจำแนก โดยแบบทดสอบที่นำมาวิเคราะห์เป็นชุดข้อสอบชนิดเลือกตอบ 4 – 5 ตัวเลือก วิชา การรักษา (Therapeutics) ที่ใช้สอบนักศึกษาหลักสูตรปริญญาเอกทางเภสัชศาสตร์ 4 ปี (PharmD) คณะเภสัชศาสตร์ ชั้นปีที่ 2 และปีที่ 3 ในช่วงปี ค.ศ. 2004 - 2005 ถึง 2007 - 2008 คณะเภสัชศาสตร์ มหาวิทยาลัยเชนแนนโดอห์ (Shenandoah university) ผลการศึกษาพบว่า การวิเคราะห์ความยากง่าย ในคำถามชนิดกรณีศึกษา (p=76.51, SD=19.2) และชนิดไม่ใช่กรณีศึกษา (p=76.86, SD=18.5) ไม่แตกต่างอย่างมีนัยสำคัญทางสถิติ (p=0.75) แต่พบว่า ดัชนีอำนาจจำแนก ในคำถามชนิดไม่ใช่กรณีศึกษา (r=0.250, SD = 0.1) สูงกว่า คำถามชนิดกรณีศึกษา(r=0.227, SD=0.1) อย่างมีนัยสำคัญทางสถิติ ($p < 0.01$) และพบว่า ในคำถามชนิดไม่ใช่กรณีศึกษากลุ่ม K-type multiple choiceมีค่าความยากง่าย (p=72.01, SD=20.4) สูงกว่า กลุ่มคำถามชนิดเลือกตอบมาตรฐาน

(p=76.28, SD=18.4) อย่างมีนัยสำคัญทางสถิติ ($p< 0.01$) แต่ดัชนีอำนาจจำแนกทั้ง 2 กลุ่ม ไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ (k-type r= 0.231, SD=0.1, standard r=0.246, SD = 0.1) ที่ p=0.21 นอกจากนี้เมื่อวิเคราะห์ข้อสอบชนิดที่มี 4 ตัวเลือก และ 5 ตัวเลือก เปรียบเทียบกัน พบว่าข้อสอบชนิด 5 ตัวเลือก (p=75.04, SD = 19.0, r=0.262, SD=0.1) ยากกว่า และมีค่าดัชนีอำนาจจำแนกสูงกว่า ชนิด 4 ตัวเลือก (p=79.16, SD=17.9, r=0.221, SD=0.1) อย่างมีนัยสำคัญทางสถิติ ($p<0.001$)

Tasdemir (2010) ได้ทำการศึกษาเพื่อเปรียบเทียบความยากง่าย อำนาจการจำแนกในการประเมินผลสัมฤทธิ์ทางการเรียนระหว่างข้อสอบแบบหลายตัวเลือกกับแบบถูกผิด รูปแบบการวิจัยเป็นแบบเชิงพรรณนา กลุ่มตัวอย่างคือ นักศึกษาระดับปริญญาตรีชั้นปีที่สี่จำนวน 252 คน ในปีการศึกษา 2007 – 2008 เครื่องมือที่ใช้เก็บข้อมูลเป็นข้อสอบปลายภาครวม 100 ข้อ โดยแบ่งเป็นข้อสอบแบบหลายตัวเลือก 50 ข้อ และแบบถูกผิดจำนวนเท่ากันคือ 50 ข้อซึ่งมีโครงสร้างเนื้อหาแบบคู่ขนานกันและผ่านการตรวจสอบจากผู้ทรงคุณวุฒิในด้านความตรงตามเนื้อหา หาค่าความเที่ยงจากความสอดคล้องภายในโดยใช้สูตร Kuder - Richardson (KR 20) ได้ .779 และ Spearman – Brown ได้ .782 ซึ่งถือว่าเป็นแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีค่าความน่าเชื่อถือสูง วิเคราะห์ข้อมูลด้วยโปรแกรมสำเร็จรูป SPSS สถิติที่ใช้คือ Paired – t test เพื่อเปรียบเทียบความแตกต่างระหว่างข้อสอบทั้งสองประเภท โดยถือว่าข้อสอบที่ดีควรมีค่าอำนาจการจำแนกมากกว่า 0.19 ขึ้นไป ส่วนข้อสอบที่มีความยากง่ายพอเหมาะมีค่าความยากง่ายประมาณ 0.5 ผลการวิจัยพบว่า

1. คะแนนเฉลี่ยที่ได้จากการทดสอบด้วยข้อสอบแบบหลายตัวเลือกและแบบถูกผิดในเพศชายและหญิงไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ

2. คะแนนเฉลี่ยที่ได้จากการทดสอบแบบหลายตัวเลือกและแบบถูกผิด 28 หัวข้อที่มีโครงสร้างเนื้อหาแบบคู่ขนานกัน มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ กล่าวคือ มีข้อสอบ 13 ข้อที่ผู้เข้าสอบทำคะแนนจากแบบทดสอบแบบหลายตัวเลือกได้สูงกว่าแบบถูกผิด ส่วนอีก 15 ข้อนั้นผู้เข้าสอบทำคะแนนจากแบบทดสอบแบบถูกผิดได้สูงกว่า ในขณะที่อีก 22 ข้อไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ

3. ความยากง่ายของแบบทดสอบแบบหลายตัวเลือกและแบบถูกผิดไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ (ข้อสอบที่มีความยากคือ $1- P \geq 0.5$) โดยพบว่าข้อสอบแบบหลายตัวเลือกมีข้อสอบที่ค่อนข้างยาก 10 ข้อ ส่วนอีก 40 ข้อค่อนข้างง่าย ส่วนข้อสอบแบบถูกผิดมี 5 ข้อที่ค่อนข้างยาก 45 ข้อค่อนข้างง่าย

4. อำนาจการจำแนกของแบบทดสอบแบบหลายตัวเลือกและแบบถูกผิด ไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ โดยข้อสอบแบบหลายตัวเลือกมีค่าอำนาจการจำแนกต่ำ มี 10 ข้อที่มีค่าอำนาจ

การจำแนกอยู่ในเกณฑ์ปกติหรือใช้ได้ 40 ข้อ ส่วนข้อสอบแบบถูกผิดมีค่าอำนาจการจำแนกต่ำ14 ข้อที่มีค่าอำนาจการจำแนกอยู่ในเกณฑ์ปกติใช้ได้ 36 ข้อ

Oluseyi & Ajeigbe (2012) ได้ทำการวิจัยเรื่องการวิเคราะห์ข้อสอบชนิดหลายตัวเลือกในวิชาเคมีพื้นฐานของมหาวิทยาลัยในไนจีเรียการวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ความยาก - ง่ายและอำนาจจำแนกของข้อสอบชนิดหลายตัวเลือกในวิชาเคมีพื้นฐาน รูปแบบการวิจัยเป็นแบบการวิเคราะห์ย้อนหลังเพื่อเปรียบเทียบหาสาเหตุ กลุ่มตัวอย่างคือนักศึกษาจำนวน 800 คน ที่ได้มาจากการสุ่มตัวอย่างแบบง่ายจากจำนวนประชากรที่เป็นนักศึกษาที่เรียนวิชาเคมี 102 ในปีการศึกษา 2008 – 2009 ข้อสอบชนิด 4 ตัวเลือกจำนวน 40 ข้อสร้างขึ้นจากเนื้อหาในรายวิชาดังกล่าว โดยอาจารย์ผู้ร่วมสอนในแต่ละหัวข้อในรายวิชานั้น ผู้วิจัยคัดเลือกนักศึกษาจำนวน 442 คนที่ได้คะแนนในเปอร์เซ็นไทล์ที่ 30 ทั้งจากคะแนนที่น้อยที่สุดและมากที่สุดมาเป็นหน่วยในการวิเคราะห์ สถิติที่ใช้คือค่าความถี่ ค่าความยาก - ง่าย และค่าอำนาจจำแนกผลการศึกษาพบว่า ค่าความยาก - ง่ายรวมทั้งฉบับเท่ากับ 0.54 ค่าอำนาจจำแนกเท่ากับ 0.32 ตัวเลือกที่นักศึกษาเลือกตอบมากที่สุด คือ ตัวเลือกที่ 4 ซึ่งพบถึง 16 ข้อ ในจำนวนข้อสอบที่ 40 ข้อ พบข้อสอบที่มีความผิดพลาดในการวัด3 ข้อ เนื่องจากอำนาจจำแนกเท่ากับ 0 ซึ่งผู้วิจัยอภิปรายว่าอาจจะเกิดจากการสร้างข้อสอบไม่ถูกต้องหรือผู้ตรวจเฉลยผิด นอกจากนี้ พบด้วยว่า ข้อสอบข้อที่ยากที่สุดจำนวน 1 ข้อ มีค่าความยาก - ง่ายเท่ากับ 0.245 และมีค่าอำนาจจำแนกเท่ากับ 0.23 ซึ่งมีนักเรียนจำนวน 76 จาก 442 คนที่ตอบได้ถูก การที่ค่าความยากง่ายของข้อสอบเท่ากับ 0.54 แสดงว่าข้อสอบไม่ยากหรือง่ายจนเกินไป ค่าอำนาจจำแนกเท่ากับ 0.32 แสดงว่า นักศึกษาที่ได้คะแนนสูงในวิชานี้เลือกคำตอบได้ถูกต้อง ผู้วิจัยสรุปในตอนท้ายว่าการสร้างข้อสอบมีความสำคัญอย่างยิ่งและควรทำตารางแสดงผลการวิเคราะห์ข้อสอบไว้ท้ายข้อเพื่อเลือกข้อที่ดีไว้ในธนาคารข้อสอบรวมทั้งควรวิพากษ์ข้อสอบก่อนนำมาใช้กับนักศึกษาด้วย

Lee & Winke (2012) ได้ทำการศึกษาความแตกต่างของการใช้ข้อคำตอบแบบ 3,4 และ5 ตัวเลือกในการทดสอบการฟังภาษาอังกฤษ จากการศึกษาโดยทำการปรับแบบทดสอบการฟังภาษาอังกฤษของ College Scholastic Ability Test (CSAT) จาก 5 ตัวเลือก เป็น 4 ตัวเลือก และ 3 ตัวเลือก โดยผู้ทรงคุณวุฒิชาวเกาหลีจำนวน 73 คน ที่มีความเชี่ยวชาญภาษาอังกฤษในด้านการพูดและการเรียนรู้ทางภาษา โดยทำการตัดตัวลวงที่เหมาะสมในการคัดออกออกจำนวน 2 รอบ และได้แบ่งกลุ่มผู้รับการทดสอบคือ นักเรียนระดับชั้นมัธยมปลายที่ศึกษาอยู่ในกรุงโซล ประเทศเกาหลี จำนวน 264 คน เป็นจำนวน 3 กลุ่ม โดยให้แต่ละกลุ่มได้ทำแบบทดสอบจำนวน 3 ชุดคือ แบบทดสอบแบบใช้ข้อคำตอบ 5, 4 และ 3 ตัวเลือกอย่างละ1 ชุด จากแบบทดสอบทั้งหมด 9 ชุด จากผลการศึกษาพบว่า แบบทดสอบ

แบบใช้ข้อคำตอบ 3 ตัวเลือกจะมีค่าคะแนนเฉลี่ยสูงกว่า แบบทดสอบที่มี 4 ตัวเลือก และ 5 ตัวเลือก อย่างมีนัยสำคัญทางสถิติ แต่ไม่พบความแตกต่างของค่าเฉลี่ยความสามารถในการจำแนกผู้สอบของข้อสอบทั้ง 3 แบบ นอกจากนี้ ค่าความเชื่อมั่นของแบบทดสอบได้แสดงให้เห็นถึงความไม่สอดคล้องกันระหว่างจำนวนตัวเลือกของแบบทดสอบ กับชุดแบบทดสอบ อาจแปลผลได้ว่าคะแนนจากการทดสอบของแบบทดสอบแต่ละแบบ มีความสัมพันธ์ในระดับต่ำกับจำนวนตัวเลือกข้อคำตอบของแบบทดสอบในการวัดทักษะทางการฟัง จากผลการศึกษานี้ทำให้พบว่าในการออกแบบทดสอบโดยการใช้หรือไม่ใช้ข้อคำตอบแบบ 3 ตัวเลือก ขึ้นอยู่กับมุมมองของผู้ออกข้อสอบ หรือผู้มีส่วนได้ส่วนเสีย และในการพัฒนาแบบทดสอบ ผู้ทำการพัฒนาแบบทดสอบควรพิจารณาใช้สถิติในการทดสอบที่หลากหลาย และมองปัจจัยจากสภาพแวดล้อมรอบด้านในการตัดสินใจสร้างแบบทดสอบว่าควรสร้างข้อคำตอบแบบตัวเลือกจำนวนเท่าใด

Caldwell & Pate (2013) ได้ทำการศึกษาเพื่อเปรียบเทียบการใช้รูปแบบคำถามตามแนวปฏิบัติที่เป็นมาตรฐานและไม่มีมาตรฐาน ข้อสอบที่มีแนวปฏิบัติที่เป็นมาตรฐาน ภายในข้อคำถามจะมีหลักการในการสร้างคำถาม ดังนี้ 1) การใช้คำถามเชิงบวก หลีกเลี่ยง คำถามเชิงลบ เช่น ยกเว้น ไม่ใช่ 2) ในแต่ละคำถามควรใช้ตัวเลือกเพียง 3 ตัวเลือก 3) การใช้ตัวเลือก “ ไม่มีข้อใดถูกต้อง” ไม่แนะนำให้ใช้บ่อย รูปแบบการวิจัยเป็นข้อสอบปรนัย 15 ข้อ คำถาม ทดสอบกับนักศึกษาเภสัชศาสตร์ โดยนักศึกษาจำนวน 55 คน ตอบแบบสอบถามที่สร้างตามแนวปฏิบัติที่เป็นมาตรฐาน เรียกว่า standard scale, และนักศึกษา 54 คน ตอบแบบสอบถามที่ไม่ได้สร้างจากแนวปฏิบัติ เรียกว่า nonstandard scale สัดส่วนของคำถามเป็นการวัดด้านความรู้ - ความจำ ร้อยละ 60 เข้าใจ ร้อยละ 20 และ นำไปใช้ ร้อยละ 20 ผลการวิจัยพบว่า ข้อสอบที่ถูกสร้างตามแนวปฏิบัติที่เป็นมาตรฐาน นักศึกษาจะได้คะแนนสูงกว่ากลุ่มที่ทำข้อสอบที่ไม่ได้มาตรฐาน โดยค่าคะแนนของผู้สอบ ข้อสอบ nonstandard scale มีความยากกว่าข้อสอบ standard scale ถึง 12.7 คะแนน แนวปฏิบัติที่แนะนำให้หลีกเลี่ยงการใช้ตัวลวง ไม่มีข้อใดถูกต้อง แสดงให้เห็นความแตกต่างอย่างมีนัยสำคัญทางสถิติ ของคะแนนเฉลี่ย ระหว่างกลุ่ม ที่ทำ ข้อสอบ standard scale และกลุ่ม ที่ทำ ข้อสอบ nonstandard scale ร้อยละ 53.6 และ 41.3 ตามลำดับ ($p<0.001$) จากผลการวิจัยสรุปได้ว่า ข้อสอบที่ไม่ได้มาตรฐานจะมีความยาก ทำให้นักศึกษาได้คะแนนน้อยกว่าข้อสอบที่เป็นมาตรฐาน แต่ไม่สามารถแยกแยะกลุ่ม ถูก/ผิด (เก่ง/อ่อน) ได้ และนักศึกษาที่เรียนอ่อนมักจะตอบผิด ดังนั้นการใช้รูปแบบการสร้างข้อคำถามและตัวลวงควรมีการทดสอบเชิงโครงสร้างของคำถาม และความสอดคล้องระหว่างข้อถูกและผิด

สรุปว่า องค์ประกอบที่สำคัญสำหรับการวิเคราะห์ข้อสอบส่วนใหญ่คือการวิเคราะห์ดัชนีความยากง่ายและค่าอำนาจจำแนก ซึ่งสอดคล้องกับวิธีการวิเคราะห์ในการวิจัยนี้

กรอบแนวคิดในการวิจัย

แสดงดังภาพที่ 2.1

ภาพที่ 2.1 กรอบแนวคิดในการวิจัย