

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) เพื่อสร้างแบบสอบผลสัมฤทธิ์ทางการเรียนวิชา การวัดและประเมินผลการศึกษา ระดับปริญญาตรี มหาวิทยาลัยบูรพา 2) เพื่อตรวจสอบความ ยาก อำนาจจำแนก ความตรง และความเที่ยงของแบบสอบผลสัมฤทธิ์ทางการเรียนวิชา การวัด และประเมินผลในชั้นเรียนที่สร้างขึ้น และ 3) เพื่อหาคะแนนจุดตัดของแบบสอบที่สร้างขึ้นโดย การกำหนดมาตรฐานด้วยวิธีบูคมาร์ค ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง โดยแบ่ง การนำเสนอออกเป็น 4 ตอน ดังนี้

ตอนที่ 1 มโนทัศน์ของมาตรฐาน และคะแนนจุดตัด

ตอนที่ 2 วิธีการกำหนดมาตรฐาน

1. วิธีการกำหนดมาตรฐานแบบดั้งเดิม

2. วิธีการกำหนดมาตรฐานแนวใหม่

ตอนที่ 3 การกำหนดคะแนนจุดตัดด้วยวิธีบูคมาร์คโดยการประยุกต์ทฤษฎี

การตอบสนองข้อสอบ

ตอนที่ 4 แบบสอบวัดผลสัมฤทธิ์ทางการเรียน

ตอนที่ 5 งานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 มโนทัศน์ของมาตรฐาน (Standards) และคะแนนจุดตัด (Cut Score)

คำว่า “มาตรฐาน” (Standards) มีนักวิชาการให้ความหมายไว้ดังนี้

พจนานุกรมลองแมน (Longman Dictionary of Contemporary English, 1999) ได้ให้ คำนิยาม มาตรฐาน (Standards) ว่า ระดับคุณภาพที่เป็นทักษะ ความสามารถ หรือความสำเร็จ ของคน หรือสิ่งของที่ได้รับการตัดสินให้เป็นที่ยอมรับในสถานการณ์หนึ่ง ๆ

เคน (Kane, 1994 cited in Cizek, Bunch & Koons, 2004) มาตรฐาน หมายถึง ระดับความสามารถในการปฏิบัติขั้นต่ำสำหรับบางจุดประสงค์

โคเฮน เคน และคร็อค (Cohen, Kane & Crooks, 1999) มาตรฐานการปฏิบัติ (Performance Standard) เป็นการระบุชนิดของการปฏิบัติ หรือความสำเร็จของผู้สอบในกลุ่มที่ แตกต่างกัน มาตรฐานการปฏิบัติเกี่ยวข้องกับขอบเขตในการจัดจำพวกที่ระบุว่า คนที่อยู่ในกลุ่ม



สูงสามารถทำได้ และคนที่อยู่ในกลุ่มต่ำไม่สามารถทำได้ มาตรฐานการปฏิบัติอยู่ในรูปภาษาที่อธิบายคุณภาพของสิ่งหนึ่งที่เราจะเป็น

กุลlickสัน (Gullickson, 2003) ให้ความหมาย มาตรฐาน (Standards) ว่า หลักการที่ตกลงโดยคนที่เกี่ยวข้องในการปฏิบัติทางวิชาชีพของการประเมิน ถ้าได้มาตรฐานก็จะเสริมให้การประเมินมีคุณภาพ และเป็นไปอย่างยุติธรรมยิ่งขึ้น

ศิริชัย กาญจนวาสี (2545) กล่าวว่า มาตรฐาน หมายถึง คุณลักษณะ หรือระดับที่ถือเป็นคุณภาพ, ความสำเร็จ หรือความเหมาะสมอันเป็นที่ยอมรับกันทางวิชาชีพโดยทั่วไป

สรุปได้ว่า มาตรฐาน หมายถึง ระดับที่บ่งบอกถึงคุณภาพของความสามารถของผู้สอบขั้นต่ำโดยระดับดังกล่าวจะแบ่งผู้สอบออกเป็นระดับ เช่น แบ่งเป็น 2 ระดับ คือ ระดับรอบรู้ และระดับไม่รอบรู้ หรือ แบ่งเป็น 4 ระดับ คือ ระดับต่ำกว่าพื้นฐาน ระดับพื้นฐาน ระดับชำนาญการ ระดับขั้นสูง เป็นต้น มาตรฐานจะมีคำอธิบายการปฏิบัติงานซึ่งจะอยู่ในรูปภาษาที่อธิบายคุณภาพของสิ่งหนึ่งที่เราจะเป็น

ส่วนคำว่า “คะแนนจุดตัด” (Cut Score) มีนักวิชาการให้ความหมายไว้ดังนี้

เยาวดี วิบูลย์ศรี (2539) ให้ความหมายคำว่า คะแนนจุดตัด หมายถึง จุด หรือระดับค่าของคะแนนระดับใดระดับหนึ่งซึ่งอยู่ระหว่าง คะแนนเต็ม และคะแนนศูนย์ ที่ผู้วัดผลใช้เป็นจุดตัดเพื่อแบ่งความสามารถของผู้สอบออกเป็น “ผู้รอบรู้” และ “ผู้ไม่รอบรู้” “ผู้รอบรู้” ในที่นี้หมายถึงผู้ที่สอบได้คะแนนสูงถึงระดับที่กำหนดไว้ตามเกณฑ์มาตรฐานนั้น ๆ ส่วน “ผู้ไม่รอบรู้” หมายถึง ผู้ที่สอบได้คะแนนต่ำกว่าระดับที่กำหนดไว้ตามเกณฑ์มาตรฐานนั้น ๆ เช่นเดียวกัน

เบอร์ก (Berk, 1980) คะแนนจุดตัด คือ จุดบนสเกลคะแนนของแบบทดสอบที่แบ่งผู้สอบออกเป็น 2 กลุ่ม คือ กลุ่มรอบรู้ และกลุ่มไม่รอบรู้

โคเฮน เคน และครุค (Cohen, Kane & Crooks, 1999) กล่าวว่า คะแนนจุดตัดเป็นขอบเขตระหว่างรอยต่อของแต่ละกลุ่มซึ่งก็คือ จุดบนสเกลคะแนนนั่นเอง ตัวอย่างเช่น มีคะแนนจุดตัด 1 จุดที่แบ่งออกเป็น 2 กลุ่มคือ กลุ่มผ่านและกลุ่มตก ซึ่งคะแนนจุดตัดนั้นเป็นการกำหนดขึ้นในทางปฏิบัติสำหรับกำหนดกลุ่มที่แตกต่างกัน

เคน (Kane, 1994 cited in Cizek; Bunch & Koons, 2004) คะแนนจุดตัด คือ จุดบนสเกลคะแนน

สรุปได้ว่า คะแนนจุดตัด หมายถึง จุดบนสเกลคะแนน เพื่อใช้แบ่งนักเรียนออกเป็นกลุ่ม ๆ ตามความสามารถของนักเรียน

สำนักงานคณะกรรมการการวิจัยแห่งชาติ
ห้องสมุดงานวิจัย
วันที่.....11 ต.ค. 2555.....
เลขทะเบียน.....245522.....
เลขเรียกหนังสือ.....

การที่จะทราบว่ามาตรฐาน ในการประเมินหนึ่ง ๆ เป็นอย่างไร จำเป็นต้องกำหนดมาตรฐาน ซึ่งวิธีการกำหนดมาตรฐานมีด้วยกันหลายวิธี สำหรับตอนที่ 2 นำเสนอวิธีการกำหนดมาตรฐานแบบดั้งเดิม และวิธีการกำหนดมาตรฐานแนวใหม่

ตอนที่ 2 วิธีการกำหนดมาตรฐาน (Standard Setting Methods)

การกำหนดมาตรฐาน (Standard Setting) ได้มีนักวิชาการให้ความหมายไว้ดังนี้

เจเกอร์ (Jaeger, 1989 cited in Kane, 1998) ได้กล่าวว่า การกำหนดมาตรฐานเป็นการออกแบบมาเพื่อให้มีการตัดสินใจที่ชัดเจนเกี่ยวกับการจัดให้ผู้สอบแต่ละคนเข้ากลุ่มหนึ่ง หรือหลายกลุ่มโดยอยู่บนพื้นฐานของคะแนนผู้สอบแต่ละคน

ซีเซค (Cizek, 1993 cited in Cizek, Bunch & Koons, 2004) กล่าวว่า การกำหนดมาตรฐาน คือ ข้อกำหนดที่มีความเหมาะสมโดยใช้กระบวนการที่เป็นเหตุเป็นผลเพื่อให้ได้ระดับการปฏิบัติ 2 ระดับ หรือมากกว่านี้

สำหรับวิธีการกำหนดมาตรฐาน ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องสามารถแบ่งออกเป็น 2 แนวใหญ่ ๆ คือ

1) วิธีการกำหนดมาตรฐานแบบดั้งเดิม วิธีแบบสอบเป็นศูนย์กลาง (Test-Center Methods) กับวิธีผู้สอบเป็นศูนย์กลาง (Examinee-Centered Methods)

2) วิธีการกำหนดมาตรฐานแนวใหม่ เป็นแนวการกำหนดมาตรฐานที่ประยุกต์วิธีการกำหนดมาตรฐานแบบดั้งเดิม และสร้างวิธีการแบบใหม่ขึ้นมา เพื่อนำไปใช้กำหนดมาตรฐานการประเมินผลการเรียนรู้ที่มีหลากหลายรูปแบบ

1. วิธีการกำหนดมาตรฐานแบบดั้งเดิม

วิธีการกำหนดมาตรฐานแบบดั้งเดิม แบ่งออกเป็น 2 วิธี คือ วิธีแบบสอบเป็นศูนย์กลาง (Test-Center Methods) กับวิธีผู้สอบเป็นศูนย์กลาง (Examinee-Centered Methods) (Jaeger, 1989 cited in Kane, 1998) เจเกอร์ได้แบ่งวิธีการกำหนดมาตรฐานเป็น 2 วิธีดังกล่าวข้างต้น การแบ่งแบบนี้ได้มีการนำมาใช้อ้างอิงในงานวิจัยเป็นจำนวนมาก (Berk, 1996; Kiplinger, 1997; Sireci & Robin, 1999; Ben-David, 2000; Green, 2000)

1.1 วิธีแบบสอบเป็นศูนย์กลาง (Test-Center Methods) ผู้ตัดสินจะพิจารณางาน และข้อสอบที่อยู่ในแบบสอบ แล้วตัดสินระดับการปฏิบัติของงานนั้นซึ่งจะชี้ถึงความสำเร็จในมาตรฐานการปฏิบัติ วิธีแบบสอบเป็นศูนย์กลางมี 4 วิธีคือ

1.1.1 วิธีแองกอฟ (Angoff Method) ผู้คิดค้นวิธีนี้คือ William H. Angoff ในปี ค.ศ. 1971 ผู้ตัดสินจะถูกถามเพื่อให้ถึงความสามารถของผู้สอบที่มีความสามารถขั้นต่ำ (Minimally Acceptable Person) หรือนักวัดผลบางคนเรียกว่า ผู้สอบคาบเส้น ซึ่งผู้สอบที่มีความสามารถขั้นต่ำคือ คนที่มีระดับความรู้ขั้นต่ำในเรื่องที่จำเป็นต้องปฏิบัติตนเอง หลังจากนั้นผู้ตัดสินก็ทำการตัดสินความเป็นไปได้ ที่เรียกว่า “ระดับการผ่านขั้นต่ำ” (Minimum Pass Level) นั่นคือ ผู้ตัดสินจะกำหนดความสามารถขั้นต่ำของผู้สอบที่มีความเป็นไปได้หรือระดับการผ่านขั้นต่ำ โดยกำหนดในรูปของเปอร์เซ็นต์การตอบข้อสอบแต่ละข้อถูก เมื่อผู้ตัดสินทำการตัดสินความเป็นไปได้ในการตอบถูกของผู้ที่มีความสามารถครบทุกข้อแล้ว ระดับการผ่านขั้นต่ำจะถูกเฉลี่ยจากผู้ตัดสินเป็นระดับการผ่านขั้นต่ำรายข้อ ระดับการผ่านขั้นต่ำรายข้อจะถูกรวมจากทุกข้อในแบบสอบเพื่อกำหนดเป็นจุดตัด วิธีการแองกอฟเป็นวิธีที่ใช้กันอย่างกว้างขวางสำหรับการกำหนดมาตรฐานในแบบสอบวัดผลสัมฤทธิ์ทางการเรียนที่ส่งผลกระทบต่ออนาคตของนักเรียนอย่างสูง (High Stakes) (Angoff, 1971 cited in Kane, 1998)

ดังนั้นเพื่อให้สอดคล้องต่อความเข้าใจถึงวิธีการแองกอฟในการหาจุดตัดคะแนน ผู้วิจัยขอสรุปขั้นตอนการดำเนินการพร้อมทั้งยกตัวอย่างประกอบเพื่อให้เห็นภาพที่ชัดเจนยิ่งขึ้น

- 1) ผู้ตัดสินนึกถึงลักษณะความสามารถของผู้สอบที่มีความสามารถขั้นต่ำ
- 2) ผู้ตัดสินพิจารณาข้อสอบทีละข้อโดยแต่ละข้อตอบคำถามที่ว่า “ผู้สอบที่มีความสามารถขั้นต่ำมีโอกาสการตอบข้อสอบข้อนั้นถูกเท่าไร”
- 3) ผู้ตัดสินทุกคนตอบคำถามตามข้อ 2 ทุกข้อ
- 4) นำเปอร์เซ็นต์การตอบถูกของผู้สอบที่มีความสามารถขั้นต่ำของข้อสอบข้อนั้นจากผู้ตัดสินทุกคนมาบวกกันแล้วหารด้วยจำนวนผู้ตัดสิน แล้วนำค่าเฉลี่ยเปอร์เซ็นต์การตอบในแต่ละข้อมาหาค่าเฉลี่ยก็จะได้คะแนนจุดตัด จากตารางที่ 2-1 จะเห็นว่าจุดตัดคะแนนคือ 75

ตารางที่ 2-1 ตัวอย่างการคำนวณหาคะแนนจุดตัดด้วยวิธีแองกอฟจากเปอร์เซ็นต์การตอบข้อสอบ
แต่ละข้อของผู้สอบที่มีความสามารถขั้นต่ำ

ข้อที่	ผู้ตัดสิน (คนที่)					ค่าเฉลี่ย
	1	2	3	4	5	
1	90	85	95	90	90	90
2	85	80	70	75	80	78
3	80	80	85	85	85	83
4	60	65	70	70	60	65
5	85	80	75	75	80	79
6	80	75	75	80	80	78
7	65	65	70	65	65	66
8	75	70	70	70	70	71
9	80	70	75	75	75	75
10	65	65	65	60	65	64
ค่าเฉลี่ย	76.5	73.5	75	74.5	75	75

จุดตัดคะแนน

ต่อมา เลวินตัน และซีกี้(Livingston and Zieky ,1982) ได้ปรับปรุงวิธีแองกอฟใหม่ มีชื่อว่า Modified Angoff Method สำหรับสาเหตุที่ต้องปรับปรุงวิธีการแองกอฟก็เนื่องจากว่าต้องการเพิ่มความเที่ยงของผู้ตัดสินในการประมาณโอกาสในการตอบข้อสอบถูกต้องโดย การเพิ่มความสอดคล้องในการตัดสินของผู้ตัดสิน (Intra-Judge Consistency) และความสอดคล้องในการตัดสินระหว่าง ผู้ตัดสิน (Inter-Judge Consistency) ซึ่งการเพิ่มความเที่ยงของผู้ตัดสินก็จะทำให้ค่าความผิดพลาดของคะแนนจุดตัดที่ได้ลดลง (Hurtz & Auerbach, 2003)

สำหรับวิธีการกำหนดจุดตัดทำโดยการกำหนดให้ผู้ตัดสินพิจารณาโอกาสการตอบถูกของผู้สอบที่มีความสามารถขั้นต่ำจำนวน 2 รอบ โดยรอบแรก ผู้ตัดสินกำหนดโอกาส

การตอบถูกของผู้สอบที่มีความสามารถขั้นต่ำในข้อสอบข้อแรก เมื่อกำหนดเสร็จก็นำผลการตอบของผู้ตัดสินแต่ละคนมาเขียนบนกระดาษ หรือกระดาษแผ่นใหญ่เพื่อให้ผู้ตัดสินทุกคนได้เห็น ถ้าผลการตอบของผู้ตัดสินคล้ายกันคือมีค่าห่างกันไม่เกิน 10% ก็ให้ผู้ตัดสินพิจารณาข้ออื่นถัดไป สำหรับในรอบที่สอง จะมีการอภิปรายร่วมกัน กรณีที่ผลการตัดสินของแต่ละคนไม่คล้ายกัน โดยจะให้ผู้ตัดสินที่ให้เปอร์เซ็นต์สูงสุด และต่ำสุดอธิบายเหตุผลที่ให้เปอร์เซ็นต์ต่ำสุด และสูงสุด จากนั้นเปิดโอกาสให้ผู้ตัดสินสามารถปรับเปลี่ยนผลการตัดสินของแต่ละคนได้ จากนั้นนำผลการตัดสินของแต่ละคนมาหาค่าเฉลี่ยก็จะได้คะแนนจุดตัด ดังตารางที่ 2-2 ตัวอย่างการคำนวณหาคะแนนจุดตัด

ตารางที่ 2-2 ตัวอย่างการคำนวณหาคะแนนจุดตัดด้วยวิธีแองกอฟที่ได้รับการปรับปรุงจาก
เปอร์เซ็นต์การตอบข้อสอบแต่ละข้อถูกของผู้สอบที่มีความสามารถขั้นต่ำ

ข้อสอบ	ผู้ตัดสิน (คนที่)											
	1	2	3	4	5	6	7	8	9	10	11	12
1	90	90	100	100	100	90	90	90	90	60	90	100
	80	90	90	100	90	90	90	90	60	70	90	90
2	80	90	90	40	100	80	100	70	80	90	100	70
	60	70	90	60	100	80	90	80	70	80	80	80
3	90	70	80	80	100	60	80	80	80	60	50	90
	90	80	90	70	80	60	70	80	80	60	60	90
4	70	60	70	80	90	80	80	70	70	60	50	90
	70	70	60	70	80	80	70	70	70	70	70	80
5	90	60	90	40	80	50	80	70	60	60	90	70
	80	70	90	60	80	60	70	70	70	70	80	70
6	60	60	80	60	70	70	80	80	60	50	70	80
	70	60	70	70	70	70	70	80	60	50	70	80

ตารางที่ 2-2 (ต่อ)

ข้อสอบ	ผู้ตัดสิน (คนที่)											
	1	2	3	4	5	6	7	8	9	10	11	12
7	90	50	80	60	60	70	70	70	70	60	80	60
	80	60	80	70	60	70	60	80	60	50	80	70
8	80	50	70	80	40	90	70	70	60	60	70	70
	70	50	80	70	50	90	70	80	70	70	70	80
9	80	70	60	70	60	80	50	60	60	30	50	60
	90	70	70	70	60	80	50	60	60	30	50	60
10	60	80	50	60	70	90	70	60	30	40	40	50
	70	80	60	70	80	90	80	70	40	50	60	60
ค่าเฉลี่ย												
	79	68	77	67	77	77	77	72	66	57	69	76
	78	70	78	71	75	77	74	77	69	63	72	72
												จุดตัดคะแนน

วิธีเองกอฟได้รับการปรับปรุงอีกหลายครั้ง ซึ่งเลคเก้ต(Reckase, 2000 a cite in Brandon, 2004) กล่าวว่า ยังไม่มีข้อยุติเกี่ยวกับนิยามของกระบวนการเองกอฟ แต่สามารถสรุปขั้นตอนกว้าง ๆ ได้ ดังนี้ เลือกผู้ตัดสิน ฝึกผู้ตัดสิน กำหนดนิยามและอธิบายระดับการปฏิบัติที่ผู้สอบควรพึงมี และผู้ตัดสินจะประมาณระดับความสามารถขั้นต่ำของผู้สอบในข้อสอบแต่ละข้อจนครบทุกข้อ หลังจากนั้นผู้ตัดสินจะได้รับข้อมูลเช่น ค่าความยาก (p-value) ซึ่งจะนำมาใช้ประกอบการอภิปรายร่วมกันในแต่ละรอบ โดยปกติประมาณ 2-3 รอบ จากนั้นจึงคำนวณหาคะแนนจุดตัด

จากที่กล่าวมาผู้วิจัยขอสรุปวิธีการเองกอฟที่ได้รับการปรับปรุง (Modified Angoff) ดังนี้

1) ผู้ตัดสินช่วยกันพิจารณาคุณลักษณะของกลุ่มผู้สอบที่มีความสามารถขั้นต่ำหรือกลุ่มผู้สอบคาบเส้นว่ามีลักษณะการตอบวิชานั้นๆอย่างไร

2) ผู้ตัดสินพิจารณาข้อสอบทีละข้อ แล้วตอบคำถามว่า กลุ่มผู้สอบที่มีความสามารถขั้นต่ำมีโอกาสการตอบถูกในแต่ละข้อ (Livingston & Ziekey, 1982) เป็นเท่าไร แต่ก็มีบางงานวิจัยตั้งคำถามว่า ข้อสอบแต่ละข้อกลุ่มผู้สอบคาบเส้นมีโอกาสตอบถูกหรือไม่ (ผู้ตัดสินจะตอบว่า ใช่ หรือ ไม่ใช่) แล้วจึงรวมคะแนนที่ผู้ตัดสินแต่ละคนที่ตัดสินในแต่ละข้อ (Ben-David, 2000) ได้มีงานวิจัยของอิมพาลา และเพลค (Impara and Plake, 1997) ได้ทำการศึกษาเปรียบเทียบวิธีการเองกอฟที่ตั้งคำถามผู้ตัดสินเกี่ยวกับเปอร์เซ็นต์ในการตอบถูกของผู้สอบกลุ่มคาบเส้น กับวิธีการเองกอฟที่ตั้งคำถามผู้ตัดสินว่า ผู้สอบคาบเส้นมีโอกาสตอบถูกในข้อสอบข้อนั้นหรือไม่ (ผู้ตัดสินจะตอบว่า ใช่ หรือ ไม่ใช่) ผลการวิจัยพบว่าวิธีการเองกอฟที่ตั้งคำถามทั้งสองแบบให้คะแนนจุดตัดที่เท่ากัน และผู้ตัดสินยังกล่าวอีกว่า วิธีการตอบแบบใช่/ ไม่ใช่ มีความสะดวกสบายกว่าวิธีการที่ให้ผู้ตัดสินตอบแบบกำหนดโอกาสในการตอบถูกในข้อสอบแต่ละข้อ

3) ผู้ตัดสินแต่ละคนร่วมกันอภิปรายผลการตัดสินของแต่ละคน โดยเฉพาะผู้ตัดสินที่มีผลตัดสินต่ำสุด และสูงสุดให้อธิบายแสดงความคิดเห็นเหตุผลที่ทำให้ผลการตัดสินต่ำสุด และสูงสุด และระหว่างการอภิปรายผู้อำนวยความสะดวกจะแสดงรูปการแจกแจงผลการตัดสินของผู้ตัดสินแต่ละคน

4) เมื่อผู้ตัดสินอภิปรายเสร็จ ก็เปิดโอกาสให้ผู้ตัดสินคะแนนอีกครั้งหนึ่ง ซึ่งการให้คะแนนอีกครั้งเป็นการเปิดโอกาสให้ผู้ตัดสินปรับเปลี่ยนคะแนน หรือยังคงคะแนนเดิมก็ได้ ผลการตัดสินที่ได้จากผู้ตัดสินแต่ละคนในรอบที่สองนำมาหาค่าเฉลี่ยก็คือ คะแนนจุดตัดนั่นเอง

จะเห็นได้ว่าวิธีการเองกอฟแบบเดิม และวิธีเองกอฟที่ปรับปรุงมีความเหมือนและความแตกต่างกันหลายประการ ผู้วิจัยจึงขอสรุปดังนี้

ความเหมือน

1) ทั้งสองวิธีต่างก็เป็นวิธีที่พิจารณาคะแนนจุดตัดจากข้อคำถามในแบบสอบ (Test-Center Methods)

2) ทั้งสองวิธีเป็นวิธีการหาคะแนนจุดตัดจากกลุ่มผู้สอบที่มีความสามารถต่ำสุด หรือผู้สอบกลุ่มคาบเส้นเพื่อหาเปอร์เซ็นต์การตอบข้อสอบข้อนั้นถูก

3) ทั้งสองวิธีใช้กับแบบสอบแบบเลือกตอบ

ความแตกต่าง

1) วิธีเองกอฟแบบเดิมพิจารณาหาคะแนนจุดตัดเพียงรอบเดียว ในขณะที่วิธีเองกอฟที่ได้รับการปรับปรุงพิจารณาหาคะแนนจุดตัด 2-3 รอบ โดยจัดให้มีการอภิปรายร่วมกันระหว่างผู้ตัดสินและเปิดโอกาสให้ผู้ตัดสินเปลี่ยนผลการตัดสินในรอบแรกได้ ซึ่งจะทำให้คะแนนจุดตัดที่ได้มีความเที่ยงตรงมากขึ้น

2) วิธีเองกอฟที่ปรับปรุงเสนอข้อมูลสารสนเทศสำหรับพิจารณาประกอบการตัดสินใจแก่ผู้ตัดสิน เช่น ค่าความยาก หรือ ผลการตัดสินของผู้ตัดสินแต่ละคน

1.1.2 วิธีของอีเบล (Ebel's Procedure) อีเบลได้กล่าวถึงจุดอ่อนของการนิยามคะแนนจุดตัด หรือคะแนนผ่านโดยการพิจารณาจากร้อยละของคะแนนรวมทำให้ขาดรายละเอียดอื่น ๆ ที่จะนำมาใช้ในการพิจารณาด้วย เนื่องจากในบางครั้งข้อสอบอาจจะมีควมยากมาก หรือ ง่ายเกินไป หรือมีอำนาจจำแนกต่ำมากเกินไปที่ผู้ออกข้อสอบคาดหวังได้ ดังนั้นการที่ผู้สอบจะสอบผ่านหรือไม่ จึงขึ้นอยู่กับคำถามในแบบสอบมากกว่าการพิจารณาจากระดับความสามารถของผู้สอบ อีเบลจึงเสนอวิธีการแก้ไขจุดอ่อนดังกล่าวโดยพิจารณาความเกี่ยวข้อง (Relevance) หรือความสำคัญ (Important) ของแต่ละข้อคำถามในแบบสอบ สำหรับในเรื่องของความเกี่ยวข้อง อีเบลได้แบ่งออกเป็น 4 ประเภทคือ มีความจำเป็น (Essential) มีความสำคัญ (Important) เป็นที่ยอมรับได้ (Acceptable) และไม่แน่ใจ (Questionable) ส่วนในเรื่องระดับความยาก อีเบลได้แบ่งออกเป็น 3 ระดับ คือ ง่าย ปานกลาง และยาก และกำหนดร้อยละที่คาดหวังที่จะให้ผ่านข้อคำถามในแต่ละกลุ่มซึ่งร้อยละที่คาดหวังดังกล่าวถือว่าเป็นปริมาณความคาดหวังของผู้ที่มีความสามารถขั้นต่ำที่จะสอบผ่าน (Ebel, 1972 cited in Kane, 1998)

วิธีของอีเบล สามารถสรุปได้ดังนี้ (กัญจนา ลินทรัตนศิริกุล, 2534)

1) ให้ผู้ตัดสินแต่ละคนพิจารณาข้อคำถาม แล้วจัดแยกเป็น 12 ประเภท (คือ ลักษณะที่เกี่ยวข้องมี 4 ลักษณะ และลักษณะความยากมี 3 ระดับ)

2) อภิปรายแต่ละข้อคำถามดังนี้

- ถามผู้ตัดสินว่ามีกี่คนที่จัดคำถามอยู่ในระดับง่าย ปานกลาง และยาก ถ้าผู้ตัดสินจัดไม่เหมือนกัน ก็ควรจะให้อธิบายเหตุผล

- ถามผู้ตัดสินว่ามีกี่คนที่แยกประเภทของข้อคำถามออกเป็น “มีความจำเป็น มีความสำคัญ เป็นที่ยอมรับ และไม่แน่ใจ” ถ้าผู้เชี่ยวชาญแยกประเภทของข้อคำถามไม่เหมือนกัน ก็ควรให้อธิบายเหตุผลเช่นเดียวกัน

- ให้โอกาสแก่ผู้ตัดสินจัดประเภทของข้อคำถามใหม่ถ้าต้องการเปลี่ยนแปลง

3) ให้ผู้ตัดสินกำหนดร้อยละของผู้สอบที่คาดหวังว่าจะตอบในแต่ละข้อคำถามได้ถูกต้องทั้ง 12 ประเภท

4) รวบรวมผลการตัดสิน

วิธีการหาคะแนนจุดตัดของอีเบลคือ นำจำนวนข้อสอบในแต่ละลักษณะคูณกับค่าร้อยละที่คาดหวังแล้วนำผลคูณที่ได้มาหาค่าเฉลี่ย ค่าเฉลี่ยที่ได้จะเป็นคะแนนจุดตัดหรือคะแนนที่ผู้มีความสามารถขั้นต่ำจะสอบผ่าน (Jaeger, 1989 cited in Kane, 1998)

1.1.3 วิธีของนิเดลสกี (Nedelsky Procedure) นิเดลสกีได้เสนอเทคนิคนี้ในปี ค.ศ. 1954 ซึ่งเป็นเทคนิคที่เหมาะสมกับข้อสอบแบบเลือกตอบโดยข้อสอบแต่ละข้อผู้ตัดสินจะตัดสินว่ามีจำนวนตัวเลือกตอบจำนวนเท่าไรที่ความสามารถของผู้สอบขั้นต่ำควรจะตอบผิด ระดับการผ่านขั้นต่ำสำหรับข้อสอบถูกคำนวณจากจำนวนตัวเลือกที่ยังคงเหลืออยู่ หลังจากตัวเลือกผิดถูกนำออกไปไม่ใช้ในการพิจารณา ถ้าผู้ตัดสินตัดสินว่า ความสามารถขั้นต่ำของผู้สอบควรกำจัดออกทั้งหมดยกเว้นตัวเลือก 3 ตัวเลือกในแบบสอบเลือกตอบในข้อหนึ่ง ระดับการผ่านขั้นต่ำในนิเดลสกี ควรเป็น $1/3$ หรือ 0.33 ระดับการผ่านขั้นต่ำในทุกข้อจะถูกนำมารวมกันเพื่อกำหนดคะแนนจุดตัด (Nedelsky, 1954 cited in Kane, 1998)

ดังนั้นหากกล่าวโดยสรุปวิธีของนิเดลสกี มีดังนี้ (กัญจนา ลินทรตันศิริกุล, 2534)

1) ให้ผู้ตัดสินแต่ละคนพิจารณาคำถามทั้งหมด และใช้ดินสอขีดลงไปที่ตัวเลือกที่ผิด ที่เห็นว่าผู้สอบจะไม่เลือก

2) อภิปรายคำถามแต่ละข้อ ตามแนวทางดังนี้

- ให้ผู้ตัดสินพิจารณาคำตอบที่ผิดก่อน และดูว่ามีผู้ตัดสินกี่ท่านที่คิดว่าผู้สอบจะไม่เลือก และมีกี่ท่านไม่เห็นด้วย

- ถ้าผู้ตัดสินมีความเห็นไม่ตรงกัน จะต้องให้แต่ละคนอธิบายเหตุผลแล้วถามผู้ตัดสินที่ไม่เห็นด้วย ซึ่งตอนนี้ยังไม่ต้องการสรุปจนกว่าผู้ตัดสินจะแน่ใจว่าได้เข้าใจความหมายของคำว่า ผู้สอบที่มีความสามารถขั้นต่ำได้ตรงกัน

- พิจารณาตัวเลือกที่เป็นตัวผิดตัวต่อไป

3) หลังจากที่มีการอภิปรายคำถามทุกข้อในลักษณะนี้แล้ว ขั้นตอนต่อมาคือจะต้องให้ผู้ตัดสินทบทวนถึงการตัดสินโดยที่ผู้ตัดสินจะต้องทำเครื่องหมายกากบาทลงบนตัวเลือกที่เขาแน่ใจว่า ผู้สอบที่มีความสามารถต่ำจะไม่เลือก

4) รวบรวมผลการตัดสิน แล้วดำเนินการเช่นเดียวกันในคำถามต่อไป วิธีของนิเดลสกีนี้อาจให้ผู้ตัดสินแต่ละคนตัดสินเองที่ได้โดยปราศจากการอภิปรายร่วมกัน แต่ปัญหาที่จะ

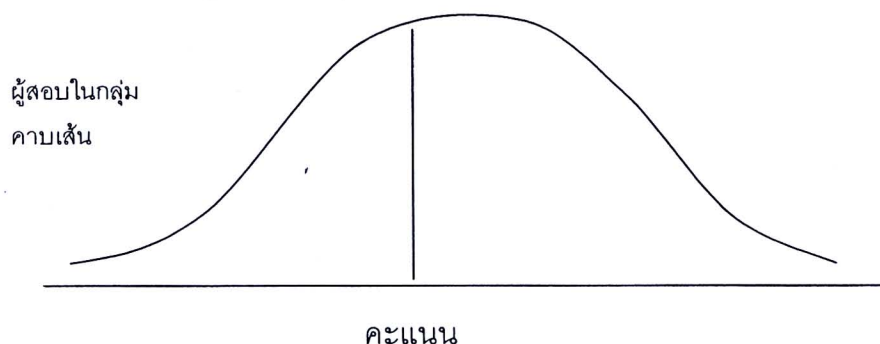
ตามมาก็คือ ผู้ตัดสินจะยังไม่คุ้นเคยกับวิธีการตัดสินเพราะฉะนั้นถ้าเป็นไปได้ก็ควรจะให้โอกาสแก่ผู้ตัดสินในการฝึกตัดสินใจ จากตัวอย่างของคำถามและอภิปรายซึ่งกันและกันก่อนที่จะมีการพิจารณาคำถามที่เหลืด้วยตนเองต่อไป

1.2 วิธีผู้สอบเป็นศูนย์กลาง (Examinee-Centered Methods) การกำหนดมาตรฐานด้วยวิธีผู้สอบเป็นศูนย์กลาง ผู้ตัดสินจะจัดจำพวกผู้สอบ จุดตัดจะถูกกำหนดโดยการกำหนดจุดที่สเกลคะแนน (Score Scale) ที่มีความคงที่มากที่สุดในกลุ่มนั้น วิธีผู้สอบเป็นศูนย์กลาง มี 2 วิธีดังนี้

1.2.1 วิธีกลุ่มคาบเส้น (Borderline-Group Method) (Livingston & Zieky, 1982 cited in Kane, 1998) ผู้ตัดสินจะระบุผู้สอบแต่ละคนที่มีความสามารถระดับคาบเส้น ซึ่งผู้สอบคาบเส้นจะเป็นที่การปฏิบัติงานของพวกเขาผ่านมาตรฐานขึ้นมาเล็กน้อย (Just Meet the Performance Standard) ผู้ตัดสินอาจเป็นครู หรือผู้ควบคุมดูแล (Supervisors) ฯลฯ ซึ่งพวกเขาเหล่านั้นจะมีประสบการณ์ต่อตัวผู้สอบแต่ละคน โดยผู้ตัดสินจะระบุผู้สอบที่มีความสามารถระดับคาบเส้น จากนั้นก็จะคำนวณมัธยฐานของคะแนนสอบของคนกลุ่มคาบเส้นก็จะได้คะแนนจุดตัดข้อดีของวิธีนี้คือ ง่ายในการนำไปใช้และง่ายในการอธิบายแต่ข้อจำกัดของวิธีนี้ก็คือ ค่อนข้างยากที่จะตัดสินว่าผู้สอบอยู่ในกลุ่มที่มีลักษณะเป็นกลุ่มคาบเส้นจริง ๆ

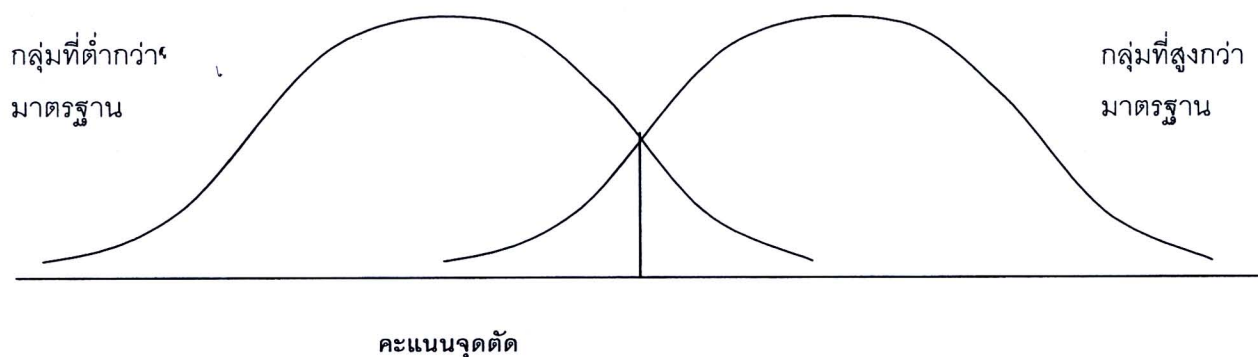
อย่างไรก็ตามหากนำวิธีนี้ไปใช้ มีขั้นตอนโดยสรุปดังนี้ (กัญญา ลินทรัตนศิริกุล, 2534)

- 1) เลือกผู้ตัดสิน
 - 2) นิยามความหมายของคำว่าระดับความรู้และทักษะของผู้สอบที่อยู่ในระดับคาบเส้น
 - 3) ระบุกลุ่มผู้สอบที่เป็นกลุ่มคาบเส้น
 - 4) เก็บรวบรวมคะแนนที่ได้จากการสอบของผู้สอบในกลุ่มคาบเส้น
 - 5) กำหนดคะแนนมัธยฐานของผู้สอบในกลุ่มคาบเส้นให้เป็นคะแนนจุดตัด
- สามารถเขียนคะแนนการแจกแจงผู้สอบในกลุ่มคาบเส้น ดังภาพที่ 2-1



ภาพที่ 2-1 คะแนนการแจกแจงผู้สอบในกลุ่มคาบเส้น

1.2.2 วิธีกลุ่มตรงข้าม (Contrasting-Group Method) (Livingston & Zieky, 1982 cited in Kane, 1998) ผู้ตัดสินจัดประเภทผู้สอบออกเป็น 2 กลุ่มคือ กลุ่มที่สูงกว่ามาตรฐาน และกลุ่มที่ต่ำกว่ามาตรฐาน วิธีกลุ่มตรงข้าม ผู้ตัดสินที่จะจัดเรียงอันดับ (Rating) คือ ครู ผู้ควบคุมดูแล หรือบางคนที่มีประสบการณ์เกี่ยวกับการจัดระดับการปฏิบัติงานของผู้สอบ หรือคนที่มีประสบการณ์การวัดเกณฑ์ภายนอก หรือประเมินแบบสอบการปฏิบัติ หลังจากแจกแจงคะแนนสำหรับสองกลุ่มเพื่อใช้ตัดสินแล้วคะแนนจุดตัดจะถูกเลือกเพื่อจำแนกความเป็นไปได้ระหว่างกลุ่มสูง และกลุ่มต่ำโดยจุดที่คะแนนการแจกแจงของผู้สอบทั้งสองกลุ่มตัดกันจะเป็นคะแนนจุดตัด ดังภาพที่ 2-2



ภาพที่ 2-2 คะแนนการแจกแจงของผู้สอบกลุ่มที่ต่ำและสูงกว่ามาตรฐาน

จากที่กล่าวมาการกำหนดมาตรฐานแบบดั้งเดิมมี 2 วิธีคือ วิธีแบบสอบเป็นศูนย์กลาง (Test-Center Methods) กับวิธีผู้สอบเป็นศูนย์กลาง (Examinee-Centered Methods) วิธีแบบสอบเป็นศูนย์กลางเป็นวิธีที่ใช้อย่างง่าย เหมาะกับแบบสอบปรนัย หรือแบบตอบสั้น ซึ่งแบบสอบลักษณะดังกล่าวทำเป็นข้อสอบจำนวนมาก ๆ ได้ แต่วิธีแบบสอบเป็นศูนย์กลางไม่เหมาะกับแบบสอบการปฏิบัติ ส่วนวิธีผู้สอบเป็นศูนย์กลาง เหมาะสำหรับการปฏิบัติในสภาพจริง (Authentic Performances) โดยวิธีนี้ผู้ตัดสินจะประเมินการปฏิบัติงานของนักเรียนแต่ละคน ซึ่งทำให้ผู้ตัดสินเสียเวลามาก (Kane, 1998) สำหรับวิธีย่อย ๆ ทั้ง 3 วิธีในแบบสอบเป็นศูนย์กลาง วิธีที่ได้รับความนิยมใช้อย่างแพร่หลายคือ วิธีของแองกอฟ (Berk, 1996) แต่เนื่องจากในช่วงประมาณปี พ.ศ. 1980 นักการศึกษาทั้งของอังกฤษและอเมริกา เริ่มตระหนักถึงความจำเป็นที่จะต้องปฏิรูปการศึกษา ความคิดที่จะเปลี่ยนแปลงในเรื่องเป้าหมายของการเรียนรู้ สาระการเรียนรู้ การจัดประสบการณ์การเรียนรู้ตลอดจนการวัดและประเมินผล ทำให้เกิดการศึกษาวิจัยเพื่อหาแนวทางในการปฏิรูปการศึกษา สำหรับการวัดและประเมินผลได้เริ่มมีการศึกษาถึงทางเลือกใหม่ของการ

ประเมิน (Alternative Assessment) อาทิเช่น การประเมินการปฏิบัติงาน (Performance Assessment). การประเมินตามสภาพจริง (Authentic Assessment) เป็นต้น (Hart, 1994 อ้างถึงใน พวงแก้ว ปุณยกกนก, 2546) ซึ่งทำให้วิธีการกำหนดมาตรฐานโดยวิธีแบบสอบเป็นศูนย์กลางเป็นวิธีการกำหนดมาตรฐานที่ไม่เหมาะสมกับการประเมินการปฏิบัติงาน และการประเมินตามสภาพจริง ดังนั้นในช่วงทศวรรษ 1990 ได้มีนักการศึกษาหลายท่านได้ปรับปรุงวิธีการกำหนดมาตรฐานเดิม หรือเสนอวิธีการกำหนดมาตรฐานใหม่เกิดขึ้นมากมาย

2. วิธีการกำหนดมาตรฐานแบบใหม่

ปัจจุบันในด้านการศึกษาและการรับรองได้ใช้การประเมินที่ประกอบด้วยข้อสอบที่มีทั้งแบบเขียนตอบ (Constructed-Response) และแบบเลือกคำตอบ (Selected-Response) (Parshall, Davey & Pashley, 2000 cited in Hambleton, Jaeger, Plake & Mills, 2000) เพื่อกำหนดมาตรฐานการเรียนรู้ของผู้สอบออกเป็นระดับ เช่น 5 ระดับคือ ตก, ต่ำกว่าพื้นฐาน, พื้นฐาน, เชี่ยวชาญ (Proficient) และก้าวหน้า (Advance) หรือถ้าแบ่งตามการรับรองสามารถแบ่งออกเป็น 2 ระดับคือ ได้รับการรับรอง และไม่ได้รับการรับรอง ส่วนการประเมินการปฏิบัติที่มีการให้คะแนนหลายระดับในงาน ๆ หนึ่ง หรือเป็นงานที่ต้องใช้หลาย ๆ ทักษะเพื่อบ่งบอกถึงความสำเร็จ สิ่งเหล่านี้นับเป็นปัญหาในวิธีการกำหนดมาตรฐาน (Standard Setting Methods) (Berk, 1986; Jaeger, 1989; Livingston & Zieky, 1982; Plake, 1998 cited in Hambleton, Jaeger, Plake & Mills, 2000) เนื่องจากวิธีการกำหนดมาตรฐานหลายวิธีในแบบดั้งเดิมไม่สามารถนำมาใช้กับการประเมินการเรียนรู้ที่ให้คะแนนมากกว่า 2 ค่าได้ (Polytomously Score) ทำให้มีการปรับ หรือพัฒนาวิธีการใหม่ขึ้นมาเพื่อใช้กับการประเมินการเรียนรู้

วิธีการกำหนดมาตรฐานแบบใหม่นี้ใช้สำหรับการประเมินที่ให้คะแนนแบบหลายค่า ซึ่งใช้ในการประเมินการเรียนรู้ (Cizek, 2001; Reckase, 2000 cited in Hambleton, Jaeger, Plake & Mills, 2000) มีด้วยกันหลายวิธีคือ

- 2.1 วิธีแองกอฟแบบปรับขยาย (Extended Angoff Method)
- 2.2 วิธีบุ๊กมาร์ค (Bookmark Method)
- 2.3 วิธีกลุ่มตรงข้าม (Contrasting Groups Method)
- 2.4 วิธีเลือกงานผู้สอบ (Examinee Paper Selections Method)
- 2.5 วิธีองค์รวม (Holistic Method)
- 2.6 วิธีเส้นภาพเด่น (Dominant Profile Method)
- 2.7 วิธีการนำนโยบายมาตัดสิน (Judgmental Policy Capturing Method)

แต่ละวิธีมีรายละเอียดดังนี้

2.1 วิธีการเอนกอฟแบบปรับขยาย (Extended Angoff Method)

กระบวนการกำหนดมาตรฐานด้วยวิธีการเอนกอฟมีมาตั้งแต่ปี ค.ศ. 1971 โดย วิลเลียม เอนกอฟ (William H. Angoff) ได้เสนอวิธีนี้ขึ้นมา และต่อมาแฮมเบลตัน และเพลค (Hambleton & Plake, 1995) ได้นำวิธีการเอนกอฟมาปรับปรุงเพื่อใช้ในการประเมินการปฏิบัติ นั่นคือ วิธีเอนกอฟแบบปรับขยายถูกพัฒนาสำหรับการประเมินการปฏิบัติที่ให้คะแนนมากกว่า 2 คะแนนเอง วิธีนี้กำหนดให้คณะผู้ตัดสินทำการประมาณคะแนนคาดหวังของผู้สอบคาบเส้น (ผู้สอบที่เกือบไม่ได้รับการรับรอง) (Borderline Candidates) ในแบบฝึกหัดแต่ละฉบับที่ให้คะแนนมากกว่า 2 คำ งานของผู้ตัดสินก็คือ กำหนดความน่าจะเป็น (เช่น คะแนนที่คาดหวัง) ที่ผู้สอบที่คาบเส้นสามารถตอบข้อสอบถูก คะแนนที่คาดหวังจากแบบฝึกหัดของผู้ตัดสินแต่ละคนสามารถนำไปใช้ในมาตรฐานการปฏิบัติในชุดของแบบฝึกหัด จากนั้นมาตรฐานของผู้ตัดสินจะถูกนำมาเฉลี่ยเพื่อให้ได้มาตรฐานตามที่ต้องการสำหรับการใช้ในการรับรองเพื่อให้เกิดความเข้าใจมากขึ้นจะขอยกตัวอย่างงานวิจัยของแฮมเบลตัน และเพลค (Hambleton & Plake, 1995)

แฮมเบลตัน และเพลค (Hambleton & Plake, 1995) กล่าวในงานวิจัยว่า วิธีนี้คณะผู้ตัดสิน (Panelists) จะประมาณคะแนนที่คาดหวังของผู้สอบที่คาบเส้น (Borderline Candidates) โดยใช้มาตรประมาณค่า 4 ระดับ สำหรับแบบฝึกหัดแต่ละฉบับซึ่งมีแบบฝึกหัดทั้งหมด 7 ฉบับที่รวมเป็นการประเมินการปฏิบัติหนึ่งงาน คณะผู้ตัดสินจะถูกถามเพื่อกำหนดน้ำหนักในแบบฝึกหัดเพื่อสะท้อนความเข้าใจถึงความสำคัญที่สัมพันธ์ของแบบฝึกหัดในการทำการตัดสิน ผ่าน-ตกในรอบสุดท้าย การตัดสินคะแนนที่คาดหวัง และน้ำหนักที่สัมพันธ์กันจะถูกทำหลาย ๆ ครั้งโดยการอภิปรายร่วมกันระหว่างผู้ตัดสินในแต่ละชั้นของกระบวนการกำหนดมาตรฐาน การอภิปรายจะช่วยให้คณะผู้ตัดสินมีฉันทมติ (Consensus) ในมาตรฐานที่ออกมา

กล่าวโดยสรุป ขั้นตอนการกำหนดมาตรฐานด้วยวิธีการเอนกอฟแบบปรับขยายตามแนวคิดของแฮมเบลตัน และเพลค (Hambleton and Plake, 1995) ดังต่อไปนี้

1. คณะผู้ตัดสินร่วมกันอภิปรายถึงลักษณะการปฏิบัติงานในแต่ละงานของผู้สอบคาบเส้นว่าควรเป็นแบบใด จนกระทั่งชัดเจน
2. ฝึกผู้ตัดสินให้เข้าใจถึงวิธีการให้คะแนน ลักษณะแบบฝึกหัดของนักเรียนแต่ละแบบฝึกหัดซึ่งแต่ละแบบฝึกหัดมีมิตีย่อย ๆ ที่ผู้ตัดสินต้องพิจารณาด้วย หลังจากนั้นผู้ตัดสินแต่ละคนกำหนดมาตรฐาน (คะแนนคาดหวังของผู้สอบคาบเส้น) ในมิตีย่อย ๆ ที่อยู่ในแบบฝึกหัดแต่ละแบบฝึกหัด จากนั้นผู้ตัดสินแต่ละคนก็กำหนดระดับความเชื่อมั่นของพวกเขาที่ได้กำหนดมาตรฐานของเขาขึ้นมา เมื่อดำเนินการเรียบร้อยแล้ว นำมาตรฐานและระดับความเชื่อมั่นของผู้ตัดสินแต่ละคนมาหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน

3. นำผู้ตัดสินทั้งหมดมาอภิปรายร่วมกันถึงการกำหนดมาตรฐานของพวกเขา ซึ่งผู้อำนวยการความสะดวกจะแสดงมาตรฐานที่ได้เจ็ลี่ยออกมาในแต่ละมิติของแบบฝึกหัดให้ตัดสินพิจารณาประกอบการอภิปราย และควรให้ผู้ตัดสินที่ให้ค่ามาตรฐานต่ำสุด และสูงสุดอภิปรายถึงการกำหนดมาตรฐานของพวกเขา แล้วให้ผู้ตัดสินคนอื่นร่วมกันอภิปรายถึงสิ่งที่ผู้ตัดสินให้ค่ามาตรฐานต่ำสุด และสูงสุดกล่าวถึง การดำเนินการอภิปรายจะดำเนินการจนกระทั่งผู้ตัดสินรู้สึกว่าการอภิปรายได้แสดงความคิดเห็นโดยสมบูรณ์แล้ว ก็จะดำเนินการในขั้นตอนต่อไป

4. ผู้ตัดสินแต่ละคนดำเนินการกำหนดมาตรฐาน (คะแนนคาดหวังของผู้สอบคาบเส้น) เป็นรอบที่ 2 และกำหนดระดับความเชื่อมั่นในมาตรฐานที่พวกเขากำหนดขึ้นมา

5. ผู้ตัดสินจะถูกถามเพื่อกำหนดน้ำหนักสัมพัทธ์ (Relative Weightings) ของแต่ละมิติโดยนำมาตรฐานที่กำหนดขึ้นประกอบการพิจารณาให้น้ำหนัก ขณะเดียวกันผู้ตัดสินก็กำหนดระดับความเชื่อมั่นที่พวกเขาให้น้ำหนักสัมพัทธ์ จากนั้นนำน้ำหนักสัมพัทธ์ของผู้ตัดสินแต่ละคนมาหาค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน

6. เมื่อผู้ตัดสินแต่ละคนกำหนดน้ำหนักสัมพัทธ์ในแต่ละมิติของแต่ละแบบฝึกหัดเรียบร้อยแล้วก็มาอภิปรายร่วมกัน โดยนำค่าเฉลี่ยของน้ำหนักสัมพัทธ์ของแต่ละมิติมาให้ผู้ตัดสินพิจารณาประกอบการอภิปราย และให้ผู้ตัดสินที่กำหนดน้ำหนักสัมพัทธ์สูงสุดและต่ำสุดอภิปรายถึงการกำหนดน้ำหนักสัมพัทธ์ของพวกเขา การอภิปรายดำเนินไปจนกระทั่งผู้ตัดสินทั้งหมดเห็นว่าการอภิปรายสมบูรณ์แล้วจึงดำเนินการขั้นต่อไป

7. ผู้ตัดสินจะถูกถามเพื่อกำหนดน้ำหนักสัมพัทธ์เป็นรอบที่ 2 ในแต่ละมิติของแต่ละแบบฝึกหัด และให้ผู้ตัดสินกำหนดระดับความเชื่อมั่นของการกำหนดน้ำหนักสัมพัทธ์

8. ผู้อำนวยการความสะดวกนำค่ามาตรฐานในแต่ละมิติของแบบฝึกหัดมาคูณกับค่าน้ำหนักสัมพัทธ์ แล้วนำค่าของแต่ละมิติมาบวกกันก็จะได้ค่ามาตรฐานรวมในแบบฝึกหัดขึ้นนั้น ตัวอย่างเช่น แบบฝึกหัดขึ้นหนึ่งมี 3 มิติ จะมีมาตรฐานแต่ละมิติคือ 3, 4 และ 2 และมีค่าน้ำหนักสัมพัทธ์ในแต่ละมิติคือ .25, .35 และ .40 ดังนั้นมาตรฐานของแบบฝึกหัดขึ้นนี้คือ $(3 \times .25) + (4 \times .35) + (2 \times .40) = 2.95$ หาค่ามาตรฐานรวมของทุกแบบฝึกหัดตามวิธีการข้างต้น

9. ผู้ตัดสินจะถูกถามเพื่อกำหนดน้ำหนักสัมพัทธ์ของแต่ละแบบฝึกหัดเช่น ถ้ามีแบบฝึกหัด 5 ฉบับ ก็จะมีน้ำหนักสัมพัทธ์ 5 ค่า (น้ำหนักสัมพัทธ์เฉลี่ยของผู้สอบแต่ละคน) และพิจารณาระดับความเชื่อมั่นถึงการกำหนดน้ำหนักสัมพัทธ์เหล่านั้น

10. ผู้อำนวยการความสะดวกนำค่ามาตรฐานของแบบฝึกหัดแต่ละขึ้น (คำนวณได้ตามข้อ 8) มาคูณกับค่าน้ำหนักสัมพัทธ์ของแบบฝึกหัดแต่ละขึ้น (จากข้อ 9) แล้วจึงนำผลคูณที่ได้ในแต่ละแบบฝึกหัดมาบวกกัน ก็จะได้ค่ามาตรฐานของแบบฝึกหัดทั้งหมดในภาพรวม



จุดแข็งและข้อบกพร่อง

จุดแข็ง วิธีการเองกอฟแบบปรับขยายเป็นวิธีที่ง่ายชัดเจน กระบวนการให้น้ำหนักทำให้ผู้ตัดสินกำหนดค่าของคำถามแต่ละข้อแตกต่างกัน การที่ให้น้ำหนักมากในบางคำถามแสดงว่า ผู้ตัดสินรู้สึกว่ามีค่าความสำคัญมาก (Hambleton, Jaeger, Plake & Mills, 2000) นอกจากนี้ผู้ตัดสินยังเกิดความรู้สึกมั่นใจเกี่ยวกับความสามารถของพวกเขาว่า มีความถูกต้องแม่นยำในขั้นตอนของกระบวนการ (Plake, Hambleton & Jaeger, 1997)

ข้อบกพร่อง วิธีการเองกอฟแบบปรับขยายอาจใช้ได้ไม่ดี เมื่อนำไปใช้กับการปฏิบัติที่เป็นธรรมชาติที่มีความยืดหยุ่นสูง (Hambleton, Jaeger, Plake & Mills, 2000) นอกจากนี้ในงานวิจัยของแฮมเบิลตัน และเพลค (Hambleton & Plake, 1995) คณะผู้ตัดสินได้กล่าวถึงปัญหาด้านความตรงของมาตรฐานรอบสุดท้าย โดยคณะผู้ตัดสินต้องการให้มีการใช้นโยบายในการกำหนดมาตรฐานที่ซับซ้อนมากกว่านี้เพื่อให้เกิดการเชื่อมโยง (Conjunctive) และ/หรือ กฎการทดแทนการตัดสินใจ (Compensatory Decision Rules) โดยวิธีการเองกอฟแบบปรับขยายเป็นการทดแทน (Compensatory) เพราะคะแนนการผ่าน (Passing Score) รอบสุดท้ายเป็นการรวมน้ำหนักของคะแนนแบบฝึกหัดแต่ละแบบฝึกหัด ดังนั้นผู้สอบที่ทำแบบฝึกหัดหลายแบบฝึกหัด ซึ่งพบว่ามีบางแบบฝึกหัดได้คะแนนการปฏิบัติต่ำ แต่ผู้สอบคนนี้ก็ยังคงผ่านการประเมิน เนื่องจากว่าการปฏิบัติของนักเรียนคนนั้นในแบบฝึกหัดอื่นที่เหลือได้คะแนนสูง สิ่งนี้เป็นหลักการของวิธีการทดแทน (Compensatory Approach)

2.2 วิธีบุ้คมาร์ค (Bookmark Method)

วิธีการบุ้คมาร์คเป็นหนึ่งในกระบวนการผังข้อสอบ (Item-Mapping Procedure) ที่มีหลายกระบวนการ (Cizek; Bunch & Koons, 2004) กระบวนการกำหนดมาตรฐานบุ้คมาร์ค (Bookmark Standard Setting) ได้รับการเสนอขึ้นเป็นครั้งแรกโดยลิวอิส มิเซล และกรีน (Lewis, Mizel & Green) ในปี ค.ศ. 1996 วิธีการบุ้คมาร์คได้รับการพัฒนาเพื่อให้ผู้ตัดสินมีกระบวนการคิด (cognitive) เพื่อตัดสินได้ง่ายขึ้น วิธีนี้ได้รับความนิยมอย่างรวดเร็ว และได้มีการใช้บุ้คมาร์คในการกำหนดคะแนนจุดตัดที่เป็นการประเมินระดับประเทศ (Large-Scale Assessment) วิธีบุ้คมาร์คถูกนำมาใช้กับแบบสอบที่ประกอบด้วยแบบเลือกตอบ และแบบเขียนตอบ

กระบวนการบุ้คมาร์คเป็นการรวมเนื้อหาข้อสอบกับผลการสอบของผู้สอบเข้าด้วยกัน กระบวนการนี้เปิดโอกาสให้ผู้ตัดสินอภิปรายและให้โอกาสเปลี่ยนคะแนนจุดตัดที่เคยกำหนดในรอบแรกเป็นคะแนนจุดตัดอันใหม่ได้ ประโยชน์อีกประการหนึ่งของบุ้คมาร์คคือ สามารถใช้กับแบบสอบที่มีรูปแบบผลสมที่ประกอบด้วย ข้อสอบแบบเลือกตอบ (ให้คะแนนแบบ 2 ค่า) และข้อสอบแบบเขียนตอบ (ให้คะแนนมากกว่า 2 ค่า) นอกจากนี้บุ้คมาร์คยังออกแบบมาให้ง่ายต่อ

การตัดสินโดยมีคู่มือจัดเรียงข้อสอบจากข้อง่ายไปข้อยาก (Mitzel, Lewis, Patz, & Green, 2001 cite in Beretvas, 2004)

คู่มือจัดเรียงข้อสอบ (Ordered Item Booklet) เป็นสิ่งสำคัญที่ใช้ในกระบวนการ บัญญัติข้อสอบ ข้อสอบจะถูกจัดเรียงตามค่าความยากจากข้อสอบง่ายสุดไปยังข้อสอบยากสุด โดยใช้ กระบวนการ IRT เพื่อให้ทราบถึงค่าความยากของข้อสอบ และความสามารถของนักเรียนที่อยู่บน มาตรฐานเดียวกัน (Same Continuum) ข้อสอบแบบเขียนตอบก็จะถูกนำมาไว้ในมาตรฐานเดียวกับ ข้อสอบแบบเลือกตอบโดยคะแนนแต่ละคะแนนในข้อสอบแบบเขียนตอบจะถูกนำมากำหนดบน มาตรฐาน แล้วจัดเรียงลำดับลงในคู่มือ (Egan, 2002; Kiplinger, 1997) ตัวอย่างเช่น ข้อสอบข้อที่ 37 เป็นข้อสอบเขียนตอบที่ให้คะแนนแบบ 0, 1, 2 ซึ่งคะแนนเหล่านี้จะถูกนำมาเป็นตัวแทนลงบน สเกลความสามารถ (Ability Scale) และในคู่มือ โดยจะมีข้อสอบข้อที่ 37.1 แทนข้อที่ให้คะแนน 1 คะแนน ซึ่งจะถูกกำหนดให้ใกล้กับจุดเริ่มต้นของคู่มือมากกว่า 2 คะแนน ขณะที่ข้อสอบข้อที่ 37.2 แทนข้อสอบที่ให้คะแนน 2 คะแนน จะถูกกำหนดให้ใกล้ท้ายเล่มของคู่มือเพราะว่า 2 คะแนนทำได้ ยากกว่า 1 ส่วนคะแนน 0 คะแนนจะไม่ถูกนำมาใช้เพราะว่า มันแสดงถึงการไม่มีคะแนนคำตอบ เป็นต้น

เมื่อเตรียมคู่มือจัดเรียงข้อสอบเรียบร้อยแล้ว ขั้นตอนต่อไปเป็นกระบวนการหา คะแนนจุดตัดโดยผู้ตัดสินแต่ละคนได้รับคู่มือจัดเรียงข้อสอบซึ่งลักษณะของคู่มือคือ มีข้อสอบ เหมือนกับแบบสอบที่ใช้สอบ เพียงแต่ว่าคู่มือมีการจัดเรียงข้อสอบใหม่ตามความยากของข้อสอบ (ข้อสอบ 1 ข้อต่อ 1 หน้า) เมื่อผู้ตัดสินได้รับคู่มือแล้ว ผู้ตัดสินก็เริ่มพิจารณาข้อสอบทีละข้อไป เรื่อย ๆ โดยพิจารณาว่าผู้สอบกลุ่มคาบเส้นมีโอกาสตอบข้อสอบข้อนั้นถูก 2 ใน 3 หรือ 67% เกณฑ์ตรงนี้เรียกว่าค่า RP (Response Probability) ค่า RP 2/3 เหมาะนำมาใช้กับวิธีการกำหนด มาตรฐานด้วยวิธีบัญญัติ (Buckendahl, Smith, Impara, & Plake, 2002; Mizel et al., 2001; Reckase, 2000; Skaggs & Tessema, 2001) โดย Huynh (1998) แสดงให้เห็นว่า โมเดล IRT แบบ 3 พารามิเตอร์ ฟังก์ชันสารสนเทศของข้อสอบ (Item Information Function) จะมีค่าสูงสุดที่ θ สำหรับ $P(\theta) = (c+2)/3$ เมื่อไม่ได้พิจารณาการเดา ($c = 0$) ค่า RP จะมีค่าเท่ากับ 2/3 นั่นคือ ฮิวส์ (Huynh) พบว่า $RP = .67$ ให้ค่าฟังก์ชันสารสนเทศข้อสอบสูงสูดนั่นเอง

สำหรับการกำหนดบัญญัติ (กำหนดคะแนนจุดตัด) ดำเนินการซ้ำกัน 3 รอบ แต่ละ รอบทำขึ้นเพื่อช่วยเพิ่มฉันทามติ และลดความแตกต่างของผู้ตัดสิน

รอบที่ 1

เป้าหมายหลักของรอบ 1 เพื่อให้ผู้ตัดสินคุ้นเคยกับคู่มือการจัดเรียงข้อสอบซึ่งเป็นการดำเนินการขั้นต้นของการกำหนดบัญญัติ ในรอบนี้ผู้ตัดสินทำงานเป็นกลุ่มย่อยโดยการ

อภิปรายข้อสอบเป็นรายข้อว่า อะไรที่ทำให้ข้อสอบข้อนี้ยากกว่าข้อสอบข้อก่อน การบรรยายการปฏิบัติที่มีระดับต่างกัน (เช่น ระดับพื้นฐาน ระดับชำนาญการ และระดับสูง) จะถูกนำมาแสดง และอภิปรายร่วมกัน จากนั้นผู้ตัดสินถูกถามเพื่ออภิปราย และตัดสินเนื้อหาว่า นักเรียนควรมีความรู้ที่จะอยู่ในระดับการปฏิบัติใด การตัดสินของผู้ตัดสินจะทำการตัดสินคะแนนจุดตัดอิสระจากกันโดยการนำที่คั่นหนังสือมาคั่นระหว่างข้อสอบที่ถูกตัดสินว่าเป็นตัวแทนของคะแนนจุดตัด ที่คั่นหนังสือ 1 อันจะถูกคั่นสำหรับแต่ละคะแนนจุดตัดที่ต้องการ ข้อสอบก่อนหน้าที่คั่นหนังสือนี้จะได้รับการพิจารณาว่า นักเรียนทั้งหมดถูกคาดหวังว่ารู้ และสามารถปฏิบัติงานได้สำเร็จที่ความน่าจะเป็นอย่างน้อย 0.67 ในทางตรงกันข้ามนักเรียนเหล่านี้ถูกคาดหวังว่าสามารถปฏิบัติงานได้สำเร็จในข้อสอบที่อยู่หลังจากข้อสอบที่คั่นหนังสือที่ความน่าจะเป็นน้อยกว่า 0.67

รอบที่ 2

กิจกรรมอย่างแรกในรอบที่ 2 เกี่ยวข้องกับผู้ตัดสินแต่ละคนทำการคั่นหนังสือลงในคู่มือจัดเรียงข้อสอบของแต่ละคน ตัวอย่างเช่น ผู้ตัดสินกลุ่มหนึ่งมี 6 คน คู่มือจัดเรียงข้อสอบของผู้ตัดสินทั้ง 6 คน มีที่คั่นหนังสือ 6 อันสำหรับแต่ละคะแนนจุดตัด ในการอภิปรายจะให้นั้นข้อสอบที่อยู่ระหว่างที่คั่นหนังสืออันแรก และอันสุดท้ายสำหรับแต่ละระดับการปฏิบัติ เมื่อการอภิปรายสิ้นสุดลงผู้ตัดสินทำการคั่นหนังสือลงในคู่มือจัดเรียงข้อสอบอีกครั้งค่าเฉลี่ยในการกำหนดบุคมาร์ค (คั่นหนังสือ) ในรอบที่ 2 สำหรับแต่ละคะแนนจุดตัด จะถูกนำมาเป็นข้อเสนอแนะของกลุ่มเพื่อใช้กำหนดคะแนนจุดตัดต่อไป

รอบที่ 3

รอบที่ 3 ผู้ตัดสินรวมกันเป็นกลุ่มใหญ่ จากนั้นเริ่มจากการแสดงข้อมูลผลกระทบ (Impact Data) ให้ผู้ตัดสินพิจารณา เปอร์เซนต์การตกของนักเรียนแต่ละระดับการปฏิบัติถูกนำมาแสดง ให้ค่าเฉลี่ยคะแนนจุดตัดของแต่ละกลุ่มที่ได้จากรอบที่ 2 กับข้อมูลของนักเรียนว่ามีการปฏิบัติอย่างไร ผู้ตัดสินอภิปรายจุดที่คั่นหนังสือในกลุ่มใหญ่ จากนั้นจึงให้ผู้ตัดสินดำเนินการตัดสินอีกครั้งอย่างอิสระจากกันโดยการให้ผู้ตัดสินคั่นหนังสือ ค่าเฉลี่ยคะแนนจุดตัดของกลุ่มใหญ่ถูกพิจารณาเพื่อเป็นคะแนนจุดตัดขั้นสุดท้ายสำหรับทำเป็นระดับการปฏิบัติ

เมื่อได้คะแนนจุดตัดในรอบสุดท้ายแล้ว ขั้นตอนต่อไปคือ การบรรยายระดับการปฏิบัติ (Performance Level Descriptors) โดยผู้ตัดสินช่วยกันเขียนการบรรยายการปฏิบัติในประเด็น ความรู้เฉพาะ ทักษะ และความสามารถของนักเรียน ณ ระดับการปฏิบัติที่กำหนดไว้ ข้อสอบก่อนที่ทำการคั่นหนังสือถูกนำมาพิจารณาเนื้อหาว่า นักเรียน ณ ระดับการปฏิบัตินี้ถูกคาดหวังว่า มีความสามารถในการตอบคำถามถูกด้วยความน่าจะเป็นอย่างน้อย 0.67 ความรู้และทักษะที่นำมาตอบคำถามได้สำเร็จในข้อสอบเหล่านี้ถูกสังเคราะห์เข้าอย่างเป็นระบบเป็นการ

บรรยายสำหรับการปฏิบัตินี้ ด้วยเหตุนี้การบรรยายระดับการปฏิบัติเกิดการขยายออกมาของกระบวนการกำหนดคะแนนจุดตัด

จุดดีและข้อบกพร่อง

จุดดี

1) วิธีนี้ผู้ตัดสินพิจารณาแบบสอบในภาพรวมโดยอยู่บนพื้นฐานการพิจารณาจากผลคะแนนของนักเรียน

2) วิธีนี้ผู้คุมารค์ถูกออกแบบมาให้ใช้กับข้อสอบแบบเลือกตอบที่ให้คะแนนแบบ 2 ค่า และข้อสอบแบบเขียนตอบที่ให้คะแนนมากกว่า 2 ค่า อย่างเท่าเทียมกันเพราะข้อสอบทุกข้ออยู่บนสเกล IRT เหมือนกัน

3) กระบวนการนี้ผู้คุมารค์อยู่บนพื้นฐานที่มีข้อมูลเชิงประจักษ์ และทำให้การตัดสินของผู้ตัดสินเป็นไปด้วยความง่าย เพราะวิธีการนี้ผู้ตัดสินจะไม่ได้ถูกถามเพื่อประมาณอย่างคาดคะเนเองถึงระดับความยากของข้อสอบแต่ละข้อในกลุ่มนักเรียนที่สมมุติขึ้นเอง แต่ผู้ตัดสินจะถูกถามเพื่อให้ค้นหาในคู่มือที่เรียงลำดับข้อสอบไว้ว่าหน้าไหนที่พวกเขาารู้สึกว่า ข้อสอบหน้านั้นมีความยากของข้อสอบเกินจากระดับที่คาดว่า นักเรียนที่มีความรู้จะสามารถทำได้ นั่นคือ ผู้ตัดสินกำหนดคะแนนจุดตัดตามที่การปฏิบัติงานของนักเรียนควรจะเป็นและตามเงื่อนไขการปฏิบัติที่นักเรียนแสดงออกมา

ข้อบกพร่อง

1) บางครั้งไม่มีข้อสอบจำนวนมากพอสำหรับพิจารณานักเรียนที่อยู่ในมาตรฐานระดับ ก้าวหน้า หรือระดับตก ซึ่งทั้งสองระดับเป็นระดับสุดโต่ง (เก่งมาก กับอ่อนมาก) ซึ่งอาจทำให้เกิด floor and ceiling effect ได้

2) วิธีการนี้ไม่ได้แสดงข้อมูลการปฏิบัติงานของนักเรียนในสเกลรายงานผลซึ่งการรายงานนี้นำมาใช้ตัดสินมาตรฐานการปฏิบัติรอบสุดท้าย ตัวอย่างเช่น ข้อสอบอาจจะกำหนดสเกลรายงานผลที่ระดับความสามารถที่นักเรียนมีโอกาสทำได้สำเร็จ 50% หรือกำหนดสเกลรายงานผลที่ระดับความสามารถของนักเรียนที่มีโอกาสทำได้สำเร็จ 65% ลักษณะแบบนี้เป็นปัญหาการกำหนดที่มีความหลากหลายซึ่งอาจส่งผลกระทบต่อมาตรฐานการปฏิบัติได้ ดังนั้นจึงต้องมีการวิจัยในประเด็นดังกล่าว

ดังนั้นเพื่อให้เกิดความชัดเจน ผู้วิจัยได้สรุปความเหมือน และความแตกต่างระหว่างวิธีการกำหนดมาตรฐานด้วยวิธีเองกอฟ และวิธีผู้คุมารค์ดังนี้

ความเหมือน

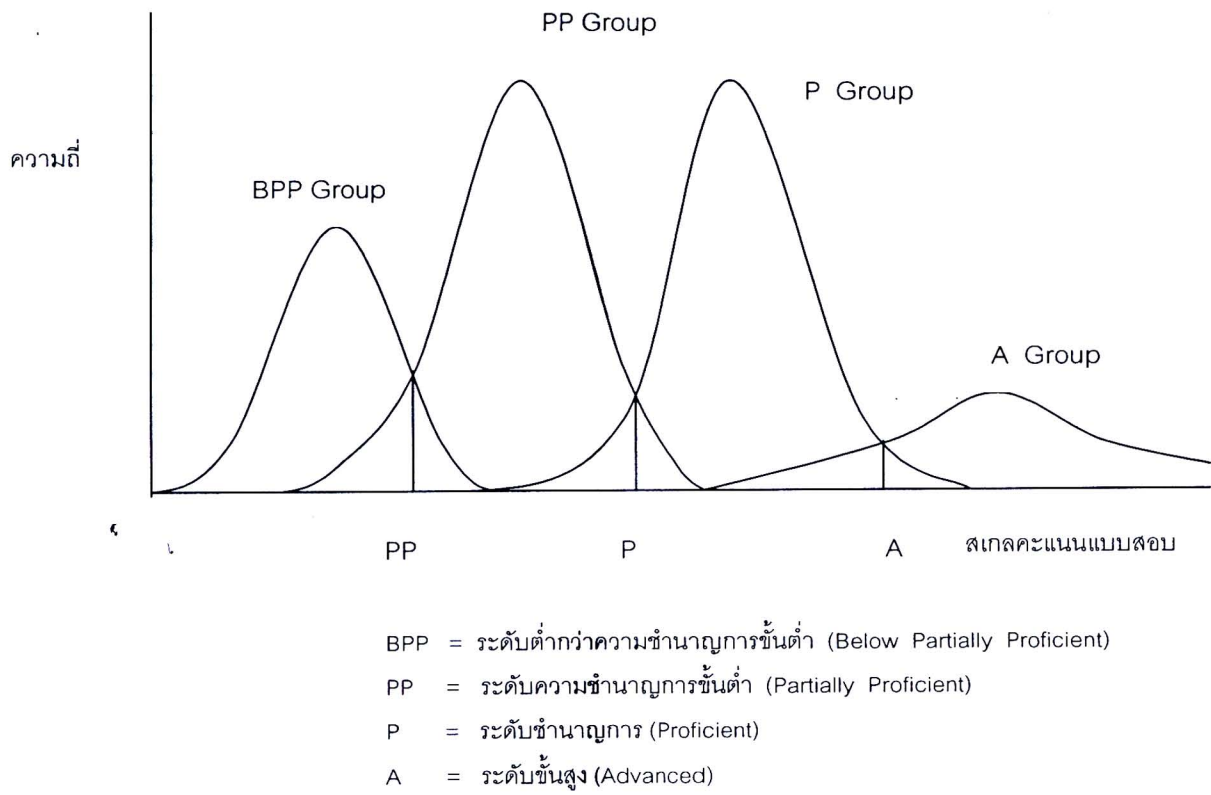
- 1) ทั้งสองวิธีต่างก็เป็นวิธีที่พิจารณาหาคะแนนจุดตัดจากข้อคำถามในแบบสอบ (test-center methods)
- 2) ทั้งสองวิธีเป็นวิธีการหาคะแนนจุดตัดจากกลุ่มผู้สอบที่มีความสามารถต่ำสุด หรือผู้สอบกลุ่มคาบเส้นเพื่อหาโอกาสการตอบข้อสอบข้อนั้นถูก

ความแตกต่าง

- 1) วิธีบุคมาร์คเป็นวิธีที่มีการจัดเตรียมสารสนเทศให้กับผู้ตัดสินเพื่อช่วยพิจารณา กำหนดคะแนนจุดตัด โดยจัดเตรียมในรูปคู่มือจัดเรียงข้อสอบตามความยาก ในขณะที่วิธีแองกอฟผู้ตัดสินพิจารณาจากตัวแบบทดสอบโดยไม่มีการจัดเรียงข้อสอบตามความยาก
- 2) วิธีบุคมาร์คสามารถนำไปใช้กับแบบทดสอบที่มีรูปแบบผสมคือ มีทั้งข้อสอบแบบเลือกตอบและข้อสอบแบบเขียนตอบอยู่ในฉบับเดียวกัน ในขณะที่วิธีการแองกอฟเหมาะกับแบบสอบแบบเลือกตอบ
- 3) วิธีการบุคมาร์คสามารถนำไปใช้หาคะแนนจุดตัดที่มีหลายระดับได้ส่วนวิธีแองกอฟไม่เหมาะที่จะนำไปใช้หาคะแนนจุดตัดที่มีหลายระดับ

2.3 วิธีกลุ่มตรงข้าม (Contrasting Groups Method)

การกำหนดมาตรฐานการปฏิบัติวิธีนี้ทำโดยการเปรียบเทียบการแจกแจงคะแนนสอบของผู้สอบโดยผู้สอบจะถูกแบ่งออกเป็นกลุ่ม ๆ ผู้ตัดสินเป็นผู้ที่ทราบว่านักเรียนแต่ละคนมีความสามารถเป็นอย่างไร (เช่น ครูทราบความสามารถทางคณิตศาสตร์ของนักเรียน) ผู้ตัดสินถูกถามเพื่อให้จำแนกนักเรียนออกเป็นกลุ่มตามมาตรฐานที่กำหนด (เช่น ทำงานต่ำกว่าระดับเกณฑ์ ทำงานได้ตามระดับเกณฑ์ ทำงานได้สูงกว่าระดับเกณฑ์) ผู้สอบถูกจัดจำพวกโดยไม่ใช้ความรู้จากคะแนนสอบของพวกเขา แต่คะแนนสอบของผู้สอบแต่ละคนจะถูกนำมาจัดจำพวกในการแจกแจงความถี่ ผลจากการแจกแจงความถี่นำมาเปรียบเทียบเพื่อตัดสินระดับของการซ้อนกัน (Overlap) คะแนนจุดตัดจะถูกจำแนกเพื่อใช้แบ่งระหว่างกลุ่มให้เป็นมาตรฐานการปฏิบัติ



ภาพที่ 2-3 คะแนนการแจกแจงกลุ่มผู้สอบค่าเส้นสำหรับระดับการปฏิบัติหลายระดับ

จุดดีและข้อบกพร่อง

ข้อดีของวิธีนี้คือ เป็นวิธีที่ต้องการให้ผู้ตัดสินทำการตัดสินนักเรียนเป็นรายบุคคลโดยจัดให้กลุ่มที่คล้าย ๆ กันอยู่ระดับเดียวกัน แต่การจัดจำพวกผู้สอบดังกล่าวมีความเป็นไปได้ที่จะผิดพลาดเพราะการแจกแจงความถี่มีการทับซ้อนกันทำให้การกำหนดมาตรฐานการปฏิบัติทำได้ยาก นอกจากนี้ครูอาจจะลังเลที่จะให้นักเรียนของพวกเขาเข้ากลุ่มต่ำ เนื่องจากครูเหล่านี้ให้เกรดนักเรียนเหล่านี้ผ่านมาแล้ว ปัญหาใหญ่อีกประการหนึ่ง คือ การทำให้การกำหนดมาตรฐานสามารถสรุปอ้างอิง (Generalizability) ไปยังประชากรผู้สอบเป็นไปได้ยาก ถ้ากลุ่มตัวอย่างผู้สอบที่ใช้ในการวิเคราะห์ไม่เป็นตัวแทนที่ดีแล้ว มาตรฐานการปฏิบัติก็จะไม่เหมาะสม (Hambleton, Jaeger, Plake & Mills, 2000)

2.4 วิธีเลือกงานผู้สอบ (Examinee Paper Selections Method)

วิธีการนี้ใช้กับการประเมินการปฏิบัติงานที่ให้คะแนนมากกว่า 2 ค่า วิธีนี้คล้ายคลึงกับวิธีกลุ่มค่าเส้น (Borderline Group Method) ซึ่งเป็นวิธีที่ใช้กับการให้คะแนนแบบ 2 ค่า

เพียงแต่ว่าคณะผู้ตัดสินพิจารณาการตอบของผู้สอบมากกว่าการจัดเรียงอันดับผู้สอบอย่างเป็นอิสระ คณะผู้ตัดสินจะเลือกชิ้นงาน (Papers) ที่พวกเขารู้สึกว่าเป็นตัวแทนงานของกลุ่มผู้สอบที่คาบเส้นสำหรับในแต่ละข้อ คะแนนเฉลี่ยจากชิ้นงานที่ถูกเลือกในแต่ละระดับมาตรฐานการปฏิบัติงานจะถูกใช้เป็น “ค่าการผ่านขั้นต่ำ” (Minimum Passing Value) สำหรับแต่ละข้อคำถาม ค่าขั้นต่ำจะถูกรวมเพื่อใช้ตัดสินมาตรฐานการปฏิบัติแต่ละมาตรฐานในแบบทดสอบ

จุดดีและข้อบกพร่อง

คณะผู้ตัดสิน (โดยเฉพาะครู) ชอบวิธีนี้เพราะพวกเขามีโอกาสในการตัดสินใจโดยมีงานของผู้สอบประกอบการพิจารณา อย่างไรก็ตามวิธีนี้นำมาใช้กับผู้สอบกลุ่มเล็ก ๆ ได้ยาก เนื่องจากการแสดงชิ้นงานจากการแจกแจงงานของผู้สอบเป็นไปได้ยาก ดังนั้นจึงต้องใช้ความพยายามอย่างมากเพื่อเลือกชิ้นงานของผู้สอบ ระบุระบบรหัส จัดระบบกล่องบรรจุชิ้นงานของผู้สอบแก่ผู้ตัดสิน และใช้ชิ้นงานจำนวนมากเพื่อให้คณะผู้ตัดสินประเมินเนื่องจากว่าไม่มีการกำหนดน้ำหนักองค์ประกอบในแต่ละส่วนให้แตกต่างกัน ซึ่งข้อความแต่ละข้อจะถูกนำไปกำหนดเป็นแต่ละมาตรฐานการปฏิบัติ แม้ว่ามีบางองค์ประกอบที่อาจมีค่าน้อยกว่าองค์ประกอบอื่น นอกจากนี้ยังไม่มีเหตุผลว่า ทำไมการให้น้ำหนักจึงไม่ได้นำมาใช้กับกระบวนการเลือกชิ้นงานซึ่งทำให้การนำไปปฏิบัติได้ไม่ดี (Hambleton, Jaeger, Plake & Mills, 2000)

2.5 วิธีองค์รวม (Holistic Method)

วิธีนี้คล้ายกับวิธีเลือกงานผู้สอบ (Examinee Paper Selection Method) ถูกเสนอโดย the National Academy of Education (1993) เพื่อกำหนดมาตรฐานของ NAGB และ ACT วิธีองค์รวม (Holistic) คณะผู้ตัดสินถูกถามเพื่อให้พิจารณางานที่สมบูรณ์ (ประกอบด้วยแบบฝึกหัดทั้งหมด หรืองานในการประเมินผลของผู้สอบ) และตัดสินว่าคู่มือเล่มไหนที่เป็นตัวแทนกลุ่มผู้สอบที่คาบเส้น (หรือผู้สอบรอบรู้และไม่รอบรู้ หรือ พื้นฐาน ขำนาญการ และก้าวหน้า) ในบางกรณีมาตรฐานค่าจะนำมาใช้สำหรับคณะผู้ตัดสินเพื่อแบ่งชิ้นงานของผู้สอบออกเป็นกลุ่ม ๆ เช่น ต่ำกว่าพื้นฐาน แยกเป็นสูง กลาง ต่ำ และสูงกว่าพื้นฐานแยกเป็นสูง กลาง ต่ำ วิธีนี้ถูกนำไปใช้เป็นทางเลือกเช่นเดียวกับวิธีแองกอฟในแบบสอบแบบเลือกตอบโดยตรวจสอบข้อวิจารณ์ว่า วิธีแองกอฟ (เน้นระดับข้อสอบ) มีจุดอ่อนตรงการปฏิบัติงานของผู้สอบ NAGB และ ACT ได้นำวิธีนี้มาทดสอบกับข้อมูลของ NAEP ในหลายวิชา ระดับชั้นเกรด 4, 8 และ 12

จุดดีและข้อบกพร่อง

ข้อดี คือ คณะผู้ตัดสินใช้งานของผู้สอบจริง ๆ นอกจากนี้คณะผู้ตัดสินยังได้พิจารณางานของผู้สอบอย่างยืดหยุ่นโดยให้ผู้ตัดสินมีความรู้สึกเมตตา (Forgiving) มากกว่าวิธีที่จัด

เรียงลำดับงานต่องาน อย่างไรก็ตามข้อคำถามที่เหลือคณะผู้ตัดสินสามารถนำมาใช้เพื่อหาข้อมูลจากผู้สอบได้ แต่ถ้าคู่มือมีความหนาหรือยาวมาก คณะผู้ตัดสินจะทำให้มีความตรง และความเที่ยงตรงในการตัดสินเกี่ยวกับคุณภาพของงานได้หรือไม่ (Hambleton, Jaeger, Plake & Mills, 2000)

2.6 วิธีเส้นภาพเด่น (Dominant Profile Method)

วิธีนี้ถูกใช้สำหรับการประเมินงานหรือแบบฝึกหัดจำนวนจำกัด แต่ละงานมีการให้คะแนนแบบหลายค่า วิธีนี้ใช้ยุทธวิธีสร้างฉันทมติ (Consensus-Building) ผู้ตัดสินจะถูกถามเพื่อให้เกิดกฎที่เป็นนโยบาย (ตัดสินใจ) เพื่อใช้กำหนดคะแนนงานหรือแบบฝึกหัดที่ใช้ประกอบการประเมินเป็นค่าต่ำสุดที่ต้องการให้ผ่าน องค์ประกอบการทดแทน (Compensatory) และองค์ประกอบที่เชื่อมโยง (Conjunctive) มีความยืดหยุ่นสำหรับกฎการตัดสินใจ ดังนั้นผู้ตัดสินสามารถสร้างกฎการตัดสินใจที่ซับซ้อนเพื่อให้มีความหมายถึงผู้สอบที่มีคุณภาพที่ผ่านการประเมิน วิธีนี้มีความซับซ้อนในการใช้มากกว่ามาตรฐานการปฏิบัติที่มีเพียงมาตรฐานเดียว

หลังจากผู้ตัดสินได้รับการฝึกเกี่ยวกับงานหรือแบบฝึกหัดที่ใช้ประกอบการประเมิน เช่น พิจารณาความหมายของคะแนนที่มีความเป็นไปได้หลาย ๆ ค่า ในแต่ละงานหรือแบบฝึกหัด จากนั้นผู้ตัดสินจะถูกถามอย่างอิสระเพื่อสร้างกฎในการตัดสินใจในการผ่านขึ้นมา กฎการตัดสินใจนี้จะถูกวิเคราะห์อย่างมีคุณภาพเพื่อใช้กำหนดงานว่ามีความเป็นตัวแทนคล้ายกับองค์ประกอบ จากนั้นข้อสรุปจากกฎการตัดสินใจจะถูกนำมาแสดงให้คณะผู้ตัดสินพิจารณาอภิปราย และปรับปรุง การใช้ในบางครั้งได้มีการเพิ่มรอบของกฎการตัดสินใจเป็นรายบุคคลขึ้นมาเพื่อมีเป้าหมายในการสร้างฉันทมติร่วมกันในผู้ตัดสินเกี่ยวกับกฎการตัดสินใจเพียงกฎเดียว บ่อยครั้งข้อมูลผู้สอบถูกนำมาให้ผู้ตัดสินพิจารณาประกอบผลกระทบของกฎการตัดสินใจหลาย ๆ แบบ กฎการตัดสินใจถูกลงคะแนนเสียงเพื่อรับรองจากผู้ตัดสิน และเกิดการปฏิบัติของผู้สอบขึ้นต่ำที่มีคุณภาพที่จะผ่านนั้นคือ เป็นเส้นภาพ (Profiles) ทั้งหมดที่มีค่าคะแนน (Score Value) แสดงถึงจุดที่เท่ากับหรือเหนือกว่าเส้นภาพขั้นต่ำซึ่งเป็นเส้นที่ถูกลงความเห็นว่าเป็นผ่าน และเส้นภาพที่มีค่าคะแนนไม่ถึง กฎการตัดสินใจก็ตกไป

จุดดีและข้อบกพร่อง

วิธีนี้ยอมให้ผู้ตัดสินมีความยืดหยุ่นสูงในการสร้างกฎการตัดสินใจ สำหรับข้อสรุปของกระบวนการ ถ้ากลุ่มมีฉันทมติร่วมกันผู้ตัดสินก็มีแนวโน้มที่จะเกิดความรู้สึกมั่นใจว่ากฎการตัดสินใจมีความเหมาะสมสำหรับใช้ในการตัดสินใจว่าผ่าน/ตก ในทางตรงข้ามวิธีนี้ไม่มีความจำเป็นว่ากลุ่มจะต้องมีฉันทมติทั้งหมด เสียงข้างน้อยก็ได้รับการพิจารณาด้วย ผู้ตัดสินบางคน

อาจจะเข้มงวดต่อกฎการตัดสินใจขั้นสุดท้ายและบางครั้งไม่เต็มใจที่จะยอมรับความคิดเห็นของกลุ่ม สถานการณ์เกิดเมื่อค่ามีจุดที่ต่ำสุด (Minimum Point Values) เช่น ไม่ยอมให้คะแนนที่ต่ำมาก ๆ จากแบบฝึกหัดเข้าสู่กฎการตัดสินใจ บางผู้ตัดสินอาจรู้สึกมั่นใจว่า การปฏิบัติที่ได้คะแนนต่ำจะไม่คงเส้นคงวากับการรับรอง ในขณะที่ผู้ตัดสินบางคนอาจรู้สึกว่าไม่มีแบบฝึกหัดยังไม่ดีพอ แต่ก็ยังพอยอมรับได้ ถ้าในภาพรวมของแบบฝึกหัดทั้งหมดมีระดับสูงเพียงพอ จะเห็นว่าข้อบกพร่องคือ กระบวนการกำหนดมาตรฐานบางครั้งจะทำให้ได้ผลอันหมดดีของกลุ่มที่ไม่ดีพอ นอกจากนี้วิธีนี้มีศักยภาพสำหรับองค์ประกอบที่เชื่อมโยง (Conjunctive Component) ในส่วนของกฎการตัดสินใจเพราะว่า คุณภาพการวัดมีจำกัด (โดยเฉพาะความเที่ยงของคะแนน) สำหรับรูปแบบของแบบฝึกหัด

2.7 วิธีการตัดสินเชิงนโยบาย (Judgmental Policy Capturing Method)

วิธีนี้เกี่ยวข้องกับการตัดสินเส้นภาพ (Profile) ของคะแนนนักเรียนที่ได้จากการทำแบบฝึกหัด วิธีการใช้การตัดสินเชิงนโยบายมีประโยชน์มากเมื่อแต่ละแบบฝึกหัดถูกให้คะแนนที่แสดงคะแนนอย่างน้อย 3 ค่า ซึ่งดีกว่าการให้คะแนนแบบผ่าน/ตก

วิธีดำเนินการคณะผู้ตัดสินต้องได้รับการฝึกอย่างเข้มข้นเพื่อให้พวกเขาเข้าใจความหมายของแต่ละคะแนนที่นักเรียนทำในแต่ละแบบฝึกหัด เมื่อแสดงเส้นภาพคะแนนนักเรียนจากทุกแบบฝึกหัดแล้ว ผู้ตัดสินก็จะสร้างรูปภาพทางความคิดของคุณภาพผู้สอบด้านการปฏิบัติ คะแนนเส้นภาพจะถูกจำแนกอย่างอิสระเพื่อกำหนดแต่ละคะแนนเส้นภาพลงในแต่ละประเภทที่ระบุข้อความมาตรฐานการปฏิบัตินั้น ๆ (เช่น ต่ำมาก ค่อนข้างต่ำ ต่ำเล็กน้อย สูงเล็กน้อย ค่อนข้างสูง สูงมาก)

ผลการตัดสินจะถูกนำมาวิเคราะห์ แต่ละประเภทจะถูกนำมากำหนดเป็นค่าตัวเลข กระบวนการทางสถิติ (เช่นการวิเคราะห์การถดถอยพหุ) จะถูกนำมาใช้เพื่อกำหนดความสัมพันธ์ทางคณิตศาสตร์ระหว่างเส้นภาพคะแนนผู้สอบ กับประเภทการปฏิบัติ (Performance Categories) ความสัมพันธ์ดังกล่าวจะกำหนดน้ำหนักเพื่อนำไปใช้กับคะแนนของผู้สอบในแต่ละแบบฝึกหัด และสำหรับคำนวณคะแนนการปฏิบัติในผู้สอบแต่ละคน

คะแนนการปฏิบัติทั้งหมดถูกคำนวณสำหรับผู้สอบแต่ละคน เส้นภาพที่เป็นคะแนนแบบฝึกหัดถูกแสดงกับผู้ตัดสิน เส้นภาพเหล่านี้ถูกเรียงลำดับคะแนนปฏิบัติสูงสุดไปยังคะแนนปฏิบัติต่ำสุด ผู้ตัดสินจะถูกถามเพื่อกำหนดคะแนนปฏิบัติต่ำสุดที่สามารถประเมินผ่านได้ ข้อเสนอแนะจากผู้ตัดสินทั้งหมดก็ถูกนำมาจัดเป็นตาราง และการแจกแจงค่ามัธยฐานจะถูกนำมากำหนดเป็นมาตรฐานการปฏิบัติ

ขั้นตอนทั้งสองถูกนำมาทำซ้ำโดยให้โอกาสผู้ตัดสินอธิบายเหตุผลจากข้อเสนอแนะเดิมของพวกเขา มาตรฐานการปฏิบัติขั้นสุดท้ายจะยึดเสียงข้างมากของผู้ตัดสินและบางครั้งมาตรฐานการปฏิบัติอาจมีการเปลี่ยนแปลงเพียงเล็กน้อย

จุดดีและข้อบกพร่อง

จุดดีคือ ผลลัพธ์สำคัญที่ได้จากการใช้การตัดสินเชิงนโยบาย (Judgmental Policy Capturing) คือ ชุดของน้ำหนักที่กำหนดความสำคัญของแบบฝึกหัด แม้ว่ามันจะมีความเป็นไปได้ที่ใช้วิธีนี้กับโมเดลที่ประมาณการกำหนดมาตรฐานการปฏิบัติสำหรับแต่ละแบบฝึกหัดในการประเมิน ซึ่งโมเดลการวิเคราะห์การถดถอยพหุจะถูกนำมาใช้บ่งชี้กับวิธีที่เป็นการทดแทน (Compensatory) วิธีนี้ผู้สอบสามารถทดแทนงานการปฏิบัติที่ได้คะแนนต่ำในบางแบบฝึกหัดโดยแสดงคะแนนปฏิบัติที่ได้คะแนนสูงในแบบฝึกหัดอื่นแทน มาตรฐานการปฏิบัติทดแทน (Compensatory Performance Standards) ได้แสดงถึงความเที่ยงของการตัดสินใจที่เพิ่มขึ้น

สำหรับข้อบกพร่องสำคัญคือ ใช้เวลาในการฝึกผู้ตัดสินเป็นระยะเวลานานและเมื่อใช้เวลานาน ค่าใช้จ่ายก็จะสูงตามไปด้วย

โมเดลการใช้การตัดสินเชิงนโยบาย ผู้ตัดสินจะมีนโยบายการกำหนดมาตรฐาน ส่วนโมเดลกระบวนการผู้ตัดสินจะทำการแยกประเภทการปฏิบัติของผู้สอบมากกว่าที่จะถามผู้ตัดสินเกี่ยวกับนโยบายของพวกเขา ซึ่งบางผู้ตัดสินจะรู้สึกหงุดหงิดกับวิธีการแบบนี้เพราะการกระทำของพวกเขาไม่ได้สะท้อนสิ่งที่พวกเขาอ้างอิงนโยบายออกมาแต่ก็มีผู้สนับสนุนวิธีนี้เห็นว่า การกระทำของผู้ตัดสินเป็นตัวบ่งชี้การอ้างอิงนโยบายของพวกเขา

จากที่กล่าวมาสรุปได้ว่า วิธีการกำหนดมาตรฐานแบบดั้งเดิม มี 2 วิธีคือ 1) วิธีแบบสอบเป็นศูนย์กลาง (Test-Center Methods) ซึ่งแบ่งเป็น 4 วิธีย่อยคือ วิธีเองกอฟฟ์ วิธีของอีเบล วิธีของนีเดลสกี 2) วิธีผู้สอบเป็นศูนย์กลาง (Examinee-Centered Methods) แบ่งเป็น 2 วิธีย่อยคือ วิธีกลุ่มคาบเส้น และวิธีกลุ่มตรงข้าม วิธีการกำหนดมาตรฐานแนวใหม่ ซึ่งเป็นวิธีการที่เหมาะสมสำหรับการประเมินผลการปฏิบัติที่มีความหลากหลายซึ่งมีด้วยกัน 7 วิธีคือ วิธีเองกอฟฟ์แบบปรับขยาย (Extended Angoff Method) วิธีบุ๊กมาร์ค (Bookmark Method) วิธีกลุ่มตรงข้าม (Contrasting Groups Method) วิธีเลือกงานผู้สอบ (Examinee Paper Selections Method) วิธีองค์รวม (Holistic Method) วิธีเส้นภาพเด่น (Dominant Profile Method) และวิธีการนำนโยบายมาช่วยตัดสิน (Judgmental Policy Capturing Method)

ตอนที่ 3 การกำหนดคะแนนจุดตัดด้วยวิธีบูคมาร์คโดยการประยุกต์ทฤษฎีการตอบสนองข้อสอบ

ผู้วิจัยสรุปขั้นตอนการกำหนดมาตรฐานบูคมาร์คจากการสังเคราะห์เอกสารงานวิจัยที่เกี่ยวข้องดังนี้ (Buckendahl, 2002; Wang, 2003; Beretvas, 2004; Lewis, Green, Mitzel, Baum & Patz, 1998; Kiplinger, 1997; Lin, 2005; Dawber & Lewis, 2005; Wisconsin Department of Public Instruction, 2004; Cizek, Bunch & Koons, 2004; Ferrara, Johnson & Chen, 2005)

1. สร้างคู่มือการจัดเรียงข้อสอบ (Ordered Item Booklet: OIB) คู่มือการจัดเรียงข้อสอบดังกล่าว มีการจัดเรียงข้อสอบตามค่าความยาก โดยจัดเรียงข้อสอบจากข้อสอบง่ายที่สุดไปยังข้อสอบยากที่สุด มีรูปแบบการจัดเรียงข้อสอบ 1 ข้อ ต่อ 1 หน้า ดังภาพที่ 2-4 ตัวอย่างรูปแบบหน้าในคู่มือจัดเรียงข้อสอบ และภาพที่ 2-5 ตัวอย่างการเรียงข้อสอบจากข้อง่ายสุดไปยังข้อยากสุดในคู่มือจัดเรียงข้อสอบ

ข้อที่ 22

35

ระดับความสามารถที่โอกาสความน่าจะเป็นในการตอบถูก 67% คือ 1.725
(นักเรียนค่อนข้างเก่ง)

การวิเคราะห์ข้อสอบแบบอิงเกณฑ์

ผู้สอบจำนวน 30 คน ทำข้อสอบข้อที่ 1 ปรากฏว่าก่อนเรียนเนื้อหาวิชาผู้สอบทำข้อสอบถูก 5 คน หลังเรียนเนื้อหาวิชา ผู้สอบกลุ่มเดิมทำข้อสอบข้อเดิมถูก 23 คน

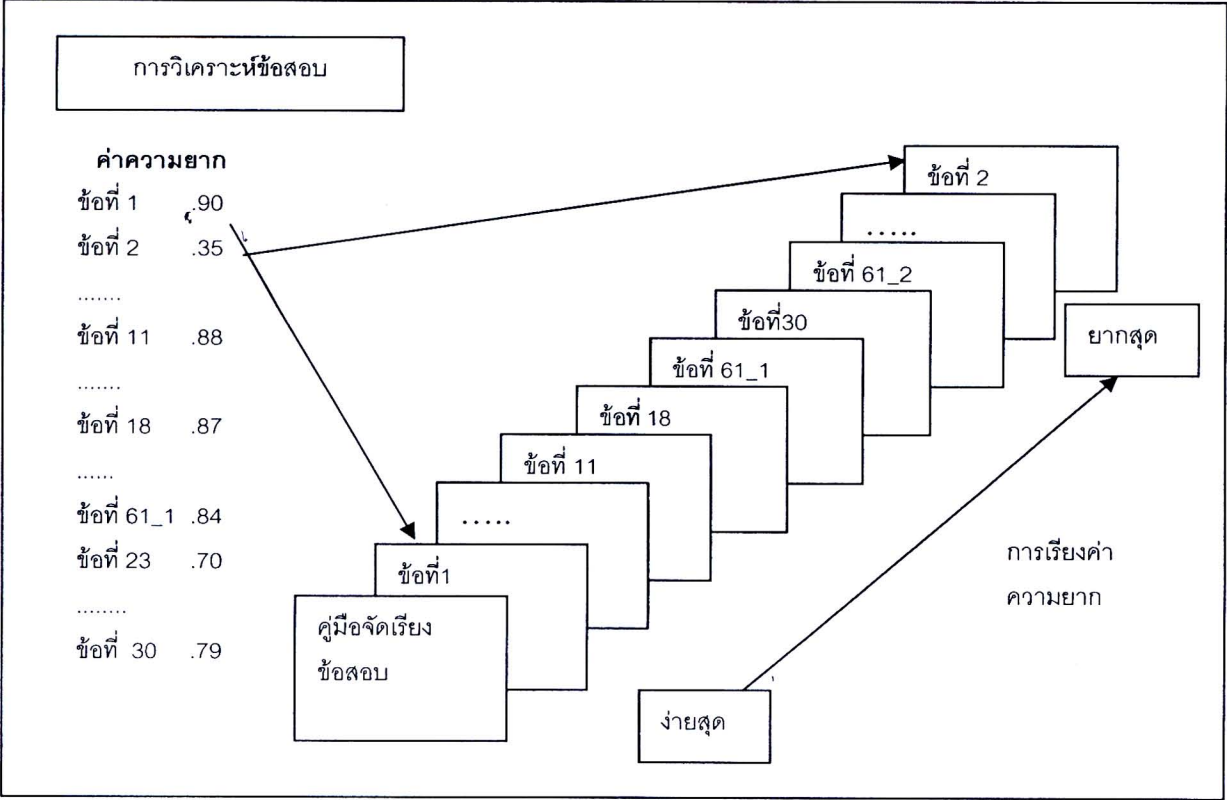
จงหาดัชนีความไวของข้อสอบ

ก. 0.17	ข. 0.60
ค. 0.76	ง. 0.84

คำตอบถูกคือ ข

ภาพที่ 2-4 ตัวอย่างรูปแบบหน้าในคู่มือจัดเรียงข้อสอบ

จากภาพที่ 2-4 ตัวเข้มบนขวา (35) เป็นตัวเลขที่แสดงเลขที่หน้าของคู่มือจัดเรียงข้อสอบ ตัวเข้มบนซ้าย (ข้อที่ 22) เป็นเลขที่ข้อสอบถัดลงมาคือ ระดับความสามารถของผู้สอบ ($\theta=1.725$) จากทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) ที่มีความน่าจะเป็นในการตอบข้อสอบข้อนี้ถูก .2/3 ข้อมูลถัดลงมาคือ ข้อมูลของข้อสอบที่แสดงข้อสอบข้อนี้เป็นเนื้อหาเรื่อง การวิเคราะห์ข้อสอบแบบอิงเกณฑ์ ส่วนบรรทัดสุดท้ายคือ คำตอบถูก



ภาพที่ 2-5 ตัวอย่างการเรียงข้อสอบจากข้อง่ายสุดไปยังข้อยากสุดในคู่มือจัดเรียงข้อสอบ

จากภาพที่ 2-5 เป็นตัวอย่างคู่มือจัดเรียงข้อสอบในงานวิจัยนี้ ประกอบด้วยข้อสอบแบบเลือกตอบจำนวน 110 ข้อ โดยเรียงลำดับจากข้อง่ายไปข้อยาก ข้อละหนึ่งหน้า และมีข้อสอบแบบอัตนัย 2 ข้อ ข้อละ 5 คะแนน โดยมีคะแนน 1, 2, 3, 4, 5 ตัวอย่างเช่น ข้อที่ 61 เป็นอัตนัย จะมีจำนวนหน้าทั้งหมด 5 หน้า โดยมีเลขที่หน้าคือ 61_1, 61_2,...,61_5 หมายความว่า ข้อสอบข้อที่ 61 ที่ระดับคะแนน 1 คะแนน, ข้อสอบข้อที่ 61 ที่ระดับคะแนน 2 คะแนน และข้อสอบข้อที่ 61 ที่ระดับคะแนน 5 คะแนน ตามลำดับ

2. ผู้ตัดสิน ร่วมสร้างนิยามความสามารถของนักเรียนใน แต่ละระดับ ว่านักเรียนแต่ละระดับควรมีความรอบรู้อย่างไร เพื่อสร้างให้เกิดความเข้าใจตรงกันในการพิจารณาตัดสินคะแนน จุดตัดในแต่ละระดับ

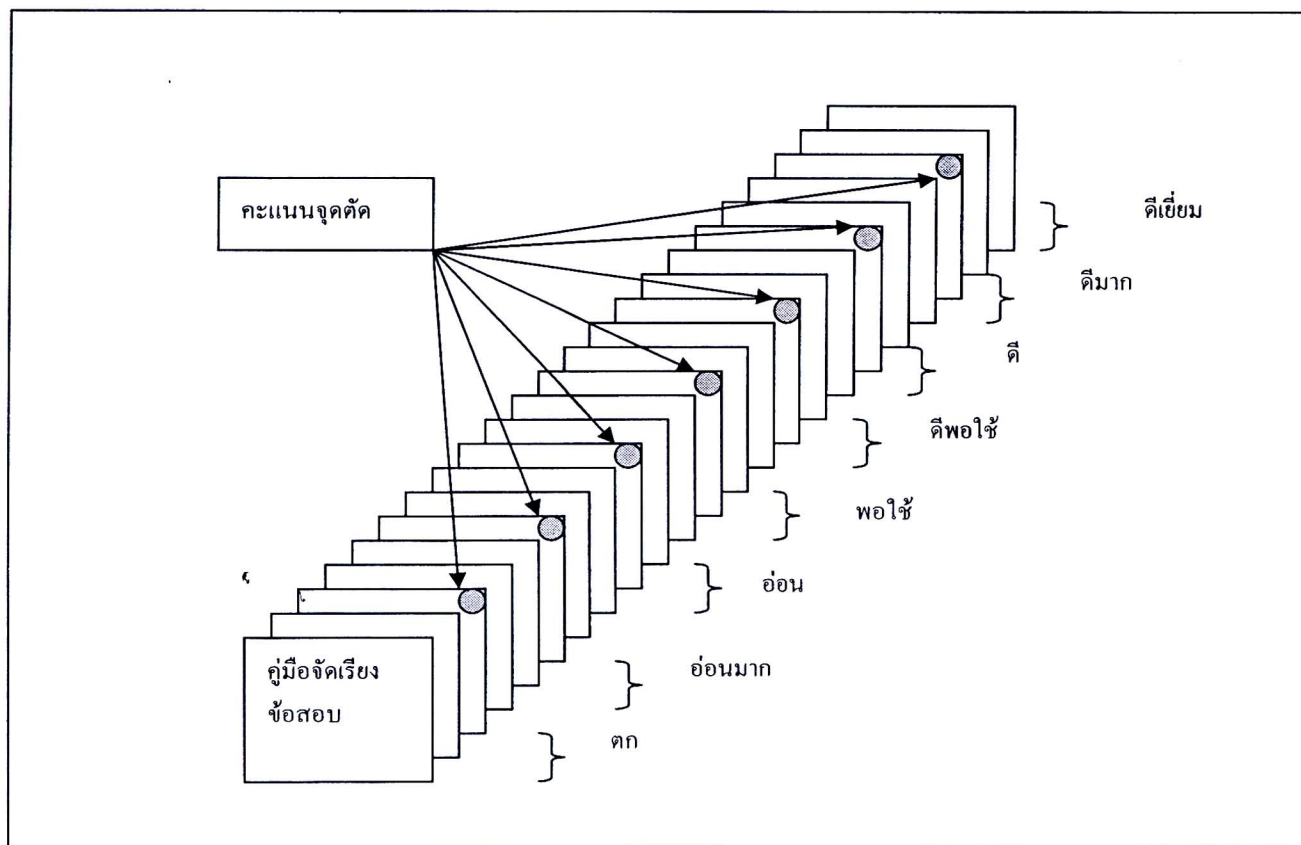
3. ผู้ตัดสินพิจารณากำหนดมาตรฐาน โดยดำเนินการพิจารณา 3 รอบ ดังนี้

รอบที่ 1 เป้าหมายหลัก คือเพื่อให้ผู้ตัดสินมีความคุ้นเคยกับคู่มือการจัดเรียงข้อสอบ ในรอบนี้จะทำงานเป็นกลุ่มย่อย ผู้ตัดสินจะทำงานอิสระจากกันนั่นคือผู้ตัดสินแต่ละคนทำการ กำหนดคะแนนจุดตัดที่ระดับพอใจเป็นจุดแรก โดยให้ผู้ตัดสินแต่ละคนพิจารณาข้อสอบทีละข้อใน คู่มือจัดเรียงข้อสอบซึ่งพิจารณาตั้งแต่หน้าแรกซึ่งเป็นข้อสอบที่ง่าย

การตอบคำถามเหล่านี้เป็นการเตรียมตัวผู้ตัดสินในการทำการตัดสินหาคะแนนจุดตัด โดยผู้ตัดสินทำการกำหนดบิคุมาร์ค (คั่นหนังสือ) บนหน้าที่ผู้ตัดสินคิดว่า “ผู้ที่คาบเส้นระดับอ่อนมากจะมีโอกาสตอบข้อสอบข้อนั้นถูก 67%” หรือมีผู้คาบเส้นระดับพอใช้ 100 คน จะมีผู้ตอบถูก 67 คน โดยผู้ตัดสินถูกฝึกให้เข้าใจว่านักเรียนที่คาบเส้นในระดับพอใช้ควรมีโอกาสตอบถูกอย่างน้อย 67% ในข้อสอบหน้าถัดไป และมีโอกาสตอบถูกสูงกว่า 67% ในหน้าก่อนหน้าที่คั่นหนังสือ

หลังจากดำเนินการกำหนดคะแนนจุดตัดที่ระดับอ่อนมากแล้ว ผู้ตัดสินดำเนินการหา คะแนนจุดตัดระดับอ่อน โดยใช้กระบวนการคล้ายกับที่กล่าวมา ในระดับนี้ผู้ตัดสินกำหนดบิคุ มาร์ค (คั่นหนังสือ) บนหน้าที่คิดว่า “ผู้ที่คาบเส้นระดับอ่อนมีโอกาสตอบข้อสอบข้อนั้นถูก 67%” จากนั้นผู้ตัดสินดำเนินการหาคะแนนจุดตัดแต่ละระดับต่อไป

เพื่อให้เห็นภาพที่ชัดเจนขึ้น ผู้วิจัยขอเสนอภาพการกำหนดบิคุมาร์ค (การคั่นหนังสือ) ดังภาพที่ 2-6



ภาพที่ 2-6 ตัวอย่างการกำหนดบุคคมารค์ (การคั่นหนังสือ) 7 ระดับ

รอบที่ 2 จัดให้มีการอภิปรายในกลุ่ม ก่อนการอภิปรายผู้ตัดสินได้รับข้อมูลเลขที่หน้าใน คู่มือจัดเรียงข้อสอบที่ผู้ตัดสินคนอื่นกำหนดบุคคมารค์ (คั่นหนังสือ) เมื่อผู้ตัดสินพิจารณาข้อมูลนั้น แล้ว ผู้ตัดสินร่วมกันอภิปรายถึงเหตุผล และความเหมาะสมของคะแนนจุดตัดที่ผู้ตัดสินแต่ละคน ตัดสินไว้ โดยเฉพาะให้ผู้ตัดสินคนที่ให้คะแนนจุดตัดสูงสุด และต่ำสุดอภิปรายเหตุผลที่ใช้คะแนน จุดตัดตรงนี้ ซึ่งการอภิปรายอาจอภิปรายในประเด็นตัวข้อสอบเกี่ยวกับระดับทักษะ และเนื้อหา ทางวิชาการที่ควรจะรอบรู้สำหรับนักเรียนในแต่ละคะแนนจุดตัด เหตุผลในการอภิปรายก็เพื่อให้ ผู้ตัดสินได้รับฟังเหตุผลของผู้ตัดสินคนอื่น การอภิปรายในกลุ่มย่อยนั้น ผู้ตัดสินแสดงความคิดเห็น และประสบการณ์ที่มีความหลากหลาย หลังจากนั้นเปิดโอกาสให้ผู้ตัดสินพิจารณาคู่มือจัดเรียง ข้อสอบอีกครั้งหนึ่งโดยนำความคิดเห็นจากผู้ตัดสินคนอื่นมาพิจารณาประกอบด้วย เพื่อหาคะแนน จุดตัดในรอบที่ 2 การพิจารณาคะแนนจุดตัดดำเนินการพิจารณาทีละระดับจนครบ เมื่อได้เลขที่ หน้าของผู้ตัดสินแต่ละคนกำหนดขึ้นมาแล้ว จากนั้นเลขที่หน้าจะถูกแปลงเป็นค่าความสามารถ

(θ) (เลขที่หน้าที่ยัดคั่นคู่มือจัดเรียงข้อสอบไว้นี้ตรงกับค่า θ เท่าไร นำค่า θ เหล่านี้มาหาค่าเฉลี่ยเพื่อนำมาประมาณความสามารถขั้นต่ำของนักเรียน)

รอบที่ 3 ผู้อำนวยการความสะดวกนำข้อมูลผลกระทบ (Impact Data) คือ เปอร์เซ็นต์ของนักเรียนที่อยู่ในแต่ละระดับ ค่าเฉลี่ยคะแนนจุดตัดจากกลุ่มย่อยทั้งสองกลุ่ม มาแสดงให้ผู้ตัดสินพิจารณา จากนั้นผู้ตัดสินอภิปรายร่วมกันโดยพิจารณาข้อสอบที่อยู่ระหว่างค่าเฉลี่ยคะแนนจุดตัดทั้งสองกลุ่ม แล้วเปิดโอกาสให้ผู้ตัดสินทำการกำหนดคะแนนจุดตัดอีกครั้งอย่างอิสระจากกัน โดยการกำหนดหน้าที่ยัดคั่นหนังสือในคู่มือจัดเรียงข้อสอบ เลขที่หน้าที่ยัดคั่นของผู้ตัดสินแต่ละคนที่กำหนดขึ้นมานั้นถูกแปลงเป็นค่าความสามารถ (θ) (เลขที่หน้าที่ยัดคั่นคู่มือจัดเรียงข้อสอบไว้นี้ตรงกับค่า θ เท่าไร แล้วนำค่า θ เหล่านี้มาหาค่าเฉลี่ยเพื่อนำมาประมาณความสามารถขั้นต่ำของนักเรียน

4. เมื่อได้มาตรฐานแต่ละระดับแล้ว จากนั้นผู้ตัดสินระดมพลังสมองเพื่อเขียนบรรยายระดับการปฏิบัติของนักเรียนในประเด็นความรู้ ทักษะ และความสามารถของนิสิต ณ ระดับคะแนนจุดตัดของระดับนั้น

สำหรับสูตรการหาค่า θ ในงานวิจัยนี้ มาจาก Wright and Stone (1979 cited in Cizek & Bunch, 2007) ได้เสนอสูตรดั้งเดิมแบบโมเดล Rasch สำหรับข้อสอบแบบเลือกตอบ ที่มีระบบการให้คะแนนเป็น 1 0 โดยกำหนดความน่าจะเป็นในการตอบถูก, $P(X=1)$ เมื่อกำหนดค่าความยาก คือ β_j และความสามารถของผู้สอบ (θ_i) ดังนี้

$$P(x = 1/\theta_i, \beta_j) = \exp(\theta_i - \beta_j) / [1 + \exp(\theta_i - \beta_j)] \quad (1)$$

เมื่อกำหนด β_j = ค่าความยากของข้อสอบ

θ_i = ความสามารถของผู้สอบ

P = ความน่าจะเป็นของการตอบถูก

exp = natural logarithm e (2.71828...)

เมื่อ แทนค่า $p = 2/3$ แล้วแก้สมการหา θ_i ดังนี้

$$\exp(\theta_i - \beta_j) / [1 + \exp(\theta_i - \beta_j)] = 2/3 \quad (2)$$

$$\exp(\theta_i - \beta_j) = 2/3 * [1 + \exp(\theta_i - \beta_j)] \quad (3)$$

$$\exp(\theta_i - \beta_j) = 2/3 + 2/3 * \exp(\theta_i - \beta_j) \quad (4)$$

$$\exp(\theta_i - \beta_j) - 2/3 * \exp(\theta_i - \beta_j) = 2/3 \quad (5)$$

$$1/3 * \exp(\theta_i - \beta_j) = 2/3 \quad (6)$$

$$\exp((\theta_i - \beta_j) = 2/3 \quad \div \quad 1/3 \quad (7)$$

$$\exp((\theta_i - \beta_j) = 2 \quad (8)$$

$$\theta_i - \beta_j = .693 \quad (9)$$

$$\theta_i = \beta_j + .693 \quad (10)$$

สำหรับสูตรการหาค่า θ จากข้อสอบแบบให้คะแนนหลายค่า เช่น ข้อสอบอัตนัยที่ใช้ในงานวิจัยนี้ มาจาก Wright and Masters (1982 cited in Cizek & Bunch, 2007) ได้เสนอสูตรตามกรอบแนวคิดของ Rasch โมเดล Partial-Credit Model (PCM) โดยยกตัวอย่างข้อสอบที่ให้คะแนนแบบ 6 ค่า (0, 1, 2, 3, 4, 5) ซึ่งการคำนวณจากสูตรข้างล่างนี้ ผู้วิจัยใช้โปรแกรม EXCEL ช่วยในการคำนวณ

$$\text{กำหนด } \pi_{nix} = \frac{\exp \sum (\theta_n - \delta_{ij})}{\sum \exp \sum (\theta_n - \delta_{ij})} \quad (11)$$

เมื่อกำหนด π_{nix} = likelihood ของผู้สอบแต่ละคน

θ_n = ความสามารถของผู้สอบ

δ_{ij} = ค่าความยากของข้อสอบข้อที่ i ณ คะแนน j

i = ข้อสอบ

j = คะแนน

เมื่อคะแนนเท่ากับ 0 ค่า $\delta_{i0} \equiv 0$ แทนค่าในสมการ 11 ได้ดังนี้

$$\sum (\theta_n - \delta_{ij}) = 0 \text{ และ } \exp \sum (\theta_n - \delta_{ij}) = 1 \quad (12)$$

$$\begin{aligned} \text{ขั้นที่ 1 } \sum (\theta_n - \delta_{ij}) &= \sum (\theta_n - \delta_{i0}) + \theta_n - \delta_{i1} \\ &= 0 + \theta_n - \delta_{i1} \\ &= \theta_n - \delta_{i1} \end{aligned} \quad (13)$$

ขั้น 2 - 4 ทำคล้ายขั้น 1 ได้สมการดังนี้

$$\text{ขั้นที่ 2 } \sum (\theta_n - \delta_{ij}) = 2\theta_n - \delta_{i1} - \delta_{i2} \quad (14)$$

$$\text{ขั้นที่ 3 } \sum (\theta_n - \delta_{ij}) = 3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} \quad (15)$$

$$\text{ขั้นที่ 4 } \sum (\theta_n - \delta_{ij}) = 4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4} \quad (16)$$

$$\text{ขั้นที่ 5 } \sum (\theta_n - \delta_{ij}) = 4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4} - \delta_{i5} \quad (17)$$

นำค่า exponential มาคูณค่า summation ของสมการข้างต้นได้ดังนี้

$$\text{ขั้นที่ 1 คะแนนเท่ากับ 0} \quad \exp(0) \quad (18)$$

$$\text{ขั้นที่ 2 คะแนนเท่ากับ 1} \quad \exp(\theta_n - \delta_{i1}) \quad (19)$$

$$\text{ขั้นที่ 3 คะแนนเท่ากับ 2} \quad \exp(2\theta_n - \delta_{i1} - \delta_{i2}) \quad (20)$$

$$\text{ขั้นที่ 4 คะแนนเท่ากับ 3} \quad \exp(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3}) \quad (21)$$

$$\text{ขั้นที่ 5 คะแนนเท่ากับ 4} \quad \exp(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4}) \quad (22)$$

$$\text{ขั้นที่ 6 คะแนนเท่ากับ 5} \quad \exp(5\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4} - \delta_{i5}) \quad (23)$$

นอกจากนี้ กรณีที่การวิเคราะห์ข้อสอบ เพื่อจัดเตรียมคู่มือการจัดเรียงข้อสอบนั้น ต้องหาค่าความสามารถ (θ) โดยการนำค่าพารามิเตอร์ความยาก อำนาจจำแนก และการเดาจากการวิเคราะห์ด้วยทฤษฎี IRT มาแทนค่าในสูตรเพื่อวิเคราะห์หาค่า θ

สำหรับสูตรหาค่า θ มาจาก Mitzel et al. (2001 cited in Cizek, Bunch & Koons, 2004) ซึ่งได้เสนอสูตรดั้งเดิมแบบโมเดล 3 พารามิเตอร์ สำหรับข้อสอบแบบเลือกตอบ ดังนี้

$$P_j(\theta) = \frac{c_j + (1 - c_j)}{1 + \exp[-1.7a_j(\theta - b_j)]} \quad (1)$$

เมื่อ

P_j	=	ความน่าจะเป็นของการตอบถูก
θ	=	ฟังก์ชันความสามารถของผู้สอบ
b_j	=	ความยากของข้อสอบ
a_j	=	อำนาจจำแนกข้อสอบ
c_j	=	threshold หรือ ตัวแปรการเดา (chance variable)
exp	=	natural logarithm e (2.71828...)

อย่างไรก็ตาม Mitzel et al. (2001 cited in Cizek, Bunch & Koons, 2004) ได้กำหนดให้ $c_j = 0$ จากสมการ (1) จะได้

$$P_j(\theta) = \frac{1}{1 + \exp[-1.7a_j(\theta - b_j)]} \quad (2)$$

เนื่องจากกำหนดข้อคำถามว่า นักเรียนที่คาบเส้นมีโอกาสตอบข้อสอบถูกต้องที่ความน่าจะเป็น .67 ซึ่งความน่าจะเป็น .67 นี้นำมาแก้สมการในสมการ (2) สำหรับค่า θ ที่ต้องการให้ตอบข้อสอบถูกต้อง ดังนี้

$$0.67 = \frac{1}{1 + e^{-1.7a_j(\theta - b_j)}}$$

$$1 + e^{-1.7a_j(\theta - b_j)} = \frac{1}{0.67}$$

$$e^{-1.7a_j(\theta - b_j)} = 1.4925 - 1$$



take ln ทั้งสองข้าง $-1.7a_j(\theta - b_j) = \ln 0.4925$

$$\theta - b_j = \frac{\ln 0.4925}{-1.7a_j}$$

$$\theta = b_j + \frac{\ln 0.4925}{-1.7a_j}$$

$$\theta = b_j + \frac{.708}{1.7a_j} \quad (3)$$

ตอนที่ 4 แบบสอบวัดผลสัมฤทธิ์ทางการเรียน

ผู้วิจัยแบ่ง ตอนที่ 4 แบบสอบวัดผลสัมฤทธิ์ทางการเรียนออกเป็น หัวเรื่องย่อยดังนี้ (สุริพร อนุศาสนนันท์, 2553)

1. ประเภทของแบบสอบผลสัมฤทธิ์ทางการเรียน
2. ขั้นตอนการสร้างแบบสอบผลสัมฤทธิ์ทางการเรียน
3. การสร้างข้อสอบแบบเลือกตอบ
4. การสร้างข้อสอบแบบเขียนตอบ

โดยมีรายละเอียดดังต่อไปนี้

1. ประเภทของแบบสอบผลสัมฤทธิ์ทางการเรียน

การจัดประเภทของแบบสอบ ได้มีเกณฑ์เพื่อแบ่งประเภทของแบบสอบออกเป็น 10 แบบคือ

ธรรมชาติของสิ่งเร้า

1) **แบบสอบที่ใช้ภาษา (Verbal Tests)** เป็นแบบสอบที่ใช้ภาษาของสังคมนั้น ๆ เพื่อเขียนสื่อความหมายกับผู้สอบ เหมาะสำหรับผู้ที่อ่านออกเขียนได้ แบบสอบประเภทนี้นำมาใช้สอบทั่ว ๆ ไป

2) **แบบสอบที่ไม่ใช้ภาษา (Non-Verbal Tests)** เป็นแบบสอบที่ไม่ใช้ภาษา แต่จะใช้รูปภาพ สัญลักษณ์ รูปทรงเรขาคณิตต่าง ๆ ของจริง ของจำลอง เหมาะสำหรับผู้ที่อ่านไม่ออกเขียนไม่ได้ เช่น ใช้วัดความพร้อมสำหรับเด็กนักเรียนระดับปฐมวัย แต่ก็นำมาใช้กับผู้อ่านออกและเขียนได้เช่นกัน ในกรณีเพื่อลดความแตกต่างทางวัฒนธรรมที่เกิดจากภาษา แบบสอบประเภทนี้จึงสามารถนำไปทดสอบได้กับคนทุกชาติทุกภาษา

เวลาที่ใช้ในการสอบ

1) **แบบสอบที่ใช้ความเร็วในการตอบ (Speed Test)** เป็นแบบสอบที่จำกัดเวลาในการตอบ โดยมีข้อสอบให้จำนวนมากและให้เวลาในการตอบช่วงสั้น ๆ เช่น ให้ข้อสอบมา 100 ข้อ ตอบภายใน 30 นาที ดังนั้นข้อสอบประเภทนี้จะเป็นข้อสอบที่ง่าย มักใช้วัดความถนัดของผู้เรียน

2) **แบบสอบไม่จำกัดเวลาในการตอบ (Power Test)** เป็นแบบสอบที่ให้เวลาตอบข้อสอบมาก หรือบางครั้งก็ไม่จำกัดเวลาในการตอบ ข้อสอบจึงค่อนข้างยาก และมีจำนวนข้อสอบไม่มาก ข้อสอบประเภทนี้จะใช้วัดการคิดขั้นสูง เช่น วิเคราะห์ สังเคราะห์ ประเมินค่า

วิธีดำเนินการสอบ

1) **แบบสอบรายบุคคล (Individual Test)** เป็นแบบสอบที่ใช้สอบทีละคนเพื่อใช้วัดพฤติกรรมที่ซับซ้อน เช่น การสอบภาคปฏิบัติที่ต้องประเมินทั้งกระบวนการ และผลงาน เพื่อสังเกตการณ์การปฏิบัติงานทุกขั้นตอน ดังนั้นวิธีนี้เหมาะสำหรับการสอบที่มีผู้สอบไม่มากนัก

2) **แบบสอบรายกลุ่ม (Group Test)** เป็นแบบสอบที่ใช้กับผู้สอบจำนวนมาก เช่น การสอบในห้องเรียน การสอบแข่งขัน สามารถจัดสอบพร้อมกันทั้งโรงเรียน และทั่วประเทศได้

วิธีการตอบ

1) **แบบสอบปากเปล่า (Oral Test)** เป็นการใช้แบบสอบที่ใช้การพูดแบบตัวต่อตัว แทนการเขียน เหมาะสำหรับผู้สอบที่อ่านไม่ออกเขียนไม่ได้ และเด็กเล็ก เนื่องจากค่อนข้างใช้เวลาในการทดสอบ ดังนั้นจึงมีข้อสอบจำนวนไม่มากนัก

2) **แบบสอบแบบเขียนตอบ (Paper-Pencil Test)** เป็นแบบสอบที่ผู้สอบเขียนคำตอบลงในกระดาษ เป็นแบบสอบที่ใช้กันอย่างแพร่หลาย

3) **แบบสอบภาคปฏิบัติ (Performance Test)** เป็นแบบสอบที่ให้ผู้สอบลงมือปฏิบัติงานจริง เช่น การปลูกต้นไม้ การถักไหมพรม การซ่อมเครื่องยนต์ เพื่อครูจะได้สังเกตกระบวนการทำงาน ตลอดจนผลงานของผู้เรียน

ลักษณะการให้คะแนน

1) **แบบสอบปรนัย (Objective Tests)** เป็นแบบสอบที่มีคำถามและคำตอบให้เลือกมาให้ ผู้เรียนก็จะเลือกข้อที่คิดว่าถูก การให้คะแนนมีความเป็นปรนัย ตรวจให้คะแนนก็ครั้งก็ได้ผลคะแนนเหมือนเดิม แบบสอบปรนัย คือ แบบเลือกหลายตอบ จับคู่ ถูก-ผิด

2) **แบบสอบอัตนัย (Subjective Test)** เป็นแบบสอบที่ให้มาเฉพาะข้อคำถาม ส่วนผู้เรียนจะต้องเขียนคำตอบเอง การให้คะแนนไม่ค่อยมีความเป็นปรนัย คะแนนที่ได้ขึ้นอยู่กับผู้ตรวจ แบบสอบอัตนัย คือ แบบเขียนตอบ แบบตอบสั้น แบบเติมคำ

ความเป็นมาตรฐานของแบบสอบ

1) **แบบสอบที่ครูสร้างขึ้นเอง (Teacher-Made Test)** เป็นแบบสอบที่ครูสร้างขึ้นมาเพื่อวัดผลการเรียนรู้ของผู้เรียน เมื่อผู้เรียนได้เรียนเนื้อหาแล้ว การสร้างแบบสอบประเภทนี้อาจจะไม่ได้มาตรฐานเนื่องจากสร้างขึ้นมาใช้ในห้องเรียน ไม่ได้วิเคราะห์คุณภาพของข้อสอบและแบบสอบเนื่องจากข้อสอบอาจจะรวบได้

2) **แบบสอบมาตรฐาน (Standardized Tests)** เป็นแบบสอบที่สร้างขึ้นอย่างมีขั้นตอน ได้มาตรฐานและได้ผ่านการวิเคราะห์ข้อสอบและแบบสอบเพื่อพัฒนาคุณภาพของข้อสอบให้มีคุณภาพ โดยทั่วไปแบบสอบมาตรฐานจะมีการหาเกณฑ์ปกติวิสัย (Norm) เพื่อใช้เปรียบเทียบความสามารถกับกลุ่ม เช่น เกณฑ์ปกติวิสัยระดับจังหวัด เกณฑ์ปกติวิสัยระดับประเทศ

การแปลความหมายคะแนน

1) **แบบสอบอิงกลุ่ม (Norm-Referenced Test)** เป็นแบบสอบที่นำคะแนนจากการสอบไปเปรียบเทียบกับกลุ่ม (เช่น คะแนนเฉลี่ยของกลุ่ม) ส่วนใหญ่มักเป็นแบบสอบที่ใช้แข่งขัน

2) **แบบสอบอิงเกณฑ์ (Criterion-Referenced Test)** เป็นแบบสอบที่มุ่งตรวจสอบความรู้ (Master) ของผู้เรียนในเนื้อหานั้น ๆ ว่าคะแนนถึงตามเกณฑ์ขั้นต่ำที่กำหนดไว้หรือไม่ ถ้าคะแนนถึงตามเกณฑ์ขั้นต่ำก็จะจัดผู้เรียนคนนั้นว่ารอบรู้ในเนื้อหานั้น สามารถเรียนเนื้อหาถัดไปหรือเรียนในระดับต่อไปได้ แต่ถ้าคะแนนสอบไม่ถึงตามเกณฑ์ขั้นต่ำที่กำหนดจะถือว่า ไม่รอบรู้ ผู้สอนก็จะจัดซ่อมเสริมต่อไป

วัตถุประสงค์ที่มุ่งวัด

1) **แบบสอบผลสัมฤทธิ์ทางการเรียน (Achievement Test)** เป็นแบบสอบที่ใช้วัดผลการเรียนรู้ที่ผู้เรียนได้เรียนเนื้อหาวิชานั้น ๆ แล้ว ว่าผู้เรียนได้เกิดการเรียนรู้มากน้อยเพียงใด

2) **แบบสอบวัดความถนัด (Aptitude Test)** เป็นแบบสอบที่ใช้วัดความสามารถที่จะเรียนรู้ในอนาคต โดยจะพยากรณ์ว่าผู้เรียนคนนี้จะมีความสามารถที่จะเรียนรู้เพียงใด แบบสอบวัดความถนัด 2 ชนิด คือ

- แบบสอบวัดความถนัดทางการเรียน (Scholastic Aptitude Test) เป็นแบบสอบที่ใช้วัดความถนัดทางวิชาการ แบบสอบจะวัด 2 ส่วนคือ ด้านภาษา และด้านคณิตศาสตร์

- แบบสอบวัดความถนัดเฉพาะ (Specific Aptitude Test) เป็นแบบสอบที่วัดความถนัดทางวิชาชีพ เช่น ความสามารถทางเครื่องยนต์ ความสามารถทางศิลปะ เป็นต้น

3) **แบบสอบวัดความสามารถ (Ability Test)** เป็นแบบสอบที่วัดความสามารถในการทำงานที่เป็นอยู่ในปัจจุบัน

ลักษณะของการใช้ประโยชน์

1) **แบบสอบระหว่างเรียน (Formative Test)** เป็นแบบสอบที่วัดเนื้อหาที่ละบทเรียน โดยการสร้างข้อสอบจะสร้างตามจุดประสงค์การเรียนรู้ ผลคะแนนที่ได้ครูจะนำมาใช้ในการปรับปรุงการเรียนการสอน เช่น การจัดสอนซ่อมเสริมแก่นักเรียนที่ไม่ผ่าน จัดกิจกรรมการเรียนรู้ให้เหมาะสมกับนักเรียนกลุ่มนั้น

2) **แบบสอบหลังการเรียนการสอน (Summative Test)** เป็นแบบสอบที่นำเนื้อหาที่เรียนทั้งหมดมาทดสอบ ข้อสอบจะวัดเนื้อหากว้าง ๆ ไม่ลงรายละเอียดเหมือนแบบสอบระหว่างเรียน ผลการสอบที่ได้จะนำไปใช้ในการตัดสินผลการเรียนรู้ของนักเรียน

ที่มาของคำตอบ

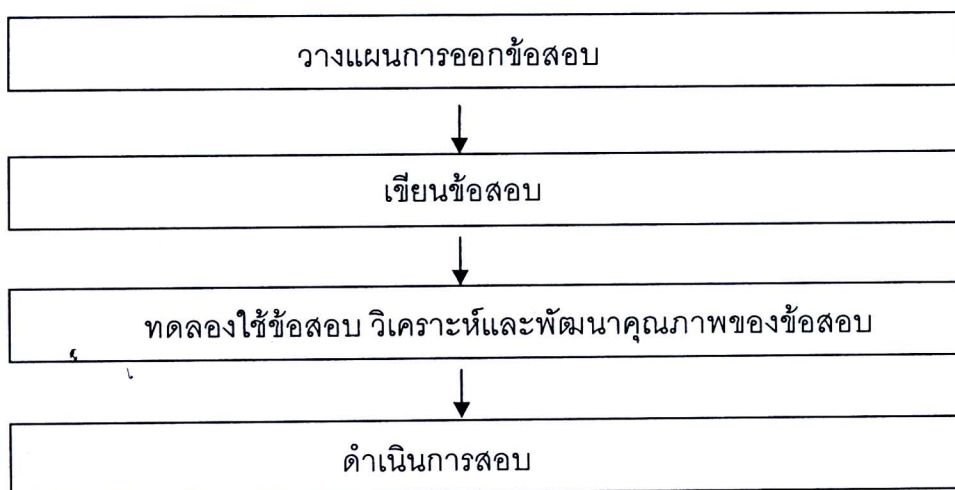
1) **แบบมีคำตอบให้เลือก (Select Response Test)** เป็นแบบสอบที่กำหนดคำตอบไว้ให้ผู้สอบเลือกคำตอบ มีด้วยกัน 3 แบบ คือ

- แบบถูกผิด (True-False)
- แบบจับคู่ (Matching)
- แบบเลือกตอบ (Multiple Choice)

2) **แบบผู้สอบเขียนคำตอบขึ้นมา (Supply Response Test)** เป็นแบบสอบที่ผู้สอบเขียนคำตอบขึ้นมาเอง อาจจะเขียนสั้น หรือยาว ตามลักษณะข้อสอบ 4 แบบดังนี้

- แบบเติมคำ (Completion)
- แบบตอบสั้น (Short Answer Items)

- แบบจำกัดคำตอบ (Restricted Response Item)
 - แบบไม่จำกัดคำตอบ (Extended Response Item)
2. ขั้นตอนการสร้างแบบสอบผลสัมฤทธิ์ทางการเรียน



ภาพที่ 2-7 ขั้นตอนการสร้างแบบสอบผลสัมฤทธิ์ทางการเรียน

รายละเอียดแต่ละขั้นตอนในการสร้างแบบสอบผลสัมฤทธิ์ทางการเรียน มีดังนี้

1) วางแผนการออกข้อสอบ

- กำหนดจุดมุ่งหมายของการทดสอบ โดยการกำหนดคุณลักษณะที่ต้องการวัด พร้อมทั้งระบุพฤติกรรมที่ต้องการให้ผู้เรียนเป็นไปตามที่คาดหวัง
- ระบุเนื้อหาที่จะทำการทดสอบ เพื่อทำการแยกเนื้อหาเป็นหัวข้อย่อย
- กำหนดประเภทของแบบสอบในแต่ละหัวข้อย่อย
- สร้างตารางวิเคราะห์แผนงานการประเมิน
- กำหนดน้ำหนักความสำคัญ หรือ สัดส่วนของข้อสอบที่จะสร้าง

2) เขียนข้อสอบ

การเขียนข้อสอบที่ดี มีดังนี้

- ควรเขียนข้อสอบให้ตรงตามเนื้อหาที่ต้องการวัด และตรงตามระดับพฤติกรรมที่ต้องการวัด เช่น จากตารางวิเคราะห์แผนงานประเมิน ต้องออกข้อสอบจำนวน 2 ข้อ เป็นเนื้อหาเรื่อง กราฟแสดงความสัมพันธ์ระหว่างปริมาณที่มีความสัมพันธ์เชิงเส้น วัดพฤติกรรมระดับการนำไปใช้ ผู้ออกข้อสอบต้องนำจุดประสงค์เชิงพฤติกรรมในเรื่องดังกล่าวมาพิจารณาใช้สร้างข้อสอบ

เช่น นักเรียนสามารถบอกได้ว่าจุดที่แทนคู่อันดับที่กำหนดให้อยู่บนกราฟเส้นตรงที่กำหนดให้ได้อย่างถูกต้อง สามารถสร้างข้อสอบได้ดังนี้

จุด (9, 12) และ (9, 36) อยู่บนเส้นตรงเส้นหนึ่ง จุดใดอยู่บนเส้นตรงเส้นนี้

ก. (200, 12)

ข. (0, 0)

ค. (9, 108)

ง. (0, 36)

- การเขียนข้อสอบควรใช้ภาษาที่ชัดเจน ไม่กำกวม นักเรียนอ่านข้อสอบแล้วเข้าใจตรงกัน

- ควรเขียนข้อสอบเกินจำนวนที่ต้องการเผื่อไว้ เพราะหลังจากวิเคราะห์ข้อสอบจะมีการตัดข้อสอบที่ไม่ได้คุณภาพออกไป

3) ทดลองใช้ข้อสอบ วิเคราะห์และพัฒนาคุณภาพของข้อสอบ นำแบบสอบที่สร้างไปทดลองใช้กับนักเรียนที่คล้ายกับกลุ่มนักเรียนที่จะสอบจริง การวิเคราะห์ข้อสอบจะหาค่าอำนาจจำแนก และค่าความยากของข้อสอบแต่ละข้อ คัดเลือกข้อสอบตามเกณฑ์ของค่าอำนาจจำแนก และค่าความยาก ส่วนข้อสอบที่ไม่เป็นไปตามเกณฑ์ดังกล่าวนำมาปรับปรุงแก้ไข หรือตัดทิ้ง

4) ดำเนินการสอบ จัดเตรียมแบบสอบสำหรับดำเนินการสอบให้เกินจำนวนผู้สอบประมาณ 5% เพื่อใช้สำรองกรณีแบบสอบบางฉบับมีปัญหาเช่น บางหน้าหาย บางหน้าไม่ชัดเจน เป็นต้น จัดเตรียมสถานที่สอบ กำหนดเวลาในการสอบที่เหมาะสมโดยขึ้นอยู่กับประเภทของข้อสอบเช่น แบบสอบประเภทใช้ความเร็วในการตอบ (Speed Test) จะจำกัดเวลาในการตอบ ส่วนแบบสอบที่ไม่จำกัดเวลาในการตอบ (Power Test) จะให้เวลาที่ผู้สอบส่วนใหญ่สามารถทำข้อสอบได้ครบ

3. การสร้างข้อสอบแบบเลือกตอบ

ข้อสอบแบบเลือกตอบ ประกอบด้วยส่วนข้อคำถาม (Stem) และส่วนตัวเลือก (Choice) โดยส่วนตัวเลือกประกอบด้วยตัวถูก (Key) และตัวลวง (Distractors)

ข้อดีของข้อสอบแบบเลือกตอบ

- 1) สามารถวัดเนื้อหาได้ครอบคลุม เพราะผู้สอบใช้เวลาในการทำข้อสอบแต่ละข้อไม่มากนัก จึงทำให้ออกข้อสอบได้หลายข้อ
- 2) ตรวจให้คะแนนง่าย และรวดเร็ว จึงเหมาะที่จะทดสอบกับการประเมินขนาดใหญ่ที่มีผู้สอบจำนวนมาก เช่น การสอบแข่งขันทั่วประเทศ

3) การตรวจให้คะแนนมีความยุติธรรม มีความเป็นปรนัย คือ ไม่ว่าใครตรวจให้คะแนน ก็ให้ผลคะแนนตรงกัน

4) สามารถสร้างให้วัดสมรรถภาพสมองด้านต่าง ๆ ตามลำดับขั้นของความรู้ได้

จุดบกพร่องของข้อสอบแบบเลือกตอบ

1) เสียค่าใช้จ่ายในการสร้างและพัฒนาสูง เนื่องจากข้อสอบแต่ละข้อใช้เวลาในการสร้างนาน และใช้กระดาษหลายหน้าในการจัดทำเป็นแบบสอบ

2) สามารถเดาคำตอบถูกได้ เนื่องจากมีการกำหนดคำตอบให้เลือก ดังนั้นเพื่อลดโอกาสในการเดาถูก จึงควรสร้างข้อสอบที่ตัวเลือกหลายตัวเลือก เช่น 5 ตัวเลือก สำหรับนักเรียนระดับมัธยมศึกษาขึ้นไป

หลักการเขียนข้อสอบแบบเลือกตอบ

1) เทคนิคการเขียนข้อคำถาม

(1) เขียนคำถามให้เป็นประโยคคำถามที่สมบูรณ์

(ไม่ดี) ภูเก็ตเป็น

ก. เกาะ

ข. ภูเขา

ค. จังหวัด

ง. ท่าจอดเรือ

จ. แหล่งท่องเที่ยว

(ดีขึ้น) ภูเก็ตเป็นชื่อของอะไร

ก. ถนน

ข. ภูเขา

ค. แม่น้ำ

ง. เมือง

จ. แหล่งท่องเที่ยว

(2) เน้นเรื่องที่จะถามให้ชัดเจนและตรงจุด

(ไม่ดี) ภาคเหนือของประเทศไทย (ดีขึ้น) ภาคเหนือของประเทศไทยมีภูมิประเทศ

ก. อากาศหนาว

ข. เป็นเทือกเขา

ค. ดอกไม้สวย

ง. เป็นที่ราบ

จ. มีวัฒนธรรมล้านนา

เป็นอย่างไร

ก. เป็นเทือกเขาสูง

ข. เป็นที่ราบลุ่ม

ค. เป็นที่ต้ำน้ำท่วมถึง

ง. เป็นที่ราบสูง

จ. เป็นทุ่งหญ้า

(3) ควรถามข้อละปัญหา

(ไม่ดี) กรุงเทพมหานครเป็นชื่อของอะไร อยู่ติดกับแม่น้ำอะไร	(ดีขึ้น) กรุงเทพมหานครอยู่ติดกับ แม่น้ำอะไร
ก. อุเรือ – เจ้าพระยา	ก. เจ้าพระยา
ข. เมือง – ท่าจีน	ข. ท่าจีน
ค. ถนน – บางปะกง	ค. บางปะกง
ง. เมือง – เจ้าพระยา	ง. ปิง
จ. ภาพ – ป่าสัก	จ. ป่าสัก

(4) ไม่ควรใช้คำถามปฏิเสธ หรือซ้อนปฏิเสธ

(ไม่ดี) ถ้าไม่กินผักบุ้งจะไม่ได้วิตามินอะไร	(ดีขึ้น) ถ้ากินผักบุ้งจะได้รับวิตามินอะไร
ก. วิตามินเอ	ก. วิตามินเอ
ข. วิตามินบี	ข. วิตามินบี
ค. วิตามินซี	ค. วิตามินซี
ง. วิตามินดี	ง. วิตามินดี
จ. วิตามินอี	จ. วิตามินอี

(5) ถามคำถามที่ให้เกิดความคิด

(ไม่ดี) ส่วนใดของพืชทำหน้าที่รับออกซิเจน	(ดีขึ้น) คน : จมูก , พืช :
ก. ใบ	ก. ใบ
ข. ดอก	ข. ดอก
ค. ราก	ค. ราก
ง. ลำต้น	ง. ลำต้น
จ. กิ่ง	จ. กิ่ง

(6) คำถามไม่ใช่คำฟุ่มเฟือย

(ไม่ดี) มนุษย์มีวิธีเพิ่มพูนสมรรถภาพ
ให้แก่ตนเองเพื่อต่อต้านโรคภัยที่
ร้ายแรงได้หลายวิธี จงบอกวิธีที่
สำคัญที่สุดที่เราควรปฏิบัติเพื่อ
ต่อต้านการติดต่อของโรค
ก. ฉีดวัคซีน
ข. อยู่ห่างผู้ป่วย
ค. กินวิตามิน
ง. กินนอนเป็นเวลา
จ. ออกกำลังกาย

(ดีขึ้น) วิธีการปฏิบัติข้อใดดีที่สุด เพื่อ
ต่อต้านโรคติดต่อ
ก. ฉีดวัคซีน
ข. อยู่ห่างผู้ป่วย
ค. กินวิตามิน
ง. กินนอนเป็นเวลา
จ. ออกกำลังกาย

(7) คำถามไม่ควรชี้แนะคำตอบ นั่นคือ มีคำบางคำในข้อคำถามที่ชี้นำไปสู่คำตอบ

(ไม่ดี) บุคคลใดมีเชื้อชาติเป็นชาว
ต่างประเทศ
ก. พระยาจักรี
ข. หมอบรัดเลย์
ค. นายทองอินทร์
ง. ขุนหลวงพะงั่ว
จ. เจ้าพระยาพระคลัง (หน)

(ดีขึ้น) บุคคลใดมีเชื้อชาติเป็นชาว
ต่างประเทศ
ก. พระยาจักรี
ข. ขุนหลวงพะงั่ว
ค. นายทองอินทร์
ง. เจ้าพระยาพระคลัง (หน)
จ. เจ้าพระยาวิชาเยนทร์

(8) ควรถามในสิ่งที่สร้างสรรค์ ไม่ควรถามที่ชี้นำไปสู่สิ่งที่ไม่ดี

(ไม่ดี) สิ่งเสพติดใดไม่ผิดกฎหมาย
ก. ฝิ่น
ข. บุหรี่
ค. กัญชา
ง. มอร์ฟีน
จ. เฮโรอีน

(ดีขึ้น) การสูบบุหรี่ก่อให้เกิดโรคอะไร
ก. ไตวาย
ข. ตับแข็ง
ค. มะเร็งตับ
ง. หลอดลมอักเสบ
จ. ถุงลมโป่งพอง

2) เทคนิคการเขียนตัวเลือก

(1) เขียนตัวเลือกให้เป็นเอกพจน์ หมายความว่า ตัวเลือกที่เป็นพวกเดียวกัน หรือประเภทเดียวกัน

(ไม่ดี) กรุงเทพมหานคร มีความสำคัญต่อประเทศในด้านใดมากที่สุด	(ดีขึ้น) กรุงเทพมหานคร มีความสำคัญต่อประเทศในด้านใดมากที่สุด
ก. เมืองหลวง	ก. เป็นศูนย์กลางค้า
ข. ที่ตั้งรัฐบาล	ข. เป็นศูนย์กลางศึกษา
ค. เป็นศูนย์กลางศึกษา	ค. เป็นแหล่งทรัพยากรทางธรรมชาติ
ง. เป็นชุมทางการสัญจร	ง. เป็นแหล่งคมนาคมขนส่ง
จ. มีประชากรหนาแน่น	จ. เป็นแหล่งท่องเที่ยว

(2) เขียนตัวเลือกให้เป็นอิสระจากกัน

(ไม่ดี) ประชากรของประเทศมีรายได้ โดยเฉลี่ยต่อเดือนประมาณเท่าไร	(ดีขึ้น) ประชากรของประเทศมีรายได้ โดยเฉลี่ยต่อเดือนประมาณเท่าไร
ก. น้อยกว่า 3,000 บาท	ก. น้อยกว่า 3,000 บาท
ข. น้อยกว่า 4,000 บาท	ข. 3,000 – 3,999 บาท
ค. ประมาณ 5,000 บาท	ค. 4,000 – 4,999 บาท
ง. มากกว่า 6,000 บาท	ง. 5,000 - 5,999 บาท
จ. มากกว่า 7,000 บาท	จ. มากกว่า 6,000 บาทขึ้นไป

(3) ความยาวของตัวเลือกควรเป็นระบบ เรียงความยาวจากน้อยไปมาก หรือมากไปน้อย

(ไม่ดี) H_2O หมายถึง	(ดีขึ้น) H_2O หมายถึง
ก. แอลกอฮอล์	ก. น้ำ
ข. ไฮโดรเจนเปอร์ออกไซด์	ข. โปตัสเซียม
ค. น้ำ	ค. แอลกอฮอล์
ง. คาร์บอนไดออกไซด์	ง. คาร์บอนไดออกไซด์
จ. โปตัสเซียม	จ. ไฮโดรเจนเปอร์ออกไซด์

(4) คำขยายประเภท “เท่านั้น ทั้งหมด ทุกที่ เสมอแน่นอน” ไม่ควรใช้กับ ตัวลง เพราะ จะทำให้เห็นผิดเด่นชัดเกินไป

(ไม่ดี) ข้อใดหมายถึงเงิน	(ดีขึ้น) ข้อใดหมายถึงเงิน
ก. สิ่งที่สังคมกำหนด	ก. ธนบัตร
ข. ธนบัตรเท่านั้น	ข. เงินเหรียญ
ค. ของที่ใช้แลกเปลี่ยนทั้งหมด	ค. สิ่งมีค่า
ง. สิ่งที่เป็นเหรียญทั้งหมด	ง. สิ่งที่สังคมกำหนด
จ. สิ่งที่มีค่ามากที่สุด	จ. ของที่ใช้แลกเปลี่ยน

(5) ด้อยเลือกไม่ควรมีประเภท “ถูกหมดทุกข้อ” “ถูกทั้ง ก และ ข” “ไม่มีข้อใดถูก”

(6) ตัวถูกไม่ควรยาวเกินไป

(7) คำตอบถูกต้องมีตัวเดียว

(8) ควรกระจายตำแหน่งตัวถูก

3) การกำหนดรูปแบบของข้อสอบเลือกตอบ

(1) แบบคำถามเดียว เป็นข้อคำถามที่มีคำถามเดียว และมีตัวเลือกเพียงชุดเดียว

ตัวอย่าง	ธาตุใดมากที่สุดในโลก
ฉ. หินปูน	
ช. ออกซิเจน	
ซ. ไฮโดรเจน	
ณ. โปตัสเซียม	

(2) แบบตัวเลือกคงที่ เป็นข้อคำถามที่ใช้ชุดตัวเลือกชุดเดียว แต่มีการกำหนดข้อคำถามไว้หลาย ๆ ข้อให้ตอบ

ตัวอย่าง คำชี้แจง คำถามจากข้อ 1 – 5 เป็นประโยคข้อความให้นักเรียนพิจารณาว่า ประโยคข้อความแต่ละข้อว่า เป็นประโยคที่ถูกต้องแล้ว หรือมีความบกพร่องของประโยคในด้านใดจากตัวเลือกต่อไปนี้

- ก. ใช้ไวยากรณ์ผิด
- ข. ใช้ถ้อยคำฟุ่มเฟือย
- ค. ใช้ภาษาไม่เหมาะสม
- ง. ใช้เหตุผลไม่เหมาะสม
- จ. ประโยคถูกต้องแล้ว

- 1) พรุ่งนี้เขาก็จะถูกส่งตัวเข้าคุกแล้ว
- 2) กิจการค้าของเขาก้าวหน้ารุ่งเรืองยิ่งขึ้น
- 3) ขออวยพรให้คู่บ่าวสาวครองรักกันจนวันตาย
- 4) ปัญหาแรงงานยืดเยื้อด้วยต่างก็อยากได้ประโยชน์
- 5) คนเป็นโรคเบาหวานไม่ควรบริโภคของมีน้ำตาลเจือปน

(3) **แบบสถานการณ์** เป็นการสร้างสถานการณ์ให้นักเรียนพิจารณาเพื่อตอบคำถาม สถานการณ์ใช้ได้หลายแบบ เช่น ข้อความ บทประพันธ์ รูปภาพ เป็นต้น

ตัวอย่าง คำชี้แจง จงใช้ข้อความนี้ตอบคำถามข้อ 1 – 2

“อกแม่นี้ให้อ่อนหวานสุดละห้อย ทั้งดาวเดือนก็เคลื่อนคล้อยลงลับไม้ สุดที่แม่จะติดตามเจ้าไปในยามนี้”

- 1) แม่มีความรู้สึกอย่างไร

- ก. ว้าเหว่
- ข. ท้อแท้
- ค. กังวล
- ง. หมดหวัง

- 2) “สุดที่แม่จะติดตาม” เพราะเหตุใด

- ก. แม่หิวและอ่อนเพลีย
- ข. มองไม่เห็นทาง
- ค. ไม่รู้ว่าอยู่ที่ใด
- ง. ยังไม่ถึงเวลา

ตัวอย่างข้อสอบเรียงตามระดับพฤติกรรมด้านพุทธิพิสัย

บลูม และคณะ ได้จำแนกสารบบจุดประสงค์การศึกษาออกเป็น 3 ด้าน คือ ด้านพุทธิพิสัย จิตพิสัย และทักษะพิสัย สำหรับด้านพุทธิพิสัยได้แบ่งพฤติกรรม

4. การสร้างข้อสอบเขียนตอบ

ข้อสอบแบบเขียนตอบ หรือข้อสอบอัตนัยเป็นข้อสอบที่ให้โอกาสผู้สอบเขียนคำตอบได้อย่างอิสระ ซึ่งการเขียนคำตอบนั้นผู้สอบต้องอาศัยทักษะการคิดวิเคราะห์ การสังเคราะห์ การสร้างสรรค์ ตลอดจนความสามารถในการเรียบเรียงถ้อยคำภาษาเขียน เพื่อเขียนออกมาเป็นคำตอบ ข้อสอบเขียนตอบแบ่งออกเป็น 2 ประเภทคือ แบบจำกัดคำตอบ (Restricted – Rspnse) และแบบไม่จำกัดคำตอบ (Etended Rspnse)

1) ข้อสอบเขียนตอบแบบจำกัดคำตอบ

เป็นข้อสอบที่จำกัดขอบเขตในการตอบ โดยมีข้อกำหนดในการตอบ เช่น จำนวนบรรทัด ประเด็นในการตอบ จำนวนข้อในการตอบ รูปแบบในการตอบเช่น เขียนเป็นตาราง รูปภาพ หรือยกตัวอย่าง จากข้อกำหนดดังกล่าวทำให้ข้อสอบประเภทนี้สามารถวัดเนื้อหาได้ครอบคลุมดีกว่าข้อสอบเขียนตอบแบบไม่จำกัดคำตอบ ตรวจให้คะแนนง่าย และมีความเชื่อมั่นสูงกว่าข้อสอบเขียนตอบแบบไม่จำกัดคำตอบ ส่วนข้อด้อยคือ อาจจำกัดความคิดของผู้สอบตามกรอบที่กำหนดให้ได้

ตัวอย่าง 1) จงอธิบายความแตกต่างระหว่างข้อสอบแบบเลือกตอบกับข้อสอบแบบ
ความเรียงมา 3 ประการ

- 2) จงบอกขั้นตอนการสร้างแบบสอบวัดผลสัมฤทธิ์ทางการเรียน
- 3) จงวิเคราะห์จุดเด่นจุดด้อยของการสอบเข้ามหาวิทยาลัยด้วยระบบ
แอดมิตชัน (เขียนไม่เกิน 1 หน้า)
- 4) จงวิเคราะห์ปัญหาวิกฤตการณ์น้ำมันแพงในปัจจุบันว่าส่งผลกระทบ
ในประเด็น การเมือง เศรษฐกิจ และสังคมของประเทศไทยอย่างไร

2) ข้อสอบเขียนตอบแบบไม่จำกัดคำตอบ

เป็นข้อสอบที่เปิดโอกาสให้ผู้สอบได้เขียนคำตอบอย่างอิสระ ผู้สอบสามารถแสดงความคิดเห็น ความคิดวิเคราะห์ ตลอดจนความคิดสร้างสรรค์ได้อย่างเต็มที่ ภายใต้กรอบเวลาที่กำหนด จึงสามารถวัดความสามารถระดับการวิเคราะห์ การสังเคราะห์และการประเมินผลได้อย่างดี แต่มีข้อด้อยในเรื่องการตรวจคำตอบเพื่อให้คะแนนจะตรวจได้ยาก และมีโอกาสที่คะแนนไม่มีความเป็นปรนัยสูงเพราะไม่มีการควบคุมทิศทางการตอบของผู้สอบ ดังนั้นก่อนที่ดำเนินการตรวจให้คะแนน ควรจัดทำเกณฑ์การคะแนนที่ชัดเจนก่อน

- ตัวอย่าง 1) จงเขียนโครงการวิจัยในประเด็น “การสำรวจความคิดเห็นเกี่ยวกับ
ปัญหาการจัดการเรียนการสอนของนิสิต และอาจารย์
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา”
- 2) ‘จงแสดงความคิดเห็นเกี่ยวกับหลักสูตรสาขาวิชาที่ท่านเรียนใน
ขณะนี้ว่ามีความเหมาะสมต่อการนำความรู้ไปประกอบวิชาชีพ
ในอนาคตหรือไม่ จงให้เหตุผลประกอบ

ข้อดี

- 1) สามารถวัดการเรียนรู้ระดับสูงได้ดี เช่น ระดับวิเคราะห์ ประเมินค่า และสร้างสรรค์
ได้ดี จึงเป็นการฝึกให้นักเรียนคิดเป็น
- 2) สร้างข้อสอบง่าย และใช้เวลาในการสร้างน้อย
- 3) เดียวกัน นักเรียนต้องมีความรู้จริงจึงจะสามารถเขียนคำตอบได้
- 4) เป็นการฝึกให้นักเรียนได้เขียนแสดงความคิดเห็น
- 5) ประหยัดค่าใช้จ่ายในการพิมพ์ข้อสอบเนื่องจากมีจำนวนข้อสอบน้อย

ข้อบกพร่อง

- 1) เนื่องจากข้อสอบข้อหนึ่งต้องใช้เวลาในการทำข้อสอบพอสมควรจึงทำให้ออก
ข้อสอบได้จำนวนน้อย ทำให้วัดได้ไม่ครอบคลุมเนื้อหา
- 2) การตรวจให้คะแนนทำได้ยาก การให้คะแนนอาจไม่คงที่ การตรวจมีโอกาสลำเอียงได้
- 3) ผู้ตรวจแต่ละคนมีมาตรฐานในการให้คะแนนแตกต่างกัน ซึ่งถ้านักเรียนมีความรู้
ความสามารถใกล้เคียงกัน แต่อาจได้คะแนนแตกต่างกัน
- 4) เสียเวลาในการตรวจให้คะแนนมาก จึงไม่เหมาะที่นำมาใช้กับการสอบขนาดใหญ่
(Large Scale) ที่มีจำนวนผู้สอบมาก

5) ถ้านักเรียนมีปัญหาในด้านการเขียนเพื่อแสดงความคิดออกมา อาจทำให้ไม่สามารถวัดผลสัมฤทธิ์ที่แท้จริงของนักเรียนคนนั้นได้

หลักการเขียนข้อสอบแบบเขียนตอบ

1) ข้อคำถามที่เป็นแบบเขียนตอบ ควรใช้เมื่อไม่สามารถวัดได้ด้วยข้อสอบปรนัย หรือวัดสมรรถภาพขั้นสูง เช่น วิเคราะห์ ประเมินค่า สร้างสรรค์ เช่น

- ให้เขียนความเรียงเรื่อง.....
- ให้วิจารณ์เรื่องราวเกี่ยวกับ.....
- ให้บอกสาเหตุหรือคาดคะเนเกี่ยวกับ.....
- ให้ระบุความแตกต่าง ระหว่าง.....
- ให้บอกข้อดี และข้อเสียเกี่ยวกับ.....
- ให้แสดงความคิดเห็น หรือวิจารณ์เกี่ยวกับ.....
- ให้วิเคราะห์สถานการณ์ที่กำหนดให้
- ให้เปรียบเทียบเรื่องราวในประเด็นที่กำหนดให้

2) เขียนคำชี้แจงเกี่ยวกับวิธีตอบให้ชัดเจน โดยระบุจำนวนข้อ เวลาที่ใช้ในการตอบ

ตัวอย่าง คำชี้แจง จงตอบคำถามต่อไปนี้ ลักษณะของข้อสอบเป็นแบบเขียนคำตอบ มีจำนวนทั้งหมด 3 ข้อ ผู้สอบควรใช้เวลาทำข้อสอบข้อละ 20 นาที รวมเวลาในการทำข้อสอบทั้งสิ้น 1 ชั่วโมง

3) ควรเขียนคะแนนเต็มในข้อสอบแต่ละข้อ

ตัวอย่าง การที่ประเทศอิสราเอลโจมตีประเทศเลบานอน จะส่งผลกระทบต่อประชาคมโลกอย่างไรบ้าง (คะแนนเต็ม 10 คะแนน)

4) เขียนคำถามให้เข้าใจ รัดกุม ชัดเจน เพื่อป้องกันไม่ให้ผู้สอบตีความหมายของข้อคำถามหลายแง่หลายมุมจนเกินไป

ไม่ดี จงเปรียบเทียบการประเมินก่อนเรียน ระหว่างเรียน และหลังเรียน

ดีขึ้น จงเปรียบเทียบความแตกต่างระหว่างการประเมินก่อนเรียน การประเมินระหว่างเรียน และการประเมินหลังเรียน ในหัวข้อต่อไปนี้

- จุดมุ่งหมาย
- เครื่องมือประเมิน
- การนำผลไปใช้

5) ข้อคำถามควรเป็นแบบจำกัดคำตอบ มากกว่าเป็นแบบไม่จำกัดคำตอบ เพื่อสะดวกในการให้คะแนน ขณะเดียวกันข้อสอบแบบจำกัดคำตอบใช้เวลาในการทำแต่ละข้อไม่มากเท่าข้อสอบแบบไม่จำกัดคำตอบ ทำให้ออกข้อสอบได้หลายข้อ โอกาสครอบคลุมเนื้อหาจึงมีมากกว่า

6) ไม่ควรให้มีการเลือกตอบเป็นบางข้อ เช่น มีข้อสอบจำนวน 5 ข้อ ให้เลือกทำ 3 ข้อ เนื่องจากข้อสอบแต่ละข้อจะวัดเนื้อหาแตกต่างกัน และความยากง่ายของข้อสอบก็แตกต่างกัน ทำให้นักเรียนเลือกทำเฉพาะข้อสอบที่นักเรียนทำได้ จึงทำให้การวัดความรู้ของนักเรียนไม่ครอบคลุมเนื้อหาทั้งหมด เพราะครูไม่ทราบได้ว่า ข้อที่นักเรียนไม่เลือกทำนั้น นักเรียนมีความรอบรู้หรือไม่รอบรู้ในเนื้อหานั้น

ข้อเสนอแนะในการตรวจให้คะแนน

1) ควรตรวจข้อสอบทีละข้อจนครบทุกคน แล้วค่อยตรวจข้อสอบข้อต่อไป เพื่อจะได้เปรียบเทียบระหว่างคำตอบของแต่ละคนทำให้การตรวจให้คะแนนมีความเที่ยงตรงมากขึ้น

2) ไม่ควรดูชื่อผู้สอบในขณะที่ตรวจให้คะแนน เพื่อป้องกันอคติทั้งทางบวก และทางลบที่มีต่อนักเรียนคนนั้น

3) ไม่ควรให้น้ำหนักคะแนนประเด็นเรื่องไวยากรณ์ การสะกด ลายมือ สูงมากเพราะประเด็นสำคัญต้องการวัดเนื้อหาที่เรียน

4) ควรมีเกณฑ์การให้คะแนนที่ชัดเจน ว่าคำตอบลักษณะใดจะได้คะแนนเท่าไร

ตัวอย่าง ข้อสอบวิชาวิทยาศาสตร์มัธยมศึกษาตอนต้น

ระดับชั้น ช่วงชั้นที่ 3

สาระการเรียนรู้ สาระที่ 3

มาตรฐาน ว.3.2(1) สังเกต สำรวจตรวจสอบ อภิปราย การเปลี่ยนแปลง สมบัติ
มวล และพลังงานของสาร เมื่อสารเกิดการละลายเปลี่ยนสถานะ และ
เกิดปฏิกิริยาเคมี รวมทั้งวิเคราะห์ปัจจัยที่ผลต่อการเปลี่ยนแปลงของ
สาร

คำถาม

ในขณะที่ฝนตก ด.ญ. อารยา ซึ่งนั่งใกล้หน้าต่างกระจก สังเกตเห็นว่าเมื่อลมหายใจ
กระทบกระจกจะเกิดฝ้าขาวเกาะบริเวณกระจก ฝ้าขาวนั้นคืออะไร
เกิดขึ้นได้อย่างไร

คำตอบที่เป็นไปได้

ฝ้าขาว คือ เมื่อไอน้ำอุ่นกระทบความเย็นจะควบแน่นหรือเปลี่ยนสถานะเป็นหยดน้ำ
ฝ้าขาว คือ หยดน้ำ เกิดขึ้นเมื่อภาวะอุณหภูมิ 2 บริเวณไม่เท่ากัน ไอน้ำรวมตัวกัน
กลายเป็นหยดน้ำ

เกณฑ์การให้คะแนน

ระดับคะแนน	แนวทางการให้คะแนน
3	ไอน้ำ เนื่องจากความแตกต่างของอุณหภูมิ อุณหภูมิภายนอกต่ำกว่า ภายใน ไอน้ำจากลมหายใจที่อุ่นเมื่อกระทบกระจกที่ภายนอกเย็นกว่าจึง กลายเป็นหยดน้ำเกาะที่กระจกในลักษณะที่เป็นฝ้าขาว
2	ตอบว่า เกิดจากไอน้ำ แต่ไม่สามารถอธิบายเหตุผลได้
1	กระจกเปื้อน หรืออื่น ๆ ที่ไม่เข้าข่าย
0	ไม่ตอบเลย

อย่างไรก็ตามแบบสอบอัตนัย ก็มีปัญหาและอุปสรรคในการใช้ ดังงานวิจัยของ จุฑา
ธรรมชาติ (2549) ได้ศึกษาและวิเคราะห์สภาพการใช้แบบสอบอัตนัยในการวัดและประเมินผลการ
เรียนรู้ของผู้เรียนระดับการศึกษาขั้นพื้นฐาน พบว่า ปัญหาและอุปสรรคในการใช้แบบสอบอัตนัย
ของครูส่วนใหญ่เกิดขึ้นจากผู้เรียนมากที่สุด คิดเป็นร้อยละ 72.42 โดยนักเรียนขาดทักษะการเขียน

อย่างถูกต้อง/ เขียนผิดไวยากรณ์ นักเรียนเขียนสะกดคำไม่ถูกต้อง และลายมือของนักเรียนอ่านลำบาก รองลงมาเกิดจากการตรวจของครู คิดเป็นร้อยละ 43.83 โดยครูเสียเวลาในการตรวจให้คะแนน/ ตรวจไม่ทันเพราะมีนักเรียนมาก ครูผู้ตรวจมีความลำเอียงขาดความยุติธรรม/ ไม่มีความเที่ยง และนักเรียนมีแนวทางในการตอบกว้างเกินไปยากสำหรับการตรวจ/ ไม่เป็นปรนัยในการตรวจ และเกิดจากข้อสอบน้อยที่สุด คิดเป็นร้อยละ 3.13 โดยครูออกข้อสอบมีจำนวนข้อน้อยไม่ครอบคลุมเนื้อหา มาก และจากการสัมภาษณ์ครูเพื่อหาแนวทางแก้ไขได้ผลการสัมภาษณ์ดังนี้

1. ครูมีการสอน และให้นักเรียนทำแบบฝึกหัดบ่อย ๆ
2. ครูฝึกให้นักเรียนอ่านคิดวิเคราะห์ควบคู่กับการฝึกเขียน
3. ครูให้คำแนะนำนักเรียนเกี่ยวกับการตอบข้อสอบอัตนัย
4. ครูพยายามใช้คำถามแบบให้มีคำตอบชัดเจน และตรวจง่าย
5. ครูกำหนดเกณฑ์การให้คะแนนที่ชัดเจนและตรวจง่าย ตรวจข้อสอบร่วมกัน ตรวจทีละข้อ หยุดพักเมื่อตรวจเป็นเวลานาน

6. ใช้แบบสอบปรนัยควบคู่กับอัตนัย

จากที่กล่าวมาแบบสอบผลสัมฤทธิ์ทางการเรียนประกอบด้วย แบบสอบเลือกตอบ แบบสอบถูกผิด แบบสอบจับคู่ แบบสอบเหล่านี้จัดเป็นแบบสอบปรนัย ส่วนแบบสอบเขียนตอบ จัดเป็นแบบแบบสอบอัตนัย Thorndlike and Hagen (1977) ได้เปรียบเทียบข้อดีและข้อบกพร่องของแบบสอบอัตนัย แบบสอบเติมคำ/ตอบสั้น และแบบสอบปรนัย ตามลักษณะของการประเมิน ดังตารางที่ 2-3

ตารางที่ 2-3 การเปรียบเทียบข้อดี และข้อบกพร่องของแบบสอบอัตนัย แบบสอบเต็มคำ/ ตอบสั้น
แบบสอบปรนัย

องค์ประกอบ	อัตนัย	เต็ม คำ/ ตอบ สั้น	ปรนัย
1. สามารถวัดความสามารถในการแก้ปัญหา	++	+	++
2. สามารถวัดความสามารถในการจัดระบบ บุรณาการ และสังเคราะห์	++	+	--
3. สามารถวัดความคิดริเริ่มหรือนวัตกรรมใหม่ในการแก้ปัญหา	++	+	--
4. สามารถแยกความสามารถในความรู้จากทักษะการเขียน การสะกดคำ และการใช้ภาษา ออกจากกัน	--	-	++
5. มีศักยภาพในการวินิจฉัยการเรียนรู้	--	+	++
6. สามารถวัดได้ครอบคลุมตามจุดประสงค์ของการเรียนการสอน	--	-	++
7. การปราศจากการเดาคำตอบ	++	++	--
8. มีความเที่ยงในการให้คะแนนระหว่างผู้ตรวจ	--	-	++
9. สามารถจำแนกตามระดับความสามารถที่ต่างกันได้	--	-	++
10. สามารถตรวจให้คะแนนโดยผู้ที่ไม่มีความรู้หรือเครื่องตรวจได้	--	-	++
11. สามารถตรวจได้อย่างรวดเร็ว	--	-	++
12. ใช้เวลาในการเขียนข้อคำถามน้อย	+	+	-

++ หมายถึง วัดได้ดี

+ หมายถึง วัดได้บ้าง

- หมายถึง วัดได้เล็กน้อย

-- หมายถึง วัดไม่ได้

ตอนที่ 5 งานวิจัยที่เกี่ยวข้อง

บาเคนดัลส (Buckendahl, 2002) ได้ทำการเปรียบเทียบการกำหนดมาตรฐานด้วยวิธีการเองกอฟ และวิธีบูคมาร์ค เพื่อเปรียบเทียบจุดตัดที่ใช้วิธีเองกอฟและวิธีบูคมาร์ค นักเรียนเกรด 7 จำนวน 448 คน เครื่องมือที่ใช้ในการวิจัยคือ แบบทดสอบคณิตศาสตร์เกรด 7 แบบเลือกคำตอบที่มีเนื้อหาทางทักษะคณิตศาสตร์ 6 อย่าง โดยมีวิธีดำเนินการดังนี้ 1) นำคณะผู้ตัดสินมาทำสัมมนาเชิงปฏิบัติการ 1 วัน เพื่อสร้างความเข้าใจในวิธีการทั้งสองวิธี 2) แบ่งกลุ่มผู้ตัดสินเป็น 2 กลุ่มคือ สำหรับวิธีเองกอฟ และวิธีบูคมาร์ค 3) แต่ละกลุ่มพิจารณาหาจุดตัด 2 รอบ โดยรอบแรกพิจารณาอย่างอิสระ รอบสองอภิปรายร่วมกันและเปิดโอกาสให้ผู้ตัดสินแต่ละคนเปลี่ยนจุดตัดของตนเองได้ 4) หาค่าเฉลี่ยจุดตัด โดยมีวิธีการวิเคราะห์ข้อมูลดังนี้ 1) หาค่าความยากโดย CTT 2) เปรียบเทียบจุดตัดและส่วนเบี่ยงเบนมาตรฐานของรอบที่ 1 และรอบที่ 2 ในแต่ละวิธี ผลการวิจัยพบว่า วิธีเองกอฟ รอบแรกได้คะแนนจุดตัด 34.92 ส่วนเบี่ยงเบนมาตรฐาน 7.79 และเปอร์เซ็นต์การตกของนักเรียน 8.9 รอบสองได้คะแนนจุดตัด 33.42 ส่วนเบี่ยงเบนมาตรฐาน 10.96 และเปอร์เซ็นต์การตก 7.6 สำหรับวิธีบูคมาร์ค รอบแรกได้คะแนนจุดตัด 33.64 ส่วนเบี่ยงเบนมาตรฐาน 11.03 เปอร์เซ็นต์การตก 7.6 รอบที่สองได้คะแนนจุดตัด 35.64 ส่วนเบี่ยงเบนมาตรฐาน 8.66 เปอร์เซ็นต์การตก 9.4 จะเห็นได้ว่า คะแนนจุดตัดในรอบสุดท้ายของทั้งสองวิธีมีความแตกต่างกันเพียงเล็กน้อย (คะแนนต่างกัน 2 คะแนน และเปอร์เซ็นต์การตกต่างกัน 1.8% ของนักเรียนที่ทำการสอบ) และพบว่าวิธีบูคมาร์ค ในรอบที่สองมีส่วนเบี่ยงเบนมาตรฐานลดลงเมื่อเทียบกับวิธีเองกอฟ ซึ่งการลดลงของส่วนเบี่ยงเบนมาตรฐานในรอบสองของวิธีบูคมาร์คนี้ แสดงให้เห็นว่าพิสัยของคะแนนจุดตัดลดลง และชี้ให้เห็นว่า การตัดสินในระหว่างผู้ตัดสิน (Inter-Judge) มีระดับสูงขึ้น

แวง (Wang, 2003) ได้นำโมเดล Rasch IRT มาใช้ในการกำหนดมาตรฐานวิธีการฝังข้อสอบ (Item Mapping) เพื่อกำหนดหามาตรฐาน หรือคะแนนจุดตัดสำหรับการสอบใบรับรองและประกาศนียบัตรด้วยวิธีฝังข้อสอบ กลุ่มตัวอย่างเป็นผู้สอบขอใบรับรองและประกาศนียบัตรโดยใช้เครื่องมือคือ แบบสอบแบบเลือกคำตอบสำหรับ professional licensure or certification examinations ซึ่งมีการสอบ 4 ครั้ง สอบครั้งที่ 1 ethics examination of a state insurance licensing program ในเดือนมิถุนายน 1999 สอบครั้งที่ 2 property examination of a state insurance licensing program ในเดือนสิงหาคม 1999 สอบครั้งที่ 3 casualty examination of a state insurance licensing program ในเดือนสิงหาคม 1999 สอบครั้งที่ 4 national nurse aide certification examination ในเดือนธันวาคม 2000 วิธีดำเนินการวิจัยมีดังนี้ 1) ผู้ตัดสินร่วมกันอธิบายลักษณะของผู้สอบที่มีความสามารถต่ำ (MCC) สำหรับขอใบรับรอง 2) วิธีเองกอฟ

ผู้ตัดสินจะถูกถามว่าเปอร์เซ็นต์ของ MCC ที่ตอบข้อสอบข้อนี้ถูกคือเท่าไร จากนั้นให้ค่าสถิติของข้อสอบพิจารณาโดยเปิดโอกาสให้ผู้ตัดสินอภิปรายและปรับเปลี่ยนจุดตัดของตนได้ จากนั้นหาค่าเฉลี่ยจุดตัด 3) วิธีฝังข้อสอบจะทำฝังข้อสอบให้ผู้ตัดสินพิจารณา และผู้ตัดสินจะถูกถามว่า “ผู้สอบกลุ่มคาบเส้น มีโอกาสตอบข้อสอบถูกอย่างน้อย .50 หรือไม่” จากนั้นจัดหา p-value และเปิดโอกาสให้อภิปรายและปรับเปลี่ยนจุดตัดของตนได้ ปกติข้อสอบหลายข้อในแต่ละคอลัมน์จะถูกเลือกเพื่อให้ผู้ตัดสินหา จันทมติของข้อสอบ กระบวนการนี้จะถูกทำซ้ำจนกระทั่งคอลัมน์ข้อสอบถูกกำหนดให้เป็นจุดที่ข้อสอบส่วนใหญ่ในคอลัมน์ได้รับจันทมติจากผู้ตัดสินส่วนใหญ่ว่ามีความน่าจะเป็นที่ ผู้สอบกลุ่มคาบเส้นจะตอบถูก .50 และระดับตรงกลางของค่าความยากของข้อสอบภายใต้คอลัมน์จะถูกกำหนดเป็นจุดตัด วิธีการวิเคราะห์ข้อมูลมีดังนี้ 1) วิเคราะห์ Rasch IRT ด้วยโปรแกรม BIGSTEPS 2) ใช้ G-theory วิเคราะห์หา Inter-judge consistency ($r \times i$) ใช้โปรแกรม GENOVA วิเคราะห์ G และ D studies ผลการวิจัยพบว่า วิธีฝังข้อสอบมีความสอดคล้องในการตัดสินระหว่างผู้ตัดสิน (Inter-Judge Consistency) มากกว่าวิธีการเองกอฟ

บิเรทวาส (Beretvas, 2004) ได้เปรียบเทียบการกำหนดค่าความยากในวิธีการบูคมาร์ค โดยการวิเคราะห์ด้วยโมเดล IRT ที่แตกต่างกัน กลุ่มตัวอย่าง ผู้วิจัยได้ใช้ฐานข้อมูลนักเรียนเกรด 4 จำนวน 63,533 คน ปี ค.ศ. 1998 ส่วนเครื่องมือการวิจัยคือ แบบสอบย่อยคณิตศาสตร์นักเรียนเกรด 4 ของ Washington Assessment of student Learning (WASL) โดยข้อสอบเป็นรูปแบบผสม 40 ข้อ แบ่งเป็นแบบเลือกตอบ 24 ข้อ และแบบเขียนตอบ 16 ข้อ ซึ่งแบ่งเป็นแบบตอบสั้น 13 ข้อ ให้คะแนนแบบ 0, 1 และ 2 และข้อสอบแสดงวิธีทำให้คะแนนเป็น 0, 1, 2, 3 และ 4 โดยมีคะแนนรวมเท่ากับ 62 คะแนน ผู้วิจัยนำผลการสอบจากฐานข้อมูลมาวิเคราะห์ข้อมูลหาค่าพารามิเตอร์ข้อสอบตามโมเดล IRT คือ โมเดล IRT ที่ให้คะแนนแบบ 2 ค่า แบบ 1 พารามิเตอร์ และ 2 พารามิเตอร์ และโมเดล IRT ที่ให้คะแนนแบบหลายค่าคือ partial credit (PC), generalized partial credit (GPC) และ graded response (GR) ซึ่งวิเคราะห์ด้วยโปรแกรม Parscale หรือ Multilog ค่าพารามิเตอร์ที่ได้มาคำนวณหา BDL (bookmark difficulty locations) ตามโมเดล IRT และ RP ที่แตกต่างกัน โดยใช้โปรแกรม Mathematica ช่วยในการคำนวณ โดยใช้โมเดล IRT และค่า RP ที่แตกต่างกัน ดังนี้ 1PL-PC RP $\frac{1}{2}$, 1PL-PC RP $\frac{2}{3}$, 1PL-PC RP $\frac{4}{5}$, 3PL-GPC RP $\frac{1}{2}$, 3PL-GPC RP $\frac{2}{3}$, 3PL-GPC RP $\frac{4}{5}$, 3PL-GR RP $\frac{1}{2}$, 3PL-GR RP $\frac{2}{3}$, 3PL-GR RP $\frac{4}{5}$ จากนั้นนำค่า BDL มาใช้พิจารณาจัดเรียงข้อสอบตามค่าความยากจากน้อยไปมาก เพื่อนำโมเดล IRT และค่า RP ที่แตกต่างกัน มาเปรียบเทียบกับกัน โดยใช้ spearman rank order correlation ซึ่งผลการวิเคราะห์จากโมเดล IRT และค่า RP ทั้ง 9 แบบ ค่าสหสัมพันธ์มีค่าระหว่าง .897 - .998 ที่ระดับนัยสำคัญ .001 โดยโมเดล 1PL-PC มีค่าสหสัมพันธ์ระหว่าง .900 - .974

โมเดล 3PL-GPC มีค่าสหสัมพันธ์ระหว่าง .933 - .991 และโมเดล 3PL-GR มีค่าสหสัมพันธ์ระหว่าง .976 - .995 และพบว่าทั้ง 3 โมเดลค่าสหสัมพันธ์ระหว่าง ค่า RP 2/3 และ 4/5 สูงที่สุดสำหรับในกระบวนการนับคะแนนจะนำ BDL ที่ใช้โมเดล 1PL-PC, 3PL-GPC, 3PL-GR ที่ RP 2/3 มาใช้ในคู่มือการจัดเรียงข้อสอบตามโมเดลการวิเคราะห์แต่ละโมเดล ผลการจัดเรียงข้อสอบพบว่าโมเดล 3PL-GPC, 3PL-GR มีการจัดเรียงข้อสอบใกล้เคียงกันมาก มีความแตกต่างกันเพียง 2 ข้อ ในขณะที่ 1PL-PC มีผลการจัดเรียงแตกต่างจากโมเดลทั้งสองแบบ

สุริพร อนุศาสนนันท์ (2550) ได้เปรียบเทียบคุณภาพของการกำหนดมาตรฐานระหว่างวิธีการเองกอฟที่ได้รับการปรับปรุงกับวิธีการนับคะแนน 2) เพื่อศึกษาผลของจำนวนผู้ตัดสิน และจำนวนครั้งที่มีการกำหนดมาตรฐานด้วยวิธีการเองกอฟที่ได้รับการปรับปรุง และวิธีการนับคะแนน กลุ่มตัวอย่างประกอบด้วย ครูผู้สอนวิชาคณิตศาสตร์ ซึ่งเป็นผู้ตัดสินจำนวน 12 คน และนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2549 จำนวน 987 คน ในจังหวัดชลบุรี เครื่องมือที่ใช้ในการวิจัยคือ แบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 1 จำนวน 4 ฉบับ รวม 100 ข้อ ผลการวิจัยพบว่า 1. แบบสอบวัดผลสัมฤทธิ์ทางการเรียน มีคะแนนจุดตัดที่กำหนดด้วยวิธีเองกอฟที่ได้รับการปรับปรุง 7 ระดับคือ ระดับดีเยี่ยม (A) ระดับดีมาก (B+) ระดับดี (B) ระดับดีพอใช้ (C+) ระดับพอใช้ (C) ระดับอ่อน (D+) และระดับอ่อนมาก (D) ครั้งที่ 1 มีค่า 72.86, 64.31, 56.18, 43.37, 33.59, 21.34 และ 12.98 ตามลำดับ ครั้งที่ 2 มีค่า 73.28, 63.74, 55.28, 45.11, 35.67, 22.94, 12.51 ตามลำดับ ครั้งที่ 3 มีค่า 73.49, 63.25, 52.82, 52.82, 41.46, 31, 19.07, 11.72 ตามลำดับ และมีคะแนนจุดตัดด้วยวิธีนับคะแนน ครั้งที่ 1 มีค่า 80.42, 65.75 47.5, 33.17, 24, 14.75, 7.83 ตามลำดับ ครั้งที่ 2 เท่ากับ 80.58, 66.67, 47.5, 33.75, 24.58, 14.75, 7.83 ตามลำดับ และครั้งที่ 3 มีค่า 80.92, 64.08, 47.33, 34, 24.08, 14.83, 7.92 ตามลำดับ 2. การกำหนดมาตรฐานด้วยวิธีนับคะแนนมีค่าความเที่ยงสูงกว่าวิธีเองกอฟที่ได้รับการปรับปรุงอย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ .01 ทั้ง 7 ระดับ 3. ค่าความตรงของการกำหนดมาตรฐาน โดยพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างระดับผลการเรียนที่นักเรียนได้รับจากโรงเรียน กับระดับมาตรฐานที่กำหนดด้วยวิธีเองกอฟที่ได้รับการปรับปรุง เท่ากับ .661 - .678 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างระดับผลการเรียนที่นักเรียนได้รับจากโรงเรียน กับระดับมาตรฐานที่กำหนดด้วยวิธีนับคะแนน เท่ากับ .533 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และเมื่อทดสอบความแตกต่างที่ระดับนัยสำคัญทางสถิติ .05 พบว่า ไม่แตกต่างกัน 4. ค่าความเที่ยงของการกำหนดมาตรฐานมีค่าเพิ่มขึ้น เมื่อจำนวนผู้ตัดสิน และจำนวนครั้งเพิ่มขึ้น