บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันวงการศึกษา ธุรกิจ อุตสาหกรรมและการทหาร ได้ใช้การสอบเพื่อคัดเลือกบุคคล ตรวจสอบความรู้ความสามารถของบุคคล สนับสนุนการเลื่อนตำแหน่ง ออกใบรับรองหรือใบอนุญาตกันอย่างแพร่หลาย และมีแนวโน้มในการนำเอาผลการสอบไปใช้ประกอบการตัดสินใจในเรื่องเหล่านี้เพิ่มมากขึ้น ทำให้ปัญหาเรื่องความไม่ยุติธรรมหรือความลำเอียงของข้อสอบหรือแบบสอบที่ใช้ในการสอบแต่ละครั้ง ได้กลายมาเป็นประเด็นสำคัญประเด็นหนึ่งในการประเมินความตรงของแบบสอบ เนื่องจากข้อสอบหรือแบบสอบที่ลำเอียงเข้าข้างผู้สอบกลุ่มย่อยบางกลุ่มของผู้เข้าสอบทั้งหมด อาจทำให้ผู้สอบกลุ่มย่อยกลุ่มนั้นได้เปรียบกว่าผู้สอบกลุ่มย่อยกลุ่มอื่น ๆ ทั้ง ๆ ที่สอบด้วยข้อสอบข้อเดียวกันหรือแบบสอบฉบับเดียวกัน ทั้งนี้ในการสอบแต่ละครั้งอาจจำแนกผู้สอบออกเป็นกลุ่มย่อย ตามลักษณะที่แตกต่างกันในด้านต่าง ๆ เป็นต้นว่า เชื้อชาติ เผ่าพันธุ์ เพศ ศาสนา ภาษา อายุ ประสบการณ์ หรือ ภูมิหลังอื่น ๆ ของผู้สอบซึ่งอาจทำให้ผู้สอบกลุ่มย่อยบางกลุ่มเกิดการเสียเปรียบ (Scheuneman and Bleistein, 1989)

การศึกษาเรื่องผลการสอบของผู้สอบกลุ่มย่อยต่าง ๆ ของผู้เข้าสอบทั้งหมดมีมานานแล้ว ส่วนเรื่องความยุติธรรมในการสอบระหว่างผู้สอบกลุ่มย่อย เพิ่งมีการศึกษากันอย่างจริงจังในช่วงปลายทศวรรษ 1960 โดยมีการเสนอวิธีการต่าง ๆ เพื่อนำไปใช้ตรวจสอบความลำเอียงของแบบสอบ (test bias) หรือ ความลำเอียงในการคัดเลือกผู้สอบ (bias in selection) ขณะเดียวกันในช่วงเวลานั้น นักพัฒนาแบบสอบก็กำลังสนใจวิธีการจำแนกข้อสอบที่ไม่เหมาะสมกับผู้สอบกลุ่มย่อยบางกลุ่มออกจากแบบสอบ ก่อนที่จะพัฒนาเป็นแบบสอบฉบับสมบูรณ์ต่อไป จึงทำให้เกิดความจำเป็นต้องพัฒนาวิธีการตรวจสอบความลำเอียงของข้อสอบ (item bias) เพื่อใช้เป็นแนวทางในการจำแนกข้อสอบที่ลำเอียงต่อผู้สอบกลุ่มย่อยบางกลุ่มออกจากแบบสอบ หรือคลังข้อสอบ ปัจจุบันการตรวจสอบความลำเอียงของข้อสอบ ได้เป็นส่วนหนึ่งของกระบวนการพัฒนาแบบสอบและการประเมินแบบสอบ เช่นเดียวกันกับการวิเคราะห์ข้อสอบและการตรวจสอบความเที่ยงของแบบสอบ (Hambleton and Others, 1993)

ในสมัยแรก ๆ ของการศึกษาเรื่อง ผลการสอบเพื่อคัดเลือกคนเข้าเรียนต่อหรือเข้าทำงาน พบดัชนีความลำเอียงปรากฏกับกลุ่มคนต่างชาติ ต่างเพศ ทำให้ต้องศึกษา "ความลำเอียงในการคัดเลือกผู้สอบ" ต่อมาเพื่อให้การศึกษาเรื่องนี้มีความชัดเจนยิ่งขึ้น จึงจำกัดให้แคบลงมาศึกษาในระดับข้อสอบ ที่เรียกว่า "ความลำเอียงของข้อสอบ" ซึ่งในปัจจุบันนักวิจัยส่วนใหญ่ใช้คำว่า "ข้อสอบทำหน้าที่ต่างกันกับกลุ่มผู้สอบย่อยต่างกลุ่ม" หรือเรียกสั้น ๆ ว่า "ข้อสอบทำหน้าที่ต่างกัน (DIFferential Item Functioning)" หรือเรียกให้สั้น ๆ ง่าย ๆ ว่า "DIF" ทั้งนี้เนื่องจากเห็นว่าเป็นคำที่มีความหมายกลาง ๆ จึงมีความเหมาะสมในเชิงวิชาการมากกว่าคำว่า "ความลำเอียง" ซึ่งเป็นภาษาที่ใช้กันในทางสังคมและมีความหมายไปในทางลบ อย่างไรก็ตาม คำสองคำนี้มีจุดเน้นที่แตกต่างกัน โดยคำว่า "ความลำเอียงของข้อสอบ" เน้นที่อิทธิพลที่สังเกตได้ของกลุ่มผู้สอบย่อยที่มุ่งศึกษา ส่วนคำว่า "ข้อสอบทำหน้าที่ต่างกัน" เน้นที่ลักษณะทางสถิติของข้อสอบที่ตรวจสอบได้ด้วยวิธีวิเคราะห์ทางสถิติ ซึ่งเป็นส่วนประกอบหนึ่งของสิ่งที่แสดงถึงความลำเอียงของข้อสอบ (Scheuneman and Bleistein, 1989 ; Angoff, 1993 ; Cole, 1993 ; Hambleton and Others, 1993 ; Holland and Wainer, 1993 ; Zieky, 1993 ; Camilli and Shepard, 1994 ; Linn and Gronlund, 1995) จากจุดเน้นนี้แสดงให้เห็นว่า วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นเงื่อนไขที่จำเป็น (necessary condition) ในการประเมินความลำเอียงของข้อสอบ แต่ไม่ใช่เงื่อนไขที่เพียงพอ (not sufficient condition) ในการประเมินความลำเอียงของข้อสอบ เนื่องจากถ้าใช้วิธีการทางสถิติตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเพียงอย่างเดียว ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันที่ได้ ยังสรุปไม่ได้ว่าข้อสอบลำเอียงหรือไม่ การประเมินความลำเอียงของข้อสอบ ยังต้องรวมถึงการใช้ผู้เชี่ยวชาญพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบสอบ ที่เรียกว่า "วิธีการตัดสินข้อสอบ (judgmental method)" ก่อนที่จะสรุปว่า ข้อสอบข้อนั้นลำเอียงหรือไม่ (Angoff, 1993 ; Linn, 1993 ; Ramsay, 1993 ; Zieky, 1993 ; Camilli and Shepard, 1994)

มีผู้ให้ความหมายเกี่ยวกับคำว่า ความลำเอียงของการวัด (measurement bias) ไว้หลายคน Millsap และ Everson (1993) ได้ให้ความหมายทั่วไปว่า ความลำเอียงของการวัด หมายถึงความไม่ถูกต้องของการวัดอย่างเป็นระบบ

Hulin, et al. (1983) กล่าวไว้ว่า ความลำเอียงของข้อสอบเกิดขึ้นเมื่อผู้สอบที่มีคุณลักษณะ (trait) ที่ต้องการวัดในปริมาณที่เท่ากัน แต่มาจากกลุ่มประชากรย่อยต่างกัน มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องต่างกัน (กรณีวัดความสามารถ) หรือมีความน่าจะเป็นในการตอบข้อสอบในทางบวกต่างกัน (กรณีวัดเจตคติ)

Dorans และ Holland (1993) กล่าวไว้ในทำนองเดียวกันว่า ข้อสอบทำหน้าที่ต่างกัน หมายถึง ความแตกต่างในการทำหน้าที่ของข้อสอบ หลังจากกลุ่มผู้สอบได้ถูกจับคู่ตามความสามารถหรือลักษณะ (attribute) ที่ข้อสอบข้อนั้นวัด

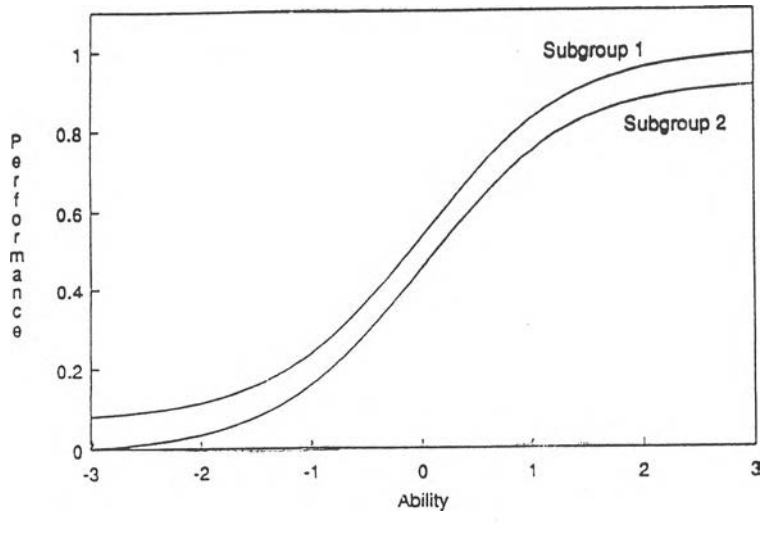
Millsap และ Everson (1993) กล่าวไว้ในทำนองเดียวกันว่า ข้อสอบทำหน้าที่ต่างกัน หมายถึง ความแตกต่างในการทำหน้าที่ของแบบสอบหรือข้อสอบ ระหว่างกลุ่มผู้สอบซึ่งถูกจับคู่ตามลักษณะ ที่วัดโดยแบบสอบหรือข้อสอบนั้น

Hambleton, et al. (1993) กล่าวไว้ว่า นิยามของคำว่า "ข้อสอบทำหน้าที่ต่างกัน" ที่ยอมรับกันอย่างกว้างขวาง ก็คือ ข้อสอบทำหน้าที่ต่างกัน (ข้อสอบที่มีศักยภาพลำเอียง) ภายใต้เงื่อนไขว่า ผู้สอบมีความสามารถเท่ากัน แต่มาจากกลุ่มประชากรย่อยต่างกัน (เช่น เพศชาย และเพศหญิง) มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นถูกไม่เท่ากัน

โดยสรุป ข้อสอบทำหน้าที่ต่างกัน หมายถึง ข้อสอบที่ผู้สอบซึ่งมีความสามารถเท่ากันในสิ่งที่ต้องการวัด มีโอกาสตอบข้อสอบข้อนั้นได้ถูกไม่เท่ากัน เนื่องจากผู้สอบอยู่ในกลุ่มประชากรย่อยต่างกัน

Mellenbergh (1982) ได้จำแนกประเภทของข้อสอบทำหน้าที่ต่างกัน ออกเป็น 2 ประเภท ได้แก่

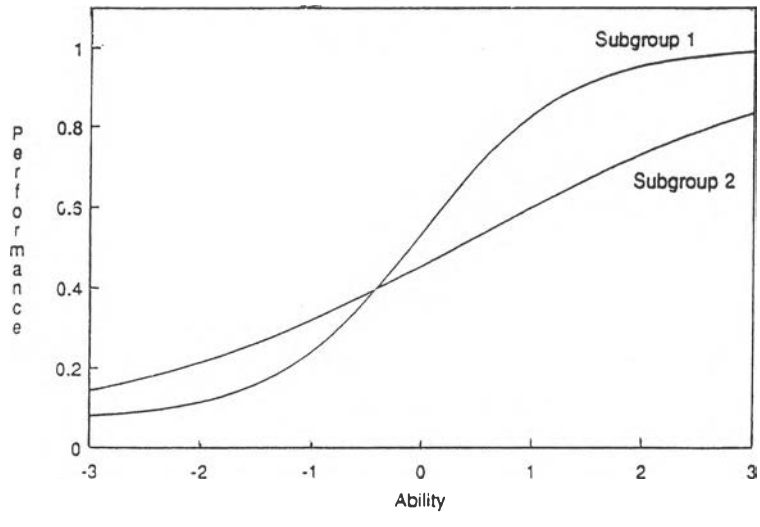
1. ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ (uniform DIF) หมายถึง ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้สอบย่อย 2 กลุ่ม คงเส้นคงวาในทุกระดับความสามารถของผู้สอบ ดังภาพที่ 1



ภาพที่ 1 ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ

จากภาพที่ 1 แสดงให้เห็นว่า ผลการตอบข้อสอบ (performance) ของผู้สอบกลุ่มย่อยสอง (subgroup 2) ต่ำกว่าผู้สอบกลุ่มย่อยหนึ่ง (subgroup 1) อย่างคงเส้นคงวา ในทุก ๆ ระดับความสามารถของผู้สอบ (ability levels)

2. ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (nonuniform DIF) หมายถึง ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้สอบย่อย 2 กลุ่ม ไม่คงเส้นคงวาในทุกระดับความสามารถของผู้สอบ ดังภาพที่ 2



ภาพที่ 2 ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ

จากภาพที่ 2 แสดงให้เห็นว่า ผลการตอบข้อสอบ (performance) ของผู้สอบกลุ่มย่อยสอง (subgroup 2) ต่ำกว่าผู้เข้าสอบกลุ่มย่อยหนึ่ง (subgroup 1) ในช่วงระดับความสามารถสูง ๆ แต่ในช่วงระดับความสามารถต่ำ ๆ ผลการตอบข้อสอบของผู้สอบกลุ่มย่อยสอง กลับสูงกว่าผู้สอบกลุ่มย่อยหนึ่ง

วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี ซึ่ง Hambleton, et al. (1993) ได้จำแนกออกเป็น 3 กลุ่มใหญ่ ๆ ดังนี้

1. กลุ่มวิธีที่ใช้ทฤษฎีการสอบแบบดั้งเดิม (Methods Using Classical Test Theory)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มนี้พัฒนามาจากหลักการของทฤษฎีการสอบแบบดั้งเดิม โดยปกติแล้วใช้คะแนนที่สังเกตได้ของผู้สอบแต่ละคนเป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อย และเปรียบเทียบค่าความยากของข้อสอบแต่ละข้อระหว่างกลุ่มผู้สอบย่อยที่สนใจ

ศึกษา วิธีการในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (analysis of variance) วิธีสหสัมพันธ์ (correlational methods) (Green and Draper, 1972 quoted in Scheuneman and Bleistein, 1989) วิธีแปลงค่าความยากของข้อสอบ (transformed item difficulty method, TID) หรือ วิธีกำหนดจุดค่าเดลต้า (delta plot method) (Angoff, 1982) การวิเคราะห์ตัวลวง (distractor analysis) (Scheuneman, 1982) วิธีสหสัมพันธ์บางส่วน (partial correlation methods) (Stricker, 1982) และวิธีการทำให้เป็นมาตรฐาน (standardization method) (Dorans and Kulick, 1983)

ข้อได้เปรียบของวิธีการในกลุ่มนี้ก็คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายในการดำเนินการไม่สูงนัก ใช้ตรวจสอบกับกลุ่มตัวอย่างขนาดเล็ก และสามารถอธิบายให้คนทั่วไปเข้าใจได้ง่าย ส่วนข้อเสียเปรียบของวิธีการในกลุ่มนี้ก็คือ ค่าสถิติของข้อสอบเปลี่ยนไปตามกลุ่มตัวอย่าง เมื่อกลุ่มตัวอย่างที่ศึกษาเปลี่ยนแปลงไป ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันที่ได้ก็อาจเปลี่ยนแปลงไปด้วย ทำให้การอ้างอิงผลการศึกษาไปยังกลุ่มประชากร อาจมีความเชื่อถือได้น้อยลง

2. กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Methods Using Item Response Theory)

วิธีการในกลุ่มนี้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามกรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบ โดยปกติแล้วใช้การเปรียบเทียบโค้งลักษณะข้อสอบ (item characteristic curves : ICCs) ของกลุ่มผู้สอบย่อยที่ศึกษาตามระดับความสามารถของผู้สอบ ถ้าโค้งลักษณะข้อสอบของกลุ่มผู้สอบย่อยสองกลุ่ม มีรูปร่างเดียวกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าโค้งลักษณะข้อสอบของกลุ่มผู้สอบย่อยสองกลุ่ม มีรูปร่างต่างกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน ค่าพารามิเตอร์ของโค้งลักษณะข้อสอบ ได้แก่ ค่าความยากของข้อสอบ (b-parameter) ค่าอำนาจจำแนกของข้อสอบ (a-parameter) และค่าการเดาข้อสอบถูก (c-parameter) วิธีการในกลุ่มนี้ ได้แก่ วิธี analysis of fit (Durovic, 1975, quoted in Hambleton and Others, 1993) วิธี difficulty shift (Wright, Mead, and Draba, 1976, quoted in Hambleton and Others, 1993) ซึ่งใช้โมเดล IRT แบบ หนึ่ง พารามิเตอร์ วิธี IRT area (Ironson and Subkoviak, 1979 ; Raju, 1988 ; 1990) และวิธี two stage (Lord, 1980) ซึ่งใช้โมเดล IRT แบบ สอง หรือสาม พารามิเตอร์ และ วิธี plot (Hambleton and Rogers, 1991, quoted in Hambleton and Others, 1993)

ข้อได้เปรียบของวิธีการในกลุ่มนี้ก็คือ การแก้ไขข้อบกพร่องของทฤษฎีการสอบแบบดั้งเดิม ทำให้ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่างที่สุ่มมาจากประชากรเดียวกัน

การประมาณค่าความสามารถของผู้สอบเป็นอิสระจากค่าความยากของแบบสอบ โมเดลทางคณิตศาสตร์ง่ายต่อการจับคู่โค้งลักษณะข้อสอบตามระดับความสามารถของผู้สอบ ทำให้สามารถศึกษาความแตกต่างของผลการตอบข้อสอบตามระดับความสามารถของกลุ่มผู้สอบย่อยได้ ไม่ต้องมีข้อตกลงเบื้องต้นเรื่องแบบสอบคู่ขนานในการหาค่าสัมประสิทธิ์ความเที่ยงของแบบสอบ และถ้าผลการตอบข้อสอบของกลุ่มผู้สอบสอดคล้องกับข้อตกลงเบื้องต้นของโมเดล IRT แล้ว วิธีในกลุ่มนี้ก็น่าจะเป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ได้ผลดีวิธีหนึ่ง เนื่องจากเป็นวิธีที่มีทฤษฎีการตอบสนองข้อสอบสนับสนุนและใช้การประมาณค่าความสามารถที่แท้จริงของผู้สอบแทนคะแนนที่สังเกตได้ ดังเช่นที่ใช้ในกลุ่มวิธีที่ใช้ทฤษฎีการสอบแบบดั้งเดิม ส่วนข้อเสียเปรียบของวิธีการในกลุ่มนี้ก็คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสลับซับซ้อน เสียค่าใช้จ่ายในการดำเนินการสูง และต้องการกลุ่มตัวอย่างขนาดใหญ่

3. กลุ่มวิธีที่ใช้วิธีไค-สแควร์ (Methods Using Chi-Square Methods)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มนี้ บางครั้งก็เรียกว่า กลุ่มวิธีไค-สแควร์ (chi-square methods) วิธีการในกลุ่มนี้ใช้ค่าสถิติไค-สแควร์แสดงการทำหน้าที่ต่างกันของข้อสอบ และใช้คะแนนของแบบสอบหรือคะแนนของแบบสอบที่ทำให้บริสุทธิ์ (purified test score)เป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อยสองกลุ่มที่ศึกษา ก่อนการเปรียบเทียบผลการตอบข้อสอบ วิธีการในกลุ่มนี้ ได้แก่ วิธีตารางการณ์จร (contingengy table method)(Scheuneman, 1975 ; 1979) วิธีตารางการณ์จรปรับขยาย (modified contingency table method)(Veale, 1977, quoted in Hambleton and Others, 1993) วิธีล็อก-ลิเนียร์ (log-linear models) (Mellenbergh, 1982) วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel method : MH) (Holland and Thayer, 1986 ; 1988) และวิธีการถดถอยโลจิสติก (logistic regression method) (Swaminathan and Rogers, 1990)

ข้อได้เปรียบของวิธีการในกลุ่มนี้ก็คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยากซับซ้อน เสียค่าใช้จ่ายในการดำเนินการไม่สูง ใช้กับกลุ่มตัวอย่างขนาดไม่ใหญ่นัก บางวิธีการมีหลักการที่ดีในการจับคู่กลุ่มผู้สอบย่อยตามความสามารถของผู้สอบ และมีการทดสอบนัยสำคัญ ส่วนข้อเสียเปรียบของวิธีการในกลุ่มนี้ก็คล้าย ๆ กับวิธีที่ใช้ทฤษฎีการสอบแบบดั้งเดิม

จะเห็นว่า วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีหลายวิธี Hambleton , et al. (1993) ได้ให้ความเห็นว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เหมาะสม ต้องใช้หลักการเปรียบเทียบผลการตอบข้อสอบของผู้สอบกลุ่มอ้างอิง (reference group) กับกลุ่มสนใจ (focal group) หลังจากการจับคู่กลุ่มผู้สอบสองกลุ่มตามความสามารถของ

ผู้สอบแล้ว ซึ่งสอดคล้องกับความเห็นของ Angoff (1993), และ Dorans และ Holland, (1993) ส่วน Mellenbergh (1989) ได้เรียกวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้หลักการจับคู่ผู้สอบของกลุ่มอ้างอิงกับกลุ่มสนใจตามความสามารถของผู้สอบว่าเป็น " วิธีตามเงื่อนไข (conditional methods) " กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบและวิธีแมนเทล-แฮนส์เซล เป็นวิธีตามเงื่อนไข ปัจจุบันจึงเป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่สำคัญ

วิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ เป็นวิธีที่มีข้อได้เปรียบในทางทฤษฎี (Holland and Thayer, 1988 ; Hambleton and Others, 1993) เนื่องจากใช้ marginal maximum likelihood estimation (MMLE) หรือ marginal bayes estimation (MBE) ในการทดสอบพารามิเตอร์ของโมเดล IRT จึงมีอำนาจทดสอบสูง แต่ข้อจำกัดที่สำคัญของวิธีนี้ก็คือ ข้อยุ่งยากในการนำไปใช้งานภาคสนาม ส่วนวิธีแมนเทล-แฮนส์เซล เป็นวิธีที่มีข้อได้เปรียบในทางปฏิบัติ (Hambleton and Others, 1993 ; Holland and Wainer, 1993) เนื่องจากสามารถคำนวณได้ง่าย เพราะใช้ตารางการณ์จรและสถิติไค-สแควร์ แต่ข้อจำกัดที่สำคัญของวิธีนี้ก็คือ การใช้คะแนนที่สังเกตได้แทนความสามารถของกลุ่มผู้สอบ ทำให้ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันอาจเปลี่ยนแปลงไป เมื่อใช้กลุ่มตัวอย่างแตกต่างกัน

วิธีแมนเทล-แฮนส์เซล เป็นวิธีที่ได้รับความนิยมและได้รับการยอมรับ หน่วยงานบริการทดสอบทางการศึกษาแห่งสหรัฐอเมริกา (Educational Testing Service : ETS) ได้แนะนำให้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล และเป็นวิธีมาตรฐานที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโครงการทดสอบสำคัญ ๆ ของหน่วยงาน วิธีนี้ได้ถูกนำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในโครงการทดสอบของหน่วยงานมากที่สุด นับตั้งแต่ปี ค.ศ. 1987 เป็นต้นมา (Ryan, 1991 ; Dorans and Holland, 1993 ; Hambleton and Others, 1993 ; Zieky, 1993 ; Mazor and Others, 1995)

เหตุผลสำคัญที่วิธีแมนเทล-แฮนส์เซลเป็นที่ยอมรับนั้น พอจะสรุปได้ดังนี้

1. การนิยามคำว่า "ข้อสอบทำหน้าที่ต่างกัน" ของวิธีแมนเทล-แฮนส์เซล กับ วิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ เหมือนกัน กล่าวคือ ข้อสอบทำหน้าที่ต่างกัน ภายใต้เงื่อนไขที่ว่าผู้สอบมีความสามารถในระดับเดียวกัน แต่อยู่ในกลุ่มประชากรย่อยต่างกัน มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นถูกต่างกัน (Clauser and Others, 1991a ; Hambleton and Others, 1993)

2. กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมนเทล-แฮนส์เซล ไม่ยุ่งยากซับซ้อน และสามารถปฏิบัติได้ง่ายกว่าวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Clauser

and Others, 1991a ; Swaminathan and Rogers, 1991 ; Dorans and Holland, 1993 ; Hambleton and Others, 1993 ; Holland and Wainer, 1993 ; Millsap and Everson, 1993 ; Mazor and Others, 1994 ; Uttaro and Millsap, 1994)

3. **เทคนิคทางสถิติเชื่อถือได้** มีการทดสอบนัยสำคัญและวัดขนาดอิทธิพล (Holland and Thayer, 1988 ; Hills, 1989 ; Clauser and Others, 1991a ; Swaminathan and Rogers, 1991 ; Millsap and Everson, 1993 ; Mazor and Others, 1994)

4. **เสียค่าใช้จ่ายในการดำเนินการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ต่ำกว่าวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ** (Hills, 1989 ; Clauser and Others, 1991b ; Dorans and Holland, 1993 ; Mazor and Others, 1994 ; Uttaro and Millsap, 1994)

5. **มีหลักการที่สามารถเข้าใจได้ง่าย** (Hambleton and Others, 1993 ; Holland and Wainer, 1993 ; Millsap and Everson, 1993 ; Haladyna, 1994)

6. **ใช้กับกลุ่มตัวอย่างขนาดเล็กกว่าวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ** (Hills, 1989 ; Scheuneman and Bleistein, 1989 ; Ryan, 1991 ; Zieky, 1993 ; Mazor and Others, 1994)

7. **ใช้กับกลุ่มผู้สอบย่อยที่ขนาดไม่เท่ากันได้** (Hills, 1989)

8. **ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันโดยวิธีแมนเทล-แฮนส์เซล กับ วิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ คล้ายคลึงกัน** (Hambleton and Rogers, 1989 ; Clauser and Others, 1991a ; Camilli and Shepard, 1994)

Uttaro และ Millsap (1994) **ได้ให้ความเห็นว่า การศึกษาความถูกต้องของวิธีแมนเทล-แฮนส์เซล ยังทำได้ไม่ทั่วถึง ส่วนใหญ่ศึกษาจากสถานการณ์จำลองเพื่อดูว่าวิธีแมนเทล-แฮนส์เซล มีอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขต่าง ๆ เพียงใด และมีความคลาดเคลื่อน ประเภทที่ I ในระดับที่ยอมรับได้หรือไม่ ยังขาดการศึกษาจากสภาพที่เป็นจริงหรือข้อมูลเชิงประจักษ์**

จากที่กล่าวมาข้างต้น ผู้วิจัยเห็นว่า วิธีแมนเทล-แฮนส์เซลเป็นวิธีตามเงื่อนไข สอดคล้องกับนิยามของข้อสอบทำหน้าที่ต่างกัน ยอมผ่อนคลายข้อตกลงเบื้องต้น เรื่องความเป็นเอกมิติ (unidimentional) **ของแบบสอบ เพียงคาดหมายได้ว่าเป็นเอกมิติ** (implicitly or approximately unidimentional) **ก็สามารถใช้คะแนนของแบบสอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบได้** (Angoff, 1993 ; Hambleton & Others, 1993) **และจากงานวิจัยพบว่า ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันระหว่างวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ** (IRT area method) **กับวิธีแมนเทล-**

แฮนส์เซล มีความสอดคล้องกันสูง (Hambleton and Rogers, 1989) จึงสนใจที่จะศึกษาวิธีแมนเทล-แฮนส์เซล ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับข้อมูลเชิงประจักษ์

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล เป็นการเปรียบเทียบโอกาสในการตอบข้อสอบถูกของผู้สอบสองกลุ่ม หลังจากการจับคู่กลุ่มผู้สอบตามความรู้ ทักษะ หรือความสามารถของผู้สอบ การจับคู่กลุ่มผู้สอบสองกลุ่มเป็นเงื่อนไขที่สำคัญ เพราะเป็นเกณฑ์ที่ใช้แทนความสามารถของกลุ่มผู้สอบสองกลุ่ม ดังนั้น ถ้าเกณฑ์การจับคู่กลุ่มผู้สอบที่ใช้แทนความสามารถของกลุ่มผู้สอบมีความถูกต้องมาก ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันก็ถูกต้องมากด้วย ตามหลักการแล้ว วิธีแมนเทล-แฮนส์เซล สามารถใช้เกณฑ์การจับคู่กลุ่มผู้สอบได้ทั้งเกณฑ์ภายในแบบสอบ (internal criterion) และเกณฑ์ภายนอกแบบสอบ (external criterion) ถ้าพิจารณากันโดยทั่ว ๆ ไปแล้ว จะเห็นว่าการใช้เกณฑ์ภายนอกแบบสอบที่มีอยู่ก่อนแล้ว น่าจะมีความเหมาะสมมากกว่าการใช้เกณฑ์ภายในแบบสอบ เพราะเกณฑ์ภายนอกแบบสอบไม่มีข้อสอบที่ต้องการตรวจสอบการทำหน้าที่ต่างกันรวมอยู่ด้วย จึงน่าจะมีความตรงสูงกว่าเกณฑ์ภายในแบบสอบ แต่ในความเป็นจริงแล้ว เกณฑ์ภายนอกแบบสอบที่มีความเหมาะสมค่อนข้างหาได้ยาก (Clauser and Others, 1993 ; Hambleton and Others, 1993) นักพัฒนาแบบสอบจึงนิยมใช้เกณฑ์ภายในแบบสอบ เนื่องจากพิจารณาเห็นว่าเป็นการตรวจสอบข้อสอบในบริบทของแบบสอบ ในทางปฏิบัติจึงใช้คะแนนของแบบสอบทั้งฉบับ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบ เพราะเกี่ยวข้องโดยตรงกับความรู้ ทักษะ หรือความสามารถที่ถูกวัดด้วยแบบสอบ สามารถตรวจสอบความเที่ยงและความตรงของแบบสอบได้ ผู้สอบทุกคนสอบภายใต้สถานการณ์เดียวกัน และสามารถตีความผลการสอบของกลุ่มผู้สอบได้อย่างมีความหมาย แต่จุดอ่อนของการใช้คะแนนของแบบสอบทั้งฉบับเป็นเกณฑ์การจับคู่กลุ่มผู้สอบก็คือ การรวมคะแนนของข้อสอบทำหน้าที่ต่างกันเข้ามาเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ ในการแก้ไขจุดอ่อนนี้ Holland และ Thayer (1988) ได้เสนอให้ใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบ 2 ขั้นตอน (two steps procedure) ดังนี้

ขั้นตอนที่ 1 การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบใช้คะแนนของแบบสอบทั้งฉบับ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อยสองกลุ่ม เพื่อวิเคราะห์ข้อสอบแต่ละข้อในแบบสอบว่าข้อสอบข้อใดทำหน้าที่ต่างกัน ถ้าพบว่าข้อสอบข้อใดทำหน้าที่ต่างกัน ก็นำเอาคะแนนของข้อสอบข้อนั้นออกจากคะแนนรวมของผู้สอบแต่ละคน แล้วใช้คะแนนรวมของข้อสอบที่เหลือเป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อยสองกลุ่ม ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ในขั้นตอนที่ 2

ขั้นตอนที่ 2 การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ใช้คะแนนรวมของแบบสอบที่นำเอาข้อสอบทำหน้าที่ต่างกันซึ่งตรวจพบ ในขั้นตอนที่ 1 ออกไปแล้ว ซึ่งเรียกคะแนนนี้ว่า "คะแนนของแบบสอบที่ทำให้บริสุทธิ์ (purified test score)" เป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อยสองกลุ่ม แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแต่ละข้อในแบบสอบซ้ำอีกครั้ง

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้เทคนิค 2 ขั้นตอนนี้ เรียกว่า การทำให้เกณฑ์การจับคู่กลุ่มผู้สอบมีความบริสุทธิ์ (purification of matching criterion)

อย่างไรก็ตาม วิธีแมนเทล-แฮนส์เซล แบบปกติ (traditional MH procedure) ดังกล่าวข้างต้น พบว่า ยังมีจุดอ่อนคือ ไม่ไวต่อการตรวจพบข้อสอบทำหน้าที่ต่างกัน แบบไม่สม่ำเสมอ (Hills, 1989 ; Scheuneman and Bleistein, 1989 ; Swaminathan and Rogers, 1990) การทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ หมายถึง ความแตกต่างของความน่าจะเป็นในการตอบข้อสอบถูก ระหว่างกลุ่มผู้สอบย่อยสองกลุ่มไม่คงเส้นคงวาในทุกระดับความสามารถของผู้สอบ นั่นก็คือ การทำหน้าที่ต่างกันแบบไม่สม่ำเสมอเกิดขึ้น เมื่อมีปฏิสัมพันธ์ระหว่างระดับความสามารถของกลุ่มผู้สอบกับความเป็นสมาชิกของกลุ่ม (Mellenbergh, 1982) ในทฤษฎีการตอบสนองข้อสอบ คำว่า "ปฏิสัมพันธ์ (interaction)" หมายถึง ความแตกต่างของค่าอำนาจจำแนกของข้อสอบ (a-parameter) โดยพิจารณาจากโค้งลักษณะข้อสอบของกลุ่มผู้สอบย่อยสองกลุ่มไม่ขนานกัน (Swaminathan and Rogers, 1990 ; Mazor and Others, 1994)

แม้ว่า มีวิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบอยู่บ้างแล้ว เช่น วิธีล็อก-ลิเนียร์ วิธีการถดถอยโลจิสติก และวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ เป็นต้น แต่ วิธีล็อก-ลิเนียร์ และ วิธีการถดถอยโลจิสติก ก็มีข้อยุ่งยากในทางปฏิบัติมากกว่าวิธีแมนเทล-แฮนส์เซล นอกจากนี้วิธีการถดถอยโลจิสติก ต้องเสียค่าใช้จ่ายในการคำนวณสูงกว่าวิธีแมนเทล-แฮนส์เซล ประมาณ 3 - 4 เท่า (Swaminathan and Rogers, 1990) ส่วนวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ ก็มีข้อยุ่งยากในทางปฏิบัติ เสียค่าใช้จ่ายในการตรวจสอบสูง และยังต้องการกลุ่มตัวอย่างขนาดใหญ่ จึงทำให้ผู้ปฏิบัติไม่นิยมนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในทางตรงกันข้ามวิธีแมนเทล-แฮนส์เซล มีความสะดวกในทางปฏิบัติ เสียค่าใช้จ่ายในการคำนวณไม่แพง และสามารถใช้ได้กับกลุ่มตัวอย่างขนาดไม่ใหญ่นัก จึงทำให้ผู้ปฏิบัตินิยมนำวิธีแมนเทล-แฮนส์เซลไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังนั้น จึงสมควรที่จะพัฒนาวิธีแมนเทล-แฮนส์เซล ให้สามารถตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบให้ได้ผลดียิ่งขึ้น

จากการศึกษาของ Swaminathan และ Rogers (1990) เรื่อง การเปรียบเทียบผลการตรวจพบข้อสอบทำหน้าที่ต่างกันด้วยวิธีการถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซล พบว่า วิธีแมนเทล-แฮนส์เซล แบบปกติ (traditional MH procedure) สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ชนิดที่โค้งลักษณะข้อสอบ (ICCs) ของกลุ่มผู้สอบสองกลุ่มตัดกันห่างจากจุดกลางของช่วงความสามารถของผู้สอบ แต่ตรวจไม่พบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ชนิดที่โค้งลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่มตัดกันใกล้ ๆ จุดกลางของช่วงความสามารถของผู้สอบ เหตุที่เป็นเช่นนี้ เพราะวิธีแมนเทล-แฮนส์เซล ให้ค่าสถิติที่มีเครื่องหมาย (signed statistic) จึงทำให้ความแตกต่างทางลบ (negative differences) ระหว่างกลุ่มผู้สอบย่อยสองกลุ่มหักล้างกับความแตกต่างทางบวก (positive differences) ระหว่างกลุ่มผู้สอบย่อยสองกลุ่ม ณ ระดับความสามารถนั้น เนื่องจากแต่ละเซลของวิธีแมนเทล-แฮนส์เซล ถูกถ่วงน้ำหนักด้วยจำนวนผู้สอบ ณ ระดับความสามารถนั้น และ โอกาสในการตอบข้อสอบถูกขึ้นอยู่กับฟังก์ชั่นการแจกแจงความสามารถของผู้สอบและระดับความยากของข้อสอบ (Mazor and Others, 1994) ดังนั้น ถ้ากลุ่มผู้สอบถูกแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มผู้สอบที่มีความสามารถสูง กับกลุ่มผู้สอบที่มีความสามารถต่ำ และแบ่งกลุ่มข้อสอบตามระดับความยากของข้อสอบเป็น 3 กลุ่ม คือ กลุ่มข้อสอบยาก กลุ่มข้อสอบยากง่ายปานกลาง และกลุ่มข้อสอบง่าย อาจทำให้สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้ดียิ่งขึ้น การปรับปรุงวิธีการวิเคราะห์ข้อมูลตามแนวความคิดนี้ ไม่น่าจะทำให้วิธีแมนเทล-แฮนส์เซลด้อยคุณค่าเรื่องความได้เปรียบในทางปฏิบัติ เรื่องนี้ Mazor, et al. (1994) ได้ศึกษาเบื้องต้นในเชิงทฤษฎีไว้บ้างแล้ว เฉพาะในประเด็นการแบ่งกลุ่มผู้สอบตามคะแนนผลการสอบของผู้สอบ ดังนั้นในงานวิจัยครั้งนี้ จึงมุ่งศึกษาประเด็นต่าง ๆ ที่ยังไม่มีผู้ใดเคยศึกษาไว้ คือการศึกษากับข้อมูลเชิงประจักษ์ การแบ่งกลุ่มผู้สอบตามค่าความสามารถของผู้สอบ และแบ่งกลุ่มข้อสอบตามค่าความยากของข้อสอบ

จากเหตุผลดังกล่าวข้างต้น ผู้วิจัยจึงสนใจที่จะศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ (traditional MH procedure) กับ วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ในการศึกษาครั้งนี้ใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามข้อแนะนำของ Hambleton, et al. (1993) เครื่องมือที่ใช้เก็บรวบรวมข้อมูล ได้แก่ แบบสอบวัดความสามารถในการอ่านภาษาไทย ชั้นมัธยมศึกษาปีที่ 1 ซึ่งผู้วิจัยสร้างขึ้น เหตุที่เลือกศึกษาวิชาภาษาไทย เพราะข้อสอบวิชาภาษามีแนวโน้มที่จะมีข้อสอบทำหน้าที่ต่างกันมากกว่าวิชาคณิตศาสตร์ (Mcpeek and Wild, 1986, quoted in Ryan, 1991 ;

Linn, 1993) และเลือกศึกษากับนักเรียนระดับชั้นมัธยมศึกษาปีที่ 1 เพราะเป็นการศึกษาในระดับที่สูงกว่าการศึกษาภาคบังคับของไทย จึงถือว่าผู้สอบมีความสามารถในการอ่านภาษาไทยแล้ว ส่วนตัวแปรจำแนกกลุ่มประชากรที่เลือกศึกษา ได้แก่ ตัวแปรเพศ เนื่องจากพิจารณาเห็นว่าเป็นตัวแปรที่รู้ชัดและสามารถแบ่งได้โดยปราศจากความคลาดเคลื่อน (Millsap and Everson, 1993 ; Zieky, 1993) จึงมีความเหมาะสมที่จะนำมาใช้ศึกษาในเชิงวิธีการกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยข้อมูลจริง ผู้วิจัยให้เพศชายเป็นกลุ่มสนใจ (focal group) และเพศหญิงเป็นกลุ่มอ้างอิง (reference group) เพราะเพศหญิงมีแนวโน้มได้เปรียบในข้อสอบที่มีเนื้อหาเกี่ยวกับภาษา (สุพัฒน์ สุกมลสันต์, 2534 ; Angoff, 1993 ; กาญจนา วัจนสุนทร, 2538) ส่วนการตรวจหาจำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอในแบบสอบวัดความสามารถในการอ่านภาษาไทย เพื่อใช้เป็นเกณฑ์สำหรับเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับ วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ นั้น ผู้วิจัยเลือกใช้วิธี IRT area เนื่องจากมีทฤษฎีการตอบสนองข้อสอบสนับสนุน และใช้การประมาณค่าความสามารถที่แท้จริงของผู้สอบ ทำให้ผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน มีความถูกต้องสูง (Holland and Thayer, 1988 ; Hambleton and Others, 1993)

วัตถุประสงค์ของการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับ วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ด้วยข้อมูลเชิงประจักษ์ และมีวัตถุประสงค์เฉพาะ ดังนี้

เพื่อเปรียบเทียบสัดส่วนการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับ วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ เมื่อใช้จำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ที่ตรวจพบตามวิธี IRT area เป็นเกณฑ์

สมมุติฐานการวิจัย

Mazor, et al. (1994) ได้ศึกษาเบื้องต้นในเชิงทฤษฎีเรื่อง การใช้วิธีแมนเทล-แฮนส์เซล ตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ พบว่า ถ้ากลุ่มผู้สอบถูกแบ่งครึ่งเป็น 2 กลุ่มตามคะแนนผลการสอบ แล้ววิเคราะห์แต่ละกลุ่มแยกจากกันอย่างอิสระ มีแนวโน้มทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้เพิ่มขึ้น ทั้งนี้ อาจเป็นเพราะว่าการแบ่งครึ่งกลุ่มผู้สอบตามคะแนนผลการสอบ จะทำให้กลุ่มผู้สอบที่ถูกแบ่งมีความเป็นเอกพันธ์มากขึ้น ทำให้ความแปรปรวนของกลุ่มอ้างอิง (Var R_{rm}) ซึ่งใช้เป็นตัวหารในสูตรของ MH-χ^2 ลดลง ส่งผลให้ค่าสถิติ MH-χ^2 ที่คำนวณได้สูงขึ้น จึงทำให้มีโอกาสตรวจพบข้อสอบทำหน้าที่ต่างกันได้เพิ่มขึ้น และการที่แบ่งความสามารถของผู้สอบออกเป็น 2 กลุ่ม ทำให้จุดที่แบ่งความสามารถของผู้สอบอยู่ใกล้เคียงกับจุดกลางของช่วงความสามารถของผู้สอบ ซึ่งเป็นบริเวณตัดกันของโค้งลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่มของข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ชนิดที่ตรวจไม่ค่อยพบ ตามวิธีแมนเทล-แฮนส์เซล แบบปกติ (Swaminathan and Rogers, 1990) และยังส่งผลให้ความแตกต่างทางลบ (negative difference) กับความแตกต่างทางบวก (positive difference) ของอัตราส่วนแต้มต่อร่วม (common odds ratio) ในแต่ละระดับความสามารถมีโอกาสหักล้างกันน้อยลง ทำให้สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ชนิดที่โค้งลักษณะข้อสอบของกลุ่มผู้สอบย่อยสองกลุ่มตัดกันใกล้ ๆ จุดกลางของช่วงความสามารถได้เพิ่มขึ้น ด้วยเหตุที่โอกาสในการตอบข้อสอบถูกตามวิธีแมนเทล-แฮนส์เซล ขึ้นอยู่กับฟังก์ชั่นการแจกแจงความสามารถของผู้สอบและระดับความยากของข้อสอบ (Mazor and Others, 1994) ดังนั้น การวิเคราะห์ข้อมูลโดยแบ่งผู้สอบตามระดับความสามารถ ออกเป็น 2 กลุ่ม คือ กลุ่มผู้สอบที่มีความสามารถสูง กับ กลุ่มผู้สอบที่มีความสามารถต่ำ แล้วแบ่งข้อสอบออกเป็น 3 กลุ่ม คือ กลุ่มข้อสอบยาก กลุ่มข้อสอบยากง่ายปานกลาง และกลุ่มข้อสอบง่าย จึงน่าจะตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ได้จำนวนมากกว่าการวิเคราะห์ข้อมูลตามวิธีแมนเทล-แฮนส์เซล แบบปกติ (traditional MH procedure) ทำให้ผู้วิจัยกำหนดสมมุติฐานการวิจัย ไว้ดังนี้

สัดส่วนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอที่ตรวจพบของวิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ สูงกว่า วิธีแมนเทล-แฮนส์เซล แบบปกติ เมื่อใช้จำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ที่ตรวจพบตามวิธี IRT area เป็นเกณฑ์

ขอบเขตของการวิจัย

1. การวิจัยครั้งนี้ใช้ข้อมูลจริงในบริบทของประเทศไทย โดยตัวแปรจำแนกกลุ่มประชากรที่ศึกษา ได้แก่ ตัวแปรเพศ

2. ตัวแปรที่ศึกษา

2.1 ตัวแปรอิสระ มี 2 ตัว คือ

2.1.1 ความสามารถของผู้สอบ แบ่งออกเป็น 2 กลุ่ม คือ

2.1.1.1 กลุ่มผู้สอบที่มีความสามารถสูง

2.1.1.2 กลุ่มผู้สอบที่มีความสามารถต่ำ

2.1.2 ความยากของข้อสอบ แบ่งออกเป็น 3 กลุ่ม คือ

2.1.2.1 กลุ่มข้อสอบยาก

2.1.2.2 กลุ่มข้อสอบยากง่ายปานกลาง

2.1.2.3 กลุ่มข้อสอบง่าย

2.2 ตัวแปรตาม คือ จำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอที่ตรวจพบ

3. เครื่องมือที่ใช้ในการวิจัยครั้งนี้ ได้แก่ แบบสอบวัดความสามารถในการอ่านภาษาไทย ชั้นมัธยมศึกษาปีที่ 1 ชนิดเลือกตอบ 4 ตัวเลือก จำนวน 75 ข้อ

ข้อจำกัดของการวิจัย

การวิจัยครั้งนี้เป็นการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับวิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ โดยมุ่งเปรียบเทียบเฉพาะสัดส่วนการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอระหว่างสองวิธีดังกล่าวข้างต้น เมื่อใช้จำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอที่ตรวจพบ ตามวิธี IRT area เป็นเกณฑ์ เท่านั้น ไม่ได้ศึกษาไปถึงเรื่อง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน (false positives) ทั้ง ๆ ที่ความจริงแล้วเป็นข้อสอบทำหน้าที่ไม่ต่างกัน

ข้อตกลงเบื้องต้น

กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ เป็นวิธีที่มีทฤษฎีการตอบสนองข้อสอบสนับสนุน และใช้การประมาณค่าความสามารถที่แท้จริงของผู้สอบ จึงเป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่มีความถูกต้องสูง (Holland and Thayer, 1988 ; Hambleton and Others, 1993) ซึ่งในกลุ่มวิธีนี้ วิธี IRT area ได้รับการยอมรับมากที่สุด (Hambleton and Others, 1993) ประกอบกับผลการศึกษาของ Kim และ Cohen (1994) พบว่า แบบจำลองของทฤษฎีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ มีความคลาดเคลื่อน ประเภทที่ I ต่ำกว่า แบบจำลอง 3 พารามิเตอร์ ดังนั้น ผู้วิจัยจึงเลือกใช้วิธี IRT area แบบ 2 พารามิเตอร์ ตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ในแบบสอบวัดความสามารถในการอ่านภาษาไทย โดยถือว่าจำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอที่ตรวจพบ เป็นเกณฑ์ที่ใช้ได้ดี สำหรับเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับ วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ

คำจำกัดความที่ใช้ในการวิจัย

แบบสอบวัดความสามารถในการอ่านภาษาไทย หมายถึง แบบสอบที่วัดความสามารถในการจับใจความของสิ่งที่อ่าน (เรื่อง บทความ หรือสถานการณ์) แสดงออกโดยการแปลความหมายของสิ่งที่อ่านจากรูปแบบหนึ่งไปสู่อีกรูปแบบหนึ่ง (ข้อความหรือตัวเลข) หรือการตีความหมายของสิ่งที่อ่าน (การอธิบายหรือสรุปผล) หรือ การประมาณแนวโน้มในอนาคตของสิ่งที่อ่าน (ทำนายผลหรือสิ่งที่อาจเกิดขึ้นตามมา) (Bloom, 1956, quoted in Linn and Gronlund, 1995)

ข้อสอบทำหน้าที่ต่างกัน หมายถึง ข้อสอบที่ผู้สอบซึ่งมีความสามารถเท่ากันในสิ่งที่ต้องการวัด มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้สอบย่อยต่างกัน คือ กลุ่มผู้สอบเพศหญิง กับ กลุ่มผู้สอบเพศชาย

วิธีแมนเทล-แฮนส์เซล แบบปกติ (traditional MH procedure) หมายถึง การวิเคราะห์หาค่าสถิติ MH-χ^2 ที่ใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการใช้กลุ่มผู้สอบทั้งหมดและใช้คะแนนรวมของแบบสอบทั้งฉบับ เป็นเกณฑ์การจับคู่ผู้สอบกลุ่มสนใจกับกลุ่มอ้างอิง

วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ หมายถึง การวิเคราะห์หาค่าสถิติ MH- χ^2 ที่ใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการใช้กลุ่มผู้สอบที่มีความสามารถสูง และใช้คะแนนรวมของแบบสอบทั้งฉบับหรือคะแนนรวมของกลุ่มข้อสอบยาก กลุ่มข้อสอบยากง่ายปานกลาง กลุ่มข้อสอบง่าย เป็นเกณฑ์ในการจับคู่ผู้สอบกลุ่มสนใจกับกลุ่มอ้างอิง หรือการใช้กลุ่มผู้สอบที่มีความสามารถต่ำ และใช้คะแนนรวมของแบบสอบทั้งฉบับหรือคะแนนรวมของกลุ่มข้อสอบยาก กลุ่มข้อสอบยากง่ายปานกลาง กลุ่มข้อสอบง่าย เป็นเกณฑ์ในการจับคู่ผู้สอบกลุ่มสนใจกับกลุ่มอ้างอิง

กลุ่มอ้างอิง (reference group : r) หมายถึง กลุ่มผู้สอบที่คาดว่าจะเป็นกลุ่มที่ได้เปรียบจากการทำหน้าที่ต่างกันของข้อสอบ กล่าวคือ เป็นกลุ่มผู้สอบที่มีโอกาสตอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องสูงกว่ากลุ่มสนใจ หลังจากจับคู่ผู้สอบกลุ่มสนใจกับกลุ่มอ้างอิงตามความสามารถ ในการวิจัยครั้งนี้ใช้กลุ่มเพศหญิง เป็นกลุ่มอ้างอิง

กลุ่มสนใจ (focal group : f) หมายถึง กลุ่มผู้สอบที่คาดว่าจะเป็นกลุ่มที่เสียเปรียบจากการทำหน้าที่ต่างกันของข้อสอบ กล่าวคือ เป็นกลุ่มผู้สอบที่มีโอกาสตอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องต่ำกว่ากลุ่มอ้างอิง หลังจากจับคู่ผู้สอบกลุ่มผู้สนใจกับกลุ่มอ้างอิงตามความสามารถ ในการวิจัยครั้งนี้ใช้กลุ่มเพศชาย เป็นกลุ่มสนใจ

กลุ่มผู้สอบทั้งหมด หมายถึง ผู้สอบทุกคนในกลุ่มอ้างอิงและกลุ่มสนใจ

กลุ่มผู้สอบที่มีความสามารถสูง หมายถึง ผู้สอบในกลุ่มอ้างอิง หรือกลุ่มสนใจที่มีค่าความสามารถ (θ) มากกว่าหรือเท่ากับค่าความสามารถเฉลี่ยของกลุ่มผู้สอบทั้งหมด

กลุ่มผู้สอบที่มีความสามารถต่ำ หมายถึง ผู้สอบในกลุ่มอ้างอิง หรือ กลุ่มสนใจที่มีค่าความสามารถ (θ) น้อยกว่าค่าความสามารถเฉลี่ยของกลุ่มผู้สอบทั้งหมด

แบบสอบทั้งฉบับ หมายถึง แบบสอบวัดความสามารถในการอ่านภาษาไทย จำนวน 75 ข้อ

กลุ่มข้อสอบยาก หมายถึง ข้อสอบที่มีค่าความยากมากกว่าหรือเท่ากับ 1.00 ($b \geq 1.00$)

กลุ่มข้อสอบยากง่ายปานกลาง หมายถึง ข้อสอบที่มีค่าความยากอยู่ระหว่างน้อยกว่า 1.00 และมากกว่า -1.00 ($-1.00 < b < 1.00$) (Clauser,et al., 1991b)

กลุ่มข้อสอบง่าย หมายถึง ข้อสอบที่มีค่าความยากน้อยกว่าหรือเท่ากับ -1.00 ($b \leq -1.00$)

จำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอที่ตรวจพบ หมายถึง จำนวนข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอที่ตรวจพบสอดคล้องกัน ระหว่างวิธี IRT area แบบ 2 พารามิเตอร์ กับ วิธีแมนเทล-แฮนส์เซล แบบปกติ หรือวิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ

ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ หมายถึง ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้สอบ 2 กลุ่ม คงเส้นคงวาในทุกระดับความสามารถของผู้สอบ

ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ หมายถึง ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้สอบ 2 กลุ่ม ไม่คงเส้นคงวาในทุกระดับความสามารถของผู้สอบ

การระบุผิดพลาดว่าข้อสอบทำหน้าที่ไม่ต่างกัน (False Negative) หมายถึง การระบุว่าข้อสอบทำหน้าที่ไม่ต่างกัน ทั้ง ๆ ที่ความจริงแล้วเป็น ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ตามวิธี IRT area ที่ใช้เป็นเกณฑ์

ประโยชน์ที่คาดว่าจะได้รับ

งานวิจัยครั้งนี้ มุ่งเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับวิธีแมนเทล- แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ โดยการใช้ข้อมูลเชิงประจักษ์ ผู้วิจัยจึงคาดว่าจะเป็นประโยชน์ ดังนี้

1. เป็นแนวทางในการพัฒนาวิธีแมนเทล-แฮนส์เซล ตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ โดยการใช้ข้อมูลเชิงประจักษ์ กับตัวแปรจำแนกกลุ่มประชากรอื่น ๆ เช่น เชื้อชาติ ศาสนา สถานภาพทางเศรษฐกิจและสังคม เป็นต้น

2. ผู้ที่สนใจเรื่อง การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สามารถนำเอาวิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับตัวแปรจำแนกกลุ่มประชากรอื่น ๆ หรือแบบสอบวิชาอื่น ๆ ได้ โดยคำนึงถึงความได้เปรียบในทางปฏิบัติ การประหยัดเวลาและค่าใช้จ่าย

3. เป็นแนวทางในการวิเคราะห์หาดัชนี ที่แสดงการทำหน้าที่ต่างกันของข้อสอบ ตามวิธีแมนเทล-แฮนส์เซล เพื่อนำไปใช้เป็นแนวทางในพิจารณาปรับปรุงเนื้อหาสาระของข้อสอบ ให้มีความยุติธรรมกับกลุ่มผู้สอบย่อยกลุ่มต่าง ๆ