บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

1. โครงการพัฒนาระบบและเครื่องมือการประเมินระดับท้องถิ่น
2. การประเมินคุณภาพการศึกษาขั้นพื้นฐาน ระดับท้องถิ่น (LAS)
3. ความเป็นมาของการศึกษาการทำหน้าที่ต่างกันของข้อสอบ
4. ความหมายของการทำหน้าที่ต่างกันของข้อสอบ
5. สาเหตุที่ทำให้ข้อสอบทำหน้าที่ต่างกัน
6. ประเภทของการทำหน้าที่ต่างกันของข้อสอบ
7. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
8. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
9. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย IRT
10. งานวิจัยที่เกี่ยวข้อง

1. โครงการพัฒนาระบบและเครื่องมือการประเมินระดับท้องถิ่น

1.1 กิจกรรม โครงการสร้างและพัฒนาเครือข่ายการประเมินระดับท้องถิ่น

สำนักทดสอบทางการศึกษาเริ่มดำเนินโครงการสร้างและพัฒนาเครือข่ายการประเมินระดับท้องถิ่น มาตั้งแต่ปีงบประมาณ 2548 โดยยึดหลักการที่ว่าความมุ่งมั่นของการจัดการศึกษาที่สำคัญประการหนึ่ง คือ ความเสมอภาคทางการศึกษาสำหรับผู้เรียนทุกคน ซึ่งหมายความว่าผู้เรียนทุกคนจะต้องประสบความสำเร็จในการเรียนรู้ตามจุดหมายของหลักสูตรหรือมาตรฐานการเรียนรู้ที่กำหนดไว้ จึงเป็นความรับผิดชอบของสถานศึกษาเขตพื้นที่และส่วนกลาง (ต้นสังกัด) ในการจัดการศึกษาที่ต้องมีการพัฒนาและจัดระบบการประเมินผลการเรียนรู้เพื่อการพัฒนาและประกันคุณภาพการศึกษาของผู้เรียน การตรวจสอบและประเมินสรุปรวมทั้งรายงานผลสัมฤทธิ์ของผู้เรียนดังกล่าวจะเป็นการสร้างความมั่นใจให้แก่ผู้เรียน ผู้ปกครองและสังคมว่าผู้เรียนทุกคนได้รับการศึกษาบรรลุเป้าหมายคุณภาพและมาตรฐานที่กำหนดไว้ การพัฒนาระบบการประเมินเพื่อประกันคุณภาพผู้เรียนจึงมีความจำเป็นและสำคัญอย่างยิ่งที่ต้องออกแบบและพัฒนาระบบ เพื่อนำไปสู่การปฏิบัติโดยเฉพาะในบริบทการเปลี่ยนแปลงการปฏิรูป

การศึกษาที่กระจายอำนาจการจัดการศึกษาไปสู่สถานศึกษาและเขตพื้นที่ แต่คงไว้ซึ่งความเป็นเอกภาพในนโยบายของหน่วยงานต้นสังกัดและกระทรวงศึกษาธิการ (สำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ, 2551)

ดังนั้น ระบบดังกล่าว จะต้องเข้าถึงผลการปฏิบัติของผู้เรียน (Performance Standard) ตามมาตรฐานการเรียนรู้ของหลักสูตร ที่สนองวัตถุประสงค์การจัดการศึกษา 3 ประการสำคัญ คือ

1. การจัดการศึกษาเพื่อให้บริการการสอนและการเรียนรู้

2. การรับรองผลสัมฤทธิ์ของผู้เรียนที่สอดคล้องกับมาตรฐานและสำเร็จการศึกษาในระดับต่าง ๆ

3. การแสดงภาระรับผิดชอบของหน่วยงานจัดการศึกษา

วัตถุประสงค์ดังกล่าวได้กำหนดโดยพระราชบัญญัติการศึกษา พุทธศักราช 2542 และกฎ กระทรวงเรื่องระบบหลักเกณฑ์และวิธีการประกันคุณภาพการศึกษาภายในสถานศึกษาระดับการศึกษาขั้นพื้นฐานและปฐมวัย ซึ่งระบุให้สถานศึกษาและหน่วยงานต้นสังกัดต้องจัดให้มีระบบประกันคุณภาพภายใน เพื่อพัฒนาคุณภาพผู้เรียนมีการตรวจสอบทบทวนคุณภาพภายในสถานศึกษาอย่างต่อเนื่อง รวมทั้งจัดทำรายงานประจำปีเสนอต่อสาธารณชนและหน่วยงานต้นสังกัด ทั้งนี้เขตพื้นที่และหน่วยงานต้นสังกัดจะต้องตรวจสอบคุณภาพการศึกษา อย่างน้อย 3 ปี / ครั้ง

การพัฒนาระบบการประเมินเพื่อประกันคุณภาพผู้เรียนจึงมีความจำเป็นและสำคัญอย่างยิ่งที่ต้องออกแบบและพัฒนาระบบเพื่อนำไปสู่การปฏิบัติ โดยเฉพาะในบริบทการเปลี่ยนแปลงการปฏิรูปการศึกษาที่กระจายอำนาจการจัดการศึกษาไปสู่สถานศึกษาและเขตพื้นที่แต่คงไว้ซึ่งความเป็นเอกภาพในนโยบายของหน่วยงานต้นสังกัดคือ สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานและกระทรวงศึกษาธิการ

ระบบการประเมินเพื่อการประกันคุณภาพการศึกษาที่มุ่งเน้นความสำเร็จตามเป้าหมายการจัดการศึกษาคือ มาตรฐานคุณภาพผู้เรียนตามหลักสูตรการศึกษาขั้นพื้นฐาน จึงควรประกอบด้วยการประเมินจากหน่วยงานส่วนกลาง (สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน) การประเมินระดับท้องถิ่น (สำนักงานเขตพื้นที่การศึกษา) และการประเมินระดับชั้นเรียน ซึ่งจะทำให้ได้ข้อมูลที่สมบูรณ์ เที่ยงตรงและน่าเชื่อถือของผู้เรียน

เพื่อสนองนโยบายทางการศึกษาและโครงสร้างความรับผิดชอบการจัดการศึกษาตามพระราชบัญญัติบริหารระบบราชการของกระทรวงศึกษาธิการ สำนักทดสอบทางการศึกษาซึ่งรับผิดชอบการพัฒนาระบบการประเมินคุณภาพการศึกษาของสำนักงานคณะกรรมการการศึกษา

ขั้นพื้นฐานจึงวางแผนเพื่อดำเนินโครงการพัฒนาระบบการประเมินระดับท้องถิ่น โดยกำหนดเป้าหมายสำคัญคือ การส่งเสริมสนับสนุนให้เขตพื้นที่มีระบบและรูปแบบการประเมินระดับท้องถิ่นเพื่อประเมินผลสัมฤทธิ์ผู้เรียนที่น่าเชื่อถือและเที่ยงตรง (reliable and valid) ซึ่งจะทำให้ได้ข้อมูลผลสัมฤทธิ์ผู้เรียนเพื่อการพัฒนาและรายงานผลที่จะนำไปสู่การตัดสินใจที่ถูกต้อง ทั้งในกำหนดนโยบายและการพัฒนาระดับผู้เรียน

1.2 วัตถุประสงค์โครงการ

1.2.1 พัฒนาระบบการประเมินระดับท้องถิ่นเพื่อการประกันคุณภาพผลสัมฤทธิ์ของผู้เรียนที่เชื่อมโยงสอดคล้องสอดคล้องตามภารกิจ ความรับผิดชอบของทุกฝ่าย เพื่อการพัฒนาและประกันคุณภาพการศึกษา

1.2.2 เพื่อพัฒนารูปแบบการประเมินระดับท้องถิ่นของสำนักงานเขตพื้นที่การศึกษาที่มีประสิทธิภาพ นำไปสู่การปฏิบัติงานร่วมกันระหว่างเขตพื้นที่และสถานศึกษา เพื่อให้ได้ข้อมูลที่เที่ยงตรงและเชื่อถือได้ นำไปสู่การพัฒนาคุณภาพผู้เรียนอย่างแท้จริง

2. การประเมินคุณภาพการศึกษาขั้นพื้นฐาน ระดับท้องถิ่น (LAS)

2.1 หลักการและแนวคิด

พระราชบัญญัติการศึกษาแห่งชาติ พ.ศ.2542 มาตรา 47 กำหนดให้มีระบบการประกันคุณภาพการศึกษา เพื่อพัฒนาคุณภาพและมาตรฐานการศึกษาในทุกระดับ และมาตรา 48 ให้หน่วยงานต้นสังกัดและสถานศึกษา จัดให้มีระบบการประกันคุณภาพการศึกษาภายในสถานศึกษาและให้ถือว่าการประกันคุณภาพภายใน เป็นส่วนหนึ่งของกระบวนการบริหารการศึกษาที่ต้องดำเนินการอย่างต่อเนื่อง โดยมีการจัดทำรายงานประจำปีเสนอต่อหน่วยงานต้นสังกัด หน่วยงานที่เกี่ยวข้อง และเปิดเผยต่อสาธารณชน เพื่อนำไปสู่การพัฒนาและมาตรฐานการศึกษาและเพื่อรองรับการประกันคุณภาพภายนอก การประเมินคุณภาพขั้นพื้นฐานจึงเป็นกระบวนการหรือวิธีการ เพื่อให้ได้ข้อมูลที่จะเป็นตัวบ่งชี้ถึงผลสำเร็จในการจัดการศึกษาซึ่งเป็นส่วนประกอบสำคัญส่วนหนึ่งในการประกันคุณภาพภายใน หลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2544 จึงกำหนดแนวทางการวัดและประเมินผลการเรียนรู้เพื่อให้ได้ข้อมูลสารสนเทศที่แสดงการพัฒนาการ ความก้าวหน้า และความสำเร็จทางการเรียนของผู้เรียน ซึ่งสถานศึกษาต้องจัดให้มีการประเมินผลการเรียนให้เป็นไปในมาตรฐานเดียวกัน ทั้งในระดับชั้นเรียน ระดับสถานศึกษาระดับเขตพื้นที่การศึกษา และระดับชาติ ข้อมูลที่ได้จาการประเมินจะนำไปใช้ในการพัฒนาคุณภาพของผู้เรียน และคุณภาพการจัดการศึกษาของสถานศึกษาแต่ละแห่ง และเพื่อเป็น

สารสนเทศรองรับการประเมินคุณภาพการศึกษาภายนอก (สำนักงานเขตพื้นที่การศึกษาเพชรบูรณ์ เขต 1, 2552, หน้า 1-10)

โครงการประเมินคุณภาพการศึกษาขั้นพื้นฐาน ระดับท้องถิ่น (LAS) ปีการศึกษา 2552 เป็นการตรวจสอบ ควบคุม กำกับดูแล และรักษาคุณภาพการศึกษาของสถานศึกษา ซึ่งสำนักงานคณะกรรมการศึกษาขั้นพื้นฐานมอบให้สำนักงานเขตพื้นที่การศึกษา รับผิดชอบประเมินนักเรียนทุกคนในชั้นประถมศึกษาปีที่ 2 ประถมศึกษาปีที่ 5 มัธยมศึกษาปีที่ 2 และชั้นมัธยมศึกษาปีที่ 5 โดยให้มีการดำเนินการสอบในระดับท้องถิ่นร่วมกันเพื่อเป็นการพัฒนาอย่างต่อเนื่อง ซึ่งผลการประเมินจะเป็นข้อมูลสำคัญในการปรับปรุง พัฒนาตนเองของผู้เรียน และการจัดการเรียนการสอนของสถานศึกษา ตามนโยบายของสำนักงานคณะกรรมการศึกษาขั้นพื้นฐานต่อไป

2.2 วัตถุประสงค์

2.2.1 เพื่อให้เขตพื้นที่การศึกษา เครือข่ายและสถานศึกษาตลอดจนคณะกรรมการดำเนินการจัดสอบได้อย่างถูกต้องตามระเบียบ มีประสิทธิภาพ และได้มาตรฐาน

2.2.2 เพื่อให้สถานศึกษาและเขตพื้นที่การศึกษา มีข้อมูลผลสัมฤทธิ์นักเรียนชั้นประถม ศึกษาปีที่ 2 ชั้นประถมศึกษาปีที่ 5 ชั้นมัธยมศึกษาปีที่ 2 และชั้นมัธยมศึกษาปีที่ 5 นำไปปรับปรุงพัฒนาผู้เรียนรายบุคคล

2.2.3 เพื่อให้ได้ข้อมูลย้อนกลับสำหรับใช้ในกระบวนการตัดสินใจ และกำหนดแผนพัฒนาคุณภาพการจัดการศึกษาระดับประเทศ เขตพื้นที่การศึกษา และระดับสถานศึกษา

2.3 ขอบเขตการประเมิน

การประเมินคุณภาพการศึกษาขั้นพื้นฐาน ระดับท้องถิ่น ปีการศึกษา 2552 จะเป็นการประเมินผลสัมฤทธิ์ทางการเรียนของผู้เรียนตามจุดมุ่งหมายของหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2544 และหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ใช้แบบทดสอบปรนัยชนิดเลือกตอบ (Multiple choices) โดยมีสาระการประเมินดังนี้

ช่วงชั้นที่ 1 (ประถมศึกษาปีที่ 2) สอบสาระการเรียนรู้คณิตศาสตร์ ภาษาไทย และวิทยาศาสตร์

ช่วงชั้นที่ 2 (ประถมศึกษาปีที่ 5) สอบสาระการเรียนรู้คณิตศาสตร์ ภาษาไทย ภาษาอังกฤษ และวิทยาศาสตร์

ช่วงชั้นที่ 3 (มัธยมศึกษาปีที่ 2) และช่วงชั้นที่ 4 (มัธยมศึกษาปีที่ 5) สอบสาระการเรียนรู้คณิตศาสตร์ ภาษาไทย วิทยาศาสตร์ ภาษาอังกฤษและ สังคมศึกษา ศาสนาและวัฒนธรรม

2.4 เครื่องมือที่ใช้ในการสอบ

ตาราง 1 แสดงรายละเอียดของเครื่องมือที่ใช้ในโครงการประเมินคุณภาพการศึกษาขั้นพื้นฐานระดับท้องถิ่น (LAS) ปีการศึกษา 2552 ของสำนักงานเขตพื้นที่การศึกษาเพชรบูรณ์ เขต 1

ชั้น	สาระการเรียนรู้		จำนวนข้อ	เวลา (นาที)
ประถมศึกษาปีที่ 2	ตอนที่ 1 คณิตศาสตร์		30	60
	ตอนที่ 2 ภาษาไทย		30	40
	ตอนที่ 3 วิทยาศาสตร์		30	40
ประถมศึกษาปีที่ 5	ฉบับที่ 1	ตอนที่ 1 คณิตศาสตร์	35	60
		ตอนที่ 2 ภาษาไทย	45	60
	ฉบับที่ 1	ตอนที่ 3 การคิดคำนวณ (ภาคปฏิบัติ)	7	30
	ฉบับที่ 2	ตอนที่ 1 วิทยาศาสตร์	40	60
		ตอนที่ 2 ภาษาอังกฤษ	40	60
มัธยมศึกษาปีที่ 2	ฉบับที่ 1	ตอนที่ 1 คณิตศาสตร์	40	60
		ตอนที่ 2 ภาษาไทย	40	60
		ตอนที่ 3 สังคมศึกษา	40	60
	ฉบับที่ 2	ตอนที่ 1 วิทยาศาสตร์	40	60
		ตอนที่ 2 ภาษาอังกฤษ	40	60
	ฉบับที่ 3 ภาคปฏิบัติ	ตอนที่ 1 คิดคำนวณ	9	30
		ตอนที่ 2 การเขียน	1	30
มัธยมศึกษาปีที่ 5	ฉบับที่ 1	ตอนที่ 1 คณิตศาสตร์	40	60
		ตอนที่ 2 ภาษาไทย	40	60
		ตอนที่ 3 สังคมศึกษา	40	60
	ฉบับที่ 2	ตอนที่ 1 วิทยาศาสตร์	40	60
		ตอนที่ 2 ภาษาอังกฤษ	40	60

3. ความเป็นมาของการศึกษาการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาคุณภาพข้อสอบของผู้สอบกลุ่มต่าง ๆ ในประชากรมีมานานแล้ว แต่ด้านความยุติธรรมของข้อสอบหรือแบบสอบระหว่างผู้สอบกลุ่มต่าง ๆ เริ่มศึกษากันอย่างจริงจังในช่วงปลายทศวรรษของปี ค.ศ.1960 มีการเสนอวิธีการต่าง ๆ เพื่อตรวจสอบความลำเอียงของข้อสอบ (Item bias) ความลำเอียงของแบบสอบ (Test bias) และความลำเอียงในการคัดเลือก (Selection bias) โดยนิยามความลำเอียงว่าเป็น ความคลาดเคลื่อนอย่างเป็นระบบ (Systematic error) ที่เกิดขึ้นจากการวัด ความพยายามของการตรวจสอบความลำเอียงดังกล่าว ดำเนินไปเพื่อจำแนกข้อสอบที่ทำหน้าที่ไม่เหมาะสมหรือไม่ยุติธรรมสำหรับปรับปรุงหรือตัดข้อสอบข้อนั้นออกจากแบบสอบ เป็นการขจัดข้อสอบที่เกิดปัญหาความยุติธรรมระหว่างกลุ่มผู้สอบกลุ่มต่าง ๆ ที่มีลักษณะบางอย่างแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ ประสบการณ์ เป็นต้น เพื่อพัฒนาแบบสอบให้มีคุณภาพเหมาะสมสำหรับนำไปใช้ทดสอบต่อไป (ศิริชัย กาญจนวาสี, 2545, หน้า103)

การศึกษาความลำเอียงของข้อสอบนั้นเริ่มมาจากการสังเกตผลการสอบคัดเลือก ซึ่งพบว่าไม่เป็นไปตามสัดส่วนสติปัญญาหรือโครงสร้างของประชากร กรณีที่ทำให้มีการตื่นตัวมากในเรื่องความลำเอียงของข้อสอบ คือ ในปี ค.ศ. 1971 มาร์โคเดอฟูนิส (Marco Kefunis) และคนอื่น ๆ ซึ่งถูกปฏิเสธจากโรงเรียนกฎหมายของมหาวิทยาลัยวอชิงตัน ได้ฟ้องร้องว่าเขาได้คะแนนสูงกว่าผู้ที่ได้รับการคัดเลือกบางคน และได้ยื่นฎีกาฟ้องร้องชาร์ล โอเดการ์ด (Charles Odegaurs) เพื่อให้พิจารณาทบทวนการคัดเลือกนักศึกษาใหม่ และจากนั้นเป็นต้นมาการพิจารณาตรวจสอบความลำเอียงของข้อสอบระหว่างผู้สอบกลุ่มย่อยก็ได้รับปฏิบัติกันจนถึงปัจจุบัน ซึ่งการตรวจสอบความลำเอียงจะทำการตรวจสอบก่อนและหลังจากการทดลองใช้ การสำรวจความลำเอียงของข้อสอบก่อนนำไปทดลองใช้จะตรวจสอบโดยผู้เชี่ยวชาญ โดยพิจารณาถึงรูปแบบของข้อสอบ เนื้อหา คำที่ใช้และอื่น ๆ เพื่อไม่ให้เกิดความลำเอียง (รัชดาพร แก้วชาฎก, 2544, หน้า 14)

ศิริชัย กาญจนวาสี (2550, หน้า 116) ได้เสนอแนวคิดเกี่ยวกับการใช้คำและความหมายโดยมีประเด็นโต้แย้งว่าความลำเอียงของข้อสอบ เป็นผลการตัดสินว่าข้อสอบมีความยุติธรรมหรือไม่ อันส่งผลต่อการบรรลุจุดมุ่งหมายของการใช้แบบสอบหรือความลำเอียงของข้อสอบเป็นสารสนเทศทางสถิติที่ได้จากข้อสอบเกี่ยวกับความสัมพันธ์ระหว่างคุณลักษณะที่ข้อสอบมุ่งวัดกับประสบการณ์ของผู้สอบกลุ่มต่าง ๆ ที่ทำการสอบ เมื่อกลุ่มผู้สอบต่างกลุ่มกันตอบข้อสอบข้อเดียวกัน ความแตกต่างที่เกิดขึ้นอาจมาจากความไม่เหมาะสมของข้อคำถาม ซึ่งสามารถเกิดขึ้นได้หลายลักษณะหรือประสบการณ์ของผู้สอบซึ่งอาจมีลักษณะพื้นฐานเดิมแตกต่างกันในหลาย

สถานการณ์จึงไม่เหมาะสมที่จะใช้คำว่า ข้อสอบลำเอียง (Biased item) เนื่องจากเป็นภาษาที่มีความหมายในเชิงลบประกอบกับเกณฑ์ที่ใช้สำหรับตัดสินความลำเอียงยังมีความคลุมเครือและค่อนข้างสับสน ดังนั้นจึงควรเปลี่ยนมาใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) ซึ่งเป็นคำที่มีความเป็นกลางและเหมาะสมกว่า

การทำหน้าที่ต่างกันของข้อสอบ (DIF) กับความลำเอียงของข้อสอบ (Item bias) มีแนวคิดที่ต่างกัน สำหรับการทำหน้าที่ต่างกันของข้อสอบ เป็นกระบวนการที่เน้นการใช้วิธีการทางสถิติสำหรับตรวจสอบ เพื่อให้ได้สารสนเทศเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบสำหรับกลุ่มผู้สอบกลุ่มย่อยที่มีลักษณะเฉพาะบางอย่างแตกต่างกัน ส่วนความลำเอียงของข้อสอบเป็นกระบวน การตัดสอนความยุติธรรมของข้อสอบ โดยนำสารสนเทศการทำหน้าที่ต่างกันของข้อสอบมาวิเคราะห์ เชิงตรรกะ (Logical analysis) โดยผู้เชี่ยวชาญพิจารณาถึงการเขียนข้อสอบ เนื้อหาสาระของข้อสอบและจุดมุ่งหมายของการวัด เพื่อระบุว่าข้อสอบข้อนั้นลำเอียงเข้าข้างกลุ่มใดหรือไม่ เพราะเหตุใดจึงเป็นการตัดสินความลำเอียงของข้อสอบ

4. ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

นักวิจัยทางการวัดผลหลายท่านได้ให้ความหมายของความลำเอียงของข้อสอบและการทำหน้าที่ต่างกันของข้อสอบ ไว้ดังนี้

Scheuneman (1979 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า 104-105) ได้ให้ความหมายของความลำเอียงของข้อสอบว่า หมายถึง สัดส่วนของผู้สอบที่ตอบข้อสอบได้ถูกต้องไม่เท่ากันในแต่ละกลุ่มประชากรที่ใช้ในการศึกษา เมื่อกลุ่มผู้สอบมีคะแนนเท่ากันและข้อสอบมีความเป็นเอกพันธ์

Rudner,Geston and Knight (1980 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า104-105) ให้ความหมายของความลำเอียงของข้อสอบว่า เป็นข้อสอบที่มีค่าความยากสัมพัทธ์สำหรับสมาชิกของผู้สอบกลุ่มหนึ่งมากกว่าสมาชิกของผู้สอบอีกกลุ่มหนึ่ง

Popham (1981 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า 104-105) กล่าวว่าความลำเอียงของข้อสอบ หมายถึง ความโน้มเอียงของข้อสอบที่เมื่อใช้คะแนนจากข้อสอบนั้นแล้วทำให้การตัดสินผลเป็นไปอย่างไม่ยุติธรรม

Hulin,Drasgow and Parson (1983 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า104-105)ให้ความหมายของความลำเอียงของข้อสอบว่า หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันสำหรับการวัดความสามารถหรือโอกาสในการตอบข้อสอบในทางบวกแตกต่างกัน

สำหรับการวัดเจตคติ เมื่อผู้สอบที่มีคุณลักษณะของการวัดในปริมาณเท่ากัน แต่มาจากกลุ่มประชากรย่อยที่แตกต่างกัน

Dorans and Kulick (1986 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า 104-105) ได้ให้ความหมายของความลำเอียงของข้อสอบว่า หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบกลุ่มหนึ่งมีค่าต่ำกว่าหรือสูงกว่าผู้สอบอีกกลุ่มหนึ่งที่มีระดับความสามารถเดียวกัน

Kederman (1990 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า 104-105) ให้ความหมายของความลำเอียงของข้อสอบว่า หมายถึง คะแนนข้อสอบของกลุ่มผู้สอบที่มีความ สามารถเท่ากันแต่มาจากต่างกลุ่มกัน มีความแตกต่างกันอย่างเป็นระบบ

Holland and Wainer (1993 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า104-105) กล่าวว่าการทำหน้าที่ต่างกันของข้อสอบ หมายถึง สารสนเทศทางสถิติของข้อสอบที่ได้จากผลการตอบของผู้สอบต่างกลุ่มกันและมีความสามารถเท่ากัน แต่มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

Camili andShepard (1994 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า104-105) กล่าวว่าการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความเป็นพหุมิติในการวัดของข้อสอบซึ่งแสดงได้จากการแจกแจงความสามารถหลัก (Primary ability) ของกลุ่มผู้สอบตั้งแต่ 2 กลุ่มขึ้นไป มีความเท่ากันแต่มีการแจกแจงความสามารถรอง (Secondary ability) แตกต่างกัน

Narayanan and Swaminathan (1996 อ้างอิงใน ศิริชัย กาญจนวาสี, 2545, หน้า104-105) ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบว่า หมายถึง ฟังก์ชันการตอบสนองข้อสอบซึ่งคำนวณจากกลุ่มผู้สอบกลุ่มย่อยที่ต่างกัน มีค่าไม่เท่ากัน

พัชรี ปิยภัณฑ์ (2531, หน้า 7) ให้ความหมายของข้อสอบที่มีความลำเอียงว่า หมายถึง แบบสอบที่นำไปทดสอบกับกลุ่มที่มีความสามารถเท่ากันแต่มีความแตกต่างกันทางด้านเพศ เชื้อชาติ ศาสนา หรือสภาพภูมิศาสตร์ และเศรษฐกิจ ทำให้บุคคลในแต่ละกลุ่มมีการเสียเปรียบในการทำข้อสอบข้อเดียวกัน

บุญชม ศรีสะอาด (2535 อ้างอิงใน สุรศักดิ์ อมรรัตนศักดิ์, 2544, หน้า 284) ได้ให้ความหมายว่า ข้อสอบที่มีความลำเอียงเป็นข้อสอบที่ให้ผลการสอบแตกต่างกันจากกลุ่มผู้สอบที่มีความสามารถไม่ต่างกัน และเสนอว่าผลการสอบที่ต่างกันนั้นมาจากองค์ประกอบบางอย่าง เช่น ภาษา เพศ วัฒนธรรม ศาสนา เชื้อชาติ สภาพทางภูมิศาสตร์ สภาพทางเศรษฐกิจและสังคม ประสบการณ์ส่วนตัว เป็นต้น

จิตสุดา ธราพร (2539 อ้างอิงใน สุรศักดิ์ อมรรัตนศักดิ์, 2544, หน้า 284) ได้ให้ความหมายของความลำเอียงของข้อสอบว่า หมายถึง ข้อสอบที่เมื่อนำไปทดสอบกับกลุ่มผู้สอบที่มีความสามารถเท่ากันภายใต้ประชากรเดียวกันแต่มีลักษณะของกลุ่มย่อยที่แตกต่างกันทางด้านใดด้านหนึ่ง เช่น เพศ เชื้อชาติ ศาสนา วัฒนธรรม สภาพทางภูมิศาสตร์ หรือเศรษฐกิจ แล้วปรากฏว่าโอกาสในการที่จะตอบข้อนั้นถูกไม่เท่ากัน หรือทำให้ในแต่ละกลุ่มย่อยมีการได้เปรียบเสียเปรียบจากการตอบข้อสอบข้อเดียวกัน

ศิริชัย กาญจนวาสี (2545, หน้า 105) ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบ (DIF) ว่าหมายถึง การที่ข้อสอบทำให้ผู้สอบจากต่างกลุ่มกันที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือมีฟังก์ชันตอบสนองข้อสอบแตกต่างกัน การทำหน้าที่ต่างกันของข้อสอบเกิดขึ้นเมื่อนำข้อสอบไปทดสอบกับผู้สอบกลุ่มย่อยต่างกัน ที่มีความสามารถหลัก(Primary ability) ระดับเดียวกันหรือมีคุณลักษณะแฝง (Latent trait) ที่ต้องการวัดเท่ากัน แต่มีความสามารถรอง(Secondary ability)แตกต่างกัน ทำให้ผู้สอบต่างกลุ่มที่นำมาจับคู่เปรียบเทียบมีโอกาสตอบข้อสอบถูกแตกต่างกัน

จากความเป็นมาและความหมายข้างต้น ดังนั้นพอสรุปได้ว่าแบบทดสอบที่มีทำหน้าที่ต่างกัน หมายถึง แบบทดสอบที่เมื่อนำไปทดสอบกับกลุ่มผู้สอบที่มีความสามารถเท่าเทียมกัน แต่มีลักษณะของกลุ่มย่อยแตกต่างกัน เช่น เพศ เชื้อชาติ ศาสนา วัฒนธรรม สภาพภูมิศาสตร์ เศรษฐกิจ หรือแม้แต่ประสบการณ์เป็นต้น แล้วปรากฏว่าความน่าจะเป็นในการตอบข้อสอบถูกไม่เท่ากันทำให้มีการได้เปรียบเสียเปรียบจากการตอบข้อสอบเดียวกัน

5. สาเหตุที่ทำให้ข้อสอบทำหน้าที่ต่างกัน

จากการวิเคราะห์การทำหน้าที่ต่างกันด้วยค่าสถิติแบบต่าง ๆ ในแบบทดสอบหลาย ๆ ฉบับ เพื่อศึกษาถึงสาเหตุที่ทำให้เกิดการทำหน้าที่ต่างกันในตัวคำถามเหล่านั้น พบว่าสาเหตุที่น่าจะก่อ ให้เกิดการทำหน้าที่ต่างกันมีหลายสาเหตุด้วยกัน ซึ่งสามารถสรุปสาเหตุที่ก่อให้เกิดการทำหน้าที่ต่างกันของข้อสอบมากที่สุดได้ดังต่อไปนี้ (Scheunerman, 1982 อ้างอิงใน รัชดาพร แก้วชาฎก, 2544, หน้า 13)

1. เดา (Guess) อาจจะเกิดจากข้อสอบนั้นยากเกินไป หรือเวลาไม่เพียงพอจะก่อให้เกิดความไม่เท่าเทียมกันในโอกาสการตอบข้อสอบถูกของผู้สอบแต่ละคน

2. ความเร็ว (Speed) หรือเวลาในการตอบจะทำให้เกิดการเดาหรือในกรณีเวลาน้อยอาจจะทำข้อสอบไม่ทันซึ่งจะมีผลกับข้อสอบข้อหลัง ๆ โดยเฉพาะในการศึกษาความลำเอียงของข้อสอบวัดความถนัด

3. ความกำกวมหรือความไม่ชัดเจน (Unclear) ของข้อคำถาม นั้นคือข้อคำถามขาดความเป็นปรนัย การใช้ภาษาถิ่นหรือใช้คำที่ไม่เป็นภาษากลางในการสื่อความหมายซึ่งจะก่อให้เกิดความลำเอียงกับกลุ่มภาษาใดภาษาหนึ่งขึ้นได้

4. ลำดับขั้นของคำถาม (Series) อาจจะเป็นสิ่งที่ก่อให้เกิดความสับสนหรือชี้แนะคำตอบข้อสอบบางข้อได้

5. สถานภาพทางสังคมหรือความเกี่ยวข้องทางสังคม (Social Implication) ก็เป็นสิ่งที่ก่อ ให้เกิดความแตกต่างระหว่างกลุ่มผู้สอบแต่ละกลุ่มได้

6. ประสบการณ์หรือการฝึกฝนของแต่ละกลุ่มย่อยเป็นสิ่งที่ก่อให้เกิดการได้เปรียบเสียเปรียบของแต่ละกลุ่มค่อนข้างจะชัดเจน

7. องค์ประกอบทางวัฒนธรรม ความเป็นอยู่ ขนบธรรมเนียมประเพณี เชื้อชาติ ศาสนาก็จะเอื้อให้กับบางกลุ่มย่อย จึงทำให้เกิดการได้เปรียบเสียเปรียบในบางเนื้อหาวิชาได้

นอกจากนี้การทำหน้าที่ต่างกันของข้อสอบอาจจะเกิดจากสาเหตุที่สำคัญ 2 ประการ คือ

1. การเลือกเนื้อหา คือ ผู้สร้างข้อสอบเลือกเนื้อหาเฉพาะส่วนใดส่วนหนึ่งมาสร้างข้อสอบทำให้ได้ข้อสอบที่มีเนื้อหาไม่ครอบคลุมและไม่ได้สัดส่วนที่สมดุลกัน

2. การสร้างข้อสอบ คือ การใช้ภาษาหรือข้อความบางอย่างที่เอื้อให้เกิดประโยชน์กับผู้สอบกลุ่มใดกลุ่มหนึ่ง

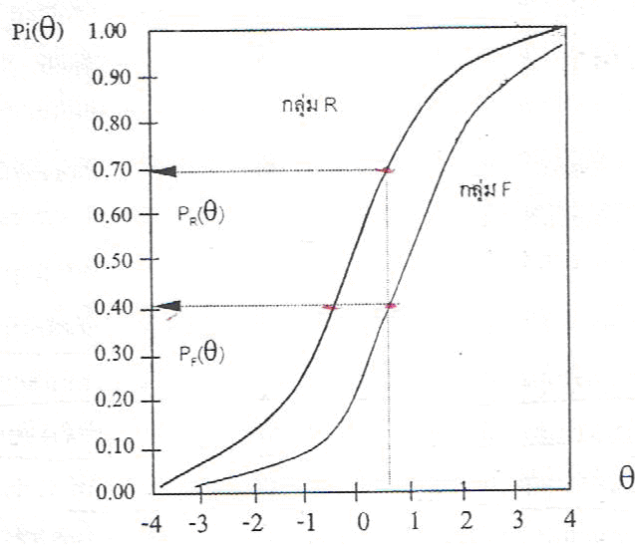
6. ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบเป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป ปกตินิยมทำการเปรียบเทียบ 2 กลุ่ม ประกอบด้วยกลุ่มแรก เรียกว่า กลุ่มเปรียบเทียบ (Focal group หรือกลุ่ม F) เป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ และกลุ่มที่สอง เรียกว่า กลุ่มอ้างอิง (Reference group หรือกลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้อง (ศิริชัย กาญจนวาสี, 2545, หน้า 106)

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะพบว่า ข้อสอบสามารถทำหน้าที่แตกต่างกัน 2 ประเภท ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และแบบอเนกรูป (Nonuniform)

6.1 ข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หมายถึง ข้อสอบที่ทำให้ผู้สอบ กลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งอย่างสม่ำเสมอในทุก

ระดับ ความสามารถ เมื่อพิจารณาโค้งคุณลักษณะของข้อสอบของผู้สอบ 2 กลุ่ม จะพบว่าไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม ดังภาพ 1



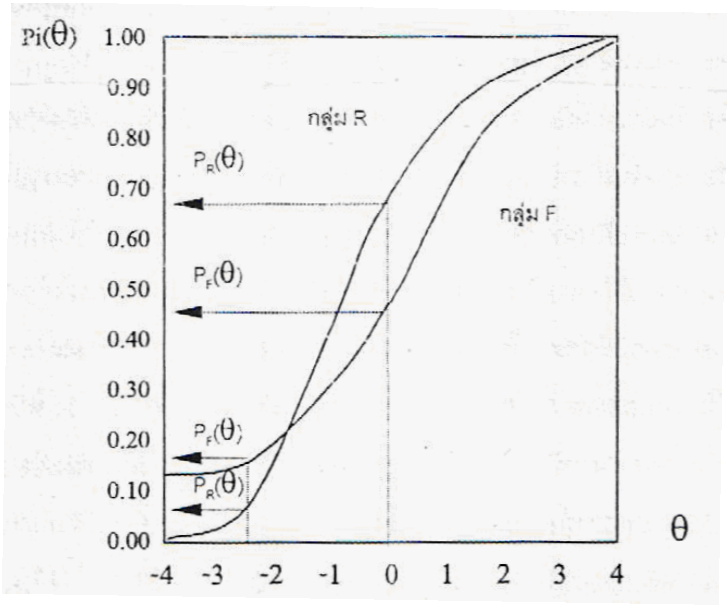
ภาพ 1 แสดงข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF)

ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) สามารถพิจารณา “ปฏิสัมพันธ์” ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูปแล้วโค้งลักษณะของข้อสอบ (Item Characteristic Curves: ICC) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน หรือมีฟังก์ชันการตอบข้อสอบต่างกัน ดังนั้นความแตกต่างระหว่างโค้งลักษณะข้อสอบทั้งสองแบบจะบ่งบอกถึงขนาดและทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งสามารถคำนวณได้โดยใช้สูตรของการคำนวณพื้นที่ของ Raju

6.2 ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มแตกต่างกันอย่างไม่สม่ำเสมอในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างกลุ่มความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม เช่น ที่ระดับความสามารถหนึ่งกลุ่มผู้สอบกลุ่ม R มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม F แต่ที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้สอบกลุ่ม F มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม R

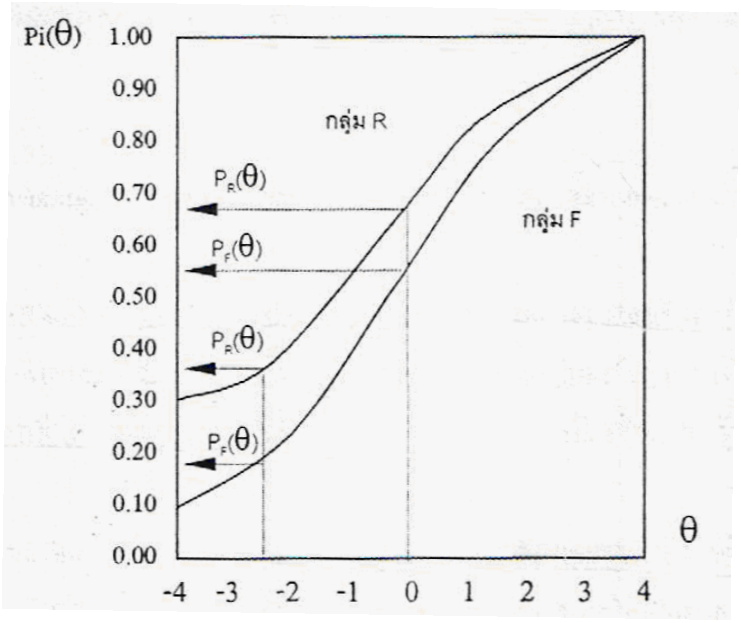
ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปสามารถจำแนกได้เป็น 2 ลักษณะ ดังนี้

1. ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้นเมื่อโค้งลักษณะข้อสอบตัดกันระหว่างช่วงความสามารถของผู้สอบหรือเรียกว่า ข้อสอบที่ทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Nondirectional DIF) ดังภาพ 2



ภาพ 2 แสดงข้อสอบที่ทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Nondirectional DIF)

2. ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์เป็นลำดับ (Ordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบต่างกันอย่างไม่สม่ำเสมอ แต่ไม่ตัดกัน หรืออาจตัดกันนอกช่วงความสามารถของผู้สอบตรงปลายสุดของช่วงความสามารถต่ำหรือสูง อาจเรียกข้อสอบลักษณะนี้ว่า ข้อสอบที่ทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirectional DIF) ดังภาพ 3



ภาพ 3 แสดงข้อสอบที่ทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirectional DIF)

7. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่ม ที่มีความสามารถหลักที่มุ่งวัดเท่ากัน แต่คาดว่าจะมีความได้เปรียบเสียเปรียบกัน โดยที่กลุ่มหนึ่งถือเป็นกลุ่มอ้างอิง ซึ่งคาดว่าน่าจะได้เปรียบในการตอบข้อสอบข้อนั้น หรือมีโอกาสตอบข้อสอบได้ถูกต้องมากกว่า ส่วนอีกลุ่มหนึ่งคือกลุ่มเปรียบเทียบ ซึ่งเป็นกลุ่มที่สนใจศึกษาและคาดว่าน่าจะเป็นกลุ่มที่เสียเปรียบ (ศิริชัย กาญจนวาสี, 2545, หน้า 108 – 111)

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำเป็นต้องจับคู่ผู้สอบตามความสามารถ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เกณฑ์การจับคู่ที่นิยมใช้กันมี 2 วิธี ดังนี้

1. เกณฑ์ภายนอก (External Criterion)

การวิเคราะห์การทำหน้าที่ต่างกันโดยใช้เกณฑ์ภายนอกนี้สามารถนำไปใช้ได้ทั้งข้อสอบ รายข้อและแบบทดสอบทั้งฉบับ โดยการใช้คะแนนจากแบบสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิค การวิเคราะห์การทดถอย (Regression analysis) เพื่อทำการเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

หลักการนี้มีจุดมุ่งหมายเพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบสอบ อื่นจากตัวแปรทำนายซึ่งเป็นคะแนนรายข้อ หรือคะแนนแบบสอบระหว่างกลุ่มอ้างอิง

และกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะใช้คะแนนรายข้อเป็นตัวแปรทำนาย แต่ถ้าเป็นการทำหน้าที่ต่างกันของแบบสอบจะใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอกอาจใช้คะแนนรวมทั้งฉบับหรือเกรดเฉลี่ยหรือผลสัมฤทธิ์ในงานที่เกี่ยวข้องของผู้สอบ สมการทำนายสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแสดงได้ดังนี้

กลุ่มอ้างอิง (R) $Y_i = A_R + B_R X_i$

กลุ่มเปรียบเทียบ (F) $Y_i = A_F + B_F X_i$

เมื่อ Y_i = คะแนนของตัวแปรเกณฑ์ภายนอก

X_i = คะแนนของตัวแปรทำนาย

A = ค่าคงที่หรือค่าตัดแกน (intercept)

B = ค่าความชัน (slope)

จากฟังก์ชันการทำนายทั้ง 2 สมการสามารถเปรียบเทียบค่าตัดแกน (A) และค่าความชัน (B) ของเส้นกราฟระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบได้ ถ้าเส้นกราฟดังกล่าวมีค่าความชันหรือค่าตัดแกนแตกต่างกันสำหรับข้อสอบใดหรือแบบสอบใด แสดงว่าข้อสอบหรือแบบสอบนั้นมีการทำหน้าที่ต่างกัน โดยเข้าข้างกลุ่มผู้สอบที่มีค่าตัดแกนหรือค่าความชันสูงกว่า

การใช้เกณฑ์ภายนอกมีข้อดี คือ เกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบและแบบสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อนตรงที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ ในทางปฏิบัติเป็นการยากที่จะหาตัวแปรเกณฑ์ภายนอกจากแบบสอบฉบับอื่นที่มีความตรงเชิงทำนายและมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ถ้าตัวแปรเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าวจะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบขาดความแม่นยำ และความสมบูรณ์

2. เกณฑ์ภายใน (Internal Criterion)

การวิเคราะห์การทำหน้าที่ต่างกันโดยใช้เกณฑ์ภายในเป็นการนำวิธีทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบ โดยเน้นการพิจารณาจากโครงสร้างภายในของแบบสอบเป็นหลัก ด้วยการวิเคราะห์ผลจากการตอบข้อสอบและความสามรถหรือคะแนนจริงของผู้สอบที่ได้จากแบบสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้สอบจากกลุ่มอ้างอิงและกลุ่ม

เปรียบเทียบที่มีความสามารถหรือคะแนนจริงที่เท่ากันว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ ค่าสถิติที่นิยมนำมาใช้พอสรุปได้ดังนี้

2.1 การทดสอบปฏิสัมพันธ์ (Interaction)

ในระยะเริ่มแรกของการศึกษาความลำเอียงของข้อสอบมีการใช้สถิติทดสอบเอฟ (F-test) จากการวิเคราะห์ความแปรปรวน (ANOVA) เพื่อทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับข้อสอบ ถ้าการทดสอบมีนัยสำคัญเป็นสัญญาณของการทำหน้าที่ต่างกันของข้อสอบ จากนั้นจึงทำการวิเคราะห์ต่อด้วยวิธีการ Post Hoc เพื่อระบุข้อสอบที่มีผลต่อการเกิดปฏิสัมพันธ์ ซึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกัน

วิธีการนี้มีข้อดีที่สามารถศึกษาผู้สอบได้หลาย ๆ กลุ่มได้สะดวก แต่มีจุดอ่อนในเรื่องการควบคุมกลุ่มต่าง ๆ ให้มีคามสามารถที่ทัดเทียมกัน ขนาดกลุ่มตัวอย่างของกลุ่มต่าง ๆ และอัตราความคลาดเคลื่อนประเภทที่ 1 จะสูงขึ้น ถ้าจำนวนข้อสอบเพิ่มมากขึ้น

2.2 การวัดความเบี่ยงเบนสัมพัทธ์ (Relative Deviation)

การคำนวณค่าความยากของข้อสอบ เช่น p , b เป็นต้น เมื่อคำนวณแยกระหว่างกลุ่มและแปลงให้เป็นค่าความยากมาตรฐาน (Δ) สามารถนำมาพล๊อตเปรียบเทียบเป็นรายข้อ ถ้าข้อใดเบี่ยงเบนไปจากแกนหลักที่คาดหมายหรือเบี่ยงเบนเกินจากความคลาดเคลื่อนมาตรฐานของค่าความยากที่กำหนด ย่อมแสดงถึงการทำหน้าที่ต่างกันของข้อสอบ รวมทั้งสามารถคำนวณค่าสหสัมพันธ์ระหว่างค่าความยากรายข้อระหว่างกลุ่ม เพื่อแสดงถึงการทำหน้าที่ต่างกันของข้อสอบ ถ้าสหสัมพันธ์เข้าใกล้ 1.00 แสดงว่าค่าความยากสัมพัทธ์ของข้อสอบมีค่าใกล้เคียงกันระหว่างกลุ่ม ดังนั้นแบบสอบวัดคุณลักษณะคล้ายกันระหว่างกลุ่ม

วิธีการนี้มีข้อดีข้อเสียคล้ายการทดสอบปฏิสัมพันธ์ นอกจากนี้ค่าความยากของข้อสอบ (p) มิใช่ตัวแทนของความยากจริงของข้อสอบ และได้รับอิทธิพลจากค่าแทรกซ้อนอื่นได้แก่ ค่าอำนาจจำแนก และความสามารถของผู้สอบ

2.3 การเปรียบเทียบน้ำหนักตัวประกอบ (Factor Loading)

การวิเคราะห์ตัวประกอบ (Factor Analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ในการตรวจสอบความตรงเชิงทฤษฎีหรือโครงสร้าง (Construct Validity) เมื่อนำการวิเคราะห์ตัวประกอบมาใช้ในการวิเคราะห์โครงสร้างของแบบสอบแยกตามกลุ่มผู้สอบ ความไม่สอดคล้องกันระหว่างน้ำหนักตัวประกอบบนคุณลักษณะสำคัญที่มุ่งวัด หรือความแตกต่างของค่าเฉลี่ยคะแนน

ตัวประกอบระหว่างกลุ่ม (Factor scores) ระหว่างกลุ่มผู้สอบย่อมสะท้อนการทำหน้าที่ต่างกันของข้อสอบและแบบสอบ

การใช้เทคนิคการวิเคราะห์ตัวประกอบเชิงสำรวจ(Exploratory Factor Analysis ; EFA) สำหรับศึกษาการทำหน้าที่ต่างกันจะมีจุดอ่อนในเรื่องความไม่สอดคล้องระหว่างน้ำหนักตัว ประกอบอาจเกิดจากความแตกต่างของความสามารถระหว่างกลุ่มได้ แนวทางที่เหมาะสมจึงควรใช้เทคนิคการวิเคราะห์ตัวประกอบเชิงยืนยัน (Confirmatory Factor Analysis ; CFA) นอกจากนี้ยังสามารถใช้ CFA สำหรับตรวจสอบความแตกต่างระหว่างกลุ่มในด้านคุณลักษณะหรือความสามารถหลัก และความสามารถรองได้อีกด้วย

2.4 การเปรียบเทียบโอกาสตอบข้อสอบถูก

การวิเคราะห์โอกาสตอบข้อสอบถูกของผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน เป็นแนวทางสำคัญที่นิยมใช้กันและเป็นที่ยอมรับในปัจจุบันสำหรับบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ มีการคำนวณค่าสถิติ 2 แนวทาง ดังนี้

2.4.1 เปรียบเทียบค่าสัดส่วนหรือความน่าจะเป็นในการตอบข้อสอบถูกของผู้สอบ ต่างกลุ่มที่มีความสามารถเท่ากัน เช่น วิธีแมนเทล – แฮนส์เซล (MH) เป็นต้น

2.4.2 เปรียบเทียบค่าฟังก์ชันการตอบสนองข้อสอบหรือโค้งลักษณะข้อสอบต่างกลุ่มที่มีระดับความสามารถเท่ากัน เป็นวิธีที่อยู่บนพื้นฐานของทฤษฎี IRT เช่น วิธีวัดความแตกต่างของพื้นที่ วิธีวัดความแตกต่างของค่าพารามิเตอร์ความยาก วิธีการทดสอบไค – สแควร์ของลอร์ด (Lord's χ^2-test) เป็นต้น

วิธีการนี้มีข้อดีที่สำคัญได้แก่ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือมีกลไกควบคุมความสามารถของผู้สอบโดยการจับคู่กลุ่มความสามารถ เพื่อทำการเปรียบเทียบณ ตำแหน่งต่าง ๆ ที่มีความสามรถเท่ากัน จึงเป็นวิธีที่ยอมรับทั่วไป แต่มีข้อจำกัดในด้านความสลับซับซ้อนของแนวคิดพื้นฐาน และการวิเคราะห์มีความจำเป็นต้องใช้โปรแกรมคอมพิวเตอร์โดยเฉพาะ

8. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF detection) จำแนกตามลักษณะการตรวจให้คะแนนได้เป็น 2 ประเภท คือ ข้อสอบที่มีการให้คะแนนแบบทวิวิภาค หรือสองค่า (Dichotomous scoring) และข้อสอบทีมีการให้คะแนนแบบพหุวิภาค หรือหลายค่า (Polytomous scoring) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแต่ละประเภท ยังสามารถจำแนกได้อีก 2 มิติ ได้แก่ มิติลักษณะของตัวแปรเกณฑ์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้คะแนนสังเกตได้ (Observed

score) และกลุ่มวิธีที่ใช้คะแนนสังเกตไม่ได้หรือคะแนนของตัวแปรแฝง (Latent variable) และมิติลักษณะของสถิติวิเคราะห์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้สถิติพาราเมตริก (Parametric approach) และกลุ่มวิธีที่ใช้สถิตินันพาราเมตริก (Nonparametric approach) รายชื่อวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่สำคัญ ๆ ดังแสดงในตาราง 1 (ศิริชัย กาญจนวาสี, 2545, หน้า 112)

ตาราง 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการตรวจให้คะแนนแบบทวิวิภาค (Dichotomous DIF) และพหุวิภาค (Polytomous DIF)

ประเภทและตัวแปรเกณฑ์	พาราเมตริก	นันพาราเมตริก
1. DIF แบบทวิวิภาค		
1.1 คะแนนที่สังเกตได้	ANOVA	TID
(Observed score)	Logistic Regression	HM
		STND
1.2 คุณลักษณะ/ตัวแปรแฝง	IRT – D^2	SIBTEST
(Latent variable)	Lord's χ^2	
	General IRTLR	
	Loglinear IRTLR	
2. DIF แบบทพหุวิภาค		
2.1 คะแนนที่สังเกตได้	ANOVA	
(Observed score)	Polytomous Logistic	Polytomous STND
	Regression	GMH
2.2 คุณลักษณะ/ตัวแปรแฝง	General IRTLR	Polytomous SIBTEST
(Latent variable)	PCM	GPCM

1. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิวิภาค

1.1 กลุ่มวิธีที่ใช้คะแนนที่สังเกตได้

วิธีในกลุ่มนี้มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) หรือกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (Non – IRT approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่

1.1.1 การวิเคราะห์ความแปรปรวน (ANOVA)

1.1.2 วิธีการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression: LR)

1.1.3 วิธีแปลงค่าความยากข้อสอบ (Transformed Item Difficulty: TID)

1.1.4 วิธีแมนเทล – แฮนส์เซล (Mantel – Haenszel: MH)

1.1.5 วิธีดัชนีมาตรฐาน (Standardization: STND) การปรับให้เป็นมาตรฐานด้วยน้ำหนักตัวประกอบ

1.2 กลุ่มวิธีที่ใช้คุณลักษณะแฝง

วิธีในกลุ่มนี้ใช้คุณลักษณะหรือตัวแปรแฝง ซึ่งวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่

1.2.1 วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT – D^2)

1.2.2 วิธีไค – สแควร์ของลอร์ด (Lord's χ^2)

1.2.3 วิธีอัตราส่วนไลค์ลิฮูดทั่วไป (General IRT Likelihood Ratio)

1.2.4 วิธีอัตราส่วนไลค์ลิฮูด ลอกลิเนียร์ (Loglinear IRT Likelihood Ratio)

1.2.5 วิธีซิปเทสท์ (SIBTEST)

2. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบพหุวิภาค

2.1 กลุ่มวิธีที่ใช้คะแนนที่สังเกตได้

2.1.1 วิธีการวิเคราะห์ความแปรปรวน (ANOVA)

2.1.2 วิธีการวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous Logistic Regression)

2.1.3 วิธีดัชนีมาตรฐานพหุวิภาค (Polytomous Standardization)

2.1.4 วิธีแมนเทล – แฮนส์เซลทั่วไป (General Mantel – Haenszel: GMH)

2.2 กลุ่มวิธีที่ใช้คุณลักษณะแฝง

2.2.1 วิธีอัตราส่วนไลค์ลิฮูดในรูปทั่วไป (General IRT Likelihood Ratio)

2.2.2 วิธีการให้คะแนนบางส่วน (Partial Credit Model: PCM)

2.2.3 วิธีซิปเทสท์พหุวิภาค (Polytomous SIBTEST)

2.2.4 วิธีการให้คะแนนบางส่วนทั่วไป (Generalized Partial Credit Model: GPCM)

กาญจนนา วัธนสุนทร (2538, หน้า 12) ได้สรุปวิธีที่นิยมใช้มี 6 วิธี ได้แก่ (1) ค่าความยากแปลง (Transformed Item Difficulty: TID) (2) วิธีวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) (3) วิธีไค – สแควร์ (Chi – Square: χ^2) (4) วิธีทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) (5) วิธีแมนเทล – แฮนส์เซล (Mantel – Haenszel: MH) และ (6) วิธีซิปเทสท์ (SIBTEST) ซึ่งแต่ละวิธีมีขั้นตอนการวิเคราะห์ ข้อดี ข้อบกพร่องดังตาราง 3

ตาราง 3 แสดงการเปรียบเทียบวิธีต่าง ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ประเด็น	TID	ANOVA	χ^2	IRT	MH	SIBTEST
1. ข้อตกลงเบื้องต้น	ผลร่วมระหว่างกลุ่มกับข้อสอบเป็นตัวบ่งชี้การทำหน้าที่ต่างกัน	ความแปรปรวนและความแปรปรวนร่วมของข้อสอบต้องเท่ากัน	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ	แบบสอบเป็นเอกมิติและโค้งลักษณะข้อสอบสามารถแสดงฟังก์ชันค่าความสามารถและโอกาสในการตอบข้อนั้นถูก	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบและมิติการวัด 2 มิติ คือ คุณลักษณะแฝงเป้าหมายและคุณลักษณะแฝงแทรกซ้อน

ตาราง 3 (ต่อ)

ประเด็น	TID	ANOVA	χ^2	IRT	MH	SIBTEST
2. สิ่งที่ทำการวิเคราะห์	ผลร่วมระหว่างการเป็นสมาชิกในกลุ่มกับการตอบถูก	ผลร่วมระหว่างการเป็นสมาชิกในกลุ่มกับการตอบถูก	ความแตกต่างของอัตราส่วนการตอบถูกตามระดับคะแนนรวม	ความแตกต่างของฟังก์ชันการตอบข้อสอบที่ระดับความสามารถเดียวกัน	ความแตกต่างของอัตราส่วนการตอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน	ความแตกต่างระหว่างคะแนนเฉลี่ยและอัตราส่วนการตอบข้อสอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน
3. สิ่งที่พิจารณาในการตัดสิน DIF	ระยะห่างของจุดเดลตาจากเส้นแกนหลัก	ความมีนัยสำคัญทางสถิติของ F-test	ความมีนัยสำคัญทางสถิติของ χ^2	พื้นที่ระหว่างโค้งลักษณะข้อสอบ	ค่าดัชนี α_{MH} และความมีนัยสำคัญทางสถิติ	ค่าดัชนี β_{SIB} และความมีนัยสำคัญทางสถิติ
4. ทฤษฎีพื้นฐาน	CTT	-	-	IRT	-	IRT ชนิดพหุมิติ
5. ข้อดี	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย	ใช้กลุ่มตัวอย่างน้อย	คำนวณง่าย มีเกณฑ์ตายตัวในการแปลผล	ให้รายละเอียดมากและไม่มีการเปลี่ยนแปลงค่าพารา - มิเตอร์	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อยประหยัด	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย ตรวจสอบ DIF ได้หลายข้อในคราวเดียวกัน
6. ข้อบกพร่อง	มีความคลาดเคลื่อนเมื่อค่า a สูงและค่า b เปลี่ยนไปตามกลุ่มผู้สอบ	การคำนวณค่อนข้างยุ่งยากและไม่มีดัชนีบอกระดับการทำหน้าที่ต่างกัน	ไม่มีเกณฑ์กำหนดตายตัวในการกำหนดช่วงคะแนนและค่า b เปลี่ยนตามกลุ่มผู้สอบ	มีการคำนวณซับซ้อนหลายรอบ แปลผลยาก ใช้กลุ่มตัวอย่างมาก ค่าใช้จ่ายสูง	ไม่มีความไวในการตรวจสอบ DIF แบบอเนกรูป	อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มสูงขึ้นเมื่อคะแนนเฉลี่ยแตกต่างกันมาก

9. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย IRT

9.1 ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT)

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) เป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในบุคคล (Latent trait or ability) กับผลการตอบข้อสอบหรือข้อคำถามโดยใช้โค้งลักษณะข้อสอบ(Item Characteristic Curve: ICC) ซึ่งมีการกำหนดลักษณะของข้อสอบด้วยพารามิเตอร์ ทฤษฎีนี้ได้อธิบายความสัมพันธ์ดังกล่าวในรูปของฟังก์ชันคณิตศาสตร์ หรือโมเดลที่แสดงความสัมพันธ์ระหว่างระดับความสามารถคุณลักษณะของข้อสอบและโอกาสของการตอบข้อสอบได้ถูกที่เรียกว่า ฟังก์ชันการตอบสนองข้อสอบ ซึ่งมีลักษณะความสัมพันธ์เป็นแบบฟังก์ชันโลจิส โค้งลักษณะข้อสอบมีหลายลักษณะขึ้นอยู่กับโมเดล (Model) หรือแบบจำลองที่ใช้อธิบายความสัมพันธ์ดังกล่าว โมเดลที่นิยมใช้กัน คือ โมเดลแบบหนึ่งพารามิเตอร์ (One-Parameter Model) โมเดลแบบสองพารามิเตอร์ (Two-Parameter Model) และโมเดลแบบสามพารามิเตอร์ (Three-Parameter Model) (ศิริชัย กาญจนวาสี, 2550, หน้า 53-54)

9.1.1 หลักการของทฤษฎีการตอบสนองข้อสอบ

ในการวิเคราะห์ข้อสอบโดยทั่วไปแล้วจะพิจารณาตามทฤษฎีดั้งเดิม (Classical Test Theory) ซึ่งเมื่อพิจารณารายข้อ (Item) จะดูจากค่าความยาก (p) และค่าอำนาจจำแนก (r) เมื่อพิจารณารวมทั้งฉบับ (Test) จะดูจากค่าความเชื่อมั่น (Reliability) และค่าความเที่ยงตรง (Validity) ซึ่งจากการพิจารณาโดยภาพรวมแล้วพบว่ายังมีจุดอ่อนอยู่หลายประการ คือ ประการแรก ค่าพารามิเตอร์ของข้อสอบแปรเปลี่ยนไปตามกลุ่มของผู้สอบที่แตกต่างกันในด้านความสามารถ (Ability) ประการที่สอง การเปรียบเทียบความสามารถของผู้สอบจำกัดอยู่ในสถานการณ์ที่ทดสอบ และประการที่สาม จะไม่สามารถบอกได้ว่าผู้เข้าสอบคนหนึ่งจะทำข้อสอบได้เพียงใดเมื่อได้เผชิญกับข้อคำถามหนึ่ง ยกเว้นเมื่อได้มีการใช้ข้อสอบนั้นแล้วกับกลุ่มตัวอย่างที่คล้ายคลึงกันกับบุคคลนั้น จากเหตุการณ์เหล่านี้ทำให้นักทดสอบทางจิตวิทยาสำรวจและพัฒนาทฤษฎี ทฤษฎีที่พัฒนาจนถึงปัจจุบัน คือ ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) โดยใช้ข้อตกลงเบื้องต้นว่าความสามารถของบุคคลการทำข้อสอบที่วัดความสามารถนั้นเป็นอย่างไร ซึ่งทฤษฎีการตอบข้อสอบมีหลักการที่สำคัญอยู่ที่การใช้ผลการตอบแบบทดสอบมาอธิบายถึงความสามารถของผู้สอบในเรื่องที่ทำการทดสอบนั้น ซึ่งจุดเด่นของการวิเคราะห์ทฤษฎีนี้ คือ ค่าพารามิเตอร์ของข้อสอบจะคงที่โดยค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าสัมประสิทธิ์การเดา (c) จะเป็นค่าที่ไม่แปรเปลี่ยนไปตามกลุ่มผู้ตอบไม่ว่าจะนำไปสอบกับผู้ใดก็ตาม เมื่อทราบลักษณะการตอบข้อสอบในแต่ละข้อคำถามของผู้เข้าสอบคนใดจะสามารถ

คำนวณหาค่าความสามารถที่แท้จริงของผู้สอบคนนั้นได้ โดยค่านี้จะสัมพันธ์โดยตรงกับคะแนนจริงซึ่งถือเป็นลักษณะของความเป็นอิสระของข้อสอบ

9.1.2 ข้อตกลงเบื้องต้นของทฤษฎีการตอบข้อคำถาม

1) ความเป็นมิติเดียวของข้อสอบ (Unidimensionality Test) เป็นแบบทดสอบที่ประกอบด้วยข้อคำถามที่เป็นเอกพันธ์ นั่นคือ แบบทดสอบนั้นจะต้องมุ่งวัดความสามารถอย่างใดอย่างหนึ่งที่มีลักษณะเด่น ๆ เพียงความสามารถเดียว หากไม่กำหนดข้อตกลงเบื้องต้นเช่นนี้จะทำให้โมเดลที่ใช้มีความสลับซับซ้อนมาก สำหรับการตรวจสอบว่าแบบทดสอบนั้นวัดในมิติเดียวหรือไม่ สามารถกระทำได้โดยเทคนิคการวิเคราะห์ทางสถิติ ได้แก่ การวิเคราะห์องค์ประกอบ (Factor Analysis) เพื่อคำนวณค่าไอเกน (Eigen Value) สำหรับศึกษาอัตราส่วนระหว่างค่าไอเกนของตัวประกอบแรกกับตัวประกอบถัดไป ถ้ามีอัตราส่วนที่สูงแสดงถึงเครื่องมือหรือแบบวัดคุณลักษณะเด่นเดียว (Single Dominant Factor) หรือทำการวิเคราะห์เชิงยืนยัน (Confirmatory Factor Analysis) เพื่อตรวจสองเชิงยืนยันว่า เครื่องมือหรือแบบสอบมุ่งวัดเพียงคุณลักษณะเดียวหรือความสามารถเดียว (ศิริชัย กาญจนวาสี, 2545, หน้า 67)

2) ความเป็นอิสระในการตอบข้อสอบ (Local Independence) เป็นการกำหนดข้อตกลงเบื้องต้นเกี่ยวกับความเป็นอิสระในการตอบข้อสอบ กล่าวคือ การตอบข้อสอบข้อใดข้อหนึ่งถูกจะไม่มีผลต่อการตอบข้อสอบข้ออื่น ๆ ในการตรวจสอบว่าข้อสอบแต่ละข้อเป็นไปตามข้อตกลงเกี่ยวกับความเป็นอิสระในการตอบข้อสอบหรือไม่ ให้สังเกตความเป็นมิติเดียวกันของแบบทดสอบถ้าแบบทดสอบมีความเป็นมิติเดียวแล้ว แบบทดสอบนั้นจะมีความเป็นอิสระในการตอบข้อสอบด้วย

3) โค้งลักษณะข้อสอบ (Item Characteristic Curve) เป็นฟังก์ชันทางคณิตศาสตร์ที่แสดงถึงความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบข้อนั้นได้ถูกต้องตรงกับระดับความสามารถที่วัดได้ โดยใช้ชุดของข้อสอบหรือแบบทดสอบนั้นและความน่าจะเป็นที่ผู้สอบจะตอบข้อสอบได้ถูกต้องขึ้นอยู่กับโค้งลักษณะข้อสบของแต่ละโมเดลที่ใช้แต่จะไม่ขึ้นกับการแจกแจงความสามารถของกลุ่มตัวอย่าง กล่าวคือ คุณสมบัติของโค้งลักษณะของข้อสอบจะไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ ฉะนั้นจึงทำให้ความน่าจะเป็นในการตอบข้อสอบถูกแต่ละข้อไม่แปรเปลี่ยนด้วย โดยผู้สอบที่มีความสามารถในระดับสูงจะมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องมากกว่าข้อสอบที่มีความสามารถในระดับต่ำ

9.1.3 พารามิเตอร์ของทฤษฎีการตอบสนองข้อสอบ

1) พารามิเตอร์ของข้อสอบ (Item Parameter) ได้แก่

ค่าความยาก (b) หมายถึง สัดส่วนของคนที่ทำข้อสอบข้อนั้นถูก หรือหมายถึงค่าที่แสดงถึงระดับความสามารถของผู้สอบ (θ) ที่จุดโค้งลักษณะข้อสอบมีความชันมากที่สุดมีค่าตั้งแต่ $-\infty$ ถึง ∞ แต่ในทางปฏิบัติจะอยู่ระหว่าง -3 ถึง +3 ค่า -3 บ่งบอกว่าข้อสอบนั้นง่ายมาก และค่า +3 แสดงว่าข้อสอบนั้นยากมาก

ค่าอำนาจจำแนก (a) หมายถึง ความสามารถของข้อสอบที่แยกผู้สอบออกเป็น 2 กลุ่ม คือ กลุ่มตอบถูกกับกลุ่มตอบผิด ในการวิจัยนี้หมายถึง ค่าที่เป็นสัดส่วนโดยตรงกับความชันของโค้งคุณลักษณะของข้อสอบ ณ จุดเปลี่ยนโค้งมีค่าตั้งแต่ $-\infty$ ถึง ∞ แต่ในทางปฏิบัติมีค่าตั้งแต่ 0 ถึง 2 เพราะค่า a ที่เป็นลบแสดงว่าข้อสอบไม่ดี ใช้ไม่ได้ต้องตัดทิ้ง ค่า 0 แสดงว่าข้อสอบไม่มีค่าอำนาจจำแนก ค่า +2 แสดงว่าข้อสอบมีค่าอำนาจจำแนกสูง ในการคัดเลือกข้อสอบ ข้อสอบที่คัดไว้จะมีค่า a ตั้งแต่ 0.3 ขึ้นไป

ค่าสัมประสิทธิ์การเดา (c) หมายถึง ความน่าจะเป็นของบุคคลหนึ่งที่ไม่มีความสามารถในการตอบข้อสอบนั้นได้ถูกต้อง เป็นค่าที่แสดงถึงโอกาสการตอบข้อสอบถูกโดยไม่มีความรู้ในเรื่องนั้น ๆ มีค่าจาก 0 ถึง 1 จะคัดเลือกเอาข้อสอบที่มีค่า c ต่ำ กว่า 0.3

2) พารามิเตอร์ของผู้สอบ ได้แก่

ระดับความสามารถของผู้สอบ (θ) หมายถึง ศักยภาพของผู้สอบที่ประมาณได้จากการตอบข้อสอบตามทฤษฎีการตอบสนองข้อสอบ มีค่าอยู่ระหว่าง -3 ถึง +3 ค่า –3 แสดงว่ามีความสามารถต่ำ และค่า +3 แสดงว่ามีความสามารถสูง

9.1.4 รูปแบบของโมเดลโลจิสติก

โมเดลโลจิสติกถูกพัฒนาขึ้นเพื่อให้สะดวกต่อการนำไปใช้จึงพัฒนาขึ้นเป็น 3 รูปแบบ ดังนี้ (ศิริชัย กาญจนวาสี, 2550, หน้า 49)

1) โมเดลโลจิสติก 1 พารามิเตอร์ (One – Parameter Logistic Model) โมเดลนี้เบอร์นบอม พัฒนาขึ้นในปี 1968 ซึ่งบังเอิญตรงกับรูปแบบของราสช์ เป็นโมเดลที่อธิบายข้อสอบด้วยค่าพารามิเตอร์เพียงค่าเดียว คือ ค่าความยากซึ่งโอกาสที่ผู้สอบจะทำข้อสอบได้ถูกหรือไม่นั้นขึ้นอยู่กับระดับความสามารถของตนเองกับระดับความยากของข้อสอบ ดังนั้นจึงถือว่าค่าการเดาเป็นศูนย์ ($c_i = 0$) และค่าอำนาจจำแนกของข้อสอบจะคงที่ทั้งฉบับ เขียนเป็นฟังก์ชันได้ดังนี้

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{\theta - b_i}} \quad , \quad i = 1, 2, 3, ..., n$$

เมื่อ $P_i(\theta)$ แทน โอกาสที่ผู้มีความสามารถ θ จะทำข้อสอบข้อที่ i ได้ถูกต้อง

θ แทน ระดับความสามารถที่แท้จริงของผู้สอบ

b_i แทน ค่าความยากของข้อสอบข้อที่ i

e แทน ค่าคงที่มีค่าเท่ากับ 2.7182818

2) โมเดลโลจิสติก 2 พารามิเตอร์ (Two – Parameter Logistic Model)

เบอร์นบอร์ม (Birnbaum) ได้พัฒนาโมเดลนี้ขึ้นมาและกำหนดให้ทุกข้อไม่มีการเดาเกิดขึ้น คือ ค่า c_i มีค่าเป็นศูนย์ทุกข้อ กล่าวคือ ผู้สอบที่มีความสามารถต่ำสุดไม่มีโอกาสที่จะทำข้อสอบถูกในข้อสอบที่มีค่าความยากสูง ซึ่งเบอร์นบอร์มได้เสนอรูปแบบของสมการดังนี้

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i^{\theta - b_i}}} \quad , \quad i = 1, 2, 3, ..., n$$

เมื่อ $P_i(\theta)$ แทน โอกาสที่ผู้มีความสามารถ θ จะทำข้อสอบข้อที่ i ได้ถูกต้อง

θ แทน ระดับความสามารถที่แท้จริงของผู้สอบ

a_i แทน ค่าอำนาจจำแนกข้อสอบข้อที่ i

b_i แทน ค่าความยากของข้อสอบข้อที่ i

e แทน ค่าคงที่มีค่าเท่ากับ 2.7182818

D แทน ค่าคงที่ซึ่งมีค่า 1.7

3) โมเดลโลจิสติก 3 พารามิเตอร์ (Three – parameter Logistic Model)

เป็นโมเดลที่พัฒนามาจาก Two-Parameter Logistic Model เพื่อให้เหมาะกับแบบทดสอบที่มีอิทธิพลจากการเดาเข้ามาแฝงอยู่ด้วย และเป็นโค้งลักษณะข้อสอบที่แสดงถึงลักษณะข้อสอบที่มีค่าพารามิเตอร์ของข้อสอบ 3 ตัว ซึ่งเบอร์นบอร์ม ได้เสนอรูปแบบของสมการดังนี้

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i^{\theta-b_i}}} \quad , \quad i = 1,2,3,...,n$$

เมื่อ $P_i(\theta)$ แทน โอกาสที่ผู้มีความสามารถ θ จะทำข้อสอบข้อที่ i ได้ถูกต้อง

θ แทน ระดับความสามารถที่แท้จริงของผู้สอบจะมีค่าอยู่ระหว่าง –3 ถึง +3 และค่า –3 แสดงว่ามีค่าความสามารถต่ำ +3 แสดงว่ามีค่าความสามารถสูง

a_i แทน ค่าอำนาจจำแนกข้อสอบข้อที่ i ซึ่งมีค่าสัดส่วนโดยตรงกับค่าความชันของโค้ง ณ จุดเปลี่ยนโค้งที่จุด $\theta = b_i$ โดยทั่วไปมักจะเลือกใช้ช่วงของค่าอำนาจจำแนกของข้อสอบอยู่ระหว่าง 0 ถึง 2

b_i แทน ค่าความยากของข้อสอบข้อที่ i เป็นระดับความสามารถของผู้สอบ ณ จุดเปลี่ยนโค้งและมีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติจะมีค่าอยู่ระหว่าง -2 ถึง +2 ค่า –2 แสดงว่าข้อสอบง่ายมาก และค่า +2 แสดงว่าข้อสอบยาก

c_i แทน ค่าการเดาของข้อสอบข้อที่ i เป็นความน่าจะเป็นหรือโอกาสของคนที่มีความสามารถต่ำจะตอบข้อสอบถูก มีค่าอยู่ระหว่าง 0 ถึง 1 โดยทั่วไปแล้วข้อสอบที่ดีจะต้องมีค่าการเดาต่ำกว่า 0.30

e แทน ค่าคงที่มีค่าเท่ากับ 2.7182818

D แทน ค่าคงที่ซึ่งมีค่า 1.7

9.1.5 ฟังก์ชันสารสนเทศของข้อสอบตามแนวทฤษฎีการตอบสนองข้อสอบ

ในการประมาณค่าความสามารถของผู้สอบด้วยวิธี Maximum Likelihood นั้นความแน่นอนของการประมาณค่าความสามารถแสดงได้ในเทอมของ Information Function โดยที่ในโมเดลคลาสสิคอลนั้นเราศึกษาเรื่องความเที่ยง (Reliability) ของคะแนนและความคลาดเคลื่อนมาตรฐานของการวัด (Standard Error of Measurement) ซึ่งค่าที่ได้จะแปรเปลี่ยนไปตามกลุ่มผู้สอบอันเป็นจุดอ่อนในการศึกษา แต่ในทฤษฎีการตอบสนองข้อสอบนั้นจะพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบ ซึ่งเป็นดัชนีบอกความแม่นยำในการประมาณค่าความสามารถที่แท้จริงได้จากผลรวมของฟังก์ชันสารสนเทศของข้อสอบโดยที่ค่าฟังก์ชันสารสนเทศ

ของข้อสอบจะแตกต่างกันไปตามค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าอำนาจจำแนก ค่าความยากและค่าการเดา เป็นต้น ค่าฟังก์ชันสารสนเทศของข้อสอบหาได้ดังนี้

$$I_i(\theta) = \frac{(P_i')^2}{P_i Q_i} \quad , \ i = 1, 2, 3, ..., n \qquad (1)$$

เมื่อ $I_i(\theta)$ แทน ค่า Item Information Functions

P_i' แทน ความชันของ ICC ที่ระดับความสามารถ θ

P_i แทน ความน่าจะเป็นที่ผู้สอบที่มีความสามารถ θ ตอบข้อสอบข้อที่ i ถูก

Q_i แทน $1 - P_i$

สำหรับในโมเดลโลจิสติกแบบ 3 พารามิเตอร์จะให้ค่าอนุพันธ์และค่าสารสนเทศของข้อสอบ ดังนี้

$$P_i = \frac{c_i + e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}} \qquad (2)$$

$$Q_i = \frac{1 - c_i}{1 + e^{1.7a(\theta - b_i)_i}} \qquad (3)$$

$$P_i' = \frac{1.7\, a_i(1 - c_i)}{e^{1.7a_i(\theta - b_i)} + 2 + e^{-1.7a_i(\theta - b_i)}} \qquad (4)$$

แทนค่า (2) , (3) และ (4) ลงใน (1) จะได้

$$I_i(\theta) = \frac{(1.7\, a_i)^2(1 - c_i)}{\left[c_i + e^{1.7a_i(\theta - b_i)}\right]\left[1 + e^{-1.7a_i(\theta - b_i)}\right]^2} \qquad (5)$$

จากสมการ (5) หมายถึง ค่าฟังก์ชันสารสนเทศของข้อสอบที่เขียนอยู่ในรูปของค่าพารามิเตอร์ข้อสอบ สำหรับค่าพารามิเตอร์ของแบบทดสอบหาได้จากผลรวมของค่าฟังก์ชันสารสนเทศข้อสอบทั้งหมดในแบบทดสอบ เขียนสมการได้ดังนี้

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

เมื่อ $I(\theta)$ แทน ค่าฟังก์ชันสารสนเทศของแบบทดสอบ Test Information Function

จากสมการ(1) จะเห็นว่าข้อสอบแต่ละข้อมี Item Information Curve ซึ่งจะขึ้นอยู่กับความชันของ ICC และความแปรปรวนของการตอบข้อสอบถูกในแต่ละข้อในแต่ละระดับความสามารถ และยิ่งความชันของ ICC มีค่ามาก ๆ ประกอบกับความแปรปรวนมีค่าน้อย ๆ Item Information Curve ที่ระดับความสามารถนั้นจะยิ่งสูงขึ้นสำหรับ Item Information Curve ที่มีค่าสูงสุด ณ ระดับความสามารถใดก็จะจำแนกระดับความสามารถผู้สอบได้ดี ณ ระดับความสามารถนั้น ดังนั้นประโยชน์ที่ได้จากประเด็นข้างต้นก็คือ ถ้ามีกลุ่มของข้อสอบอยู่ชุดหนึ่งที่สามารถทราบ Information Curve ของแต่ละข้อ เราก็จะสามารถสร้างแบบทดสอบฉบับหนึ่งให้มี Test Information Curve ณ ระดับหนึ่งของความสามารถตามที่เราต้องการได้ และนั้นหมายถึงว่าเราสามารถสร้างฉบับแบบทดสอบให้เป็นไปตามจุดมุ่งหมายของการสอบได้ เช่น ถ้าต้องการได้แบบทดสอบคัดเลือก ก็ต้องเลือกใช้ข้อสอบที่มีความสูงสุดของโค้งที่ระดับความสามารถสูง ๆ ซึ่งก็คือ เลือกข้อสอบที่จะให้ได้ Test Information Curve สูงที่ระดับความสามารถสูง ๆ เป็นต้น

9.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย IRT โมเดล

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย IRT โมเดล ประกอบด้วย 2 ขั้นตอนที่สัมพันธ์กันได้แก่ ขั้นตอนแรก ทำการวัดขนาดของการทำหน้าที่ต่างกันของข้อสอบ (Measurement of DIF) และขั้นตอนที่สอง ทำการทดสอบทางสถิติของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Statistical test of DIF) คำถามหลักของการตรวจสอบ DIF ด้วย IRT โมเดล คือ “โค้งลักษณะข้อสอบของประชากรต่างกลุ่มมีความแตกต่างกันหรือไม่” จึงต้องวัดขนาดของความแตกต่าง และทดสอบทางสถิติว่ามีความแตกต่างอย่างมีนัยสำคัญระหว่างกลุ่มหรือไม่ (ศิริชัย กาญจนวาสี, 2550, หน้า 130)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย IRT โมเดล ที่นิยมกันมีดังนี้

9.2.1 วิธีวัดความแตกต่างของพื้นที่ (Area measures: AREA)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวัดพื้นที่จะเปรียบเทียบฟังก์ชันการตอบข้อสอบระหว่างกลุ่ม โดยการคำนวณค่าประมาณพื้นที่ระหว่างโค้งลักษณะข้อสอบ จากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน ซึ่งมีสูตรทั่วไปที่ใช้ในการคำนวณพื้นที่ดังนี้

$$A = f_s\left[P_R(\theta) - P_F(\theta)\right]$$

เมื่อ A = ดัชนีพื้นที่

f_s = ฟังก์ชัน f ในช่วง s ซึ่ง $s = (\theta_L, \theta_H)$ โดยที่ L และ H เป็นค่าสูงสุดและต่ำสุดตามลำดับ

$P_R(\theta)$ = โอกาสของการตอบข้อสอบถูกที่ระดับความสามารถ θ จากผู้สอบกลุ่มอ้างอิง

$P_F(\theta)$ = โอกาสของการตอบข้อสอบถูกที่ระดับความสามารถ θ จากผู้สอบกลุ่มเปรียบเทียบ

ในการคำนวณพื้นที่สามารถเลือกคำนวณได้ทั้งชนิดเครื่องหมาย (Signed area) หรือชนิดไม่คิดเครื่องหมาย (Unsigned area) และคำนวณพื้นที่แบบต่อเนื่อง (Continuous integration) หรือการประมาณค่าแบบไม่ต่อเนื่อง (Discrete approximation)

การคำนวณในระยะแรกเป็นแบบไม่ต่อเนื่องและไม่คิดเครื่องหมาย รูดเดอร์ (1977 อ้างอิงใน ศิริชัย กาญจนวาสี, 2550, หน้า 131) ได้เสนอสูตรไว้ดังนี้

$$R = \sum_{j=-3}^{3} \left|D_j\right| \Delta$$

เมื่อ $D_j = P_R(\theta) - P_F(\theta)$

$\Delta = .005$

ดัชนี R นี้คำนวณแบบไม่คิดเครื่องหมาย เมื่อนำค่าสัมบูรณ์ออกก็จะเปลี่ยนเป็นดัชนีชนิดคิดเครื่องหมาย ซึ่งอาจมีเครื่องหมายเป็นบวกหรือลบก็ได้ ต่อมา ลินและคณะ (1981 อ้างอิงใน ศิริชัย กาญจนวาสี, 2550, หน้า 131) ได้เสนอดัชนีชนิดไม่คิดเครื่องหมายเรียกว่า รากกำลังสองของความแตกต่างเฉลี่ย (Root Mean Squared Different: RMSD) ระหว่างฟังก์ชันการตอบสนองข้อสอบ โดยแบ่งช่วงระดับความสามารถระหว่าง -3 ถึง +3 ออกเป็น 600 ช่วง ดังนี้

$$RMSD = \sqrt{\frac{1}{600}\sum_{j=1}^{n}[P_R(\theta) - P_F(\theta)]^2}$$

รูดเดอร์ เกทสัน และไนท์ (Ruder, Getson & Knigkt, 1980, หน้า 213) ได้เสนอสูตรการคำนวณพื้นที่แบบต่อเนื่องชนิดคิดเครื่องหมาย (SIGNED – AREA) และไม่คิดเครื่องหมาย (UNSIGNED – AREA) ดังนี้

$$SIGNED\ AREA = \int[P_R(\theta) - P_F(\theta)]d\theta$$

$$UNSIGNED\ AREA = \sqrt{\int[P_R(\theta) - P_F(\theta)]^2 d\theta}$$

ดัชนี SIGNED – AREA ถ้ามีเครื่องหมายเป็น + แสดงว่ากลุ่มอ้างอิงทำข้อสอบได้ดีกว่า แต่ถ้าเครื่องหมายเป็น - แสดงว่ากลุ่มเปรียบเทียบทำได้ดีกว่า สำหรับดัชนี UNSIGNED – AREA ที่มีค่ามากกว่า SIGNED – AREA เป็นสัญญาณแสดงว่า โค้งลักษณะข้อสอบของกลุ่มทั้งสองตัดกัน

วิธีการวัดความแตกต่างของพื้นที่เป็นวิธีที่ทำความเข้าใจได้ง่ายสามารถวาดภาพแสดงได้ชัดเจน แต่มีจุดอ่อนที่มิได้สนใจระบุช่วง θ ที่มีความแตกต่างกันมากและมีปัญหาด้านความน่าเชื่อ ถือของค่าที่คำนวณได้ ถ้ากลุ่มทั้งสองมีค่าพารามิเตอร์ c แตกต่างกัน วิธีการวัดความแตกต่างของพื้นที่ซึ่งนิยมใช้กันมี 2 วิธี คือ วิธีวัดพื้นที่ของราจู (Ralu, 1990, หน้า 197 – 207) และวิธีการวัดพื้นที่ของคิมและโคเฮน (Kim & Cohen, 1991 , หน้า 270) ซึ่งมีรายละเอียดดังนี้

1) วิธีการวัดพื้นที่ของราจู

ราจู (Ralu, 1990, หน้า 197– 207) ได้เสนอสูตรการคำนวณพื้นที่ในช่วงเปิดของ θ ชนิดคิดเครื่องหมาย (Open interval signed area or Exact Signed: ESA)

และพื้นที่ชนิดไม่คิดเครื่องหมาย (Open interval unsigned area or Exact Unsigned: EUA) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ทฤษฎี IRT โมเดลโลจิสติก แบบ 1, 2 และ 3 พารามิเตอร์ สูตรทั่วไปที่ใช้ในการคำนวณพื้นที่ทั้งสองทั้งชนิดมีลักษณะดังนี้

$$ESA = \int_{-\infty}^{+\infty} [P_R(\theta) - P_F(\theta)] d\theta$$

$$EUA = \int_{-\infty}^{+\infty} |P_R(\theta) - P_F(\theta)| d\theta$$

ในการทดสอบว่าข้อสอบที่นำมาตรวจสอบทำหน้าที่ต่างกันหรือไม่ ราจูได้เสนอให้นำดัชนีการวัดพื้นที่มาทดสอบนัยสำคัญโดยใช้สถิติ z ภายใต้สมมติฐานการแจกแจงแบบปกติ ในการทดสอบดังกล่าวแบ่งออกเป็น 2 ลักษณะ ดังนี้

การทดสอบนัยสำคัญของ ESA

นำดัชนี ESA ของข้อสอบข้อที่ i ไปทดสอบความแตกต่างกับ 0 โดยใช้สถิติ z ดังนี้

$$Z_i(ESA) = \frac{\hat{b}_{iF} - \hat{b}_{iR}}{\sqrt{Var(\hat{b}_{iF}) + Var(\hat{b}_{iR})}}$$

การทดสอบนัยสำคัญของ EUA

ในการทดสอบนัยสำคัญของดัชนี EUA สามารถแบ่งออกเป็น 2 กรณี คือ กรณีที่ 1 เมื่อ $a_{iR} = a_{iF}$ จะทดสอบเหมือนกับดัชนี ESA สำหรับกรณีที่ 2 $a_{iR} \neq a_{iF}$ จะนำดัชนี EUA ของข้อสอบข้อที่ i ไปทดสอบความแตกต่างกับ 0 โดยใช้สถิติ z ดังนี้

$$Z_i(H) = \frac{H_i}{\sqrt{Var(H_i)}}$$

2) วิธีการวัดพื้นที่ของ Kim และ Cohen

คิมและโคเฮน (Kim & Cohen, 1991, หน้า 270) ได้พัฒนาสูตรการคำนวณพื้นที่ในช่วงปิด ชนิดคิดเครื่องหมาย (Close – interval Signed Area: CSA) และพื้นที่

ชนิดไม่คิดเครื่องหมาย (Close – interval Unsigned Area: CUA) โดยคำนวณพื้นที่ระหว่างฟังก์ชันการตอบสนองข้อสอบ (IRTs) บนช่วงความสามารถ $[\theta_1 , \theta_2]$ซึ่งมีสูตรในรูปทั่วไปดังนี้

$$CSA = \int_{\theta_1}^{\theta_2} [P_R(\theta) - P_F(\theta)]d\theta$$
$$= S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)$$

$$CUA = \int_{\theta_1}^{\theta_2} |P_R(\theta) - P_F(\theta)|d\theta$$
$$= |S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)|$$

สำหรับการทดสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวัดพื้นที่ของคิมและโคเฮน (1991, หน้า 270 – 278) ไม่ได้เสนอวิธีการทดสอบนัยสำคัญ แต่จะนำขนาดของพื้นที่ไปเปรียบเทียบกับเกณฑ์ที่กำหนดขึ้น

9.2.2 วิธีวัดความแตกต่างของค่าพารามิเตอร์ความยาก (b parameter difference)

ดัชนีที่ง่ายที่สุดที่สามารถสะท้อนความแตกต่างระหว่างฟังก์ชันการตอบสนองข้อสอบ (IRFs) คือ การพิจารณาความแตกต่างระหว่างค่าพารามิเตอร์ b ระหว่างกลุ่มอ้างอิง และกลุ่มเปรียบเทียบ ด้วยดัชนีความแตกต่างระหว่างค่าความยากที่คิดเครื่องหมาย เมื่อมีการควบคุมค่า θ (Signed B Difference controlling for θ: SBD - θ)

$$SDB-\theta = b_F - b_R$$
$$= \Delta b$$

ค่า Δb ซึ่งเป็นความแตกต่างของค่าพารามิเตอร์ความยากระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ถ้า Δb มีค่าเป็นบวก แสดงว่า กลุ่มอ้างอิงทำข้อสอบได้ดีกว่า แต่ถ้า Δb มีค่าเป็นลบ แสดงว่า กลุ่มเปรียบเทียบทำข้อสอบได้ดีกว่า

สำหรับการทดสอบนัยสำคัญของความแตกต่างระหว่างค่าพารามิเตอร์ความยากระหว่าง 2 กลุ่ม สำหรับการทดสอบ $H_0 : \Delta b = 0$ มีสถิติสำหรับการทดสอบดังนี้ (ศิริชัย กาญจนวาสี, 2550, หน้า 133 – 134)

$$d = \frac{\Delta b}{S_{\Delta b}}$$

เมื่อ d = สถิติทดสอบซึ่งมีการแจกแจงแบบปกติ (z)

Δb = $b_F - b_R$

$S_{\Delta b}$ = $\sqrt{S_F^2 + S_R^2}$

S_F = ความคลาดเคลื่อนมาตรฐานของ b_F

S_R = ความคลาดเคลื่อนมาตรฐานของ b_R

วิธีวัดความแตกต่างของค่า b อาจจะไม่เหมาะสมสำหรับโมเดล 1 - พารามิเตอร์ เพราะมีข้อจำกัดที่ c = 0 และ a เท่ากัน วิธีนี้จึงเหมาะสมสำหรับใช้โมเดล 2 - พารามิเตอร์ และ 3 - พารามิเตอร์

9.2.3. วิธีการทดสอบไค – สแควร์ ของ ลอร์ด (Lord's χ^2 - test)

ลอร์ด (Lord, 1980 , หน้า 72) ได้เสนอวิธีการทดสอบไค – สแควร์ เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย IRT โมเดล 3 – พารามิเตอร์ โดยขยายแนวคิดของการทดสอบความแตกต่างของ b ให้ครอบคลุมถึงการทดสอบความแตกต่างระหว่าง a พร้อมกันไป หลักการทดสอบด้วยวิธีนี้เป็นการทดสอบความแตกต่างระหว่างค่าพารามิเตอร์ของข้อสอบตาม ฟังก์ชันการตอบสนองข้อสอบ (IRTs) จากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

ในการวิเคราะห์เริ่มต้นด้วยการประมาณค่าพารามิเตอร์ a_i , b_i และ c_i จากผู้สอบทั้งสองกลุ่มรวมกัน จากนั้นกำหนดให้ค่าพารามิเตอร์ c_i เป็นค่าเดิมที่คงที่ สำหรับประมาณค่าพารามิเตอร์ a_i และ b_i สำหรับแต่ละกลุ่ม จากนั้นปรับค่า a_i และ b_i ของแต่ละกลุ่มให้อยู่บนสเกลเดียวกันเพื่อนำพารามิเตอร์ a_i และ b_i ของแต่ละข้อมาเปรียบเทียบระหว่างกลุ่มโดยใช้สถิติทดสอบไค – สแควร์ ที่มีระดับชั้นของความเป็นอิสระเท่ากับ 2 ดังนี้

H_0 : $b_{iF} = b_{iR}$ และ $a_{iF} = a_{iR}$

H_1 : $b_{iF} \neq b_{iR}$ และ $a_{iF} \neq a_{iR}$

โดยสถิติทดสอบไค – สแควร์ คำนวณได้ด้วยสูตร ดังนี้

$$\chi^2 = V_i' S_i^{-1} V_i$$

เมื่อ V_i' = เวกเตอร์ความแตกต่างของค่าประมาณพารามิเตอร์ของข้อสอบข้อที่ i ระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ $(a_{iF} - a_{iR}, b_{iF} - b_{iR})$

S_i = เมตริกซ์ความแปรปรวน – ความแปรปรวนร่วมของค่าความแตกต่างของค่าประมาณพามิเตอร์ $(a_{iF} - a_{iR})$ และ $(b_{iF} - b_{iR})$ ซึ่งมีขนาด 2 X 2

S_i^{-1} = อินเวอรส์ของเมตริกซ์ S_i

ผลการทดสอบพิจารณาได้จาก การเปรียบเทียบค่า χ^2 ที่คำนวณได้กับค่าวิกฤตของ χ^2 ที่ df = 2 ถ้า χ^2 ที่คำนวณได้ทีค่ามากกว่าค่าวิกฤตแสดงว่าการทดสอบมีนัยสำคัญ นั่นคือข้อสอบข้อนั้นทำหน้าที่ต่างกัน

10. งานวิจัยที่เกี่ยวข้อง

ชัชชัย เผ่าพงษ์ (2527) ได้ศึกษาการวิเคราะห์ความลำเอียงของข้อสอบจากแบบวัดความถนัดทางการเรียนด้านคณิตศาสตร์และภาษา ระดับชั้นมัธยมศึกษาตอนต้น การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อ วิเคราะห์ความลำเอียงของแบบทดสอบมาตรฐานที่มีผลต่อกลุ่มนักเรียนชายและหญิง ซึ่งกำลังศึกษาอยู่ในระดับชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2524 การวิเคราะห์ความลำเอียงของข้อสอบกระทำโดยเปรียบเทียบโค้งลักษณะข้อสอบด้วยวิธีแบบจำลองลอจิสติกแบบคำนวณค่าพารามิเตอร์ของข้อสอบ 3 ตัว กลุ่มตัวอย่างที่ใช้แบ่งออกเป็น 4 กลุ่ม สองกลุ่มแรกเป็นนักเรียนชาย 1,610 คน และนักเรียนหญิง 1,337 คน ซึ่งทำการวิเคราะห์ด้วยแบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์ และสองกลุ่มหลังเป็นนักเรียนชาย 1,316 คน และนักเรียนหญิง 985 คน ซึ่งทำการวิเคราะห์ด้วยแบบทดสอบวัดความถนัดทางการเรียนด้านภาษา ผลการวิจัยได้ดังนี้ จากแบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์ จำนวน 30 ข้อ มีข้อสอบที่ลำเอียงต่อนักเรียนชายโดยเฉพาะ 7 ข้อ และลำเอียงต่อนักเรียนหญิงโดยเฉพาะ 2 ข้อ ข้อสอบที่มีความลำเอียง 2 ช่วงมาตราความสามารถมี 14 ข้อ และข้อสอบที่มีความลำเอียง 3 ช่วงมาตราความสามารถมี 7 ข้อ จากจำนวนข้อสอบทั้งหมดที่มีความลำเอียงจัดเป็นข้อสอบที่มีความลำเอียงต่ำ 25 ข้อ ปานกลาง 3 ข้อ และมีความลำเอียงสูง 2 ข้อ และจากแบบทดสอบวัดความถนัดทางการเรียนด้านภาษา จำนวน 40 ข้อ มีข้อสอบที่ลำเอียงต่อนักเรียนชายโดยเฉพาะ 3 ข้อ และลำเอียงต่อนักเรียนหญิงโดยเฉพาะ 8 ข้อ ข้อสอบที่มีความลำเอียง 2 ช่วงมาตราความสามารถมี 25 ข้อ และข้อสอบที่มีความลำเอียง 3 ช่วงมาตราความสามารถมี 4 ข้อ

จากจำนวนข้อสอบทั้งหมดที่มีความลำเอียงจัดเป็นข้อสอบที่มีความลำเอียงต่ำ 31 ข้อ ปานกลาง 7 ข้อ และมีความลำเอียงสูง 2 ข้อ ผลของการทดสอบระหว่างกลุ่มนักเรียนชายและนักเรียนหญิงด้วยแบบทดสอบมาตรฐาน 2 ฉบับนี้ มีขีดจำกัดเรื่องการเปรียบเทียบผลการสอบที่ได้จากนักเรียน 2 กลุ่มนี้ ไม่สามารถจะนำมาตัดสินรวมกันได้ ข้อสอบหลายข้อมีความลำเอียงทางการทดสอบสูงมาก ทำให้นักเรียนชายและนักเรียนหญิงที่มีระดับความสามารถเท่ากันตอบข้อสอบเดียวกันถูกด้วยความน่าจะเป็นที่แตกต่างกัน เพื่อให้ได้แบบทดสอบที่มีคุณภาพสูงขึ้น จึงควรปรับปรุงข้อสอบที่มีดัชนีความลำเอียงทางการทดสอบมากกว่า 0.40 ขึ้นไป ให้มีความยุติธรรมทางการทดสอบต่อกลุ่มนักเรียนชายและหญิงเท่ากัน

ทัศนีย์ พรีมนตรี (2530) ได้วิเคราะห์ความลำเอียงของแบบทดสอบวิชาคณิตสาสตร์ตามโครงการตรวจสอบคุณภาพทางการศึกษาของสำนักทดสอบทางการศึกษากรมวิชาการ กระทรวงศึกษาธิการ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526 จำนวน 7,036 คน โดยเปรียบเทียบจำนวนข้อสอบมีความลำเอียงระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในภาคภูมิศาสตร์ 5 ภาค ได้แก่ ภาคเหนือ ภาคกลาง ภาคใต้ ภาคตะวันออกเฉียงเหนือ และภาคตะวันออก ด้วยวิธีวิเคราะห์ 3 วิธี คือ วิธีกำหนดจุดค่าเดลตา วิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ และวิธีทดสอบความแตกต่างระหว่างกลุ่มด้วยสถิติไค – สแควร์ ในโมเดลล็อกลีเนีย 2 โมเดล คือ โมเดลที่ไม่มีพารามิเตอร์ผลร่วมระหว่างระดับคะแนนกับกลุ่มและโมเดลที่ไม่มีพารามิเตอร์ของผลหลักที่เกิดจากกลุ่ม ผลการวิเคราะห์พบว่า ข้อสอบที่ให้ผลลำเอียงจากวิธีโค้งลักษณะข้อสอบที่ใช้ 3 พารามิเตอร์มีจำนวนข้อมากที่สุด และในแต่ละการวิเคราะห์พบว่า มีข้อสอบที่ให้ผลลำเอียงตรงกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในทุกภาค

สุรศักดิ์ อมรรัตนศักดิ์ (2531) ได้ศึกษาดัชนีความลำเอียงของข้อสอบ โดยพิจารณาถึงสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีวิเคราะห์ความลำเอียงของข้อสอบ 4 วิธี คือ (1) วิธีวิเคราะห์ความแปรปรวน (2) วิธีแปลงค่าความยากของข้อสอบ (3) วิธีโค้งลักษณะข้อสอบ 1 พารามิเตอร์ และ (4) วิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ จากแบบทดสอบที่ใช้สอบแข่งขันเพื่อบรรจุเข้ารับเป็นข้าราชการครูในปี พ.ศ.2529 ซึ่งมี 4 ฉบับ คือ ฉบับที่ 1 แบบทดสอบวิชาเอก ฉบับที่ 2 แบบทดสอบวิชาการศึกษาและกฎหมาย ฉบับที่ 3 แบบทดสอบวิชาภาษาไทย ฉบับที่ 4 และเปรียบเทียบความแตกต่างของผลการคัดเลือกก่อนและหลังการศึกษาความลำเอียงของข้อสอบในด้านจำนวนผู้ได้รับการคัดเลือกสัดส่วนชาย : หญิง ที่ได้รับการคัดเลือก และความเที่ยงของแบบสอบ ผลการวิจัยสรุปได้ว่า

1. วิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ ค้นพบข้อสอบลำเอียงมากที่สุด รองลงมาได้แก่ วิธีวิเคราะห์ความแปรปรวน วิธีที่ตรวจค้นข้อสอบได้น้อยที่สุด ได้แก่ วิธีแปลงค่าความยากของข้อสอบ

2. ทั้ง 4 วิธีมีความสัมพันธ์กันทางบวกอย่างมีนัยสำคัญที่ระดับ .001 โดยมีค่าระหว่าง .7535 - .9921

3. การใช้คะแนนดิบและคะแนนรวมอื่น ๆ อีก 5 วิธี มีจำนวนผู้ได้รับคัดเลือกแตกต่างกันประมาณร้อยละ 4 ถึง 24 ส่วนการใช้คะแนนมาตรฐานทีปกติรวมกับคะแนนแปลงแบบอื่น ๆ อีก 4 วิธี มีจำนวนผู้ได้รับคัดเลือกแตกต่างกันร้อยละ 4 ถึง 23

4. เมื่อตัดข้อที่มีความลำเอียงออก พบว่า สัดส่วนหญิงและชายที่ได้รับการคัดเลือกมีความใกล้เคียงกัน และค่าความเที่ยงของแบบสอบลดลงเล็กน้อย

เรวดี อินทสะระ (2539) ได้ศึกษาความเที่ยงเชิงพยากรณ์ของแบบสอบคัดเลือกเพื่อเข้าศึกษาในระดับชั้นปีที่ 1 ประเภทรับตรง ปีการศึกษา 2538 ของมหาวิทยาลัยสงขลานครินทร์ ด้วยวิธีการวิเคราะห์ความลำเอียง 3 วิธี คือ วิธีการใช้ทฤษฎีการตอบข้อสอบ วิธีแมนเทล – แฮนส์เซล และ วิธี SIBTEST ระหว่างเพศและวิธีการให้คะแนนที่ต่างกัน

ผลการวิจัยพบว่า การวิเคราะห์ข้อสอบที่มีความลำเอียงจำนวนมากที่สุด คือ วิธีทฤษฎีการตอบข้อสอบที่มี 3 พารามิเตอร์

ทิพย์รัตน์ ผลบุญ (2540) ได้ศึกษาดัชนีความลำเอียงของข้อสอบวัดผลสัมฤทธิ์ทางการเรียนในวิชาคณิตศาสตร์ ซึ่งเกิดจากตัวแปรเพศและประเภทของโรงเรียน ด้วยวิธีวิเคราะห์ 3 วิธี คือ วิธีไค – สแควร์ วิธีทฤษฎีการตอบสนองข้อสอบที่มี 3 พารามิเตอร์ และวิธีแมนเทล – แฮนส์เซล เปรียบเทียบความลำเอียงของข้อสอบแต่ละข้อและจำนวนข้อที่มีต่อตัวแปรเพศ และประเภทของโรงเรียน ด้วยวิธีวิเคราะห์ความลำเอียงทั้ง 3 วิธี และศึกษาสัมประสิทธิ์สหสัมพันธ์ระหว่างดัชนีความลำเอียงทั้ง 3 วิธี เครื่องมือที่ใช้เป็นแบบสอบวัดผลสัมฤทธิ์ทางการเรียนในวิชาคณิตศาสตร์ ค 102 จำนวน 30 ข้อ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ของกลุ่มโรงเรียนมัธยมศึกษาตอนต้น จังหวัดกำแพงเพชร จำนวน 1,600 คน

ผลการวิจัยพบว่า วิธีวิเคราะห์ความลำเอียงโดยวิธีทฤษฎีการตอบสนองข้อสอบที่มี 3 พารามิเตอร์ ตรวจพบจำนวนข้อสอบที่ลำเอียงมากที่สุด โดยพบข้อสอบที่ลำเอียงตามตัวแปรเพศจำนวน 12 ข้อ ตามตัวแปรโรงเรียนจำนวน 12 ข้อ รองลงมาคือ วิธีแมนเทล – แฮนส์เซล โดยพบข้อสอบที่ลำเอียงตามตัวแปรเพศจำนวน 10 ข้อ ตามตัวแปรโรงเรียนจำนวน 11 ข้อ ส่วนวิธีไค – สแควร์ เป็นวิธีที่พบข้อสอบที่ลำเอียงน้อยที่สุด โดยพบข้อสอบที่ลำเอียงตามตัวแปรเพศ

จำนวน 1 ข้อ ตามตัวแปรโรงเรียนจำนวน 2 ข้อ ตามลำดับ ในการเปรียบเทียบความลำเอียง พบว่า ลำเอียงเข้าข้างเพศชายมากกว่าเพศหญิง และลำเอียงเข้าข้างโรงเรียนสามัญศึกษามากกว่าโรงเรียนขยายโอกาสทางการศึกษา ส่วนความสัมพันธ์ของความลำเอียงของวิธีวิเคราะห์แต่ละข้อ เมื่อวิเคราะห์ตามตัวแปรเพศมีความสัมพันธ์เป็นบวกและเมื่อวิเคราะห์ตามตัวแปรโรงเรียน พบว่าวิธีไค – สแควร์ มีค่า -0.3910 วิธีทฤษฎีการตอบสนองข้อสอบที่มี 3 พารามิเตอร์ มีค่า 0.2065 และวิธีแมนเทล – แฮนส์เซลมีค่า 0.2532

เพ็ญนภา สุขสม (2540) ได้ทำการศึกษาเพื่อเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบ ด้วยวิธีวิเคราะห์ 3 วิธี คือ วิธีแปลงความยากของข้อสอบ วิธีทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ และวิธีแมนเทล– แฮนส์เซล ตามตัวแปรเพศและที่ตั้งของโรงเรียน และเพื่อศึกษาความสัมพันธ์ระหว่างดัชนีความลำเอียงทั้ง 3 วิธี โดยใช้แบบทดสอบประเมินคุณภาพและวัดผลปลายภาคเรียนวิชาภาษาไทย จำนวน 50 ข้อ เป็นเครื่องมือในการเก็บรวบรวมข้อมูล กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2538 ของโรงเรียนในสังกัดสำนักงานเขตการประถมศึกษาจังหวัดนครศรีธรรมราช จำนวน 2,400 คน ผลการวิจัยพบว่า

1. วิธีวิเคราะห์ความลำเอียงแต่ละวิธีตรวจพบจำนวนข้อสอบที่ลำเอียงแตกต่างกัน วิธีแมนเทล– แฮนส์เซลเป็นวิธีที่ตรวจพบข้อสอบที่ลำเอียงมากที่สุด รองลงมาคือ วิธีทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์และวิธีแปลงความยากของข้อสอบตรวจพบข้อสอบที่ลำเอียงน้อยที่สุด

2. ดัชนีความลำเอียงทั้ง 3 วิธี มีความสัมพันธ์กันทางบวกอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.001 ซึ่งหมายความว่า วิธีวิเคราะห์ความลำเอียงทั้ง 3 วิธี ให้ผลสอดคล้องกันเมื่อพิจารณาตามตัวแปรพบว่า ความสัมพันธ์มีค่าระหว่าง 0.4011 ถึง 0.6662 โดยที่วิธีทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์มีความสัมพันธ์กับวิธีแมนเทล– แฮนส์เซลสูงกว่าวิธีแปลงความยากของข้อสอบ และเมื่อพิจารณาตามตัวแปรที่ตั้งของโรงเรียน พบว่า ความสัมพันธ์มีค่าระหว่าง 0.5675 ถึง 0.7847 โดยวิธีทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ มีความสัมพันธ์กับวิธีแมนเทล– แฮนส์เซลสูงกว่าวิธีแปลงความยากของข้อสอบเช่นเดียวกัน

วรรณา รอดตัว (2544) เปรียบเทียบผลของวิธีวิเคราะห์ความลำเอียงของข้อสอบที่แตกต่างกัน 3 วิธี โดยมีวัตถุประสงค์เพื่อค้นหาความลำเอียงของแบบทดสอบวิชาคณิตศาสตร์ 1 ซึ่งเป็นแบบทดสอบที่ใช้ในการสอบคัดเลือกเข้าศึกษาต่อระดับปริญญาตรี ในสถาบันอุดมศึกษาของรัฐและเอกชน ที่สังกัดทบวงมหาวิทยาลัย ประจำปีการศึกษา 2542 โดยวิธีวิเคราะห์

ความลำเอียงของข้อสอบด้วยวิธีค่าอำนาจจำแนกของข้อสอบ วิธีแมนเทล– แฮนส์เซลและวิธีซิปเทสท์ เปรียบเทียบผลการวิเคราะห์ความลำเอียงและศึกษาความสัมพันธ์ของวิธีวิเคราะห์ความลำเอียงทั้ง 3 วิธี กลุ่มตัวอย่างที่ใช้เป็นผู้เข้าสอบคัดเลือกเพื่อเข้าศึกษาต่อระดับปริญญาตรี ในสถาบัน อุดมศึกษาของรัฐและเอกชน ประจำปีการศึกษา 2542 จำนวน 2,540 คน โดยใช้วิธีสุ่มอย่างง่ายมาจากผู้เข้าสอบคัดเลือกทั้งสิ้นจำนวน 99,562 คน ผู้วิจัยนำผลการสอบที่ได้มาค้นหาความลำเอียงด้วยวิธีค่าอำนาจจำแนกของข้อสอบ วิธีแมนเทล– แฮนส์เซลและวิธีซิปเทสท์ ตามตัวแปรเพศและเขตที่ตั้งของสถานศึกษา หาค่าสัมประสิทธิ์สหสัมพันธ์ของดัชนีความลำเอียงที่ได้จากการวิเคราะห์โดยใช้สูตรเพียร์สัน โพรดักโมเมนต์ และเปรียบเทียบจำนวนข้อสอบที่ลำเอียงที่ได้จากการวิเคราะห์ด้วย 3 วิธีนี้ ผลการวิจัยพบว่า

1. ผลการวิเคราะห์ความลำเอียงของข้อสอบรายข้อด้วยวิธีวิเคราะห์ 3 วิธี ระหว่างตัวแปรเพศและเขตที่ตั้งของสถานศึกษา จากข้อสอบทั้งหมด 28 ข้อ พบว่า เขตที่ตั้งของสถานศึกษาพบจำนวนข้อสอบที่ลำเอียงสูงกว่าเพศ ดังนี้ 1.) วิธีค่าอำนาจจำแนกของข้อสอบพบข้อสอบที่ลำเอียงตามเพศ จำนวน 3 ข้อ ลำอียงตามเขตที่ตั้งของสถานศึกษา จำนวน12 ข้อ 2.) วิธีแมนเทล –แฮนส์เซล พบข้อสอบที่ลำเอียงตามเพศ จำนวน3 ข้อ ลำเอียงตามเขตที่ตั้งของสถานศึกษาจำนวน 8 ข้อ 3.) วิธีซิปเทสท์ พบข้อสอบที่ลำเอียงตามเพศจำนวน4 ข้อ ลำเอียงตามเขตที่ตั้งของสถานศึกษา จำนวน 6 ข้อ

2. ค่าสหสัมพันธ์ของดัชนีความลำเอียงของข้อสอบ จากการวิเคราะห์ด้วยวิธีค่าอำนาจจำแนกของข้อสอบ วิธีแมนเทล - แฮนส์เซลและวิธีซิปเทสท์ ระหว่างตัวแปรเพศและเขตที่ตั้งของสถานศึกษา พบว่า วิธีแมนเทล – แฮนส์เซลและวิธีซิปเทสท์ มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในตัวแปรเพศและมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ.01 ในตัวแปรเขตที่ตั้งของสถานศึกษา

3. เปรียบเทียบจำนวนข้อสอบที่ลำเอียงจากการใช้วิธีวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีค่าอำนาจจำแนกของข้อสอบ วิธีแมนเทล- แฮนส์เซลและวิธีซิปเทสท์ ตามตัวแปรเพศและเขตที่ตั้งของสถานศึกษา พบว่า วิธีการวิเคราะห์ความลำเอียงของข้อสอบทั้ง3 วิธี พบจำนวนข้อสอบที่ลำเอียงไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

สุกัญญา ทองนาค (2549) ได้ศึกษาการวิเคราะห์ความลำเอียงของข้อสอบเข้าศึกษาต่อประเภทโควตา ของมหาวิทยาลัยเชียงใหม่ การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อ วิเคราะห์ข้อสอบเข้าศึกษาต่อประเภทโควตา ของมหาวิทยาลัยเชียงใหม่ วิชาภาษาไทย วิชาสังคมและวิชาสามัญ 1 จำแนกตามกลุ่มเพศ ที่ตั้ง และขนาดของโรงเรียน โดยวิเคราะห์ 3 วิธี คือ วิธีการวิเคราะห์

ความแปรปรวน วิธีแมนเทล – แฮนส์เซล และวิธีไค – สแควร์ เพื่อเปรียบเทียบจำนวนข้อสอบที่มีความลำเอียง และเพื่อวิเคราะห์ความของการวิเคราะห์ความลำเอียงของข้อสอบระหว่างวิธีวิเคราะห์ทั้ง 3 วิธี สรุปผลการวิจัยได้ดังนี้

1. ผลการวิเคราะห์ความลำเอียงของข้อสอบของข้อสอบแต่ละวิชา จากการวิเคราะห์ 3 วิธี ได้ข้อสรุปดังนี้

1.1 ข้อสอบวิชาภาษาไทย จำนวน 100 ข้อ เมื่อวิเคราะห์ความลำเอียงจำแนกตามกลุ่มเพศ ปรากฏว่า ข้อสอบมีความลำเอียงจำนวน 24 – 65 ข้อ เมื่อจำแนกตามกลุ่มที่ตั้งของโรงเรียน ข้อสอบมีความลำเอียงจำนวน 24 – 65 ข้อ และเมื่อจำแนกตามกลุ่มขนาดของโรงเรียน ข้อสอบมีความลำเอียงจำนวน 42 – 62 ข้อ

1.2 ข้อสอบวิชาสังคม จำนวน 100 ข้อ เมื่อวิเคราะห์ความลำเอียงจำแนกตามกลุ่มเพศ ปรากฏว่า ข้อสอบมีความลำเอียงจำนวน 15 – 55 ข้อ เมื่อจำแนกตามกลุ่มที่ตั้งของโรงเรียน ข้อสอบมีความลำเอียงจำนวน 40 – 63 ข้อ และเมื่อจำแนกตามกลุ่มขนาดของโรงเรียน ข้อสอบมีความลำเอียงจำนวน 49 – 62 ข้อ

1.3 ข้อสอบวิชาสามัญ 1 จำนวน 100 ข้อ เมื่อวิเคราะห์ความลำเอียงจำแนกตามกลุ่มเพศ ปรากฏว่า ข้อสอบมีความลำเอียงจำนวน 25 – 54 ข้อ เมื่อจำแนกตามกลุ่มที่ตั้งของโรงเรียน ข้อสอบมีความลำเอียงจำนวน 48 – 80 ข้อ และเมื่อจำแนกตามกลุ่มขนาดของโรงเรียน ข้อสอบมีความลำเอียงจำนวน 35 – 74 ข้อ

2. ผลการเปรียบเทียบจำนวนข้อสอบที่มีความลำอียงโดยใช้วิธีการวิเคราะห์ความแปรปรวน วิธีแมนเทล – แฮนส์เซล และวิธีไค – สแควร์ เมื่อจำแนกตามกลุ่มเพศ ที่ตั้ง และขนาดของโรงเรียน ในข้อสอบวิชาภาษาไทย วิชาสังคมและวิชาสามัญ 1 พบว่า วิธีการวิเคราะห์ทั้ง 3 วิธี ให้ผลจำนวนข้อของแบบทดสอบที่ลำเอียงแตกต่างกันอย่างมีนัยสำคัญที่ระดับ .01 โดยที่วิธีแมนเทล–แฮนส์เซลตรวจพบจำนวนข้อสอบที่ลำเอียงในแต่ละวิชาสูงที่สุด รองลงมาคือวิธีไคสแควร์

3. ผลการวิเคราะห์ความสอดคล้องของการวิเคราะห์ความลำเอียงของข้อสอบระหว่างวิธีการวิเคราะห์ความแปรปรวนวิธี วิเคราะห์โดยวิธีแมนเทล – แฮนส์เซลกับวิธีไคสแควร์ พบว่า วิเคราะห์โดยวิเคราะห์โดยวิธีแมนเทล – แฮนส์เซลกับวิธีไคสแควร์ มีความสอดคล้องกันอย่างมีนัยสำคัญที่ระดับ .01 เมื่อจำแนกตามกลุ่มเพศ ที่ตั้ง และขนาดของโรงเรียน ในข้อสอบวิชาภาษาไทย วิชาสังคมและวิชาสามัญ 1

เมอร์ส และกรอสเซน (Merz & Grossen, 1979 อ้างอิงใน เพ็ญพนา สุขสม, 2540, หน้า 8) ได้ใช้ทฤษฎีโค้งลักษณะข้อสอบในการหาความลำเอียงของข้อสอบ โดยใช้ข้อมูลที่ผู้วิจัยสร้าง

ขึ้นและใช้วิธีวิเคราะห์ความลำเอียง 7 วิธี คือ วิธีแปลงค่าความยาก 2 วิธี วิธีไค – สแควร์ 2 วิธี และวิธีโค้งลักษณะข้อสอบ 3 วิธี ผลการวิจัยพบว่าวิธีโค้งลักษณะข้อสอบที่มีพารามิเตอร์ 3 ตัว เป็นวิธีที่ดีที่สุด รองลงมาได้แก่ วิธีไค – สแควร์ และหนึ่งในสองวิธีการแปลงค่าความยากของข้อสอบเป็นวิธีการที่ใช้ได้

สับโคเวียค และคณะ (Subkoviak et al, 1984 อ้างอิงใน ทิพย์รัตน์ ผลบุญ, 2540, หน้า 41) ได้ทำการวิเคราะห์ความลำเอียงของข้อสอบ 3 วิธี คือ วิธีโค้งลักษณะข้อสอบชนิด 3 พารามิเตอร์ วิธีไค – สแควร์ และวิธีแปลงค่าความยาก เครื่องมือที่ใช้ในการวิจัยเป็นแบบสอบ 4 ตัวเลือก วัดความรู้เกี่ยวกับคำศัพท์จำนวน 50 คำ ซึ่งแบบสอบจะประกอบด้วยคำศัพท์ภาษาอังกฤษมาตรฐานจำนวน 40 ข้อ ศัพท์แสลงชาวผิวดำ 10 ข้อ กลุ่มตัวอย่างเป็นนักเรียนผิวดำ 1,008 คน นักเรียนผิวขาว 1,021 คน ผลการวิจัยพบว่า โค้งลักษณะข้อสอบชนิด 3 พารามิเตอร์ เป็นวิธีที่มีประสิทธิภาพสูงที่สุด รองลงมาคือ วิธีไค – สแควร์ ส่วนวิธีแปลงค่าความยากเป็นวิธีที่มีข้อจำกัดแต่ก็สามารถนำไปใช้ได้ในกรณีที่ไม่มีเครื่องคอมพิวเตอร์