บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน โดยนำเสนอเป็นลำดับ ดังหัวข้อต่อไปนี้

1. แนวคิดการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
   1.1 ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ
   1.2 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ
   1.3 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีถดถอยโลจิสติก
3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์และซิปเทสท์ปรับใหม่
4. แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน
   4.1 แนวคิดเกี่ยวกับแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน
   4.2 แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6

แนวคิดในการตรวจสอบทำหน้าที่ต่างกันของข้อสอบ

1. ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ

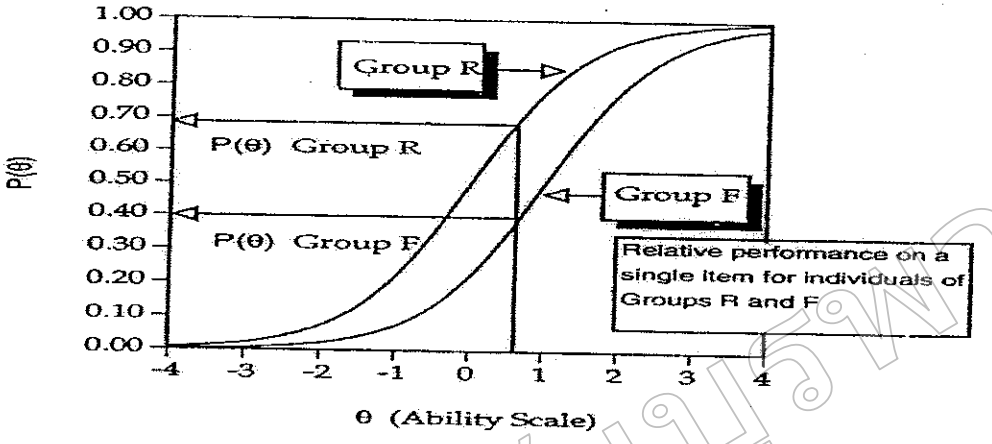
การศึกษาเรื่องผลการสอบของกลุ่มผู้เข้าสอบย่อย ของผู้เข้าสอบทั้งหมดมีมานานแล้ว แต่เรื่องความยุติธรรมในการสอบระหว่างผู้เข้าสอบย่อยกลุ่มต่าง ๆ เพิ่งมีการศึกษาอย่างจริงจังในช่วงปลายทศวรรษ 1960 โดยมีการเสนอวิธีการต่าง ๆ เพื่อนำไปใช้ตรวจสอบความลำเอียงของแบบทดสอบ (Test-Bias) หรือความลำเอียงในการคัดเลือกผู้เข้าสอบ (Selection-Bias) ขึ้นหลายวิธี ขณะเดียวกันในช่วงเวลานั้นนักพัฒนาแบบทดสอบก็สนใจวิธีการจำแนกข้อสอบที่ไม่เหมาะสมกับผู้เข้าสอบบางกลุ่มออกจากแบบทดสอบ ก่อนจะพัฒนาเป็นแบบทดสอบฉบับสมบูรณ์ ทำให้จำเป็นต้องพัฒนาวิธีการตรวจสอบความลำเอียงของข้อสอบ (Item-Bias) เพื่อใช้เป็นแนวทางในการจำแนกข้อสอบที่ลำเอียงกับผู้เข้าสอบบางกลุ่มที่มีลักษณะบางอย่างแตกต่างกัน เช่น เชื้อชาติศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ ประสบการณ์ เป็นต้น เพื่อพัฒนาแบบทดสอบให้มีคุณภาพเหมาะสมสำหรับนำไปใช้ทดสอบต่อไป

ในสมัยแรก ๆ ของการศึกษาเรื่องผลการสอบ เพื่อคัดเลือกคนเข้าศึกษาต่อหรือเข้าทำงาน ปรากฏดัชนีความลำเอียงกับกลุ่มคนต่างชาติ ต่างเพศ ทำให้ต้องมีการศึกษา "ความลำเอียงในการคัดเลือกผู้เข้าสอบ (Selection-Bias)" ต่อมาเพื่อให้การศึกษาเรื่องนี้มีความชัดเจนยิ่งขึ้น จึงได้ศึกษาในระดับข้อสอบ (Item-level) ที่เรียกว่า "ความลำเอียงของข้อสอบ (Item Bias)" ซึ่งในปัจจุบันนักวิจัยส่วนใหญ่ใช้คำว่า "ข้อสอบทำหน้าที่ต่างกันกับกลุ่มผู้เข้าสอบย่อยต่างกลุ่ม" หรือเรียกสั้น ๆ ว่า "ข้อสอบทำหน้าที่ต่างกัน (Differential Item Functioning: DIF)" ทั้งนี้เนื่องมาจากเห็นว่าเป็นคำที่มีความหมายกลาง ๆ จึงมีความเหมาะสมในเชิงวิชาการมากกว่าคำว่า "ความลำเอียง (Bias)" ซึ่งเป็นคำที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ อย่างไรก็ตาม คำสองคำนี้มีจุดเน้นที่แตกต่างกัน โดยคำว่า ความลำเอียงของข้อสอบ เน้นที่อิทธิพลที่สังเกตได้ของกลุ่มผู้เข้าสอบย่อยที่มุ่งศึกษา ส่วนคำว่า ข้อสอบทำหน้าที่ต่างกัน เน้นที่ลักษณะทางสถิติของข้อสอบที่ตรวจสอบได้ด้วยวิธีวิเคราะห์ทางสถิติ ซึ่งเป็นส่วนประกอบหนึ่งของสิ่งที่แสดงถึงความลำเอียงของข้อสอบ (Scheuneman & Bleistein, 1989; Angoff, 1993; Hambleton & Others, 1993; Zieky, 1993; Camilli & Shepard, 1994 อ้างถึงใน วลีมาศ แซ่อึ้ง, 2543) จากจุดเน้นนี้แสดงให้เห็นว่า วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นเงื่อนไขจำเป็น (Necessary Condition) ในการประเมินความลำเอียงของข้อสอบ แต่ถ้าใช้เฉพาะวิธีการทางสถิติอย่างเดียว ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันที่ได้ ก็ไม่อาจสรุปได้ว่าข้อสอบข้อนั้นลำเอียงหรือไม่ เนื่องจากการประเมินความลำเอียงของข้อสอบยังต้องรวมถึงการใช้วิธีให้ผู้เชี่ยวชาญพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบทดสอบ ที่เรียกว่า "วิธีการตัดสินข้อสอบ (Judgmental Method)" ก่อนที่เราจะสรุปว่าข้อสอบข้อนั้นลำเอียงหรือไม่ (Angoff, 1993; Linn, 1993; Ramsay, 1993; Zieky, 1993; Camilli & Shepard, 1994 อ้างถึงใน วลีมาศ แซ่อึ้ง, 2543)

2. ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

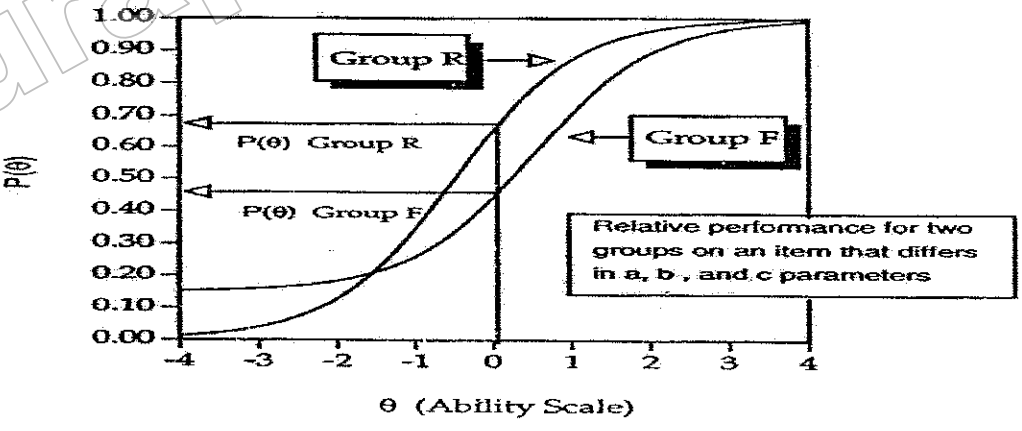
การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ พบว่าข้อสอบสามารถทำหน้าที่แตกต่างกันได้ 2 ประเภท (Mellenbergh, 1982) ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และแบบอเนกรูป (Nonuniform)

2.1 ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform) หมายถึง ข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่ง มีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งอย่างสม่ำเสมอกัน ในทุกระดับความสามารถ เมื่อพิจารณาโค้งลักษณะข้อสอบของผู้สอบ 2 กลุ่ม จะพบว่าไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม (Group Membership) ดังแสดงในภาพที่ 2



ภาพที่ 2 โค้งลักษณะข้อสอบของการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป

2.2 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform) หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มแตกต่างกันอย่างไม่สม่ำเสมอกัน ในทุกระดับความสามารถ เมื่อพิจารณาโค้งลักษณะข้อสอบของผู้สอบ 2 กลุ่ม จะพบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม (Group Membership) เช่น ที่ระดับความสามารถหนึ่งกลุ่มผู้สอบกลุ่มอ้างอิง (Reference Group; R) มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่มเปรียบเทียบ (Focal Group; F) แต่ที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้สอบกลุ่มเปรียบเทียบมีโอกาสในการตอบข้อสอบถูก มากกว่าผู้สอบกลุ่มอ้างอิง



ภาพที่ 3 โค้งลักษณะข้อสอบของการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป

ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory; IRT) สามารถพิจารณา "ปฏิสัมพันธ์" ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป แล้วโค้งลักษณะข้อสอบ (Item Characteristic Curves; ICCs) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบเหมือนกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแล้วโค้งลักษณะข้อสอบจะไม่ขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบต่างกัน ดังนั้นความแตกต่างระหว่างลักษณะข้อสอบทั้งสองแบบจะบ่งบอกถึงขนาด และทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน

3. หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF Detection) เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่ม ที่มีความสามารถหลัก (Primary Ability) ที่มุ่งวัดเท่ากัน แต่คาดว่าจะมีความได้เปรียบหรือเสียเปรียบกัน โดยกลุ่มหนึ่งถือเป็น กลุ่มอ้างอิง (Reference Group; R) ซึ่งคาดว่าน่าจะได้เปรียบในการตอบข้อสอบข้อนั้น หรือมีโอกาสตอบข้อสอบได้ถูกต้องมากกว่า ส่วนอีกกลุ่มหนึ่ง คือ กลุ่มเปรียบเทียบ (Focal Group; F) ซึ่งเป็นกลุ่มที่สนใจศึกษาและคาดว่าน่าจะเป็นกลุ่มที่เสียเปรียบ

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำเป็นต้องจับคู่ (Matching) ผู้สอบตามความสามารถ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

เกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมีสองวิธี ดังนี้

3.1 เกณฑ์ภายนอก (External Criterion)

การวิเคราะห์การทำหน้าที่ต่างกันโดยใช้คะแนนจากแบบทดสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิคการวิเคราะห์การถดถอย (Regression Analysis) เพื่อทำการเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์ กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

หลักการนี้มีจุดมุ่งหมาย เพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบทดสอบอื่นจากตัวแปรทำนายซึ่งเป็นคะแนนรายข้อ หรือคะแนนแบบทดสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะใช้คะแนนรายข้อเป็นตัวแปรทำนาย แต่ถ้าเป็นการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบ จะใช้คะแนนรวมของแบบทดสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอก อาจใช้คะแนนรวมทั้งฉบับหรือเกรดเฉลี่ย หรือผลสัมฤทธิ์ในงานที่เกี่ยวข้องของผู้สอบ (Cronbach, 1970)

3.2 เกณฑ์ภายใน (Internal Criterion)

การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบ โดยเน้นการพิจารณาจากโครงสร้างภายในของแบบทดสอบเป็นหลัก ด้วยการวิเคราะห์ผลการตอบข้อสอบ และความสามารถหรือคะแนนจริงของผู้สอบที่ได้จากแบบทดสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบ หรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ การวิเคราะห์ในลักษณะนี้นิยมใช้ค่าสถิติต่าง ๆ เป็นตัวบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ ค่าสถิติทดสอบที่นิยมนำมาใช้พอสรุปได้ดังนี้ 1) การทดสอบปฏิสัมพันธ์ (Interaction) 2) การวัดความเบี่ยงเบนสัมพัทธ์ (Relative Deviation) 3) การเปรียบเทียบน้ำหนักตัวประกอบ (Factor Loading) 4) การเปรียบเทียบโอกาสตอบข้อสอบถูก

วิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี แฮมเบิลตันและคณะ (Hambleton et al., 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539, หน้า 4-6) จำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบออกเป็น 3 กลุ่มใหญ่ ๆ ดังนี้

3.2.1 กลุ่มวิธีที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม (Methods Using Classical Test Theory: CTT)

วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มนี้พัฒนาจากหลักการของทฤษฎีการทดสอบแบบดั้งเดิม โดยปกติแล้วใช้คะแนนที่สังเกตได้ของผู้เข้าสอบแต่ละคนเป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อย และเปรียบเทียบค่าความยากของข้อสอบแต่ละข้อระหว่างกลุ่มผู้เข้าสอบย่อยที่สนใจศึกษา วิธีการในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (Analysis of Variance) วิธีสหสัมพันธ์ (Correlation Methods) (Green & Draper, 1972 cited in Scheuneman, 1979) วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty Method; TID) หรือ วิธีการกำหนดจุดเดลต้า (Delta Plot Method) (Angoff, 1978) การวิเคราะห์ตัวลวง (Distracter Analysis) (Scheuneman, 1979) วิธีสหสัมพันธ์บางส่วน (Partial Correlation Methods) และวิธีการทำให้เป็นมาตรฐาน (Standardization Method) (Dorans & Kulick, 1986)

ข้อได้เปรียบของวิธีในกลุ่มนี้ คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายไม่สูงนัก ใช้ตรวจสอบกับกลุ่มตัวอย่างขนาดเล็กได้และสามารถอธิบายความทั่วไปเข้าใจได้ง่าย ส่วนข้อเสียเปรียบก็คือ ค่าสถิติของข้อสอบเปลี่ยนไปตามกลุ่มตัวอย่าง เมื่อกลุ่มตัวอย่างเปลี่ยนไปผลการตรวจพบข้อสอบทำหน้าที่ต่างกันก็เปลี่ยนไป ทำให้การอ้างอิงผลการศึกษาไปยังกลุ่มประชากรอาจมีความเชื่อถือได้น้อยลง

3.2.2 กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Methods Using Item Response Theory: IRT)

วิธีการในกลุ่มนี้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามกรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบโดยปกติแล้วใช้การเปรียบเทียบโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยตามระดับความสามารถของผู้เข้าสอบ ถ้าโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่ม มีรูปร่างเหมือนกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่มมีรูปร่างต่างกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน ค่าพารามิเตอร์ของโค้งลักษณะข้อสอบ ได้แก่ ค่าความยากของข้อสอบ (b-Parameter) ค่าอำนาจจำแนกของข้อสอบ (a-Parameter) และค่าการเดาข้อสอบถูก (c-Parameter) วิธีการในกลุ่มนี้ ได้แก่ วิธี Analysis of Fit (Durovic, 1975 cited in Hambleton et al., 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539) วิธี Difficulty Shift (Wright, Mead, & Draba, 1976 cited in Hambleton et al., 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539) ซึ่งใช้โมเดล IRT แบบหนึ่งพารามิเตอร์ วิธี IRT Area (Ironson & Subkoviak, 1979; Raju, 1988; 1990 อ้างถึงใน เสรี ชัดแช้ม, 2539) และวิธี Two Stage (Lord, 1980) ซึ่งใช้โมเดล IRT แบบสองหรือสามพารามิเตอร์ และวิธี Plot (Hambleton & Rogers, 1991 cited in Hambleton et al., 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539)

ข้อได้เปรียบของวิธีการในกลุ่มนี้คือ การแก้ไขข้อบกพร่องของทฤษฎีการตอบข้อสอบแบบดั้งเดิมทำให้ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่างที่สุ่มมาจากประชากรเดียวกัน การประมาณค่าความสามารถของผู้สอบเป็นอิสระจากค่าความยากของแบบทดสอบ โมเดลทางคณิตศาสตร์ง่ายต่อการจับคู่โค้งลักษณะข้อสอบตามระดับความสามารถของผู้เข้าสอบย่อยได้ ไม่ต้องมีข้อตกลงเบื้องต้นเรื่องแบบทดสอบคู่ขนานในการหาค่าสัมประสิทธิ์ความเที่ยงของแบบทดสอบ ส่วนข้อเสียเปรียบของวิธีการในกลุ่มนี้ คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สลับซับซ้อน เสียค่าใช้จ่ายในการวิเคราะห์ข้อมูลสูง และต้องใช้กับกลุ่มตัวอย่างขนาดใหญ่

3.2.3 กลุ่มวิธีที่ใช้วิธีไค-สแควร์ (Methods Using Chi-square Methods)

วิธีในกลุ่มนี้ใช้ค่าไค-สแควร์ (Chi-square Methods) เป็นดัชนีแสดงการทำหน้าที่ต่างกันของข้อสอบ และใช้คะแนนของแบบทดสอบ ที่ทำให้บริสุทธิ์เป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อย ๆ ก่อนการเปรียบเทียบผลการตอบข้อสอบ วิธีการในกลุ่มนี้ ได้แก่ วิธีตารางการณ์จร (Contingency Table Method) (Scheuneman, 1979) วิธีตารางการณ์จรปรับขยาย (Modified Contingency Table Method) (Veale, 1977 cited in Hambleton et al., 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539) วิธีล็อก-ลิเนียร์ (Log-Linear Models) (Mellenbergh, 1982) วิธีแมนเทล-แฮนส์เซล

(Mantel-Haenszel Method: MH) (Holland & Thayer, 1986 cited in Holland & Wainer, 1993) และวิธีการถดถอยโลจิสติก (Logistic Regression Methods) (Swaminathan & Rogers, 1990 cited in Rogers & Swaminathan, 1993)

ข้อได้เปรียบของวิธีในกลุ่มนี้คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายในการวิเคราะห์ข้อมูลไม่สูง ใช้ได้กับกลุ่มตัวอย่างขนาดไม่ใหญ่ และบางวิธีมีหลักการที่ดีในการจับคู่กลุ่มผู้เข้าสอบย่อยตามความสามารถของผู้สอบ และมีการทดสอบนัยสำคัญ ส่วนข้อเสียเปรียบของวิธีในกลุ่มนี้ก็คล้าย ๆ กับกลุ่มที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีถดถอยโลจิสติก (Logistic Regression)

วิธีถดถอยโลจิสติก เป็นวิธีที่พัฒนามาจากวิธีลอกลิเนียร์ (Loglinear) ของเมลเลนเบอร์ก (Mellenberg, 1982) ซึ่งเสนอโดยสวามินาธานและโรเจอร์ (Swaminathan & Rogers, 1990 cited in Rogers & Swaminathan, 1993) วิธีนี้อยู่บนพื้นฐานของโมเดลการวิเคราะห์สมการถดถอยโลจิสติก ซึ่งเป็นโมเดลที่มีพื้นฐานเป็นแบบจำลองที่สามารถเพิ่มตัวแปรความสามารถ และปฏิสัมพันธ์เข้าไปในสมการได้ และสอดคล้องกับธรรมชาติของการทำหน้าที่ต่างกันของแบบทดสอบได้ดี จึงสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูปได้

โมเดลถดถอยโลจิสติก เป็นโมเดลที่มีความยืดหยุ่นและใช้ได้ง่าย ในการวิเคราะห์สมการถดถอยโลจิสติก หรือเรียกว่าการวิเคราะห์โลจิส (Logit Analysis) เป็นการวิเคราะห์สมการทำนาย เมื่อต้องการศึกษาผลของตัวแปรทำนายที่มีต่อตัวแปรเกณฑ์ซึ่งมีลักษณะเป็นทวิวิภาค (Dichotomous Variable) โดยใช้ฟังก์ชันโลจิสติก (Logistic Function) ในการแสดงความสัมพันธ์ระหว่างค่าของตัวแปรทำนาย กับค่าความน่าจะเป็นของการเกิดเหตุการณ์ตามตัวแปรเกณฑ์

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยแบ่งสมการสำหรับผู้สอบออกเป็นสองกลุ่ม สมการมาตรฐานของโมเดลการถดถอยโลจิสติก สำหรับคำนวณโอกาสของการตอบข้อสอบถูก เป็นดังนี้

$$P(U_{ij}=1/\theta_{ij})=\frac{e^{(\beta_{ij}+\beta_{ij}\theta_{ij})}}{1+e^{(\beta_{ij}+\beta_{ij}\theta_{ij})}} \quad , i = 1, 2,..., n_j : j = 1, 2$$

เมื่อ U_{ij} คือ โอกาสของการตอบข้อสอบถูกของบุคคลที่ i ในกลุ่ม j

θ_{ij} คือ ความสามารถ (คะแนนรวม) ของผู้สอบที่ i ในกลุ่ม j

β_{0j} คือ พารามิเตอร์จุดตัด (Intercept Parameter)

β_{1j} คือ ค่าพารามิเตอร์ความชัน (Slope Parameter)

ถ้า $\beta_{01} = \beta_{02}$ และ $\beta_{11} = \beta_{12}$ เป็นลักษณะโค้งถดถอยโลจิสติก สำหรับผู้สอบสองกลุ่มเหมือนกันแสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (No DIF)

$\beta_{01} \neq \beta_{02}$ และ $\beta_{11} = \beta_{12}$ เป็นลักษณะโค้งถดถอยโลจิสติก สำหรับผู้สอบสองกลุ่มเท่าเทียมกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF)

$\beta_{01} = \beta_{02}$ และ $\beta_{11} \neq \beta_{12}$ เป็นลักษณะโค้งถดถอยโลจิสติก สำหรับผู้สอบสองกลุ่มไม่เท่าเทียมกันแสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF)

จากสมการถดถอยโลจิสติกดังกล่าว สามารถเขียนเป็นสมการใหม่ในรูปการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ได้ดังนี้

$$P(U_{ij}=1/\theta_{ij})=\frac{e^{Z_{ij}}}{1+e^{Z_{ij}}}$$

โดยที่ $$Z_{ij} = \tau_0 + \tau_1\theta_{ij} + \tau_2 G_j + \tau_3\theta_{ij}G_j$$

เมื่อ

$p(U_{ij}=1/\theta_{ij})$ คือ โอกาสในการตอบข้อสอบถูกของผู้สอบที่ i ในกลุ่ม j ซึ่งมีความสามารถ θ

θ_{ij} คือ ความสามารถของผู้สอบ (คะแนนรวม) คนที่ i ในกลุ่มที่ j

G_j คือ กลุ่มผู้สอบ ($G = 1$ สมาชิกกลุ่ม 1 หรือกลุ่มเปรียบเทียบ, $G = 2$ สมาชิกกลุ่ม 2 หรือกลุ่มอ้างอิง)

$\theta_{ij}G_j$ คือ ปฏิสัมพันธ์ของตัวแปรอิสระ 2 ตัว คือ θ_{ij} กับ G_j

τ_0 คือ พารามิเตอร์จุดตัด

τ_1 คือ ผลของความสามารถของผู้สอบ

τ_2 คือ ความแตกต่างระหว่างกลุ่มในการตอบข้อสอบถูก

τ_3 คือ ปฏิสัมพันธ์ระหว่างกลุ่มกับความสามารถผู้สอบ

โมเดลการถดถอยโลจิสติกข้างต้น สามารถเปลี่ยนเป็นโมเดลเชิงเส้นในเมทริกซ์โลจิท (Logist Metric) ซึ่งจะอยู่ในรูป Log ของอัตราส่วนของโอกาสในการตอบข้อสอบถูกต่อโอกาส

ในการตอบข้อสอบผิด ดังนี้

$$log\left[\frac{p}{1-p}\right]=z_{ij}=\tau_0+\tau_1\theta_{ij}+\tau_2 G_j+\tau_3\theta_{ij}G_j$$

จากโมเดลดังกล่าว เทอม $\theta_i G_j$ เป็นผลคูณของตัวแปรอิสระ θ_i และ G_j ส่วนพารามิเตอร์ τ_2 และ τ_3 สอดคล้องกับเทอมของพารามิเตอร์ในสมการของโมเดลการถดถอยโลจิสติก ดังนี้

$$\tau_2 = \beta_{01} - \beta_{02}$$

$$\tau_3 = \beta_{11} - \beta_{12}$$

ในการตัดสินว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูปหรืออเนกรูป พิจารณาพารามิเตอร์ τ_2 และ τ_3 กล่าวคือ ถ้า $\tau_2 \neq 0$ และ $\tau_3 = 0$ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป และถ้า $\tau_3 \neq 0$ และ $\tau_2 = 0$ หรือไม่ก็ได้แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป

สำหรับการประมาณค่าพารามิเตอร์ตามโมเดลโลจิสติก ของข้อสอบแต่ละข้อของโมเดล Z_{ij} ใช้วิธีประมาณค่าด้วยวิธีความควรจะเป็นสูงสุด (Maximum Likelihood Estimation; MLE) ซึ่งเขียนในรูปฟังก์ชันได้ดังนี้

$$L(U_{ij} / \theta) = \prod_{i=1}^{n}\prod_{j=1}^{k} P\left(U_{ij}\right)^{U_{ij}}\left[1 - P\left(U_{ij}\right)\right]^{1-U_{ij}}$$

โดยที่ k และ n แทนขนาดกลุ่มตัวอย่าง และความยาวของแบบทดสอบตามลำดับ สำหรับค่าประมาณของพารามิเตอร์โดยใช้วิธีความควรจะเป็นสูงสุด มีการแจกแจงแบบปกติของตัวแปรพหุในรูปเชิงเส้นกำกับ (Asymptotically Multivariate Normal) ซึ่งมีค่าเฉลี่ยของเวคเตอร์ τ และ เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมในรูป Σ ในขณะที่ Σ^{-1} เป็นเมทริกสารสนเทศกำหนดดังนี้

$$\Sigma^{-1} = -E\left[\frac{\partial^2}{\partial\tau_r\partial\tau_s} lnL\right] \quad ; \quad r, s = 0, 1, 2, 3$$

เมื่อ E และ lnL แทนค่าความคาดหมายของเมทริกซ์และลอกริทึมของฟังก์ชันไลค์ลิฮูด ตามลำดับ ดังนั้นการแจกแจงของการประมาณค่าพารามิเตอร์ด้วยวิธี MLE จะอยู่ในรูป ดังนี้

$$\tau \sim N(\tau, \Sigma)$$

โดยที่ $\tau = \lfloor \tau_0, \tau_1, \tau_2, \tau_3 \rfloor$ ส่วนความคลาดเคลื่อนมาตรฐานเชิงเส้นกำกับของค่าประมาณของ τ_s $(S = 0,1,2,3)$ เมื่อ S เป็นสมาชิกแนวเส้นทแยงมุมของ Σ สามารถคำนวณได้จากสูตร ดังนี้

$$SE(\hat{\tau}_s) = \sqrt{\Sigma^2}$$

ในการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบจะทดสอบสมาชิกของ τ_s ซึ่งสมมติฐานที่สนใจคือ $H_0: \tau_2 = 0$ และ $H_0: \tau_3 = 0$ สมมติฐานทั้งสองสามารถทดสอบพร้อม ๆ กันไป ดังนี้

$H_0: C_\tau = 0$

$H_1: C_\tau \neq 0$

โดยที่ C เป็นเมทริกซ์ขนาด 2×4 ดังนี้

$$C = \begin{vmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

การทดสอบนัยสำคัญของสมมติฐานจะใช้สถิติไค-สแควร์ ที่ระดับชั้นความเป็นอิสระเท่ากับ 2 ($df = 2$) ดังนี้

$$\chi^2 = \tau' C' \left(\hat{C \Sigma C'} \right)^{-1} C \hat{\tau}'$$

ถ้า χ^2 มีค่ามากกว่า $\chi^2_{(\alpha, 2)}$ แสดงว่า ปฏิเสธสมมติฐานศูนย์ นั่นคือ ข้อสอบทำหน้าที่ต่างกัน (DIF) นั่นเอง

นอกจากนี้ คามิลลิ และเชฟเพอร์ด (Camilli & Shepard, 1994) ได้เสนอวิธีทดสอบสมมติฐานเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบโดยเปรียบเทียบระหว่างโมเดลการถดถอยโลจิสติกแบบ 2 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 วิเคราะห์นัยสำคัญของปฏิสัมพันธ์

Model 1 $\Psi_{ij} = \tau_0 + \tau_1 X_{ij} + \tau_2 G_j + \tau_3 (G_j X_{ij})$

เมื่อ ψ_{ij} คือ ค่าโลจิท หรือค่าลอกอัตราส่วนโอกาสตอบถูกต่อโอกาสตอบผิดของผู้สอบ i ในกลุ่มที่ j

X_{ij} คือ คะแนนรวมของผู้สอบ i ในกลุ่ม j

G_j คือ กลุ่มผู้สอบ

Model 2 $\Psi_{ij} = \tau_0 + \tau_1 X_{ij} + \tau_2 G_j$

การทดสอบความแตกต่างระหว่างโมเดล 1 และโมเดล 2 ด้วยการทดสอบไค-สแควร์ที่ระดับชั้นความเป็นอิสระเท่ากับ 1($df=1$) ถ้าพบความแตกต่างอย่างมีนัยสำคัญ หรือ τ_3 มีนัยสำคัญทางสถิติ แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) แต่ถ้า τ_3 ไม่มีนัยสำคัญให้ดำเนินการขั้นตอนที่ 2

ขั้นตอนที่ 2 วิเคราะห์นัยสำคัญของกลุ่มผู้สอบ

Model 2 $\Psi_{ij} = \tau_0 + \tau_1 X_{ij} + \tau_2 G_j$

Model 3 $\Psi_{ij} = \tau_0 + \tau_1 X_{ij}$

ทดสอบความแตกต่างระหว่างโมเดล 2 และโมเดล 3 ด้วยการทดสอบไค-สแควร์ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 ($df=$) ถ้าพบความแตกต่างอย่างมีนัยสำคัญหรือ τ_2 มีนัยสำคัญทางสถิติ แสดงว่าข้อสอบข้อนั้นทำหน้าที่แตกต่างกันแบบเอกรูป (Uniform DIF) ซึ่งจะให้ผลคล้ายกับ MH_{DIF}

งานวิจัยที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก ดูลิทเทิล และเคลียรี่ (Doolittle & Cleary, 1987) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบต่อเพศในแบบทดสอบ ACT (Assessment Mathematic Test) ฉบับการคำนวณทางคณิตศาสตร์ เรขาคณิต และคณิตศาสตร์การให้เหตุผล จำนวน 8 ฉบับ กลุ่มตัวอย่างที่ใช้เป็นเพศชายและเพศหญิง จำนวน 1,300 - 1,400 คน โดยเป็นเพศหญิงประมาณร้อยละ 55

วิเคราะห์ข้อมูลด้วยดัชนีของลินน์ และฮาร์นิสช์ (Lin & Harnisch, 1981 cited Doolittle & Cleary, 1987) ซึ่งเป็นโมเดลโลจิสติกแบบ 3 พารามิเตอร์ ถ้าค่า Z ที่ได้เป็นบวกแสดงว่าง่ายสำหรับกลุ่มเปรียบเทียบ และถ้าเป็นลบแสดงว่ายากสำหรับกลุ่มเปรียบเทียบ ผลการวิจัยพบว่า ค่าดัชนี Z มีค่าเป็นลบสำหรับข้อสอบที่วัดด้านเรขาคณิต การใช้เหตุผลเชิงพีชคณิต และเลขคณิตในทุกฉบับ แสดงว่าง่ายสำหรับเพศชาย

สวามินาธาน และโรเจอร์ (Swaminathan & Rogers, 1990 cited in Rogers & Swaminathan, 1993) ได้เปรียบเทียบระหว่างวิธีถดถอยโลจิสติก และวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรู โดยศึกษาในสถานการณ์จำลอง 6 เงื่อนไข คือ ขนาดกลุ่มตัวอย่างสองระดับ (25 คน ต่อกลุ่ม และ 500 คนต่อกลุ่ม) และความยาวของแบบทดสอบ 3 ระดับ (40, 60 และ 80 ข้อ) ซึ่งในแบบทดสอบแต่ละชุดประกอบด้วยสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันจำนวน 20% โดยครึ่งหนึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและอีกครึ่งหนึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป สำหรับผลการตอบข้อสอบทั้งหมดจำลองโดยใช้โปรแกรม DATAGEN โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ในการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปจะกำหนดให้พารามิเตอร์อำนาจจำแนกระหว่างผู้สอบ 2 กลุ่มมีค่าเท่ากัน ในขณะที่พารามิเตอร์ความยากจะมีค่าแปรเปลี่ยนสำหรับการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปจะกำหนดให้พารามิเตอร์ความยากระหว่างผู้สอบ 2 กลุ่มมีค่าเท่ากัน ส่วนพารามิเตอร์อำนาจจำแนกจะมีค่าแปรเปลี่ยนในการควบคุมขนาดของการทำหน้าที่ต่างกันของข้อสอบจะใช้พื้นที่ระหว่างโค้งลักษณะข้อสอบ ซึ่งคำนวณโดยใช้สูตรของราจู (Raju, n.d. cited in Rogers & Swaminathan, 1993)

ผลการศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป วิธีถดถอยโลจิสติกมีอำนาจการทดสอบสูงกว่าวิธีแมนเทล-แฮนส์เซล ส่วนการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปทั้ง 2 วิธีมีอำนาจการทดสอบเท่าเทียมกัน สำหรับปัจจัยของขนาดกลุ่มตัวอย่าง พบว่า เมื่อขนาดกลุ่มตัวอย่าง 250 คน ต่อกลุ่มผู้สอบทั้ง 2 วิธีมีความถูกต้องแม่นยำในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปประมาณ 75% และเมื่อขนาดกลุ่มตัวอย่าง 500 คน ต่อกลุ่มผู้สอบทั้ง 2 วิธีมีความถูกต้องแม่นยำในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปประมาณ 100% สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า วิธีแมนเทล-แฮนส์เซล มีอัตราความคลาดเคลื่อนร้อยละ 1 ในขณะที่วิธีถดถอยโลจิสติก มีอัตราความคลาดเคลื่อนระหว่างร้อยละ 1 ถึง 6 เมื่อพิจารณาถึงค่าใช้จ่ายในการวิเคราะห์ ปรากฏว่าวิธีถดถอยโลจิสติกมีค่าใช้จ่ายสูงกว่าวิธีแมนเทล-แฮนส์เซล ประมาณ 3 - 4 เท่า

โรเจอร์ และสวามินาธาน (Rogers & Swaminathan, 1993) ได้ศึกษาเปรียบเทียบวิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป โดยศึกษาในสถานการณ์จำลอง 2 ครั้ง คือ ครั้งที่ 1 ศึกษาการแจกแจงของสถิติที่ใช้ทดสอบ และครั้งที่ 2 ศึกษาอำนาจการทดสอบ ภายใต้การจำลองข้อมูล 2 เงื่อนไข ($2 \times 2 \times 2 \times 2 \times 2$) ความเหมาะสมของข้อมูลกับโมเดล 2 ระดับ (เหมาะสมและไม่เหมาะสม) ขนาดกลุ่มตัวอย่าง 2 ระดับ (250 คนต่อกลุ่มและ 500 คนต่อกลุ่ม) ขนาดความยาวของแบบทดสอบ 2 ระดับ (40 ข้อและ 80 ข้อ) โค้งของการแจกแจงคะแนนการสอบ 2 ระดับ (ปกติและเบ้ซ้าย) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 2 ระดับ (15% และ 0%) ในแต่ละเงื่อนไขจะจำลองข้อสอบที่ทำหน้าที่ต่างกัน 2 ประเภท คือ แบบเอกรูปและแบบอเนกรูป โดยข้อสอบที่ทำหน้าที่ต่างกันในแต่ละประเภทและแต่ละเงื่อนไขจะมีขนาดของข้อสอบที่ทำหน้าที่ต่างกันเท่ากับ .2, .4, .6 และ .8 ซึ่งขนาดของข้อสอบดังกล่าวคำนวณจากพื้นที่ระหว่าง IRFs สำหรับการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปจะมีข้อสอบ 4 ลักษณะ คือ a ต่ำกับ b สูง, a ปานกลางกับ b ต่ำ, a ปานกลางกับ b สูง และ a สูงกับ b สูง ส่วนการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปจะมีข้อสอบ 4 ลักษณะ b ต่ำกับ a ต่ำ, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง และ b สูงกับ a ต่ำ

ผลการศึกษากรณีการศึกษาการแจกแจงของสถิติ พบว่า วิธีถดถอยโลจิสติก และวิธีแมนเทล-แฮนส์เซล ให้ผลการตรวจสอบเป็นไปตามที่คาดไว้เกือบทุกเงื่อนไข เมื่อข้อสอบยากมากและค่าอำนาจจำแนกสูง การแจกแจงของสถิติที่ใช้ทดสอบวิธีถดถอยโลจิสติก จะไม่เป็นไปตามที่คาดไว้ สำหรับผลการศึกษาอำนาจการทดสอบ พบว่า วิธีการถดถอยโลจิสติก และวิธีแมนเทล-แฮนส์เซล มีอำนาจการทดสอบเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป วิธีถดถอยโลจิสติกมีอำนาจการทดสอบสูงกว่าวิธีแมนเทล-แฮนส์เซล สำหรับปัจจัยของขนาดกลุ่มตัวอย่าง ลักษณะของข้อสอบ และขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน จะมีผลต่ออำนาจการทดสอบของวิธีการถดถอยโลจิสติก และวิธีแมนเทล-แฮนส์เซล กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างและขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีค่าเพิ่มขึ้นจะมีผลทำให้อำนาจการทดสอบของทั้ง 2 วิธีดังกล่าวมีค่าสูงสุด แต่เมื่อลักษณะของข้อสอบมีค่าความยากปานกลางจะมีผลทำให้อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปของวิธีแมนเทล-แฮนส์เซลมีค่าต่ำสุด

นารายาน และสวามินาธาน (Narayanan & Swaminathan, 1996) ได้ศึกษาเปรียบเทียบ

วิธีแมนเทล-แฮนส์เซล (MH) วิธีโคร-ซิบ (CRO-SIB) และวิธีถดถอยโลจิสติก (LR) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป โดยศึกษาในสถานการณ์จำลองภายใต้เงื่อนไข 384 เงื่อนไข (4 x 2 x 3 x4 x 4) ตามปัจจัยที่แปรเปลี่ยนดังนี้ ขนาดกลุ่มตัวอย่าง 4 ระดับ (กลุ่มอ้างอิงขนาด 500 คน และ 1,000 คน กลุ่มเปรียบเทียบขนาด 200 คน และ 500 คน) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ระดับ (0%, 10% และ 20%) ความแตกต่างของการแจกแจงค่าความสามารถ 2 ระดับ (แบบเท่ากันและแบบไม่เท่ากัน) ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน 4 ระดับ (พื้นที่ระหว่าง IRFs มีค่าเท่ากับ .4, .6, .8, และ 1.0) ลักษณะของข้อสอบ 4 ระดับ (b ต่ำกับ a สูง, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง และ b สูงกับ a ต่ำ) สำหรับความยากของแบบทดสอบใช้เพียง 40 ข้อ ข้อมูลที่ใช้ในการศึกษาจำลองตามโมเดลโลจิสติกแบบ 3 พารามิเตอร์ โดยใช้โปรแกรม DATAGEN สำหรับดัชนีการตรวจสอบด้วยวิธี MH และวิธี LR ใช้โปรแกรม DICHODIF ส่วนวิธี CRO-SIB ใช้โปรแกรม CSIBTEST แล้วใช้การวิเคราะห์ ANOVA แบบ 5 ทิศทาง เพื่อทดสอบผลกระทบของปัจจัยที่ตรวจสอบด้วยวิธีทั้งสาม โดยทดสอบที่ระดับนัยสำคัญ .05 และ .01

ผลการศึกษาพบว่า วิธีโคร-ซิปและวิธีการถดถอยโลจิสติกมีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปภายใต้เกือบทุกเงื่อนไข ในขณะที่วิธีแมนเทล-แฮนส์เซลไม่มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปที่มีมีทิศทางหรือมีปฏิสัมพันธ์แบบไม่เป็นลำดับ สำหรับปัจจัยของขนาดกลุ่มผู้สอบ ลักษณะของข้อสอบ และขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีผลต่ออำนาจการทดสอบของทั้ง 3 วิธี กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างและขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีค่าเพิ่มขึ้นจะมีผลทำให้อำนาจการทดสอบของทั้ง 3 วิธีมีค่าเพิ่มมากขึ้น และเมื่อลักษณะของข้อสอบมีค่าอำนาจจำแนกสูงแล้วอำนาจการทดสอบของวิธโคร-ซิป และวิธีถดถอยโลจิสติกมีค่าเพิ่มมากขึ้นแต่เมื่อข้อสอบมีค่าความยากสูงหรือต่ำแล้วอำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลจะมีค่าเพิ่มขึ้น สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า วิธีโคร-ซิป และวิธีถดถอยโลจิสติกมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่า วิธีแมนเทล-แฮนส์เซล ภายใต้เกือบทุกเงื่อนไข นอกจากนี้ยังพบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธีมีค่าสูง เมื่อลักษณะของข้อสอบมีค่าอำนาจจำแนกสูง

คลาสเซอร์, นันกาสเตอร์, และสวามินาธาน (Clauser, Nmngester, & Swaminathan, 1996 cited in Narayanan & Swaminathan, 1996) ได้ทำการศึกษาเกี่ยวกับการค้นหาความลำเอียงภายใต้เงื่อนไขตัวแปรคะแนนการสอบ และภูมิหลังของการศึกษาของกลุ่มตัวอย่าง โดยใช้วิธีถดถอยโลจิสติกในการวิเคราะห์ความลำเอียง โดยกลุ่มตัวอย่างที่ใช้ในการศึกษาครั้งนี้เป็นนักศึกษา

วิชาแพทยศาสตร์ปีที่ 1 ในโปรแกรมของการฝึกภาคปฏิบัติที่ฝึกปฏิบัติต่างกัน 4 กลุ่ม คือ กลุ่มเภสัชศาสตร์, กลุ่มศัลยแพทย์, กลุ่มกุมารเวชศาสตร์และกลุ่มสูตินารี แบ่งเป็นเพศชาย 1,000 คน และเพศหญิง 1,000 คน ซึ่งผลการศึกษาพบว่า จำนวนข้อสอบที่มีความลำเอียงมีจำนวนน้อยอย่างมีนัยสำคัญทางสถิติภายใต้เงื่อนไขคะแนนรวมและตัวแปรทางด้านพื้นฐานการศึกษากับคะแนนรวมหรืออย่างใดอย่างหนึ่ง

แวง และเลนนี่ (Wang & Lane, 1996 cited in Narayanan & Swaminathan, 1996) ได้ทำการศึกษาเกี่ยวกับการค้นหาความลำเอียงของแบบทดสอบภาคปฏิบัติวิชาคณิตศาสตร์ โดยใช้วิธีการวิเคราะห์ความลำเอียงแตกต่างกัน 3 วิธี คือ วิธี Logistic Regression (LOG) วิธี Logistic Discriminant Function Analysis (LDFD) และวิธี HW3 กับแบบทดสอบภาคปฏิบัติวิชาคณิตศาสตร์จำนวน 36 ข้อ ซึ่งมีอยู่ 4 ฟอร์ม (ฟอร์ม A ถึง D) กลุ่มตัวอย่างที่ใช้ในการศึกษาค้นคว้าครั้งนี้เป็นนักเรียนเกรด 6-7 ในปี 1991 ถึง 1992 จำนวน 1,782 คน พบว่า วิธี HW3 ค้นพบความลำเอียงของข้อสอบที่มีความลำเอียง 7 ข้อ ที่ระดับนัยสำคัญทางสถิติ .05 และ 2 ข้อ ที่ระดับนัยสำคัญทางสถิติ .005 วิธี Logistic Discriminant Function Analysis (LDFD) ค้นพบข้อสอบที่มีความลำเอียงแบบอันยูนิฟอร์ม (Nonuniform) จากแบบทดสอบจำนวน 1 ข้อ และ 6 ข้อ จากการวิเคราะห์แบบยูนิฟอร์ม (Uniform) ที่ระดับนัยสำคัญทางสถิติ .05 และ 2 ข้อ ในการค้นหาแบบยูนิฟอร์ม (Uniform DIF) ที่ระดับนัยสำคัญทางสถิติ .05 และไม่พบข้อสอบที่มีความลำเอียงในการค้นหาแบบนอนยูนิฟอร์ม (Nonuniform DIF) ส่วนวิธี Logistic Regression (LOG) ในการค้นหาแบบ Nonuniform DIF ไม่พบข้อสอบที่มีความลำเอียง และในการค้นหาแบบ Uniform DIF พบข้อสอบที่มีความลำเอียง 9 ข้อที่ระดับนัยสำคัญทางสถิติ .05 และ 3 ข้อ ที่ระดับนัยสำคัญทางสถิติ .005 ในด้านเปอร์เซ็นต์ของการค้นหาความลำเอียงของข้อสอบทั้ง 3 วิธีจากแบบทดสอบ QCAI พบว่ามีความสัมพันธ์กันอยู่ในระดับสูง โดยมีช่วงพิสัยอยู่ระหว่าง 88% ถึง 97% ทั้งในการค้นหาแบบเอกรูปและแบบอเนกรูป ที่ระดับนัยสำคัญทางสถิติ .05

รัชรินทร์ มุคดา (2539) ได้ศึกษาเปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป โดยศึกษาจากข้อมูลที่จำลองขึ้นด้วยโปรแกรม IRTDATA เงื่อนไขที่ศึกษา ได้แก่ 1) กลุ่มความสามารถผู้สอบ 3 ระดับ คือ กลุ่มผู้สอบที่มีความสามารถ สูง ปานกลาง และต่ำ 2) ค่าความยากของข้อสอบ 3 ระดับ คือ สูง ปานกลาง และต่ำ 3) ค่าอำนาจจำแนกของข้อสอบ 3 ระดับ คือ กลุ่มข้อสอบที่มีค่าอำนาจจำแนก สูง ปานกลาง และต่ำ รวมเงื่อนไขที่ศึกษาทั้งหมด 27 เงื่อนไข ผลการวิจัยพบว่า

1. โดยภาพรวมวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก มีประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปเท่ากัน ในกลุ่มผู้สอบที่มีความสามารถสูง ปานกลาง และต่ำ

2. ในกลุ่มผู้สอบที่มีความสามารถสูง ข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันแบบอเนกรูปมากที่สุดเป็นข้อสอบที่มีค่าความยากสูง ค่าอำนาจจำแนกสูง

3. ในกลุ่มผู้สอบที่มีความสามารถปานกลาง ข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันแบบอเนกรูปมากที่สุดเป็นข้อสอบที่มีค่าความยากปานกลาง ค่าอำนาจจำแนกสูง

4. ในกลุ่มผู้สอบที่มีความสามารถต่ำ ข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันแบบอเนกรูปมากที่สุดเป็นข้อสอบที่มีค่าความยากต่ำ ค่าอำนาจจำแนกสูง

นพมาศ พิพัฒนสุข (2541) ได้ศึกษาเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติกในแบบทดสอบชนิดพหุมิติ เมื่อใช้เกณฑ์จับคู่เปรียบเทียบแตกต่างกัน 3 เกณฑ์ ได้แก่ คะแนนรวม คะแนนแบบทดสอบย่อย และคะแนนหลายแบบทดสอบย่อย โดยข้อมูลที่ใช้ในการศึกษาเก็บรวบรวมจากผลการตอบแบบทดสอบความสามารถทางคณิตศาสตร์ชั้นประถมศึกษาปีที่ 6 ที่ผู้วิจัยสร้างขึ้น กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ในสังกัดสำนักงานการประถมศึกษากรุงเทพมหานคร จำนวน 1,076 คน วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS/PC+ วิเคราะห์คุณภาพของข้อสอบโดยใช้โปรแกรม IRT (BAY) และ CTIA ตรวจสอบความตรงเชิงโครงสร้างด้วยการวิเคราะห์องค์ประกอบเชิงยืนยัน โดยใช้โปรแกรม LISREL 8.50 และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรม SIBTEST และ SPSS/PC+

ผลการวิจัยพบว่า

1. วิธี วิธีแมนเทล-แฮนส์เซล เมื่อใช้คะแนนรวมเป็นเกณฑ์จับคู่เปรียบเทียบ ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 15 ข้อ (20%) และเมื่อใช้คะแนนแบบทดสอบย่อยเป็นเกณฑ์จับคู่ตรวจพบ จำนวน 14 ข้อ (18.67%) สำหรับวิธีถดถอยโลจิสติกเมื่อใช้คะแนนรวมเป็นเกณฑ์จับคู่ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากที่สุดจำนวน 20 ข้อ (26.67%) รองลงมาคือ ใช้คะแนนแบบทดสอบย่อยเป็นเกณฑ์จับคู่ตรวจพบ จำนวน 17 ข้อ (22.67%) และตรวจพบน้อยที่สุดคือเมื่อใช้คะแนนหลายแบบทดสอบย่อยเป็นเกณฑ์จับคู่ตรวจพบ จำนวน 13 ข้อ (17.33%)

2. วิธีแมนเทล-แฮนส์เซล มีประสิทธิภาพมากกว่าวิธีถดถอยโลจิสติก ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบชนิดพหุมิติเมื่อใช้เกณฑ์จับคู่คะแนนรวมและมีประสิทธิภาพไม่แตกต่างกันเมื่อใช้เกณฑ์จับคู่คะแนนแบบทดสอบย่อย

3. วิธีถดถอยโลจิสติกเมื่อใช้เกณฑ์จับคู่เปรียบเทียบคะแนนหลายแบบทดสอบย่อย มีความเหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบชนิดพหุมิติ

ทองอยู่ สาระ (2543) ได้ศึกษาเปรียบเทียบอำนาจการตรวจสอบ และการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป ระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีถดถอยโลจิสติก โดยใช้ความยาวของแบบทดสอบและขนาดกลุ่มตัวอย่างแตกต่างกัน กลุ่มตัวอย่างที่ใช้ในการศึกษาครั้งนี้เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2542 ในโรงเรียนสังกัดกรมสามัญศึกษา ส่วนกลาง กลุ่มที่ 3 จำนวน 3,242 คน เครื่องมือที่ใช้เป็นแบบทดสอบวัดความสามารถทางสมองที่ผู้วิจัยสร้างขึ้นตามแนวโครงสร้างของโอติส-เลนนอน ซึ่งเป็นแบบทดสอบเลือกตอบห้าตัวเลือก จำนวน 80 ข้อ วัดความสามารถทั่วไปสามด้าน คือ ความเข้าใจด้านภาษา เหตุผลด้านภาษา และเหตุผลด้านภาพ ในการศึกษาครั้งนี้ผู้วิจัยได้สุ่มแบบทดสอบความยาวสามขนาด คือ 20, 40 และ 60 ข้อ และสุ่มกลุ่มตัวอย่างขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบจำนวน 1 กลุ่ม คือ 300: 100, 300: 200, 300: 300, 500:100, 500: 200, 500: 300, 1,000: 100, 1,000: 200, 1,000: 300 และ 1,000: 1,000 รวมเงื่อนไขที่ศึกษาทั้งหมด 30 เงื่อนไข (3 ขนาดความยาวแบบทดสอบ × 10 ขนาดกลุ่มตัวอย่าง)

ผลการศึกษาพบว่า

1. อำนาจการตรวจสอบ และการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันทั้งแบบเอกรูปและแบบอเนกรูป ระหว่างวิธี แมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก ภายใต้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างที่ศึกษาเกือบทุกเงื่อนไข มีค่าไม่แตกต่างกัน

2. ความยาวของแบบทดสอบไม่มีผลต่ออำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก ทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป

3. ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบในการตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก เกือบทุกเงื่อนไขของการศึกษาทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป และแบบอเนกรูป กล่าวคือเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการตรวจสอบจะมีค่าเพิ่มขึ้น แต่พบว่าขนาดกลุ่มตัวอย่างไม่มีผลต่อการจำแนกผิดพลาด ในเกือบทุกเงื่อนไขที่ศึกษา ทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป

อารี วัชรโสตถิกุล (2543) ได้ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้รูปแบบต่างกัน คือ รูปแบบคะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ด้วยวิธีการตรวจสอบต่างกัน คือ วิธีซิปเทสท์และวิธีถดถอยโลจิสติก แล้วทำการคัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ เพื่อเปรียบเทียบค่าความเที่ยง กลุ่มตัวอย่าง

เป็นนักเรียนระดับชั้นมัธยมศึกษาปีที่ 2 สังกัดกรมสามัญศึกษา กรุงเทพมหานคร ส่วนกลาง กลุ่ม 4 จำนวน 994 คน จำแนกเป็นนักเรียนชาย 480 คน และนักเรียนหญิง 514 คน เครื่องมือที่ใช้เป็นแบบทดสอบวัดผลสัมฤทธิ์วิชาคณิตศาสตร์ ชนิดเลือกตอบ 5 ตัวเลือก จำนวน 5 ข้อ ผลการศึกษาพบว่า

1. จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีซิปเทสท์และวิธีถดถอยโลจิสติกแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

2. จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีซิปเทสท์และวิธีถดถอยโลจิสติกแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในรูปแบบรวมทั้งฉบับ และแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติในรูปแบบ แยกตามเนื้อหา และรูปแบบแยกตามระดับพฤติกรรม

3. ความเที่ยงของแบบทดสอบหลังคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้รูปแบบการตรวจสอบต่างกัน แล้วทำการปรับขยายค่าความเที่ยงของแบบทดสอบให้มีจำนวนข้อเท่ากันด้วยสูตรสเปียร์แมน-บราวน์แล้ว วิธีซิปเทสท์ พบว่า ค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทดสอบเป็นรายคู่ต่อ พบว่าความเที่ยง แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ระหว่างรูปแบบแยกตามเนื้อหา และแยกตามระดับพฤติกรรม เมื่อตรวจสอบ โดยวิธีถดถอยโลจิสติก พบว่า ค่าความเที่ยงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

4. ความเที่ยงของแบบทดสอบหลังคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้รูปแบบการตรวจสอบต่างกัน แล้วทำการปรับขยายค่าความเที่ยงของแบบทดสอบให้มีจำนวนข้อเท่ากันด้วยสูตรสเปียร์แมน-บราวน์แล้ว เมื่อตรวจสอบโดยใช้รูปแบบคะแนนรวมทั้งฉบับ พบว่า ค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เมื่อตรวจสอบโดยใช้รูปแบบแยกตามเนื้อหาพบว่า ค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเมื่อตรวจสอบโดยใช้รูปแบบแยกตามระดับพฤติกรรม พบว่า ค่าความเที่ยงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์และซิปเทสท์ปรับใหม่

วิธีซิปเทสท์ (Simulatancous Item Bias Test; SIBTEST) พัฒนาโดยเชียลีและสเตาท์ (Shealy & Stout, 1993) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) การทำหน้าที่ต่างกันของแบบทดสอบ (DTF) และการทำหน้าที่ต่างกันของกลุ่มผู้สอบ (Differential Bundle Functioning; DBF) วิธีนี้สามารถวิเคราะห์ได้ทั้งในแบบทดสอบเอกมิติ (Unidimensional Test) และแบบทดสอบพหุมิติ (Multidimensional Test) (Stout, Li, & Nandakumar, 1997 cited in Shealy & Stout, 1993) วิธีซิปเทสท์ใช้สถิติทดสอบแบบนันพาราเมตริก (Nonparmetric) ซึ่งพัฒนามาบนพื้นฐานของทฤษฎี IRT ชนิดพหุมิติแต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบ หรือ

การประมาณค่าความสามารถแฝง วิธีซิปเทสท์ถูกออกแบบมาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว (Unidirectional DIF) ดังนั้นจึงไม่มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่มีทิศทาง (Nonunidirectional DIF) (Li & Stout, 1996 cited in Shealy & Stout, 1993) จุดเด่นของวิธีซิปเทสท์ คือ สามารถคำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่ายและไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ ทั้งยังใช้สถิติทดสอบนัยสำคัญ (Narayanan & Swaminathan, 1996) นอกจากนี้ยังสามารถนำไปประยุกต์ใช้กับกาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (Polytomous DIF) (Chang, Mazzeo, & Roussos, 1995; Narayanan & Swaminathan, 1996)

ในการศึกษาการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ของแบบทดสอบเอกมิติถือว่าข้อสอบในแบบทดสอบจะต้องมุ่งวัดคุณลักษณะหรือความสามารถแฝงเพียงลักษณะเดียว ความสามารถแฝงสามารถจำแนกเป็น ความสามารถเป้าหมายที่ต้องการวัด (Target Ability; θ) กับความสามารถแทรกซ้อนที่ไม่ใช่เป้าหมายของการวัด (Nuisance Ability; η) ตัวอย่างเช่น แบบทดสอบคำศัพท์ในวิชาภาษาอังกฤษ ข้อสอบบางข้ออาจถามความรู้สำหรับผู้ชายเป็นพิเศษ เช่น ความรู้เกี่ยวกับกีฬา ในขณะที่ข้อสอบบางข้ออาจถามความรู้เกี่ยวกับผู้หญิงโดยเฉพาะ เช่น ความรู้เกี่ยวกับงานในบ้าน จากสถานการณ์ดังกล่าวทักษะความรู้เกี่ยวกับคำศัพท์ในวิชาภาษาอังกฤษเป็นความสามารถเป้าหมายที่ต้องการวัด (θ) ส่วนความรู้ทางด้านกีฬาและงานในบ้านเป็นความสามารถแทรกซ้อน ที่ไม่ใช่เป้าหมายของการวัด (η_1 และ η_2) ข้อสอบทุกข้อในแบบทดสอบจะวัดความสามารถเป้าหมาย ส่วนข้อสอบบางข้อทำหน้าที่ต่างกันจะวัดความสามารถเป้าหมายและความ สามารถแทรกซ้อน (Nandakumar, 1993)

ถ้าให้ฟังก์ชันการตอบสนองข้อสอบ (IRF) ข้อที่ i ซึ่งขึ้นอยู่กับความสามารถ θ เพียงอย่างเดียวแทนด้วย $P_i(\theta)$ ส่วน IRF ข้อที่ i ที่ขึ้นอยู่กับความสามารถทั้ง θ และ η แทนด้วย $P_i(\theta,\eta)$ ฟังก์ชันการตอบสนองของข้อสอบดังกล่าวแบบ 3 พารามิเตอร์จะเป็น ดังนี้ (Shealy & Stout, 1993)

$$P_i(\theta) = C_i + \frac{(1-C_i)}{1+\exp\left[-1.7a_{i\theta}(\theta-b_{i\theta})\right]} \quad , i = 1, \ldots, N$$

$$P_i(\theta,\eta) = C_i + \frac{(1-C_i)}{1+\exp\left\{-1.7\left[a_{i\theta}(\theta-b_{i\theta})+a_{i\eta}(\eta-b_{i\eta})\right]\right\}} \quad , i = 1, \ldots, N$$

ดังนั้นฟังก์ชันความน่าจะเป็นอย่างมีเงื่อนไขของแบบแผนการตอบข้อสอบทั้งฉบับเป็นดังนี้

$$P\left[U/\left(\Theta=\theta,\eta=\eta\right)\right]=\prod_{i=1}^{N}P_i\left(\theta,\eta\right)^{U_i}\left[1-P_i\left(\theta,\eta\right)\right]^{1-U_i}$$

เชียลีและสเตาท์ (Shealy & Stout, 1993) ได้ใช้ Marginal IRFs อธิบายการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

$$M_{ig}(\theta)=\int_n P_i(\theta,\eta)f_g\left(\eta/\theta\right)d\eta$$

เมื่อ $M_{ig}(\theta)$ คือ Marginal IRF สำหรับความสามารถเป้าหมายที่ต้องการวัด (θ) ของผู้สอบกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบ

$P_i(\theta,n)$ คือ IRF ของข้อสอบข้อที่ i

$f_g\left(\eta/\theta\right)$ คือ การแจกแจงแบบมีเงื่อนไขของกลุ่มผู้สอบ

การเปรียบเทียบ Marginal IRF ระหว่างกลุ่มอ้างอิง (R) กับกลุ่มเปรียบเทียบ (F) จะทำให้ทราบถึงทิศทางของการได้เปรียบหรือเสียเปรียบ กล่าวคือ ถ้า $M_{iF}(\theta)<M_{iR}(\theta)$ ทุกค่าของ (θ) แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มอ้างอิง และถ้า $M_{iF}(\theta)>M_{iR}(\theta)$ ทุกค่าของ (θ) แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบ การทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียวอาจเรียกอีกอย่างหนึ่งว่า "การทำหน้าที่ต่างกันแบบไม่ตัดกัน" (Noncrossing DIF)

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามวิธีซิบเทสท์ จะแบ่งข้อสอบออกเป็นสองชุดย่อย (Subtests) คือ (1) ชุดแบบทดสอบที่มีความตรง (Valid Subtests) หรือชุดแบบทดสอบที่ใช้ในการจับคู่เปรียบเทียบ (Matching Subtest) แบบทดสอบชุดนี้ประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน (2) ชุดแบบทดสอบที่ต้องการศึกษา (Studied Subtests) ประกอบด้วยข้อสอบที่สงสัยว่าทำหน้าที่ต่างกัน ถ้าแบบทดสอบชุดแรกมีจำนวน n ข้อ (ข้อที่ 1 ถึง n) แล้วแบบทดสอบชุดที่สองจะมีจำนวน $N-n$ ข้อ (ข้อที่ $n+1$ ถึง N) เมื่อ N เป็นจำนวนข้อสอบทั้งหมด

ฟังก์ชันการตอบสนองข้อสอบของแบบทดสอบที่ต้องการศึกษา จากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ กำหนดในรูปฟังก์ชัน Marginal ดังนี้

$$M_{SR}(\theta) = \sum_{i=n+1}^{N} M_{iR}(\theta)$$

$$M_{SF}(\theta) = \sum_{i=n+1}^{N} M_{iF}(\theta)$$

เมื่อ $M_{SR}(\theta)$ คือ ผลรวมของ Marginal IRFs ของข้อสอบที่ต้องการศึกษาจากผู้สอบกลุ่มอ้างอิง ณ ระดับความสามารถ θ

$M_{SF}(\theta)$ คือ ผลรวมของ Marginal IRFs ของข้อสอบที่ต้องการศึกษาจากผู้สอบกลุ่มเปรียบเทียบ ณ ระดับความสามารถ θ

ขนาดของความแตกต่างระหว่าง $M_{SR}(\theta)$ และ $M_{SF}(\theta)$ แสดงปริมาณความเข้มของการทดสอบ การทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว หรือการทำหน้าที่ต่างกันแบบไม่ตัดกันจากชุดแบบทดสอบที่ต้องการศึกษา ณ ระดับความสามารถ θ ซึ่งสามารถคำนวณตลอดช่วงความสามารถ θ ของผู้สอบ ด้วยการอินทิเกรท ดังนี้

$$\beta_{uni} = \int_{\theta} [M_{SR}(\theta) - M_{SF}(\theta)] f_p(\theta) d\theta$$

เมื่อ β_{uni} คือ ดัชนีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว

$f_p(\theta)$ คือ ฟังก์ชันความหนาแน่นของโอกาสการแจกแจงความสามารถ θ ของผู้สอบทั้งสองกลุ่ม

ดัชนี β_{uni} ที่คำนวณได้จากสูตรดังกล่าว นำมาทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว ดังนี้

$H_0 : \beta_{uni} = 0$

$H_1 : \beta_{uni} > 0$

สมมติฐานอื่น (H_1) มีลักษณะทิศทางเดียว ซึ่งใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบที่เข้าข้างผู้สอบกลุ่มอ้างอิง สำหรับค่าประมาณของ β_{uni} คำนวณได้จากคะแนนรวมของแบบทดสอบชุดที่มีความตรง และแบบทดสอบชุดที่ต้องการศึกษา ซึ่งกำหนดด้วยสัญลักษณ์ ดังนี้

$$X = \sum_{i=1}^{n} U_i$$

$$Y = \sum_{i=n+1}^{N} U_i$$

เมื่อ X คือ คะแนนรวมของแบบทดสอบชุดที่มีความตรง

Y คือ คะแนนรวมของแบบทดสอบชุดที่ต้องการศึกษา

U_i คือ ผลการตอบข้อสอบข้อที่ i (ตอบถูกได้ 1 คะแนน และตอบผิดได้ 0 คะแนน)

นำคะแนนรวมของแบบทดสอบชุดที่มีความตรง (X) เป็นเกณฑ์ในการจับคู่ผู้ที่มีความสามารถระดับเดียวกัน แล้วคำนวณคะแนนเฉลี่ยรายข้อจากผลการตอบข้อสอบของแบบทดสอบชุดที่ต้องการศึกษาระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่มีความสามารถระดับเดียวกันมาจับคู่เปรียบเทียบกัน ณ $X = k$ โดยเขียนในรูปสัญลักษณ์ได้ ดังนี้

$$\bar{Y}_{Rk} - \bar{Y}_{Fk} \quad ; k = 0, 1, 2 \ldots, n$$

เมื่อ $\bar{Y}_{Rk}$ คือ ค่าเฉลี่ยคะแนนรายข้อ จากแบบทดสอบชุดที่ต้องการศึกษาของผู้สอบกลุ่มอ้างอิง ซึ่งได้คะแนน $X = k$

$\bar{Y}_{Fk}$ คือ ค่าเฉลี่ยคะแนนรายข้อ จากแบบทดสอบชุดที่ต้องการศึกษาของผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้คะแนน $X = k$

k คือ คะแนนรวมจากแบบทดสอบชุดที่มีความตรง

ค่า $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ ดังกล่าวเป็นความแตกต่างของผลการตอบข้อสอบในแบบทดสอบชุดที่ต้องการศึกษา ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน ถ้า $\bar{Y}_{Rk} - \bar{Y}_{Fk} = 0$ ทุกคะแนน k แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน แต่ถ้า $\bar{Y}_{Rk} - \bar{Y}_{Fk} > 0$ ทุกคะแนน k แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยจะเข้าข้างผู้สอบกลุ่มอ้างอิง ค่าความแตกต่างของผลการตอบข้อสอบสามารถประมาณค่าในรูป β_{uni} ได้ดังนี้

$$\hat{\beta}_{uni} = \sum_{k=0}^{n} \hat{P}_k \left(\bar{Y}_{Rk} - \bar{Y}_{Fk} \right)$$

โดยที่ $$\hat{P}_k = \frac{\left(J_{Rk} + J_{Fk} \right)}{\sum_{k=0}^{n} \left(J_{Rk} + J_{Fk} \right)}$$

เมื่อ $\hat{P}_k$ คือ สัดส่วนของจำนวนผู้สอบกลุ่มเปรียบเทียบและกลุ่มอ้างอิงที่ได้คะแนนรวม $X = k$ จากจำนวนผู้สอบทั้งหมด

J_{Fk} คือ จำนวนผู้สอบกลุ่มเปรียบเทียบที่ได้คะแนนรวม $X = k$

J_{Rk} คือ จำนวนผู้สอบกลุ่มอ้างอิงที่ได้คะแนนรวม $X = k$

สำหรับการทดสอบสมมติฐานศูนย์ของ No DIF ใช้สถิติ β_{uni} ดังนี้

$$\beta_{uni} = \frac{\hat{\beta}_{uni}}{\hat{\sigma}\left(\hat{\beta}_{uni}\right)}$$

โดยที่

$$\hat{\sigma}\left(\hat{\beta}_{uni}\right) = \sqrt{\sum_{k=0}^{n} \hat{P}_k^2 \left[\frac{1}{J_{Rk}} \hat{\sigma}^2(Y|k,R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y|k,F)\right]}$$

$\hat{\sigma}\left(\hat{\beta}_{uni}\right)$ คือ ค่าประมาณความคลาดเคลื่อนมาตรฐานของ β_{uni}

$\hat{\sigma}^2(Y|k,g)$ คือ ค่าประมาณความแปรปรวนของคะแนนจากแบบทดสอบชุดที่ต้องการศึกษาสำหรับผู้สอบกลุ่ม g (R หรือ F) ซึ่งมีคะแนนเท่ากับ k

J_{gk} คือ จำนวนผู้สอบกลุ่ม g (R หรือ F) ซึ่งตอบแบบทดสอบชุดที่มีความตรงแล้ว ได้คะแนนรวม $X = k$

β_{uni} มีการแจกแจงในลักษณะปกติมาตรฐาน *[N(0,1)]* จึงสามารถทดสอบนัยสำคัญด้วยสถิติทดสอบ Z ถ้าผลการทดสอบปรากฏว่า $\beta_{uni} > Z_a$ แสดงว่า การทดสอบมีนัยสำคัญ จึงปฏิเสธ H_0 นั่นคือ ข้อสอบที่นำมาตรวจสอบทำหน้าที่ต่างกัน โดยจะเข้าข้างผู้สอบกลุ่มอ้างอิงเมื่อ β_{uni} มีค่าเป็นบวก และจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบเมื่อ β_{uni} มีค่าเป็นลบ

สถิติที่ใช้ในการทดสอบสมมติฐานสำหรับสรุปอ้างอิง การทำหน้าที่ต่างกันของข้อสอบดังกล่าวมักจะมีปัญหาในกรณีที่มีความแตกต่างของการแจกแจงความสามารถเป้าหมาย ระหว่างกลุ่มผู้สอบ กล่าวคือ ถ้าผู้สอบกลุ่มอ้างอิงมีความสามารถเป้าหมายสูงกว่าผู้สอบกลุ่มเปรียบเทียบจะเกิดผลกระทบทำให้สถิติ β_{uni} มีค่าเฟ้อ (Inflate) หรือมีค่าสูงผิดปกติ ถึงแม้ว่าในความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน จึงควรแก้ไขความแตกต่างของการแจกแจงความสามารถเป้าหมาย

ด้วยการปรับแก้ค่าการถดถอย เพื่อขจัดผลกระทบดังกล่าว โดยปรับแก้ค่า $\bar{Y}_{Rk}$ และ Y_{Fk} เป็นรายคู่ ก่อนการคำนวณ β_{uni}

วิธีซิปเทสท์ปรับใหม่ ปรับปรุงขั้นตอนการวิเคราะห์ข้อมูล เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในการวิเคราะห์จะนำจำนวนผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมาแบ่งกลุ่มผู้สอบออกเป็น 2 กลุ่มตามระดับความสามารถ คือ กลุ่มผู้สอบที่มีความสามารถสูงและกลุ่มผู้สอบที่มีความสามารถต่ำ โดยใช้ค่าเฉลี่ยของคะแนนรวมเป็นเกณฑ์ในการแบ่งกลุ่มผู้สอบ แล้ววิเคราะห์ด้วยดัชนีการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ในกลุ่มผู้สอบทั้งสอง โดยวิเคราะห์แยกกันคนละกลุ่ม ซึ่งมีสูตรการคำนวณและขั้นตอนเหมือนกับวิธีซิปเทสท์

งานวิจัยที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์และวิธีซิปเทสท์ปรับใหม่

แอ็คเคอร์แมน (Ackerman, 1992) ได้ศึกษาเกี่ยวกับการทำหน้าที่เบี่ยงเบนของข้อสอบและความตรงของข้อสอบ โดยใช้กรอบความคิด MIRT ศึกษาโดยการจำลองข้อมูล 2 ชุด ชุดละ 1,000 คน ข้อมูลทั้ง 2 ชุด มีการแจกแจงความสามารถจำนวน 2 มิติแตกต่างกัน (Different Underlying Two-Dimensional Ability Distribution) ประมาณค่าพารามิเตอร์ จากการปรับเทียบข้อสอบของ ACT (Form 26 A of the ACT Assessment Programs Math Usage Test) ตามที่เร็คเคส (Reckase) เสนอไว้ในปี ค.ศ. 1985 ภายหลังการปรับเทียบข้อสอบ โดยใช้แบบจำลอง M2PL เลือกกลุ่มข้อสอบที่มีความเที่ยงตรง (Valid Sector) และกลุ่มข้อสอบที่สงสัยว่าจะมี DIF (Suspected of Being Bias) เพื่อนำไปจำลองข้อมูลเพื่อศึกษาในกรณีต่าง ๆ ดังนี้

กรณีที่ 1: Equal θ, η Distributions: No DIF, No Impact

กรณีที่ 2: Unequal θ Means: Uniform DIF

กรณีที่ 3: Unequal η Means: Uniform DIF

กรณีที่ 4: Unequal η Variance: Nonuniform DIF

กรณีที่ 5: Unequal $\rho_{\theta,\eta}$: Nonuniform DIF

ใช้วิธีการตรวจสอบ DIF 2 วิธี คือ แมนเทล-แฮนส์เซล และซิปเทสท์ ผลจากการศึกษาพบว่าวิธีแมนเทล-แฮนส์เซลและซิปเทสท์ สามารถตรวจสอบ DIF ในกรณีต่าง ๆ ได้ดี แต่วิธีแมนเทล-แฮนส์เซล ต้องแปลความหมายด้วยความระมัดระวัง ผู้วิจัยให้ข้อเสนอแนะว่าผู้สร้างแบบทดสอบและผู้ใช้แบบทดสอบควรกำหนดทักษะที่ต้องการวัดให้ชัดเจน และกำจัดข้อสอบที่วัดทักษะที่ไม่ต้องการออกไป พร้อมทั้งกล่าวว่าข้อสอบตั้งแต่ 2 ข้อขึ้นไปทำให้เกิดความเป็นหลายมิติ (Multidimensionality) และควรประมาณค่าพารามิเตอร์โดยใช้แบบจำลองหลายมิติ จึงจะทำให้การแปลความหมายผลที่ได้มีความถูกต้อง

เชียร์ลี่และสเตาท์ (Shealy & Stout, 1993) ได้เสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ โดยใช้วิธีซิปเทสท์ ซึ่งเป็นวิธีที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ เป็นวิธีการที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งกรณีที่แบบทดสอบมีความลำเอียงข้อเดียวและกรณีที่มีความลำเอียงหลายข้อ ได้ทำการศึกษาประสิทธิภาพในการตรวจสอบโดยการเปรียบเทียบกับวิธีแมนเทล-เฮนส์เซล ในกรณีที่มีข้อสอบลำเอียงข้อเดียว โดยใช้ข้อมูลผลการตอบจากแบบสอบวิชาคณิตศาสตร์ของ ACT และแบบทดสอบของ ASVAB ขนาดกลุ่มตัวอย่างของกลุ่มอ้างอิง และกลุ่มเปรียบเทียบต่างกัน คือ (3000, 3000), (3000, 1000), (1500, 1500), (1000, 500), (500, 500) และ (500, 250)

ผลการศึกษาพบว่า ในกรณีที่มีข้อสอบลำเอียงข้อเดียวทั้งวิธีแมนเทล-เฮนส์เซลและวิธีซิปเทสท์มีอำนาจในการตรวจสอบดีเท่าเทียมกันได้และวิธีซิปเทสท์มีประสิทธิภาพดีในการตรวจสอบกรณีที่มีข้อสอบที่ลำเอียงหลาย ๆ ข้อ แม้ว่ามีขนาดความลำเอียงค่อนข้างน้อย และผลการศึกษาพบว่าทั้งวิธีวิธีซิปเทสท์และแมนเทล-เฮนส์เซลจะมีประสิทธิภาพดีเมื่อใช้กับแบบทดสอบที่มีความยาวพอสมควร ($\geq$ 25 ข้อ)

นารายานาน และ สวามินาธาน (Narayanan & Swaminathan, 1994) ได้ศึกษาเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ระหว่างวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ โดยใช้ข้อมูลจำลองภายใต้การจัดกระทำ 5 ปัจจัย คือ 1) ขนาดกลุ่มตัวอย่าง 9 ระดับ (กลุ่มอ้างอิงจำนวน 300 คน 500 คนและ 1,000 คน กลุ่มเปรียบเทียบจำนวน 100 ค น 200 คน และ 300 คน) 2) ความแตกต่างของการแจกแจงค่าความสามารถ 2 ระดับ (แบบเท่ากันและไม่เท่ากัน) 3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 2 ระดับ (10% และ 20%) 4) ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน 4 ระดับ (พื้นที่ระหว่าง IRFs มีค่า .4, .6, .8 และ 1.0) และ 5) ลักษณะของข้อสอบ 6 ระดับ (ลักษณะ b ต่ำกับ a ปานกลาง, b ต่ำ a สูง, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง, b สูงกับ a ต่ำ, b สูงกับ a ปานกลาง) ข้อสอบที่จำลองมีความยาว 40 ข้อ ดังนั้นในการศึกษาครั้งนี้จะต้องจำลองข้อมูลทั้งหมด 1,296 เงื่อนไข ($9 \times 3 \times 2 \times 4 \times 6$)

ผลการศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ขนาดกลุ่มตัวอย่าง สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันและลักษณะของข้อสอบเป็นปัจจัยที่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ ถ้าการแจกแจงค่าความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ มีค่าเท่ากันแล้ววิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์จะมีประสิทธิภาพเท่ากัน แต่ถ้าการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าไม่เท่ากันแล้ววิธีซิปเทสท์จะมีประสิทธิภาพสูง

กว่าวิธีแมนเทล-แฮนส์เซล นอกจากนี้ยังพบว่า เมื่อขนาดกลุ่มอ้างอิง หรือกลุ่มเปรียบเทียบมีจำนวน 300 คน วิธีทั้งสองก็มีอำนาจการทดสอบเพียงพอในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า ขนาดของกลุ่มตัวอย่างและลักษณะของข้อสอบไม่มีผลกระทบต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ ถ้าการแจกแจงค่าความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเท่ากันแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์มีค่าสูงกว่าวิธีแมนเทล-แฮนส์เซลเล็กน้อย และถ้าความแตกต่างของการแจกแจงค่าความสามารถ มีค่าเพิ่มขึ้นจะมีผลทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 2 วิธีมีค่าเพิ่มขึ้นด้วย

รุสโซ และสเตาต์ (Roussos & Stout, 1996) ได้ศึกษาในสถานการณ์จำลองของผลกระทบของกลุ่มตัวอย่างขนาดเล็ก และค่าพารามิเตอร์ของข้อสอบที่มีต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล โดยจำลองข้อมูล 2 ครั้ง ครั้งที่ 1 เพื่อศึกษากลุ่มตัวอย่างขนาดเล็ก โดยใช้ขนาดกลุ่มตัวอย่าง 4 ระดับ (100, 200, 500 และ 1,000 คน) และความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ 3 ระดับ (0, 0.5, และ 1.0) ใช้แบบทดสอบจำนวน 25 ข้อ ซึ่งนำมาจากชุดแบบทดสอบ Armed Services Vocational Aptitude Battery (ASVAB) ประมาณค่าพารามิเตอร์ใช้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ สำหรับการจำลอง ครั้งที่ 2 เพื่อศึกษาค่าพารามิเตอร์ของข้อสอบ โดยใช้ขนาดกลุ่มตัวอย่าง 3 ระดับ (500 คน, 1,000 คน และ 3,000 คน) และความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ 2 ระดับ (0 และ 1.0) ส่วนแบบทดสอบใช้จากการศึกษาครั้งที่ 1 โดยเลือกค่าพารามิเตอร์อำนาจจำแนก 3 ระดับ (0.4, 1.0, 2.5) พารามิเตอร์ความยาก 5 ระดับ (-1.5, -0.5, 0, 0.5 และ 1.5) พารามิเตอร์การเดา 1 ระดับ (0.20) เมื่อ $d_T = 0.0$ และพารามิเตอร์การเดา 3 ระดับ (0.20, 0.10 และ 0.05) เมื่อ $d_T = 1.0$ สำหรับขนาดกลุ่มตัวอย่างระหว่างกลุ่มเปรียบเทียบและกลุ่มอ้างอิงจะใช้จำนวนเท่ากันทั้งสองครั้ง

ผลการศึกษา ครั้งที่ 1 เมื่อศึกษากลุ่มตัวอย่างขนาดเล็ก พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซลมีค่าไม่แตกต่างกัน สำหรับผลการศึกษาครั้งที่ 2 เมื่อศึกษาค่าพารามิเตอร์ของข้อสอบ พบว่า เมื่อความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบไม่แตกต่างกัน ($d_T = 1.0$) แล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์จะมีค่าต่ำกว่าวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขเกือบทุกเงื่อนไขของการตรวจสอบ

ดักกลาส รุสโซและสเตาต์ (Douglas, Roussos, & Stout, 1996 cited in Roussos & Stout, 1996) ได้ตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาข้อสอบพร้อม ๆ กัน

ครั้งละหลาย ๆ ข้อ โดยเสนอวิธีการตรวจสอบหมวดข้อสอบที่คาดว่าจะแสดงการทำหน้าที่ต่างกัน เมื่อข้อสอบเหล่านั้น ถูกนำมาพิจารณาพร้อมกัน 2 วิธี คือ ใช้เฉพาะความคิดเห็นของผู้เชี่ยวชาญ และใช้วิธีการทางสถิติและตรวจสอบซ้ำด้วยความเห็นของผู้เชี่ยวชาญ พร้อมทั้งแสดงตัวอย่างการตรวจสอบการทำหน้าที่ต่างกัน 3 ตัวอย่าง ดังนี้ ตัวอย่างที่ 1 เลือกหมวดข้อสอบโดยใช้ความเห็นของผู้เชี่ยวชาญอย่างเดียว ให้ผู้เชี่ยวชาญชาย 3 คนและหญิง 1 คน พิจารณาแบบทดสอบย่อยการใช้เหตุผลเชิงตรรกศาสตร์ (Logical Reasoning Subtest) ที่ได้จากการดำเนินการสอบในเดือนธันวาคม ค.ศ. 1991 จำนวน 49 ข้อ ในการตรวจสอบให้เพศชายเป็นกลุ่มอ้างอิงและเพศหญิงเป็นกลุ่มเปรียบเทียบ ผู้เชี่ยวชาญจัดหมวดข้อสอบเป็น 8 หมวด และพิจารณาว่าหมวดใดน่าจะให้ประโยชน์แก่เพศชายหรือเพศหญิงหลังจากนั้นวิเคราะห์ด้วยโปรแกรมคอมพิวเตอร์สำเร็จรูป SIBTEST ใช้กลุ่มตัวอย่างเพศชาย 3,000 คน เพศหญิง 3,000 คน พบว่าความคิดเห็นของผู้เชี่ยวชาญและผลการวิเคราะห์ด้วยวิธีซิปเทสท์ สอดคล้องกัน จึงทำการวิเคราะห์ปริมาณ DIF ของแต่ละข้อ พบว่าแต่ละข้อแสดง DIF ด้วยปริมาณที่น้อยมาก จึงไม่มีการคัดเลือกข้อสอบออกจากแบบทดสอบ ตัวอย่างที่ 2 ใช้วิธีการทางสถิติในการตรวจสอบหมวดข้อสอบ คือวิธี HCA (Agglomerative Hireachical Cluster Analysis) (Jain & Kubes, 1988 cited in Roussos & Stout, 1996) และ DIMTEST (Nandakumar, 1993; Shealy & Stout, 1993) แล้วจึงตรวจสอบด้วยวิธีซิปเทสท์ในการตรวจสอบใช้ข้อสอบ National Assessment of Educational Progress (NAEP) จำนวน 36 ข้อ ใช้กลุ่มตัวอย่างกลุ่มละ 50 คน ผลการตรวจสอบพบว่า การตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบประสบความสำเร็จเป็นอย่างดี ตัวอย่างที่ 3 ใช้แบบทดสอบความเข้าใจการอ่าน ที่สอบในเดือนธันวาคม ค.ศ. 1991 ซึ่งเป็นส่วนหนึ่งของการทดสอบ LSAT มีข้อสอบ 28 ข้อ แบ่งตามบทความที่กำหนดให้อ่านเป็น 4 บทความ แต่ละบทความมีข้อสอบ 5-8 ข้อ แล้วทำการตรวจสอบหมวดข้อสอบซ้ำโดยใช้ HCA และ DIMTEST หลังจากนั้นจึงวิเคราะห์ด้วย SIBTEST ใช้กลุ่มตัวอย่างกลุ่มละ 1,000 คน พบว่าทั้ง 4 บทความแสดงการทำหน้าที่ต่างกันของหมวดข้อสอบ ข้อสอบ 2 หมวดให้ประโยชน์แก่เพศชายและ 2 หมวดให้ประโยชน์แก่เพศหญิง

ชาง, มัสสิโอ, และรอสอส (Chang, Mazzeo, & Roussos, 1995) ได้ทำการศึกษาเกี่ยวกับการค้นหาความลำเอียงของแบบทดสอบที่ตรวจให้คะแนนแบบหลายค่า (Polytomously) ใช้วิธี SIBTEST โดยใช้ข้อมูลจำลอง มีวัตถุประสงค์เพื่อเปรียบเทียบวิธี Modified SIBTEST กับวิธี Mantel (MH) และวิธี Standradized Mean Difference (SMD) ในการศึกษาครั้งนี้ใน 2 ครั้ง คือ

การศึกษาครั้งที่ 1 ทำการศึกษาเปรียบเทียบวิธีทั้ง 2 วิธีการ ภายใต้เงื่อนไขของการศึกษาแบบ MH และแบบ SMD จะมีการค้นหาความลำเอียงที่ดีหรือไม่ ผลลัพธ์ของการศึกษาพบว่าวิธี Modified SIBTEST ใช้ในทางปฏิบัติที่เหมาะสมกว่าวิธีอื่น แต่วิธี MH และวิธี SMD ใช้ได้ดี

กับข้อมูลหยาบ ๆ

การศึกษาครั้งที่ 2 ใช้ข้อมูลภายใต้เงื่อนไขวิธีการทดสอบหาความลำเอียงของคะแนนที่สังเกตได้ ของแบบทดสอบที่มีการตรวจให้คะแนนแบบ 0, 1 (Dichotomous) ผลลัพธ์ของการศึกษาครั้งที่ 2 พบว่า ภายใต้เงื่อนไขทั้งหมด วิธี Modified SIBTEST ใช้ได้ดีภายใต้ความคลาดเคลื่อนประเภทที่ 1 มากกว่ากระบวนการอื่น ๆ หรือวิธีการอื่น ๆ

ชาง และคณะ (Chang et al., 1995) ได้ทำการศึกษาผลของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบหลายค่า โดยประยุกต์วิธีซิปเทสท์มาใช้เปรียบเทียบกับวิธีแมนเทล-แฮนส์เซลและวิธี SMD แบ่งการศึกษาออกเป็น 2 ตอน ตอนที่ 1 ใช้ข้อมูลจำลองภายใต้เงื่อนไขเดียวกับงานวิจัยของ ซวิค, ไทยเยอร์ และมาสสิโอ (Zwick, Thyer, & Mazzeo, 1993 cited in Chang et al., 1995) เพื่อทำการเปรียบเทียบวิธีซิปเทสท์แบบประยุกต์กับวิธีแมนเทล-แฮนส์เซลและวิธี SMD ผลที่ได้พบว่าวิธีซิปเทสท์มีประสิทธิภาพในการตรวจสอบค่อนข้างดีแต่วิธีแมนเทล-แฮนส์เซลและวิธี SMD มีประสิทธิภาพค่อนข้างดีกว่า ในการศึกษา ตอนที่ 2 ใช้ข้อมูลจำลอง คือ ข้อสอบที่ศึกษามีอำนาจจำแนกแตกต่างกัน 11 ค่า ตั้งแต่ 15 ถึง 2.00 ขนาดกลุ่มตัวอย่างต่างกัน คือ 500 1,000 และ 2,000 คน ข้อสอบมีความยาว 24 ข้อ สำหรับวิธีซิปเทสท์ และ 25 ข้อ สำหรับวิธีแมนเทล-แฮนส์เซลและวิธี SMD ผลการศึกษาสรุปว่า ทั้งวิธีแมนเทล-แฮนส์เซลและวิธี SMD มีค่าความคลาดเคลื่อนประเภทที่ 1 ค่อนข้างสูงเมื่อพารามิเตอร์ของค่าอำนาจจำแนกของข้อสอบที่ศึกษามีค่าต่างจากค่าเฉลี่ยของอำนาจจำแนกของแบบทดสอบที่มีความตรงและอัตราการปฏิเสธ (Refection Rates) ของทั้ง 3 วิธีจะมีค่าสูงขึ้นเมื่อค่าอำนาจจำแนกสูงขึ้น

ซวิค, ไทยเยอร์ และมาสสิโอ (Zwick, Thyer, & Mazzeo, 1997 cited in Chang et al., 1995) ได้ทำการศึกษาเกี่ยวกับการค้นหาความลำเอียงจากแบบทดสอบที่มีการตรวจให้คะแนนแบบหลายค่า โดยใช้วิธีวิเคราะห์ 5 วิธีด้วยกัน คือ SMD จำนวน 2 วิธี ได้แก่ SMD-H, SMD-M วิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ 2 วิธี ได้แก่ Standard SIBTEST และวิธี Modified SIBTEST โดยใช้เงื่อนไขแบบทดสอบที่มีการตรวจให้คะแนนแบบ 0, 1 จำนวน 50 ข้อ และแบบทดสอบที่มีการตรวจให้คะแนนแบบหลายค่า จำนวน 18 ข้อ ใช้กับกลุ่มตัวอย่าง 1,000 คน แบ่งเป็นกลุ่มอ้างอิง 500 คนและกลุ่มเปรียบเทียบ 500 คน โดยมีการจับคู่ระหว่างข้อสอบแบบให้คะแนน 2 ค่า กับแบบให้คะแนนหลายค่า โดยใช้โมเดลโลจิสติก 3 พารามิเตอร์ จากข้อมูลจำนวน 75 ข้อ ที่มีค่าอำนาจจำแนกในช่วงช่วง .74 ถึง 1.0 ค่า ความยากอยู่ระหว่าง-1.95 ถึง 1.95 ค่าการเดาอยู่ที่ .15 และสถานการณ์ที่จำลองขึ้นมากับกลุ่มตัวอย่างทั้ง 2 กลุ่มที่มีการแจกแจงดัชนีการปฏิบัติที่ดี ดูจาก SMD เมื่อความแตกต่างของค่าเฉลี่ยของกลุ่มที่มีความคลาดเคลื่อนมาตรฐานเป็น 1 วิธีการ Modified SIBTEST เป็นการวัดผลกระทบของขนาดกลุ่มตัวอย่างค่อนข้างดี จาก 5 วิธีการ ในทาง

ปฏิบัติจะไม่สามารถมองเห็นความแตกต่างของกลุ่มย่อยทั้ง 2 กลุ่ม จะมีการแจกแจงเบ้ไปด้านใดด้านหนึ่ง เมื่อกลุ่มมีการแจกแจงที่แตกต่างกันและข้อสอบที่ศึกษามีค่าอำนาจจำแนกสูง วิธีการ SIBTEST จะแสดงให้เห็นดีกว่าวิธีอื่นในรูปแบบความคลาดเคลื่อนประเภทที่ 1 กว่าวิธี SMD และวิธีแมนเทล-แฮนส์เซล จากแบบทดสอบตอบสั้น ๆ ค่าอำนาจจำแนกจาก 5 วิธี จะไม่เหมือนกันเป็นความแตกต่างในการวิเคราะห์ความลำเอียงและองค์ประกอบอื่น ๆ วิธีการที่ใช้กันบ่อย ๆ ในการประยุกต์ใช้หาความลำเอียงจากแบบทดสอบแบบ Polytomous เมื่อมีการสอบตัวแปรที่ใช้ คือการจับคู่กันระหว่างการสอบแบบฝึกหัด อัตราความคลาดเคลื่อนประเภทที่ 1 จากวิธีแมนเทล-แฮนส์เซลและวิธี SMD จะเกี่ยวกับเงื่อนไขของเกณฑ์ วิธีการในปัจจุบันของ SIBTEST ไม่สามารถใช้กับการจับคู่ได้ นั่นคือ ใช้กับแบบทดสอบแบบถูก-ผิดไม่ได้ งานวิจัยในปัจจุบันจึงไม่นิยมใช้กัน

ไรอัน, แคเธอรีนและชิว (Ryan, Katherine, & Shuwan, 2001 cited in Chang et al., 1995) ได้ศึกษาเพื่อตรวจสอบการทำหน้าที่เบี่ยงเบนของหมวดข้อสอบเมื่อมีการเปลี่ยนแปลงตำแหน่งของข้อสอบใช้แบบทดสอบวิชาคณิตศาสตร์ระดับอุดมศึกษาชั้นปีที่ 1 จำนวน 40 ข้อ จัดทำเป็น 2 ฟอร์ม ฟอร์มที่ 1 เรียงลำดับข้อสอบตามเนื้อหาจากง่ายไปยาก ฟอร์มที่ 2 เรียงลำดับข้อสอบแบบสุ่ม กลุ่มตัวอย่างที่ตอบข้อสอบฟอร์ม 1 เป็นเพศชายจำนวน 546 คน เพศหญิง 520 คน กลุ่มตัวอย่างที่ตอบข้อสอบ ฟอร์ม 2 เป็นเพศชายจำนวน 554 คน เพศหญิงจำนวน 511 คน วิเคราะห์โดยใช้โปรแกรมคอมพิวเตอร์สำเร็จรูป SIBTEST พบว่าการเปลี่ยนแปลงตำแหน่งข้อสอบในแบบทดสอบไม่มีผลต่อการตรวจสอบการทำหน้าที่เบี่ยงเบนของหมวดข้อสอบ

กาญจนา วัธนสุนทร (2537) ได้พัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศด้วยข้อมูลเชิงประจักษ์สำหรับดัชนี 4 ตัว คือ พื้นที่ระหว่างโค้งการตอบข้อสอบชนิดคิดเครื่องหมาย (SA) และไม่คิดเครื่องหมาย (UA) จากวิธีทฤษฎีการตอบข้อสอบโมเดล 2 พารามิเตอร์ ดัชนีแอลฟา (α_{MH}) จากวิธีแมนเทล-แฮนส์เซลและเบต้า (β_{SIB}) จากวิธี SIBTEST โดยใช้ข้อมูลการตอบข้อสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัย ปีการศึกษา 2535 ในความยาวแบบทดสอบ 20, 30 และ 40 ข้อ สำหรับวิชาคณิตศาสตร์ และ 50, 60, 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ ใช้กลุ่มผู้สอบ 6 ขนาด คือ 100, 200, 400, 600, 800 และ 1,000 คน

การพัฒนาเกณฑ์กระทำโดยคำนวณค่าดัชนีทั้ง 4 ตัว จากข้อมูลการตอบข้อสอบของผู้สอบเพศเดียวกัน เพศละ 50 ค่า สำหรับแต่ละความยาวแบบทดสอบและขนาดผู้สอบ จากนั้นนำค่าดัชนีที่ได้ทั้งหมดมาวิเคราะห์ค่าเฉลี่ยและกำหนดเกณฑ์จากค่าเฉลี่ย 2 ลักษณะ คือ เกณฑ์ที่กำหนดจากค่าเฉลี่ย ซึ่งรวมค่าดัชนีทุกข้อโดยไม่พิจารณาความแตกต่างในด้านความยาวแบบทดสอบและขนาดของผู้สอบ และเกณฑ์ที่กำหนดจากค่าเฉลี่ยที่พิจารณาถึงความยาวของ

แบบทดสอบและขนาดผู้สอบด้วย จากนั้นนำเกณฑ์ที่กำหนดไว้ไปตัดสินค่าดัชนีที่ได้จากการวิเคราะห์ระหว่างผู้สอบเพศหญิงและชาย พบว่าความสอดคล้องของการตัดสินภายในดัชนีเดียวกันมีความไม่คงที่ข้ามขนาดผู้สอบ อย่างไรก็ตามความสอดคล้องมีแนวโน้มสูงขึ้น ที่ขนาดผู้สอบตั้งแต่ 600 คนขึ้นไป ข้อค้นพบที่สำคัญจากการศึกษามีดังต่อไปนี้

1. เกณฑ์ที่พัฒนาจากข้อมูลเชิงประจักษ์เพื่อใช้ในการตัดสินความลำเอียงของข้อสอบระหว่างผู้สอบหญิงและชาย คือ

1.1 $|SA| > .80$ และ $UA > .05$ กรณีความยาวแบบทดสอบต่ำกว่า 50 ข้อ

1.2 $|SA| > .40$ และ $UA > 1.20$ กรณีความยาวแบบทดสอบ 50 ข้อ ขึ้นไป

1.3 $.60 > \alpha_{MH} > 1.40$ และ $|\beta_{SIB}| > .60$ ทุกความยาวของแบบทดสอบ

และทุกขนาดผู้สอบทั้งนี้ควรใช้ขนาดผู้สอบอย่างน้อย 800 คน สำหรับดัชนี SA และ UA และ 600 คน สำหรับดัชนี α_{MH} และ β_{SIB}

2. การตรวจค้นข้อสอบลำเอียงทางเพศมีความไม่คงที่ข้ามขนาดผู้สอบ และความยาวของแบบทดสอบ

3. ความสอดคล้องในการตรวจค้นข้อสอบลำเอียงภายในวิธีเดียวกันข้ามขนาดผู้สอบค่อนข้างต่ำ แต่จะสูงขึ้นที่ขนาดผู้สอบตั้งแต่ 600 คน ขึ้นไป

4. ข้อสอบลำเอียงวิชาคณิตศาสตร์ส่วนใหญ่ลำเอียงเข้าข้างผู้สอบชายและวิชาภาษาอังกฤษลำเอียงเข้าข้างผู้สอบหญิงเมื่อใช้ดัชนี SA และ α_{MH} แต่ดัชนี β_{SIB} ให้ผลตรงข้าม

เรวดี อินทสะระ (2539) ได้ศึกษาความตรงเชิงพยากรณ์ของแบบทดสอบคัดเลือกที่วิเคราะห์ความลำเอียงต่อเพศด้วยวิธีใช้ทฤษฎีการตอบคำถาม (IRT) วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ การตัดสินผลการสอบที่คิดคะแนนมาตรฐานทีปกติ และคะแนนน้ำหนักความสามารถ และสาเหตุความลำเอียงของข้อสอบ โดยศึกษาความลำเอียงของข้อสอบคัดเลือกเข้าศึกษาในชั้นปีที่ 1 ประเภทรับตรง ปีการศึกษา 2538 ของมหาวิทยาลัยสงขลานครินทร์ในวิชาภาษาไทย ก วิชาสังคมศึกษา ก วิชาภาษาอังกฤษ กข วิชาละ 8,127 (ชาย 2,722 คน หญิง 5,405 คน) วิชาภาษาไทย กข วิชาสังคมศึกษา กข และวิชาภาษาอังกฤษ กขค วิชาละ 5,415 คน (ชาย 1,454 คน หญิง 3,961 คน) ความตรงเชิงพยากรณ์ศึกษาจากคะแนนสอบคัดเลือกกับเกรดภาคเรียนที่ 1 ปีการศึกษา 2538 ของนักศึกษาที่ได้รับการคัดเลือกจากประเภทรับตรง สายวิทยาศาสตร์ 763 คน และสายศิลปศาสตร์ 281 คน และสาเหตุความลำเอียงของข้อสอบจากการระบุสาเหตุของนักวัดผลการศึกษาหรืออาจารย์ผู้สอน จำนวน 50 คน และนักศึกษาที่เรียนในสาขาวิชานั้น ๆ วิชาละ 50 คน

ผลการศึกษาพบว่า วิธีการตรวจสอบความลำเอียงทั้ง 3 วิธี ตัดสินจำนวนข้อสอบ ที่ลำเอียงแตกต่างกันในวิชาภาษาไทย ก ฉบับที่ 2 และวิชาสังคมศึกษา ก ฉบับที่ 1 ที่ระดับนัยสำคัญ

ทางสถิติที่ .05 นอกนั้นต่างกันที่ระดับนัยสำคัญทางสถิติที่ .01 โดยที่วิธีใช้ทฤษฎีตอบสนองข้อสอบ ตัดสินจำนวนข้อสอบที่ลำเอียงมากที่สุด ความสัมพันธ์ของลำดับที่ของการสอบไม่ว่าจะคิดคะแนนมาตรฐานทีปกติ หรือคิดคะแนนน้ำหนักความสามารถและใช้ข้อสอบจำนวนทั้งหมดหรือใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงต่างมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

ประสิทธิภาพในการทำนายผลสัมฤทธิ์ทางการเรียนในสายวิทยาศาสตร์ การคิดคะแนนน้ำหนักความสามารถทั้งใช้ข้อสอบจำนวนทั้งหมดและใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงมีประสิทธิภาพในการทำนายสูงกว่าการคิดคะแนนมาตรฐานทีปกติ ส่วนสายศิลปศาสตร์ การคิดคะแนนมาตรฐานทีปกติ ทั้งที่ใช้ข้อสอบจำนวนทั้งหมดและใช้เฉพาะข้อสอบที่ปราศจากความลำเอียง มีประสิทธิภาพในการทำนายสูงกว่าการคิดคะแนนน้ำหนักความสามารถ และสาเหตุของความลำเอียงของข้อสอบต่อเพศทั้งชายและหญิงเกิดจากข้อสอบเป็นข้อคำถามที่ผู้สอบได้รับการฝึกฝนเฉพาะจะมีโอกาสตอบถูกมากกว่าเป็นเรื่องราวที่กลุ่มนั้น ๆ สนใจและเป็นข้อสอบที่ถามความจำ

จิตติมา วรรณศรี (2539) ได้ศึกษาเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีซิปเทสท์ โดยใช้ข้อมูลจำลองจากโปรแกรม IRTDATA เงื่อนไขที่ศึกษาได้แก่ (1) ความยาวของแบบทดสอบ 3 ขนาด คือ 30, 60 และ 90 ข้อ (2) ขนาดกลุ่มตัวอย่าง 3 ขนาด คือ 200, 600 และ 1,000 คน โดยแต่ละขนาดมีอัตราส่วนระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ 4 อัตราส่วน คือ 1: 1, 1: 0.9, 1: 0.75 และ 1: 0.5 รวมเงื่อนไขที่ศึกษาทั้งหมด 36 เงื่อนไข

ผลการวิจัยพบว่า

1. วิธีแมนเทล-แฮนส์เซลกับวิธีซิปเทสท์มีประสิทธิภาพเท่าเทียมกัน ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ทุกขนาดกลุ่มตัวอย่างและทุกอัตราส่วน ภายใต้ความยาวแบบทดสอบเดียวกัน เมื่อขนาดกลุ่มตัวอย่าง 200 และ 600 คน สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 50 และเมื่อขนาดกลุ่มตัวอย่าง 1,000 คน สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 100 โดยส่วนมากวิธีซิปเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากกว่าวิธีแมนเทล-แฮนส์เซลเล็กน้อย

2. วิธีแมนเทล-แฮนส์เซลกับวิธีซิปเทสท์ มีประสิทธิภาพเท่าเทียมกัน ที่ทุกระดับความยาวของแบบทดสอบ โดยพบว่าเมื่อใช้แบบทดสอบที่มีความยาวปานกลาง (60 ข้อ) มีประสิทธิภาพในการตรวจสอบดีที่สุด

พรรณี จินตมาศ (2540) ได้ศึกษาผลการวิเคราะห์ความลำเอียงต่อเพศของข้อสอบจากแบบทดสอบคณิตศาสตร์โจทย์ปัญหา โดยจำแนกในแต่ละวิธีวิเคราะห์ความลำเอียง 3 วิธี คือ วิธีแปลงค่าความยาก วิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ และในแต่ละขนาดกลุ่มผู้สอบ คือ ขนาดกลุ่มผู้สอบ 500 คน และขนาดกลุ่มผู้สอบ 1,000 คน โดยเปรียบเทียบจำนวนข้อสอบที่มีความลำเอียงและเปรียบเทียบค่าความเที่ยงแบบแบ่งครึ่งฉบับของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว ในการศึกษาครั้งนี้มีกลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ภาคเรียนที่ 2 ปีการศึกษา 2539 ของโรงเรียนสังกัดกรมสามัญศึกษา ส่วนกลาง จำนวน 2,200 คน ซึ่งเลือกมาโดยการสุ่มแบบแบ่งชั้น มีขนาดโรงเรียนเป็นชั้นและโรงเรียนเป็นหน่วยการสุ่ม

ผลการวิจัยพบว่า เมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 500 คน วิธีซิปเทสท์พบข้อสอบที่มีความลำเอียงมากที่สุด และวิธีแปลงค่าความยากพบข้อสอบที่มีความลำเอียงน้อยที่สุด โดยจำนวนข้อสอบที่ลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติทุกวิธีวิเคราะห์และเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 1,000 คน วิธีแมนเทล-แฮนส์เซลพบข้อสอบที่มีความลำเอียงมากที่สุด วิธีแปลงค่าความยากไม่พบข้อสอบที่ลำเอียง โดยจำนวนข้อสอบที่ลำเอียงจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากกับวิธีแมนเทล-แฮนส์เซลและวิธีแปลงค่าความยากกับวิธีซิปเทสท์แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนับสำคัญทางสถิติ สำหรับจำนวนข้อสอบที่ลำเอียงจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากระหว่างกลุ่มผู้สอบขนาด 500 คน และกลุ่มผู้สอบขนาด 1,000 คน จะมีจำนวนข้อสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

ค่าความเที่ยงของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 500 คน จากการวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่าความยากกับวิธีซิปเทสท์มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติและเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 1,000 คน ค่าความเที่ยงของแบบทดสอบหลังคัดเลือกข้อสอบที่ลำเอียงออกแล้วจากการวิเคราะห์ความลำเอียง 3 วิธี มีค่าแตกต่างกัน อย่างไม่มีนัยสำคัญทางสถิติ สำหรับค่าความเที่ยงของแบบทดสอบหลังคัดเลือกข้อสอบที่ลำเอียงออกจากการวิเคราะห์ด้วยวิธีซิปเทสท์ระหว่างกลุ่มผู้สอบ 500 คนและกลุ่มผู้สอบ 1,000 คน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

วลีมาศ แซ่อึ้ง (2543) ได้ศึกษาอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก ข้อมูลที่ใช้ในการศึกษาจำลองภายใต้

โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่าการเดา (c) คงที่ แล้วจัดกระทำข้อมูลตามปัจจัย 4 ตัว คือ (1) ลักษณะของข้อสอบที่มีค่าความยาก (b) และอำนาจจำแนก (a) ระดับต่ำ ปานกลาง และสูง จำนวน 9 ลักษณะ (2) ความยาวของแบบทดสอบ 2 ระดับ (3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบทดสอบ 3 ระดับและ (4) ขนาดกลุ่มตัวอย่าง 6 ระดับ รวมข้อมูลที่ศึกษาทั้งหมดจำนวน 324 เงื่อนไข แล้วนำข้อมูลของแต่ละเงื่อนไขมาคำนวณค่าอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ผลการวิจัยพบว่า

1. อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปของวิธีซิปเทสท์ปรับใหม่และวิธีการถดถอยโลจิสติกมีค่าเท่าเทียมกันภายใต้เกือบทุกเงื่อนไข และทั้งสองวิธีดังกล่าวมีอำนาจการทดสอบสูงกว่าวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล ภายใต้เกือบทุกเงื่อนไข

2. อัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปของวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติกมีค่าอยู่ภายในเกณฑ์ของอัตราความคลาดเคลื่อนประเภทที่ 1 ที่ระดับ 10% ภายใต้เกือบทุกเงื่อนไข

วิภา จำมั่น (2544) ได้ศึกษาเปรียบเทียบผลการตรวจสอบข้อสอบที่ลำเอียงของแบบทดสอบวัดความสามารถด้านภาษา เมื่อตรวจสอบด้วยวิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ในกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลางและกลุ่มข้อสอบที่มีความง่ายสูง โดยเปรียบเทียบข้อสอบที่มีความลำเอียงและค่าความเที่ยงของแบบทดสอบหลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 1 ปีการศึกษา 2544 ของโรงเรียนสังกัดกรมสามัญศึกษา จังหวัดลพบุรี จำนวน 1,041 คน ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น

ผลการวิจัยพบว่า

1. ข้อสอบที่มีความลำเอียง เมื่อตรวจสอบด้วยวิธีซิปเทสท์ ระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างมีนัยสำคัญทางสถิติระดับ .05 ส่วนข้อสอบที่มีความลำเอียง ระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำกับกลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และระหว่างกลุ่มข้อสอบที่มีระดับความง่ายปานกลางกับกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ เมื่อตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล พบว่าข้อสอบที่มีความลำเอียง

ระหว่างกลุ่มผู้สอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างไม่มีนัยสำคัญ

2. จำนวนข้อสอบที่มีความลำเอียง ระหว่างการตรวจสอบด้วยวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล ในกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

3. ค่าความเที่ยงของแบบทดสอบหลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว ระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

สิริรัตน์ วิภาสศิลป์ (2545) ได้ศึกษาเปรียบเทียบวิธีซิปเทสท์และวิธีดีเอฟไอที ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบ หมวดข้อสอบ และแบบทดสอบ จากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ ในเงื่อนไขความยาวของแบบทดสอบ 30, 40 และ 50 ข้อ กลุ่มตัวอย่างขนาด 50, 100, 200, 500 และ 1,000 คน กลุ่มตัวอย่างในการศึกษาได้มาจากการสุ่มแบบใส่คืนจากประชากรเทียมซึ่งกำหนดจากนักเรียนชายและนักเรียนหญิงชั้นมัธยมศึกษาปีที่ 1 ในจังหวัดนนทบุรี แต่ละขนาดสุ่มกลุ่มตัวอย่าง 50 ครั้ง เครื่องมือที่ใช้ในการวิจัยเป็นแบบทดสอบวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างขึ้นซึ่งเป็นแบบปรนัยเลือกตอบ 5 ตัวเลือก จำนวน 50 ข้อ มีข้อสอบที่ผู้เชี่ยวชาญพิจารณาว่าเป็นข้อสอบที่แสดงการทำหน้าที่เบี่ยงเบนต่อเพศชาย จำนวน 16 ข้อ หลังจากเก็บรวบรวมข้อมูลแล้วคัดเลือกข้อสอบตามสัดส่วนในตารางกำหนดข้อสอบ จัดเป็นแบบทดสอบที่มีความยาว 40 และ 30 ข้อ แล้วตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบ หมวดข้อสอบ และแบบทดสอบในเงื่อนไขต่าง ๆ ด้วยโปรแกรมคอมพิวเตอร์สำเร็จรูป SIBTEST และ DFIT จากนั้นนำผลที่ได้ไปเปรียบเทียบความถูกต้องและการระบุผิดพลาดในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบด้วยวิธีเดียวกันและต่างวิธีโดยการวิเคราะห์ความแปรปรวนแบบตัวแปรพหุ คำนวณความสอดคล้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบด้วยวิธีทั้งสอง แล้วเปรียบเทียบความถูกต้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนของหมวดข้อสอบและแบบทดสอบด้วยวิธีซิปเทสท์และวิธีดีเอฟไอทีโดยใช้สถิติ *Z-test*

ผลการวิจัยสรุปได้ดังนี้

1. เมื่อแบบทดสอบประกอบด้วยข้อสอบ 30, 40 และ 50 ข้อ กลุ่มตัวอย่างขนาด 50, 100 และ 200 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนข้อสอบด้วยวิธีซิปเทสท์ไม่แตกต่างกัน กลุ่มตัวอย่างขนาด 500 และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบ

การทำหน้าที่เบี่ยงเบนของข้อสอบด้วยวิธีซิปเทสท์สูงกว่ากลุ่มตัวอย่างขนาด 50, 100 และ 200 คน แต่การระบุผิดพลาดในการตรวจสอบสูงกว่าด้วย เมื่อตรวจสอบด้วยวิธีดีเอฟไอที พบว่า กลุ่มตัวอย่างขนาด 50, 100, 200, 500 และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบไม่แตกต่างกัน

2. ทุกเงื่อนไขความยาวแบบทดสอบและกลุ่มตัวอย่างขนาดแตกต่างกัน วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่เบี่ยงของข้อสอบน้อยกว่าวิธีดีเอฟไอทีและพบว่าความสอดคล้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบด้วยวิธีทั้งสองมีค่าต่ำกว่าร้อยละ 1

3. วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนของหมวดข้อสอบมากกว่าวิธีดีเอฟไอที เมื่อแบบทดสอบมีข้อสอบ 30 ข้อ กลุ่มตัวอย่าง 1,000 คน และเมื่อแบบทดสอบมี 40 ข้อ กลุ่มตัวอย่างขนาด 500 คน

4. วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนของแบบทดสอบมากกว่าวิธีดีเอฟไอที เมื่อแบบทดสอบมีข้อสอบ 50 ข้อ กลุ่มตัวอย่างขนาด 100, 200 และ 1,000 คน

แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน (Achievement Test)

1. แนวคิดของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน (เยาวดี วิบูลย์ศรี, 2545, หน้า 14-27)

1.1 ธรรมชาติของแบบทดสอบวัดผลสัมฤทธิ์ (Achievement Test)

แบบทดสอบวัดผลสัมฤทธิ์ (Achievement Test) เป็นเครื่องมือสำหรับช่วยให้ครูสามารถตัดสินผลสัมฤทธิ์ของนักเรียนได้อย่างมีประสิทธิภาพ เพราะเป็นวิธีการประเมินพฤติกรรมของนักเรียนที่มีความเป็นอิสระได้มากกว่าวิธีอื่น ๆ แบบทดสอบวัดผลสัมฤทธิ์ที่ใช้ในโรงเรียนมุ่งวัดความรู้ในแต่ละวิชาและทักษะต่าง ๆ โดยมีวัตถุประสงค์พื้นฐานที่สำคัญ 2 ประการ คือ ประการแรก เพื่อเป็นเครื่องมือในการวัดผลสัมฤทธิ์ทางการเรียน อันเป็นข้อมูลที่ได้รับสำหรับการประเมินผลการเรียนการสอนเป็นรายบุคคล ประการที่สอง เพื่อเป็นการตรวจสอบความสามารถของนักเรียนแต่ละคน ซึ่งแตกต่างกันโดยธรรมชาติ เมื่อความสามารถของแต่ละคนมีความแตกต่างกันทั้งเด็กและผู้ใหญ่ ดังนั้นการที่จะให้บุคคลต่าง ๆ ได้รับการพัฒนาความสามารถเฉพาะตนที่มีอยู่อย่างเหมาะสม จึงก่อให้เกิดการพัฒนาแบบทดสอบวัดผลสัมฤทธิ์ที่มีประสิทธิภาพขึ้น เพื่อใช้เป็นเครื่องมือในการวัดระดับความสามารถของบุคคล หรือเพื่อจำแนกความสามารถของบุคคลที่แตกต่างกัน ทำให้เราสามารถศึกษาให้สอดคล้องกับระดับความสามารถของบุคคลต่าง ๆ ได้ โดยทั่วไปแล้วการที่โรงเรียนได้นำแบบทดสอบวัดผลสัมฤทธิ์มาใช้ในการวัดและประเมินผล

ก็เพื่อที่จะค้นหาระดับการเรียนรู้และทักษะต่าง ๆ ของนักเรียนที่เกิดขึ้นหลังจากการเรียนการสอน หรือจากประสบการณ์ทางอ้อมของกระบวนการเรียนการสอน รวมทั้งจากความรู้และทักษะอื่น ๆ ซึ่งอาจจะเกิดขึ้นในอนาคตที่สามารถนำไปใช้ได้ด้วย

1.2 ประเภทของแบบทดสอบวัดผลสัมฤทธิ์ เราสามารถจำแนกตามมิติต่าง ๆ ได้หลายมิติ ดังนี้

1.2.1 จำแนกตามขอบข่ายของเนื้อหาวิชาที่วัด ขอบข่ายเนื้อหาวิชาของแบบทดสอบ วัดผลสัมฤทธิ์นั้น อาจกำหนดให้กว้างได้ เช่น กำหนดเนื้อหาวิชาเกี่ยวกับประวัติศาสตร์ไทย โดยทั่วไป หรือจำกัดให้แคบลง เช่น กำหนดเนื้อหาวิชาเฉพาะที่เกี่ยวกับศึกเก้าทัพของประวัติ ศาสตร์ไทย เป็นต้น ตามปกติแล้ว ยังไม่มีมาตรฐานอ้างอิงสากลที่จะนำไปใช้ในการกำหนดเนื้อหาวิชาสำหรับแบบทดสอบวัดผลสัมฤทธิ์ ผู้ใช้แบบทดสอบเท่านั้นที่จะต้องกำหนดเนื้อหาวิชาขึ้นเอง โดยให้สอดคล้องกับวัตถุประสงค์ของการสอบ ผู้สร้างแบบทดสอบสามารถที่จะพัฒนาแบบทดสอบให้มีเนื้อหาได้ตามขอบข่ายที่ต้องการ สำหรับแบบทดสอบมาตรฐานวัดผลสัมฤทธิ์ซึ่งจัดพิมพ์ในต่างประเทศจำนวนมากมักจะรวมแบบทดสอบฉบับต่าง ๆ ไว้เป็นชุด (Batteries) แต่ละชุดจะครอบคลุมเนื้อหาวิชาสำหรับระดับชั้นเรียนที่ต่าง ๆ กัน ทั้งนี้ก็เพื่อให้โรงเรียนทั้งหลายสามารถวัดผลสัมฤทธิ์ของนักเรียนได้อย่างต่อเนื่องตามความเจริญงอกงามทางวิชาการในแต่ละยุคแต่ละสมัย ตัวอย่างเช่น แบบทดสอบวัดผลสัมฤทธิ์ทั่วไป (General Achievement Test Batteries) หรือเรียกย่อ ๆ ว่า GATB ซึ่งเป็นแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานที่นิยมใช้กันในระดับประถมศึกษาและมัธยมศึกษาเป็นส่วนใหญ่ ทั้งนี้เพราะเนื้อหาที่เหมาะสมกับชั้นดังกล่าวเป็นอย่างดี เช่น ในระดับประถมศึกษานั้นลักษณะของแบบทดสอบจะประกอบด้วยเนื้อหาซึ่งสัมพันธ์กับทักษะสามประการ (3R) คือ การอ่าน (Reading) การเขียน (Writing) และการคำนวณ (Arithmetic) ส่วนในระดับมัธยมศึกษาแบบทดสอบ GATB มีความมุ่งหมายเพื่อวัดผลสัมฤทธิ์โดยทั่วไป รวมทั้งเพื่อวัดพื้นฐานความรู้ของแต่ละวิชา เช่น พื้นฐานความรู้ทางวิชาสังคมศาสตร์ พื้นฐานความรู้ของวิชาวิทยาศาสตร์ธรรมชาติ เป็นต้น นอกจากนั้นก็จะเน้นการวัดความสามารถด้านการอ่าน ความเข้าใจในสาขาวิชา รวมทั้งความคิดเกี่ยวกับจำนวนและคำศัพท์ทั่วไป ซึ่งเป็นการวัดในด้านการพัฒนาความสามารถเชิงวิชาการ ดังนั้นแบบทดสอบ GATB ในระดับนี้จึงมักจะใช้ในการสอบคัดเลือกบุคคลเข้าศึกษาต่อในสถาบันอุดมศึกษาหรือในการสอบคัดเลือกเข้าแข่งขันชิงทุนประเภทต่าง ๆ เป็นต้น

1.2.2 จำแนกตามลักษณะหน้าที่ทั่วไปของแบบทดสอบ ซึ่งแบ่งออกเป็น 3 ลักษณะ ดังนี้

1.2.1.1 แบบทดสอบเพื่อการสำรวจผลสัมฤทธิ์ (Survey Tests) เป็นแบบทดสอบวัดผลสัมฤทธิ์ที่ทำหน้าที่ในการสำรวจความสามารถทั่ว ๆ ไปของนักเรียน โดยประเมินความรู้ในเนื้อหาวิชาหรือทักษะต่าง ๆ เพื่อแสดงระดับความสามารถของนักเรียน ดังนั้นแบบทดสอบวัดผลสัมฤทธิ์จึงมักจะครอบคลุมเนื้อหาทั้งในระดับกว้างและระดับทั่วไป โดยถือคะแนนรวมที่ได้จากแบบทดสอบเป็นตัวบ่งชี้ถึงระดับความสามารถที่วัดได้

1.2.1.2 แบบทดสอบเพื่อวินิจฉัยผลสัมฤทธิ์ (Diagnostic Tests) เป็นแบบทดสอบวัดผลสัมฤทธิ์ที่ทำหน้าที่วินิจฉัยเกี่ยวกับจุดเด่นและจุดด้อยขององค์ประกอบสำคัญทางด้านทักษะต่าง ๆ ของนักเรียน ตัวอย่างเช่น แบบทดสอบเพื่อวินิจฉัยผลสัมฤทธิ์ด้านภาษาอาจจะรวมแบบทดสอบย่อยหลายชุด ซึ่งครอบคลุมเนื้อหาเกี่ยวกับการรู้จักใช้คำ ความเข้าใจเกี่ยวกับถ้อยคำต่าง ๆ รวมทั้งคำศัพท์ ตลอดจนอัตราการอ่าน ความเข้าใจในเรื่องที่อ่าน การจำแนกเสียงและการจำแนกพยางค์ ฯลฯ คะแนนที่ได้จากแต่ละองค์ประกอบของแบบทดสอบวินิจฉัยดังกล่าว จะช่วยให้นักจิตวิทยาหรือครูสามารถจะตัดสินได้ว่า อะไรคือจุดบกพร่องของผู้สอบ ซึ่งจะช่วยให้สามารถสอนเสริมในส่วนของเนื้อหาวิชาหรือทักษะที่ยังขาดอยู่ได้อย่างมีประสิทธิภาพ

1.2.1.3 แบบทดสอบเพื่อวัดความพร้อม (Readiness Tests) เป็นแบบทดสอบวัดผลสัมฤทธิ์ซึ่งทำหน้าที่ในการวัดทักษะที่จำเป็นสำหรับการเรียนในชั้นที่สูงขึ้น แบบทดสอบเพื่อวัดความพร้อมใช้สำหรับทำนายการกระทำในอนาคต จึงทำหน้าที่เป็นเครื่องมือในการวัดความถนัดไปในตัวด้วย ตัวอย่างทั่ว ๆ ไปของแบบทดสอบวัดความพร้อม เช่น แบบทดสอบวัดความพร้อมในการอ่าน ซึ่งจะใช้แบบทดสอบเมื่อนักเรียนจบชั้นอนุบาล หรือชั้นเตรียมประถมศึกษาปีที่ 1 เพื่อจะตัดสินว่าเด็กเหล่านั้นได้เรียนรู้ทักษะที่จำเป็นสำหรับการอ่านเพื่อเตรียมพร้อมจะเข้าเรียนต่อไปในชั้นเรียนปกติของการศึกษาในระบบได้อย่างเหมาะสมหรือไม่ เพียงใด

1.2.3 จำแนกตามคำตอบที่ใช้ โดยทั่วไปแล้วแบบทดสอบวัดผลสัมฤทธิ์ส่วนใหญ่ที่ใช้กันมักจะเป็นประเภทข้อเขียน และที่ใช้กันค่อนข้างมาก คือแบบทดสอบภาคปฏิบัติ (Performance Test) ซึ่งเป็นแบบทดสอบที่ต้องการให้นักเรียนหรือผู้เข้าสอบได้สาธิตทักษะของเขาเอง เป็นต้นว่า ให้แสดงทักษะในการแก้ไขเครื่องยนต์กลไกที่ไม่ทำงาน หรือให้แสดงทักษะในการเล่นดนตรี เป็นต้น

สำหรับแบบทดสอบประเภทข้อเขียนนั้น ยังแยกออกได้อย่างกว้าง ๆ อีก 2 ระดับ คือ (1) ระดับการเลือกคำตอบจากที่กำหนดไว้แล้ว (Recognition) และระดับ (2) ระดับของการเขียนคำตอบจากความรู้หรือความทรงจำที่มีอยู่เดิม (Recall) ในแบบทดสอบระดับที่ 1 แต่ละข้อจะมีคำตอบที่ตายตัว และจะประกอบด้วยตัวเลือกหลาย ๆ ตัวที่เป็นไปได้รวมอยู่ใน

คำตอบที่เกี่ยวข้อง ผู้เข้าตอบจะต้องตัดสินใจเลือกคำตอบอย่างรอบคอบและถูกต้องให้สอดคล้องกับชนิดของคำถามที่ระบุไว้ ตัวอย่างของข้อสอบระดับนี้ เช่น แบบทดสอบหลายตัวเลือก (Multiple Choice) แบบทดสอบประเภทถูก-ผิด (True-False) และแบบทดสอบประเภทจับคู่ (Matching)

ส่วนแบบทดสอบระดับที่ 2 ซึ่งต้องใช้ความรู้และความทรงจำที่มีอยู่เดิมมาเขียนตอบนั้นลักษณะของคำตอบอาจจะไม่ตายตัว ขึ้นอยู่กับเหตุผลและความถูกต้องในเชิงวิชาการผสมผสานกับความคิดริเริ่มสร้างสรรค์ของผู้เข้าสอบเป็นสำคัญ แบบทดสอบระดับนี้ ได้แก่ แบบทดสอบประเภทเติมคำหรือข้อความในช่องว่าง (Completion) แบบทดสอบประเภทตอบสั้น (Short Answer) และแบบทดสอบประเภทความเรียง

1.3 แบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน

เป็นแบบทดสอบที่สร้างขึ้นโดยกลุ่มผู้เชี่ยวชาญมากกว่าที่จะสร้างขึ้นโดยบุคคลใดบุคคลหนึ่งเพียงบุคคลเดียวเท่านั้น ตามปกติแล้วผู้สร้างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานมักจะประกอบด้วยผู้เชี่ยวชาญทางด้านการวัดและประเมินผล รวมทั้งผู้เชี่ยวชาญในสาขาวิชานั้น ๆ ตลอดจนครูในโรงเรียนต่าง ๆ ซึ่งมีบทบาทในการกำหนดขอบข่ายเนื้อหาวิชาที่ต้องการทดสอบให้เหมาะสม แบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน ไม่จำเป็นต้องครอบคลุมเนื้อหาและทักษะที่มีในหลักสูตร เนื้อหาและทักษะของแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานส่วนมากมักจะได้จากตำราเรียนและความคิดเห็นของผู้เชี่ยวชาญทางด้านหลักสูตร เนื้อหาโดยทั่วไปจะเป็นความรู้และทักษะในระดับที่กว้าง ๆ เพื่อให้สามารถนำไปใช้กับนักเรียนโรงเรียนต่าง ๆ ได้ สำหรับขั้นตอนในการสร้างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน จะต้องมีการวางแผนสร้างอย่างมีระบบ คือ มีการระบุหลักการและเหตุผลของการสร้างแบบทดสอบ มีการกำหนดวัตถุประสงค์ของการสร้างที่ชัดเจน มีการทดลองใช้แบบทดสอบที่สร้างขึ้น เพื่อตรวจสอบความเป็นมาตรฐาน โดยการวิเคราะห์ระดับความยากง่าย และอำนาจจำแนกของข้อสอบ มีการหาค่าความตรง (Validity) และความเที่ยง (Reliability) ของแบบทดสอบ พร้อมทั้งพัฒนาตารางปกติวิสัย (Norm Table) เพื่อใช้ในการเปรียบเทียบ มีการกำหนดเวลาของการทดสอบและวิธีดำเนินการสอบ ตลอดจนมีคู่มือประกอบการใช้แบบทดสอบซึ่งจะระบุจุดมุ่งหมายของแบบทดสอบ ประสิทธิภาพของแบบทดสอบ รวมทั้งวิธีการใช้แบบทดสอบและวิธีการตรวจหรือวิธีการให้คะแนน พร้อมทั้งตารางปกติวิสัยของกลุ่ม

สำหรับการสร้างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานนั้น มีจุดมุ่งหมายเพื่อนำไปใช้เปรียบเทียบความสามารถของนักเรียนแต่ละคน หรือเปรียบเทียบระหว่างชั้นเรียนต่าง ๆ หรือระหว่างระบบของโรงเรียนต่าง ๆ กับกลุ่มประชากรที่กว้างขึ้น อันถือว่าเป็นกลุ่มปกติวิสัย

ของนักเรียนที่ได้เรียนรู้ในสาขาวิชานั้นมาแล้ว

1.4 ความแตกต่างระหว่างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานกับแบบทดสอบผลสัมฤทธิ์ที่ครูสร้างขึ้นสามารถจำแนกความแตกต่างที่ชัดเจนได้ 5 ประการ ดังนี้

1.4.1 การจำกัดของเนื้อหาวิชาที่สอบ แบบทดสอบวัดผลสัมฤทธิ์มาตรฐานจะสุ่มเนื้อหาสำหรับนำมาสอบในระดับที่กว้างและทั่วไป เพื่อใช้กับโรงเรียนต่าง ๆ ตลอดจนมีการกลั่นกรองเนื้อหาในการสร้างข้อกระทงโดยผู้เชี่ยวชาญทางเนื้อหาและหลักสูตร สำหรับแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้นจะเน้นเนื้อหาเฉพาะที่เกี่ยวกับการเรียนการสอนในชั้นเรียน ครูจะทำหน้าที่เป็นผู้เชี่ยวชาญซึ่งอาจจะประกอบด้วยครูคนเดียวหรือคณะครูก็ได้เป็นผู้กำหนดเนื้อหาที่เหมาะสมในการสอบ

1.4.2 การทดลองใช้แบบทดสอบ แบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน เมื่อสร้างขึ้นแล้วจะต้องมีการทดลองใช้ เพื่อทำการวิเคราะห์ประสิทธิภาพของแบบทดสอบด้วยค่าสถิติต่าง ๆ ต่อจากนั้นก็จะรายงานในคู่มือการใช้แบบทดสอบ เช่น ค่าความตรง ความเที่ยง ระดับความยากง่าย และอำนาจจำแนกของข้อกระทง ในทำนองตรงกันข้าม สำหรับแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้น โดยทั่วไปจะไม่มีการทดลองใช้เพื่อวิเคราะห์ค่าสถิติในการตรวจสอบประสิทธิภาพของแบบทดสอบมาก่อน

1.4.3 วิธีดำเนินการสอบ แบบทดสอบวัดผลสัมฤทธิ์มาตรฐานโดยปกติจะต้องมีคู่มืออธิบายวิธีดำเนินการสอบอย่างเป็นมาตรฐาน เช่น วิธีการตอบ เวลาในการสอบ ฯลฯ ผู้ใช้แบบทดสอบต้องปฏิบัติตามอย่างเคร่งครัด สำหรับแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้นโดยทั่วไปจะไม่มีคู่มือประกอบการใช้ เพราะตัวครูเองจะเป็นผู้กำหนดมาตรฐานในการปฏิบัติเกี่ยวกับวิธีการสอบ

1.4.4 วิธีการให้คะแนน แบบทดสอบวัดผลสัมฤทธิ์มาตรฐานต้องมีคำเฉลยสำหรับการตรวจให้คะแนนตามที่ระบุอยู่ในคู่มือการใช้แบบทดสอบ ส่วนแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้นนั้น ครูจะเป็นผู้ให้คะแนนตามมาตรฐานที่กำหนดขึ้นเอง

1.4.5 ตารางปกติวิสัยเพื่อการเปรียบเทียบ โดยปกติแล้วแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานจะมีการนำไปใช้กับกลุ่มอ้างอิง หรือที่เรียกว่า Norm Group เพื่อทำตารางปกติวิสัย (Norm Table) ไว้ในคู่มือของการใช้แบบทดสอบมาตรฐาน โดยมีจุดมุ่งหมายให้ผู้ใช้แบบทดสอบสามารถนำไปใช้ในการตีความสำหรับคะแนนสอบที่ได้รับ รวมทั้งใช้เป็นตารางเพื่อการเปรียบเทียบของคะแนนดังกล่าวด้วย ส่วนแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้นจะมีเพียงคะแนนของกลุ่มผู้เข้าสอบด้วยกัน ซึ่งอาจใช้เปรียบเทียบได้เฉพาะภายในกลุ่มเท่านั้น

2. แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 (สำนักทดสอบทางการศึกษา, 2546 ก, หน้า 1-9)

กระทรวงศึกษาธิการเห็นความสำคัญและความจำเป็นในการรักษาและยกระดับคุณภาพการศึกษาของสถานศึกษาต่าง ๆ ให้มีมาตรฐานทัดเทียมกัน จึงกำหนดให้มีการประเมินคุณภาพการศึกษาระดับชาติขึ้นในทุก ๆ ปี โดยจะประเมินในปลายปีการศึกษา โดยมอบหมายให้สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ดำเนินการจัดสอบวัดผลสัมฤทธิ์ทางการเรียนของนักเรียนชั้นประถมศึกษาปีที่ 6 ให้ได้ผลการประเมินที่น่าเชื่อถือและสามารถบ่งชี้คุณภาพการศึกษาระดับชาติ ระดับเขตพื้นที่การศึกษา ระดับสถานศึกษา และระดับผู้เรียนเป็นรายบุคคลได้อย่างสมเหตุสมผล

2.1 วัตถุประสงค์ของการสอบวัดผลสัมฤทธิ์ทางการเรียน

2.1.1 เพื่อนำผลไปใช้ในการกำกับ ดูแล ติดตาม และประเมินการจัดการศึกษาในระดับประเทศ สำหรับการพัฒนาคุณภาพการศึกษา

2.1.2 เพื่อนักเรียนได้นำผลจากการสอบวัดไปพัฒนาความรู้ ความสามารถและการศึกษาต่อของนักเรียนในระดับที่สูงขึ้น

2.2 ขอบข่ายการประเมิน

ดำเนินการประเมินนักเรียนทุกคนในสังกัดคณะกรรมการการศึกษาขั้นพื้นฐานและสังกัดอื่นที่จัดการศึกษาในระบบ ตามหลักสูตรประถมศึกษา พุทธศักราช 2521 (ฉบับปรับปรุง พุทธศักราช 2533) ทุกจังหวัดทั่วประเทศ ได้แก่ สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน, สำนักงานคณะกรรมการการศึกษาเอกชน, สำนักงานคณะกรรมการการอุดมศึกษา, สำนักการศึกษากรุงเทพมหานคร, สำนักบริหารการศึกษาท้องถิ่น, กองบังคับการตำรวจตระเวนชายแดน, สำนักงานตำรวจแห่งชาติ, สำนักงานพัฒนาการศึกษาและนันทนาการ

2.3 ลักษณะทั่วไปของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน

เป็นแบบทดสอบที่สร้างและพัฒนาขึ้นจนเป็นแบบทดสอบมาตรฐาน เพื่อใช้สำหรับประเมินผลการเรียนรู้ของนักเรียนที่ได้สะสมมาตลอดการเรียน ตามหลักสูตรประถมศึกษา พ.ศ. 2521 (ฉบับปรับปรุง พ.ศ. 2533) ของกระทรวงศึกษาธิการ

2.4 โครงสร้างของแบบทดสอบ

โครงสร้างของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน ชั้นประถมศึกษาปีที่ 6 มีดังนี้

2.4.1 เนื้อหาสาระของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนชั้นประถมศึกษาปีที่ 6 ประกอบด้วย เนื้อหาตามหลักสูตรประถมศึกษา พ.ศ. 2521 (ฉบับปรับปรุง พ.ศ. 2533) วิชาที่สอบในครั้งนี้ ได้แก่ วิชาภาษาไทย คณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ

2.4.2 แบบทดสอบแต่ละวิชาเน้นวัดด้านความรู้ความเข้าใจ ความคิดวิเคราะห์ และทักษะเชิงกระบวนการเฉพาะวิชา

2.5 การเตรียมการสอบ

2.5.1 การแต่งตั้งคณะกรรมการคุมสอบ

ให้สำนักงานเขตพื้นที่การศึกษาแต่งตั้งคณะกรรมการคุมสอบสลับโรงเรียนภายในเขตพื้นที่การศึกษานั้น ๆ โดยการสอบแต่ละห้องต้องให้มีคณะกรรมการคุมสอบสองคน

2.5.2 การเตรียมตัวนักเรียน

ก่อนการดำเนินการสอบให้ผู้บริหารโรงเรียน และครู-อาจารย์ ชี้แจงทำความเข้าใจกับนักเรียนให้ตระหนักถึงประโยชน์และความสำคัญของการสอบ ตั้งใจตอบแบบทดสอบให้เต็มความสามารถที่แท้จริง ทั้งนี้ข้อมูลที่ได้จากการประเมินจะเกิดประโยชน์ต่อตัวนักเรียนและส่วนรวม ดังนี้

2.5.2.1 ประโยชน์ต่อนักเรียน นักเรียนจะได้รับประโยชน์จากการสอบวัดผลสัมฤทธิ์ทางการเรียนของนักเรียนในระดับชาติ ทำให้เกิดความชำนาญในการคิด ได้เห็นรูปแบบของแบบทดสอบมาตรฐาน เป็นการเพิ่มพูนประสบการณ์ในการสอบ และคะแนนที่ได้จะเป็นข้อมูลประกอบในการพิจารณาเข้าศึกษาต่อของนักเรียน รวมทั้งได้รับทราบความสามารถของตนว่ามีจุดเด่นด้านใด ควรปรับปรุง และพัฒนาด้านใด เพราะในการสอบครั้งนี้เป็นการสอบวัดความสามารถหลายด้าน คือ ภาษาไทย คณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ผลจากการสอบต้องมาบันทึกลงในเอกสารหลักฐานการศึกษา

2.5.2.2 ประโยชน์ต่อสถานศึกษา คณะครู-อาจารย์ นำผลการประเมินไปใช้เป็นข้อมูลในการแนะแนวทางในการศึกษา ให้กับผู้เรียนและพัฒนาการจัดการเรียนการสอนให้ได้คุณภาพมาตรฐานยิ่งขึ้น

2.5.2.3 ประโยชน์ต่อประเทศ รัฐบาลและหน่วยงานที่เกี่ยวข้องกับการจัดการศึกษานำผลการประเมินไปใช้เป็นข้อมูลในการกำหนดนโยบายทางการศึกษา และส่งเสริมสนับสนุนให้สถานศึกษา และหน่วยงานที่เกี่ยวข้องพัฒนาคุณภาพการศึกษาให้ได้มาตรฐานทัดเทียมนานาอารยประเทศ

2.6 แนวปฏิบัติกรณีจำเป็นในการดำเนินการสอบ

กระทรวงศึกษาธิการ ต้องการให้นักเรียนในชั้นที่ประเมินได้รับการประเมินทุกคนพร้อมกันทั่วประเทศ เพื่อความเป็นมาตรฐานในการดำเนินงาน และผลการประเมิน มีความตรงเป็นที่น่าเชื่อถือ แต่ในกรณีที่โรงเรียนไม่สามารถดำเนินการสอบในวันที่กำหนด อาจเนื่องมาจากเหตุจำเป็นต้องปฏิบัติ ถ้าไม่ดำเนินการจำทำให้ราชการเสียหาย หรือเหตุสุดวิสัยอื่น โรงเรียนอาจ

เลื่อนการสอบออกไปได้ โดยเสนอประธานคณะกรรมการดำเนินงานระดับเขตพื้นที่การศึกษาพิจารณา แล้วแจ้งสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานทราบ

การดำเนินการสอบให้เป็นไปตามระเบียบการสอบของกระทรวงศึกษาธิการ อย่างเคร่งครัด กรณีที่นักเรียนขาดสอบให้โรงเรียนพิจารณาถึงสาเหตุและความจำเป็น แล้วจึงพิจารณาให้นักเรียนเข้าสอบชดเชย โดยเก็บรักษาข้อสอบเป็นความลับ และดำเนินการสอบตามระเบียบการสอบของกระทรวงศึกษาธิการ

2.7 การดำเนินการหลังการสอบ

การดำเนินการหลังการสอบแบ่งเป็น 3 ส่วน ดังนี้

2.7.1 เมื่อกรรมการดำเนินการสอบ เก็บและตรวจสอบความเรียบร้อยของกระดาษคำตอบแล้ว ให้นำส่งคณะกรรมการดำเนินการสอบระดับโรงเรียน จากนั้นโรงเรียนรวบรวมกระดาษคำตอบทั้งหมดส่งคณะกรรมการที่สำนักงานเขตพื้นที่การศึกษากำหนดไว้ สำหรับส่วนกลางให้รวบรวมกระดาษคำตอบส่งที่สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานกรุงเทพมหานคร

2.7.2 โรงเรียนรวบรวมแบบทดสอบทั้งหมดส่งคืนคณะกรรมการที่สำนักงานเขตพื้นที่การศึกษาแต่งตั้ง เพื่อดำเนินการ ย่อย/ เผาทำลายต่อไป สำหรับส่วนกลาง นำส่งที่โรงพิมพ์คุรุสภาลาดพร้าว

2.7.3 คณะกรรมการดำเนินงานระดับเขตพื้นที่การศึกษา รวบรวมกระดาษคำตอบทั้งหมดส่งสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการทันที เนื่องจากสำนักทดสอบทางการศึกษาจะต้องตรวจกระดาษคำตอบให้ทั้งหมด ซึ่งมีปริมาณมากให้เสร็จสิ้นโดยเร็ว เพื่อให้โรงเรียนได้ทำบันทึกลงในเอกสารหลักฐานการศึกษาได้ทันก่อนโรงเรียนปิดภาคเรียน