วิธีการทางสถิติที่ใช้ตรวจสอบข้อสอบทำหน้าที่ต่างกัน

อาจารย์เสรี ชัดแช้ม

ภาควิชาวิจัยและวัดผลการศึกษา
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

ปัจจุบันวงการศึกษา ธุรกิจอุตสาหกรรม และการทหาร ใช้การสอบเพื่อคัดเลือกบุคคล ตรวจสอบความรู้ความสามารถของบุคคล สนับสนุนการเลื่อนตำแหน่ง และการออกใบรับรองหรือใบอนุญาตกันอย่างแพร่หลาย มีการนำผลการสอบของผู้เข้าสอบไปใช้ประกอบการพิจารณาตัดสินใจในเรื่องต่างๆ เหล่านี้เพิ่มมากขึ้น ทำให้ปัญหาเรื่องความไม่ยุติธรรมหรือความลำเอียงของข้อสอบหรือแบบสอบที่ใช้ในการสอบแต่ละครั้ง เป็นประเด็นสำคัญที่ใช้พิจารณาความตรงของแบบสอบ เนื่องจากข้อสอบหรือแบบสอบที่ลำเอียงเข้าข้างกลุ่มผู้เข้าสอบย่อยบางกลุ่มของผู้เข้าสอบทั้งหมด อาจทำให้ผู้เข้าสอบกลุ่มย่อยกลุ่มนั้นได้เปรียบผู้เข้าสอบกลุ่มย่อยกลุ่มอื่นๆ ทั้งๆ ที่สอบด้วยข้อสอบเดียวกันหรือแบบสอบฉบับเดียวกัน ผู้เข้าสอบในการสอบแต่ละครั้ง อาจจำแนกออกเป็นกลุ่มย่อยๆ ได้ ตามลักษณะที่แตกต่างกันในด้าน เชื้อชาติ เผ่าพันธุ์ เพศ ศาสนา ภาษา อายุ ประสบการณ์ ทุพพลภาพ หรือภูมิหลังอื่นๆ ของผู้เข้าสอบ ที่ทำให้กลุ่มผู้เข้าสอบย่อยๆ บางกลุ่มเกิดการเสียเปรียบ

การศึกษาเรื่องผลการสอบของกลุ่มผู้เข้าสอบย่อยของผู้เข้าสอบทั้งหมดมีมานานแล้ว แต่เรื่องความยุติธรรมในการสอบระหว่างผู้เข้าสอบกลุ่มย่อยๆ เพิ่งมีการศึกษากันอย่างจริงจังในช่วงปลายทศวรรษ 1960 โดยมีการเสนอวิธีการต่างๆ เพื่อนำไปใช้ตรวจสอบความลำเอียงของแบบสอบ (test-bias) หรือความลำเอียงในการคัดเลือกผู้เข้าสอบ (selection-bias) ขึ้นหลายวิธี ในระยะเวลาเดียวกันในช่วงเวลานั้น นักพัฒนาแบบสอบมีแนวโน้มที่จะใช้วิธีการจำแนกข้อสอบที่ไม่เหมาะสมกับผู้เข้าสอบกลุ่มย่อยบางกลุ่มออกจากแบบสอบก่อนจะพัฒนามาเป็นแบบสอบฉบับสมบูรณ์ ทำให้จำเป็นต้องพัฒนาวิธีการตรวจสอบความลำเอียงของข้อสอบ (item bias) ขึ้น เพื่อใช้เป็นแนวทางในการจำแนกข้อสอบที่ลำเอียงกับผู้เข้าสอบบางกลุ่มออกจากแบบสอบหรือคลังข้อสอบ ปัจจุบันการตรวจสอบความลำเอียงของข้อสอบเป็นส่วนหนึ่งของกระบวนการพัฒนาและการประเมินแบบสอบ เช่นเดียวกับการวิเคราะห์ข้อสอบและการตรวจสอบความเที่ยงของแบบสอบ

ในสมัยแรกๆ ของการศึกษาเรื่อง ผลการสอบเพื่อคัดเลือกคนเข้าศึกษาต่อหรือเข้าทำงานปรากฏชัดนี้ความลำเอียงกับกลุ่มคนต่างชาติ ต่างเพศ ทำให้ต้องมีการศึกษา *"ความลำเอียงในการคัดเลือกผู้เข้าสอบ (selection bias)"* ต่อมาเพื่อให้การศึกษาเรื่องนี้

มีความชัดเจนยิ่งขึ้น จึงได้ศึกษาในระดับข้อสอบ (item-level) ที่เรียกว่า *"ความลำเอียงของข้อสอบ (item bias)"* ซึ่งในปัจจุบันนักวิจัยส่วนใหญ่ใช้คำว่า *"ข้อสอบทำหน้าที่ต่างกันกับกลุ่มผู้เข้าสอบย่อยต่างกลุ่ม"* หรือ เรียกสั้นๆ ว่า *"ข้อสอบทำหน้าที่ต่างกัน (DIFferential Item Functioning: DIF)"* ทั้งนี้ เนื่องจากเห็นว่าเป็นคำที่มีความหมายกลางๆ จึงมีความเหมาะสมในเชิงวิชาการมากกว่าคำว่า *"ความลำเอียง (bias)"* ซึ่งเป็นคำที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ อย่างไรก็ตาม คำสองคำนี้มีจุดเน้นที่แตกต่างกัน โดยคำว่า ความลำเอียงของข้อสอบ เน้นที่อิทธิพลที่สังเกตได้ของกลุ่มผู้เข้าสอบย่อยที่มุ่งศึกษา ส่วนคำว่า ข้อสอบทำหน้าที่ต่างกัน เน้นที่ลักษณะทางสถิติของข้อสอบที่ตรวจสอบได้ด้วยวิธีวิเคราะห์ทางสถิติ ซึ่งเป็นส่วนประกอบหนึ่งของสิ่งที่แสดงถึงความลำเอียงของข้อสอบ (Scheuneman & Bleistein, 1989 ; Angoff, 1993 ; Hambleton & Others, 1993 ; Holland & Wainer, 1993 ; Zieky, 1993 ; Camilli & Shepard, 1994) จากจุดเน้นนี้แสดงให้เห็นว่า วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นเงื่อนไขจำเป็น (necessary condition) ในการประเมินความลำเอียงของข้อสอบ แต่ถ้าใช้เฉพาะวิธีการทางสถิติเพียงอย่างเดียว ผลการตรวจพบข้อสอบหน้าที่ต่างกันที่ได้ ก็ไม่อาจสรุปได้ว่าข้อสอบข้อนั้นลำเอียงหรือไม่ เนื่องจากการประเมินความลำเอียงของข้อสอบยังต้องรวมถึงการใช้วิธีให้ผู้เชี่ยวชาญพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบสอบ ที่เรียกว่า *"วิธีการตัดสินข้อสอบ (judgmental method)"* ก่อนที่เราจะสรุปว่าข้อสอบข้อนั้นลำเอียงหรือไม่ (Angoff,1993 ; Linn, 1993 ; Ramsay, 1993 ; Zieky, 1993 ; Camilli & Shepard, 1994)

อย่างไรก็ตาม ในความเป็นจริงแล้ว วิธีการทางสถิติ หรือ วิธีการตัดสินข้อสอบวิธีใดๆ ก็ไม่สามารถตรวจสอบ *"ความลำเอียง(bias)"* ได้ แต่วิธีการเหล่านี้ถูกนำมาใช้เพื่อพิจารณาว่า ข้อสอบแต่ละข้อในการสอบแต่ละครั้ง ทำหน้าที่ในทิศทางที่เหมือนกันหรือใช้กับกลุ่มผู้เข้าสอบย่อย ตั้งแต่ 2 กลุ่ม ได้หรือไม่เท่านั้น

Holland และ Thayer (1988) ได้เรียกชื่อกลุ่มผู้เข้าสอบย่อยที่นำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

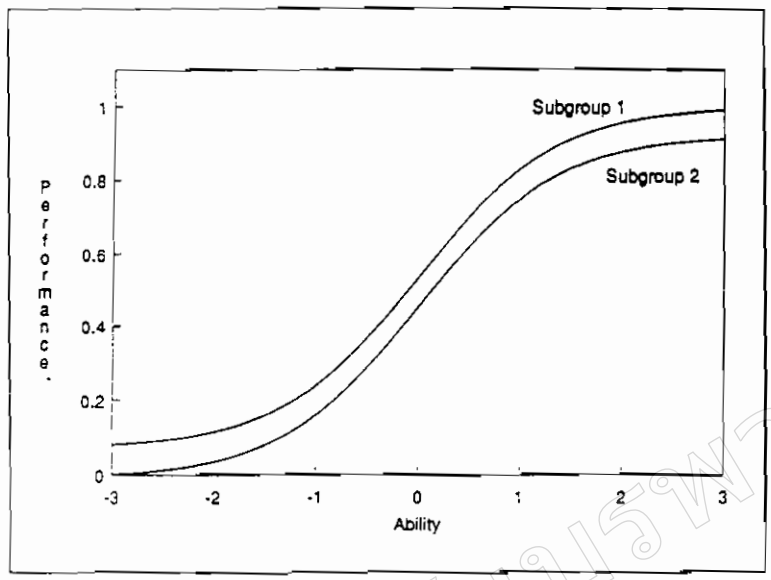
1. กลุ่มสนใจ (focal group) เป็นกลุ่มผู้เข้าสอบย่อยที่เชื่อว่าจะเสียเปรียบ ในกรณีที่ข้อสอบทำหน้าที่ต่างกัน

2. กลุ่มอ้างอิง (reference group) เป็นกลุ่มผู้เข้าสอบย่อยที่ใช้เป็นมาตรฐานในการเปรียบเทียบกับกลุ่มสนใจ เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น กลุ่มสนใจ ได้แก่ ผู้เข้าสอบผิวดำ ในขณะที่กลุ่มอ้างอิง ได้แก่ ผู้เข้าสอบผิวขาว เป็นต้น

มีผู้ให้ความหมายของคำว่า *"ข้อสอบทำหน้าที่ต่างกัน" (Differential Item Functioning : DIF)* ไว้หลายความหมาย แต่ความหมายที่เป็นที่ยอมรับกันอย่างกว้างขวางก็คือ ข้อสอบทำหน้าที่ต่างกันภายใต้เงื่อนไขผู้เข้าสอบมีความสามารถเท่ากัน แต่มาจากกลุ่มผู้เข้าสอบย่อยต่างกัน มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นถูกไม่เท่ากัน (Hambleton & Others, 1993)

Mellenbergh (1982) ได้จำแนกประเภทของข้อสอบทำหน้าที่ต่างกัน ออกเป็น 2 ประเภท ได้แก่

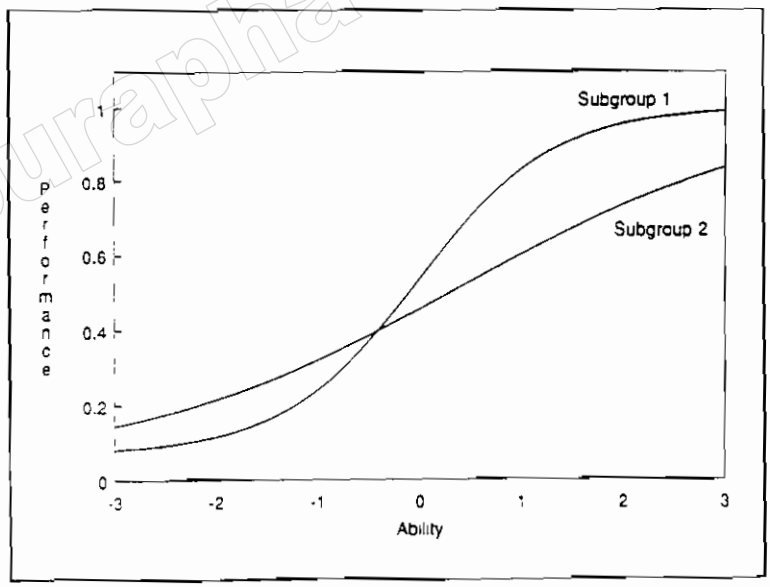
1. ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ (uniform DIF) หมายถึง ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้เข้าสอบย่อย 2 กลุ่ม คงเส้นคงวาในทุกระดับความสามารถของผู้เข้าสอบ ดังภาพที่ 1



ภาพที่ 1 ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ

จากภาพที่ 1 แสดงให้เห็นว่า ผลการตอบข้อสอบ (performance) ของผู้เข้าสอบกลุ่มย่อยที่สอง (subgroup 2) ต่ำกว่าผู้เข้าสอบกลุ่มย่อยที่หนึ่ง (subgroup 1) ในทุกๆ ระดับความสามารถของผู้เข้าสอบ (ability)

2 ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (nonuniform DIF) หมายถึง ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้เข้าสอบย่อย 2 กลุ่ม ไม่คงเส้นคงวาในทุกๆ ระดับความสามารถของผู้เข้าสอบ ดังภาพที่ 2



ภาพที่ 2 ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ

จากภาพที่ 2 แสดงให้เห็นว่า ผลการตอบข้อสอบ (performance)ของผู้เข้าสอบกลุ่มย่อยที่สอง (subgroup 2) ต่ำกว่าผู้เข้าสอบกลุ่มย่อยที่หนึ่ง (subgroup 1) ในช่วงระดับความสามารถ (ability) สูงๆ แต่ในช่วงระดับความสามารถ (ability) ต่ำๆ ผลการตอบข้อสอบ (performance) ของผู้เข้าสอบกลุ่มย่อยที่สอง (subgroup 2) กลับสูงกว่าผู้เข้าสอบกลุ่มย่อยที่หนึ่ง (subgroup 1)

การแบ่งกลุ่มวิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แบ่งได้หลายวิธี แต่ในบทความนี้ผู้เขียนใช้แนวทางของ Hambleton และคณะ (1993) ซึ่งจำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบออกเป็น 3 กลุ่มใหญ่ๆ ดังนี้

I. กลุ่มวิธีที่ใช้ทฤษฎีการสอบแบบดั้งเดิม (Methods Using Classical Test Theory)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มนี้พัฒนามาจากหลักการของทฤษฎีการสอบแบบดั้งเดิม โดยปกติแล้วใช้คะแนนที่สังเกตได้ของผู้เข้าสอบ (observed scores) แต่ละคนเป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อยและเปรียบเทียบค่าความยากของข้อสอบแต่ละข้อระหว่างกลุ่มผู้เข้าสอบย่อยเหล่านั้น วิธีการในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (analysis of variance) วิธีสหสัมพันธ์ (correlational methods) (Green & Draper, 1972, quoted in Scheuneman Bleistein,1989) วิธีแปลงค่าความยากของข้อสอบ (transformed item difficulty method) หรือ วิธีกำหนดจุดค่าเดลต้า (delta-plot method) (Angoff, 1982) การวิเคราะห์ตัวลวง (distractor analysis) (Scheuneman, 1982) วิธีสหสัมพันธ์บางส่วน (partial correlation methods)(Stricker, 1982) และวิธีการทำให้เป็นมาตรฐาน (standardization method)(Dorans & Kulick, 1983 , 1986)

ข้อได้เปรียบของวิธีในกลุ่มนี้คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายไม่สูงนัก ใช้ตรวจสอบกับกลุ่มตัวอย่างขนาดเล็กได้ และสามารถอธิบายให้คนทั่วไปเข้าใจได้ง่าย ส่วนข้อเสียเปรียบก็คือ ค่าสถิติของข้อสอบเปลี่ยนไปตามกลุ่มตัวอย่าง เมื่อกลุ่มตัวอย่างเปลี่ยนไป ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันก็เปลี่ยนไป ทำให้การอ้างอิงผลการศึกษาไปยังกลุ่มประชากรอาจมีความเชื่อถือได้น้อยลง

II. กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Methods Using Item Response Theory)

วิธีการในกลุ่มนี้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามกรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบ โดยปกติแล้วใช้การเปรียบเทียบโค้งลักษณะข้อสอบ (item characteristic curves: ICCs) ของกลุ่มผู้เข้าสอบย่อยตามระดับความสามารถของผู้เข้าสอบ ถ้าโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่ม มีรูปร่างเหมือนกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่มมีรูปร่างต่างกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน ค่าพารามิเตอร์ของโค้งลักษณะข้อสอบ ได้แก่ ค่าความยากของข้อสอบ (item difficulty, b-parameter) ค่าอำนาจจำแนกของข้อสอบ (item discrimination, a-parameter) และค่าการเดาข้อสอบ (pseudo guessing parameter, c-parameter) วิธีการในกลุ่มนี้ได้แก่ Analysis of fit method (Durovic, 1975, quoted in Hambleton & Others 1993) Difficulty shift method (Wright, Meac, & Draba, 1976, quoted

in Hambleton & Others, 1993) IRT area method (Ironson & Subkoviak, 1979) Two-stage method (Lord, 1980) และ Plot method (Hambleton & Rogers, 1991, quoted in Hambleton & Others, 1993)

ข้อได้เปรียบของวิธีการในกลุ่มนี้คือ การแก้ไขข้อบกพร่องของทฤษฎีการสอบแบบดั้งเดิม ทำให้ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่างที่สุ่มมาจากประชากรเดียวกัน การประมาณค่าความสามารถของผู้สอบเป็นอิสระจากค่าความยากของแบบสอบ (test difficulty) โมเดลทางคณิตศาสตร์ง่ายต่อการจับคู่โค้งลักษณะข้อสอบตามระดับ ความสามารถของผู้เข้าสอบ ทำให้สามารถศึกษาความแตกต่างของผลการตอบข้อสอบตามระดับความสามารถของกลุ่มผู้เข้าสอบย่อยได้ ไม่ต้องมีข้อตกลงเบื้องต้นเรื่องแบบสอบคู่ขนานในการหาค่าสัมประสิทธิ์ความเที่ยงของแบบสอบ (reliability coefficient) และถ้าผลการตอบข้อสอบของกลุ่มผู้เข้าสอบสอดคล้องกับข้อตกลงเบื้องต้นของโมเดล IRT (Item Response Theory) แล้ว ก็น่าจะเป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้ผลดี เนื่องจากเป็นวิธีการที่มีทฤษฎีการตอบสนองข้อสอบสนับสนุน และใช้การประมาณค่าความสามารถที่แท้จริงของผู้เข้าสอบ (true ability estimates) แทนคะแนนที่สังเกตได้ (observed score) ดังเช่นที่ใช้ในกลุ่มวิธีที่ใช้ทฤษฎีการสอบแบบดั้งเดิม ส่วนข้อเสียเปรียบของวิธีการในกลุ่มนี้ คือกระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสลับซับซ้อน เสียค่าใช้จ่ายในการวิเคราะห์ข้อมูลสูง และต้องใช้กับกลุ่มตัวอย่างขนาดใหญ่

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

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในกลุ่มนี้ บางครั้งก็เรียกว่ากลุ่มวิธีไค-สแควร์ (chi-square methods) วิธีในกลุ่มนี้ใช้ค่าไค-สแควร์เป็นดัชนีแสดงการทำหน้าที่ต่างกันของข้อสอบ และใช้คะแนนของแบบสอบ (test score) หรือคะแนนของแบบสอบที่ทำให้บริสุทธิ์ (purified test score) เป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อยๆ ก่อนการเปรียบเทียบผลการตอบข้อสอบ วิธีในกลุ่มนี้ได้แก่ วิธีตารางการณ์จร (contingency table method) (Scheuneman, 1975 , 1979) วิธีตารางการณ์จรปรับขยาย (modified contingency table method) (Veale,1977, quoted in Hambleton & Others, 1993) วิธีล็อก-ลิเนียร์ (log-linear methods) (Mellenbergh, 1982) วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel method) (Holland & Thayer, 1986 , 1988) และ วิธีการถดถอยโลจิสติก (logistic regression method) (Swaminathan & Rogers, 1990)

ข้อได้เปรียบของวิธีในกลุ่มนี้คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายในการวิเคราะห์ข้อมูลไม่สูง ใช้ได้กับกลุ่มตัวอย่างขนาดไม่ใหญ่นัก และบางวิธีมีหลักการที่ดีในการจับคู่กลุ่มผู้เข้าสอบย่อยตามความสามารถของผู้เข้าสอบและมีการทดสอบนัยสำคัญ ส่วนข้อเสียเปรียบของวิธีในกลุ่มนี้ก็คล้ายๆ กับกลุ่มวิธีที่ใช้ทฤษฎีการสอบแบบดั้งเดิม (Methods Using Classical Test Theory)

จากที่กล่าวมาจะเห็นว่า วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี ในบทความนี้ผู้เขียนเสนอเฉพาะวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่พิจารณาเห็นว่าน่าจะนำมาใช้ในประเทศเรา โดยจะกล่าวอย่างกว้างๆ ในประเด็นเรื่องหลักการ ข้อดี ข้อด้อยของแต่ละวิธี

และความสัมพันธ์กับวิธีอื่น ส่วนการนำแต่ละวิธีไปใช้ตรวจสอบการทำหน้าที่ต่างกัน ผู้อ่านสามารถศึกษาจากเอกสารที่อ้างอิงไว้ท้ายนี้

1. วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty Method, TID)

วิธีแปลงค่าความยากของข้อสอบ หรือเรียกอีกชื่อหนึ่งว่า วิธีกำหนดจุดค่าเดลต้า (delta-plot method) ผู้พัฒนาวิธีนี้ที่สำคัญ ได้แก่ Angoff (1982) หลักการสำคัญของวิธีนี้คือ การคำนวณค่าความยากของข้อสอบ (ค่า p) แต่ละข้อจากกลุ่มผู้เข้าสอบย่อยสองกลุ่มแยกจากกัน แล้วแปลงค่า p ที่ได้ (แต่ละข้อมีค่า p 2 ค่า) ให้เป็นค่าความยากมาตรฐานของข้อสอบ (ค่า delta) ซึ่งมีค่าเฉลี่ยเท่ากับ 13 (mean = 13) และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 4 (sd = 4) หลังจากนั้นนำค่าเดลต้าแต่ละคู่ไปกำหนดจุดคู่ลำดับบนกราฟ โดยปกติแล้วจุดคู่ลำดับของค่าเดลต้าบนกราฟของกลุ่มผู้เข้าสอบย่อยสองกลุ่มที่มีความสามารถเท่าเทียมกัน จะกระจายเป็นรูปวงรีรอบๆ แนวเส้นแกนหลัก (เส้นตรงที่ทำมุม 45 องศา ผ่านจุดกำเนิดของกราฟ) ซึ่งแสดงว่าข้อสอบยากกับกลุ่มผู้เข้าสอบย่อยสองกลุ่มเท่ากัน แต่ถ้ากลุ่มผู้เข้าสอบย่อยที่นำมาเปรียบเทียบกันมีความสามารถแตกต่างกัน จุดคู่ลำดับของค่าเดลต้าบนกราฟจะกระจายเบี่ยงเบนออกไปจากแนวเส้นแกนหลัก ซึ่งแสดงว่าข้อสอบยากกับกลุ่มผู้เข้าสอบย่อยสองกลุ่มไม่เท่ากัน ดัชนีแสดงข้อสอบทำหน้าที่ต่างกันพิจารณาจาก 1) ค่าระยะห่างตั้งฉากจากคู่ลำดับของค่าเดลต้าของข้อสอบแต่ละข้อไปยังเส้นแกนหลักของรูปวงรี 2) ส่วนเบี่ยงเบนมาตรฐานของค่าระยะห่างนี้ 3) ความแตกต่างระหว่างค่าเดลต้าของกลุ่มผู้เข้าสอบย่อยสองกลุ่ม ส่วนรายละเอียดของหลักเกณฑ์การพิจารณาข้อสอบทำหน้าที่ต่างกันมีผู้เสนอไว้แตกต่างกันไป

ข้อดีของวิธีนี้คือ การคำนวณไม่ยุ่งยาก เสียค่าใช้จ่ายไม่แพง เข้าใจง่าย และใช้กับกลุ่มตัวอย่างขนาดเล็กได้ ส่วนข้อด้อยก็คือ ถ้ากลุ่มผู้เข้าสอบย่อยสองกลุ่มที่นำมาเปรียบเทียบมีความสามารถต่างกัน อาจส่งผลให้ค่าความยากของข้อสอบซึ่งใช้เป็นดัชนีแสดงข้อสอบทำหน้าที่ต่างกันเชื่อถือได้น้อยลง หรือในกรณีที่ข้อสอบมีค่าอำนาจจำแนกต่ำเกินไป จะมีผลทำให้ความแตกต่างของค่าความยากของข้อสอบผิดปกติ หรือในกรณีที่ขนาดของกลุ่มผู้เข้าสอบย่อยต่างกัน ก็ทำให้ค่าเดลต้าที่นำมาเปรียบเทียบมีความเที่ยงไม่เท่ากัน อย่างไรก็ตามได้มีผู้เสนอเทคนิคต่างๆ เพื่อแก้ไขข้อด้อยเหล่านี้ไว้หลายคน

วิธีนี้ได้รับคำแนะนำให้ใช้กับกลุ่มตัวอย่างขนาดเล็ก หรือกรณีที่หาวิธีอื่นที่เหมาะสมไม่ได้ หรือกรณีที่ต้องการหาสารสนเทศเบื้องต้นหรือเพื่อการวิจัย อย่างไรก็ตามถ้าผู้ที่นำไปใช้แปลผลการตรวจสอบการทำหน้าที่ของข้อสอบอย่างระมัดระวัง ก็เป็นวิธีที่ได้ประโยชน์วิธีหนึ่ง (Scheuneman & Bleistein, 1989)

2. วิธีการทำให้เป็นมาตรฐาน (Standardization Method)

Dorans และ Kulick (1983 ; 1986) เป็นผู้พัฒนา วิธีการทำให้เป็นมาตรฐาน (standardization method) ขึ้น โดยพื้นฐานแล้ววิธีนี้เป็นวิธีเชิงบรรยายและไม่มีการทดสอบนัยสำคัญ หลักการสำคัญของวิธีการนี้ คือ การเปรียบเทียบการถดถอยระหว่างคะแนนข้อสอบ กับ คะแนนแบบสอบ (item-test regression) ของผลการตอบข้อสอบกลุ่มฐาน (base group) กับกลุ่มสนใจ (focal group) โดยปกติแล้วกลุ่มฐานหรือกลุ่มอ้างอิง เป็นผู้เข้าสอบกลุ่มใหญ่กว่ากลุ่มสนใจ วิธีการนี้คล้ายกับวิธีตารางการณ์จร และวิธี

แมนเทล-แฮนส์ เซล ตรงที่ใช้หลักความแตกต่างระหว่างสัดส่วนการตอบข้อสอบถูกที่ควรจะเป็นกับที่สังเกตได้ ระหว่างผู้เข้าสอบกลุ่มฐานและกลุ่มสนใจในแต่ละระดับชั้นคะแนน นอกจากนี้ วิธีนี้ยังคล้ายกับกลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ ตรงที่ใช้หลักการถดถอยระหว่างคะแนนข้อสอบกับคะแนนแบบสอบ ซึ่งคล้ายกับโค้งลักษณะข้อสอบ (item characteristic curve) ในโมเดล IRT

วิธีการทำให้เป็นมาตรฐานใช้การประมาณค่าความน่าจะเป็นในการตอบข้อสอบถูกของผู้เข้าสอบกลุ่มฐาน และกลุ่มสนใจ ในแต่ละระดับชั้นคะแนน การแบ่งระดับชั้นคะแนนอาจแบ่งตามคะแนนรวมของแบบสอบฉบับนั้น

วิธีนี้ให้ดัชนีข้อสอบทำหน้าที่ต่างกัน 2 ค่า คือ ดัชนีชนิดคิดเครื่องหมาย (signed index) และดัชนีชนิดไม่คิดเครื่องหมาย (unsigned index) การคำนวณดัชนีทั้งสองค่าใช้หลักการถ่วงน้ำหนักโดยการทำให้คะแนนของกลุ่มฐานและกลุ่มสนใจเป็นคะแนนมาตรฐาน ในทางปฏิบัติการถ่วงน้ำหนักใช้จำนวนผู้เข้าสอบกลุ่มสนใจ (focal group) ในแต่ละระดับชั้นคะแนน เนื่องจากความแตกต่างระหว่างความน่าจะเป็นในการตอบข้อสอบถูกของผู้เข้าสอบกลุ่มฐานกับกลุ่มสนใจจะมีน้ำหนักมากที่สุด เมื่อถ่วงน้ำหนักจากจำนวนผู้เข้าสอบกลุ่มสนใจ

ดัชนีชนิดคิดเครื่องหมาย (signed index) คือ ความแตกต่างของค่าความยากของข้อสอบแต่ละข้อระหว่างกลุ่มฐานกับกลุ่มสนใจ เมื่อทำให้เป็นมาตรฐานแล้ว (STD P-DIF) ส่วนดัชนีชนิดไม่คิดเครื่องหมาย (unsigned index หรือ root-mean-weighted-squared difference: RMWSD) คือ ความแตกต่างของค่ารากกำลังสองถ่วงน้ำหนักเฉลี่ยระหว่างผู้เข้าสอบกลุ่มฐานกับกลุ่มสนใจ เมื่อคำนวณโดยไม่คิดเครื่องหมายบวกและลบ ซึ่งอาจใช้การวิเคราะห์ความคลาดเคลื่อน (residual) ของการตัดกันของโค้งลักษณะข้อสอบ (ICCs) (ปฏิสัมพันธ์ระหว่างความเป็นสมาชิกของกลุ่มผู้เข้าสอบ กับ ระดับความสามารถของผู้เข้าสอบ) หรือการวิเคราะห์ส่วนที่เหลือ (remaining) หลังการปรับค่า STD P-DIF

ข้อดีของวิธีนี้ก็คือ คำนวณง่าย เสียค่าใช้จ่ายในการคำนวณไม่สูงนัก เข้าใจได้ง่าย สามารถนำไปใช้อธิบายธรรมชาติของข้อสอบทำหน้าที่ต่างกันได้ และการจับคู่ผู้เข้าสอบใช้การแบ่งคะแนนของแบบสอบทั้งฉบับออกเป็นช่วงละหนึ่งหน่วยคะแนน (unit interval) เช่นเดียวกับวิธีแมนเทล-แฮนส์เซล ส่วนข้อด้อยก็คือ ต้องใช้กับกลุ่มตัวอย่างค่อนข้างใหญ่

วิธีนี้ค่อนข้างใหม่และน่าสนใจ ปัจจุบันได้พัฒนาให้สามารถใช้ตรวจสอบการทำหน้าที่ต่างกันกับตัวเลือกทุกตัว ซึ่งเรียกว่าวิธี Complehensive DIF หรือ CDIF ทำให้เป็นประโยชน์ในการตรวจสอบการทำหน้าที่ต่างกันของตัวลวง (differential functioning of distractor) และยังนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับแบบสอบที่วัดความเร็วในการทำข้อสอบ (speedness test) ได้อีกด้วย (Dorans & Others, 1992)

วิธีนี้ใช้กับกลุ่มตัวอย่างค่อนข้างใหญ่ ทั้งนี้เพื่อแก้ปัญหาความคลาดเคลื่อนในการสุ่มตัวอย่าง (sampling error) โดยเฉพาะในกรณีที่ต้องการหาค่าดัชนีชนิดไม่คิดเครื่องหมาย

3. วิธีหาพื้นที่ระหว่างโค้งลักษณะข้อสอบ (IRT Area Method)

กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบที่นิยมใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้แก่ วิธี IRT Area (Ironson & Subkoviak, 1979) ซึ่งโดยปกติแล้วใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 2 หรือ 3 พารามิเตอร์ หลักการสำคัญของวิธีนี้คือ

การเปรียบเทียบโค้งลักษณะข้อสอบของกลุ่มผู้สอบย่อยสองกลุ่มและใช้พื้นที่ระหว่างโค้งลักษณะข้อสอบนั้นเป็นดัชนีแสดงข้อสอบทำหน้าที่ต่างกัน การประมาณค่าพารามิเตอร์ของกลุ่มผู้เข้าสอบย่อยแต่ละกลุ่มแยกออกจากกัน หลังจากนั้นจึงแปลงค่าพารามิเตอร์เหล่านั้นให้อยู่ในมาตรเดียวกัน (common metric) เพื่อนำมาเปรียบเทียบกัน กลุ่มวิธีนี้มีข้อตกลงเบื้องต้นที่สำคัญ คือ คะแนนผลการตอบข้อสอบต้องมาจากแบบสอบที่มุ่งวัดลักษณะที่สำคัญของผู้สอบเพียงลักษณะเดียว (unidimensional) และโมเดลโลจิสติก แบบ 2 หรือ 3 พารามิเตอร์ สามารถใช้แทนข้อมูลผลการตอบข้อสอบได้อย่างพอเพียง

วิธีนี้ยอมรับว่า ข้อสอบที่มีโค้งลักษณะข้อสอบแตกต่างกันระหว่างกลุ่มผู้สอบย่อยสองกลุ่ม แสดงว่าข้อสอบทำหน้าที่ต่างกัน ได้มีผู้สร้างดัชนีที่แสดงความแตกต่างของโค้งลักษณะข้อสอบและทดสอบนัยสำคัญของความแตกต่างนี้ขึ้นหลายวิธีอย่างไรก็ตาม วิธีในกลุ่มนี้แตกต่างกันไม่ชัดเจนนัก (Scheuneman & Bleistein, 1989)

วิธีที่เสนอโดย Rudner (1977, quoted in Scheuneman & Bleistein, 1989) ประมาณค่าพารามิเตอร์ของข้อสอบแยกกันเป็น 2 กลุ่ม แล้วแปลงค่าที่ได้ให้อยู่ในสเกลเดียวกัน ส่วนวิธี Two-Stage ที่เสนอโดย Lord (1980) ในครั้งแรกประมาณค่าพารามิเตอร์ของข้อสอบโดยใช้ผู้เข้าสอบทั้งหมดเพื่อประมาณ ค่าการเดาข้อสอบ (c-parameter) หลังจากนั้นจึงประมาณค่า a และ b พารามิเตอร์ของกลุ่มผู้เข้าสอบย่อยสองกลุ่มแยกจากกัน แล้วจึงแปลงค่าที่ได้ให้อยู่ในสเกลเดียวกัน และใช้ค่าไค-สแควร์ ในการเปรียบเทียบค่า a และ b พารามิเตอร์ของกลุ่มผู้เข้าสอบย่อยสองกลุ่ม อย่างไรก็ตามวิธีทั้งสองนี้ก็มีแหล่งความคลาดเคลื่อนที่เกิดจากการปรับเทียบสเกล

ในกรณีข้อมูลสอดคล้องกับข้อตกลงเบื้องต้นของโมเดล IRT วิธีนี้มีข้อดี คือ มีทฤษฎีที่น่าเชื่อถือรองรับ และใช้การประมาณค่าความสามารถที่แท้จริงของผู้เข้าสอบ แทน คะแนนที่สังเกตได้ ในการศึกษาโดยสถานการณ์จำลอง (simulation study) เพื่อประเมินผลการใช้วิธีการต่างๆ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มักพบว่าวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ดีกว่าวิธีอื่นๆ อย่างไรก็ตาม ผลการวิจัยที่ได้อาจจะขึ้นอยู่กับการสร้างข้อมูลจำลองให้สอดคล้องกับโมเดล IRT มากเกินไปก็ได้ ส่วนข้อด้อย คือ การคำนวณยุ่งยาก เสียค่าใช้จ่ายในการคำนวณสูง และต้องการกลุ่มตัวอย่างขนาดใหญ่ ส่วนในประเด็นเรื่อง ขนาดของกลุ่มตัวอย่างอย่างน้อยที่สุดเท่าใด ขึ้นอยู่กับความสอดคล้องระหว่างข้อมูล กับข้อตกลงเบื้องต้นและประเภทของโมเดล IRT ที่ใช้ เช่น ถ้าจะใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ให้ได้ผลแล้วอาจต้องใช้กลุ่มตัวอย่าง ขนาด 1,000 คนต่อกลุ่ม และแบบสอบ 40 ข้อ เป็นต้น ในบางกรณีก็ไม่สามารถประมาณค่าการเดาข้อสอบ (c-parameter) ได้ ต้องกำหนดค่าการเดาข้อสอบ เป็นค่าร่วมกัน (common value) สำหรับข้อสอบทุกๆ ข้อ

4. วิธีตารางการณ์จร (Contingency Table Method)

วิธีนี้เสนอโดย Scheuneman (1979) มีหลักการที่สำคัญคือ ถ้าข้อสอบทำหน้าที่ไม่ต่างกัน ผู้เข้าสอบที่มีความสามารถเท่ากันมีโอกาสที่จะตอบข้อสอบถูกเท่ากัน โดยไม่คำนึงถึงความเป็นสมาชิกของกลุ่มผู้เข้าสอบย่อยๆ การคำนวณค่าดัชนีใช้ตารางการณ์จร แบบ 2 ทาง (ความเป็นสมาชิกของกลุ่มผู้เข้าสอบย่อย X ระดับความสามารถของผู้เข้าสอบ) โดยระดับความสามารถของผู้เข้าสอบแทน

ด้วยช่วงของคะแนนจากแบบสอบทั้งฉบับ หรือแบบสอบย่อยซึ่งมีข้อสอบที่ต้องการตรวจสอบรวมอยู่ด้วย

ต่อมาวิธีนี้ได้ใช้ค่าไค-สแควร์ (chi-square) เป็นดัชนีแสดงการทำหน้าที่ต่างกันของข้อสอบ ค่าไค-สแควร์คำนวณโดยการแบ่งคะแนนที่ได้จากแบบสอบออกเป็นช่วงๆ ซึ่งใช้แทนระดับความสามารถของผู้เข้าสอบ โดยปกติจะแบ่งคะแนนจากแบบสอบออกเป็น 3-5 ช่วง แล้วคำนวณสัดส่วนการตอบข้อสอบถูกของกลุ่มผู้เข้าสอบย่อยแต่ละกลุ่มเปรียบเทียบกันในแต่ละช่วงคะแนน

วิธีการนี้อาจขยายไปใช้กับกลุ่มผู้เข้าสอบย่อยหลายๆ กลุ่มพร้อมๆ กันได้ เช่น ถ้าต้องการเปรียบเทียบกลุ่มผู้เข้าสอบย่อย 4 กลุ่ม และข้อสอบมีค่าดัชนีสูงเหมือนกันในกลุ่มผู้เข้าสอบย่อย 3 กลุ่มแสดงว่ากลุ่มผู้เข้าสอบย่อยที่เหลือแตกต่างกันออกไป หรืออาจกล่าวได้ว่า กลุ่มผู้เข้าสอบย่อยที่เหลือได้เปรียบหรือเสียเปรียบในข้อสอบข้อนั้น

วิธีตารางการณ์จรมีข้อดี คือ คำนวณง่าย เสียค่าใช้จ่ายไม่แพง เหมาะกับกลุ่มตัวอย่างขนาดเล็ก และสามารถตรวจสอบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้ มักใช้เป็นทางเลือกในกรณีที่ มีข้อจำกัดในด้านอุปกรณ์การคำนวณ ส่วนข้อด้อยก็คือ ค่าสถิติที่ได้ไม่คงที่ ขึ้นอยู่กับเทคนิคการเลือกคะแนนที่ใช้แทนระดับความสามารถของผู้เข้าสอบ

5. วิธีล็อก-ลิเนียร์ (Log-linear Methods)

วิธีล็อก-ลิเนียร์พัฒนามาจากวิธีไค-สแควร์แบบดั้งเดิม ผู้พัฒนาวิธีนี้ที่สำคัญ ได้แก่ Mellenbergh (1982) หลักการสำคัญก็คือ การทดสอบความแตกต่างของสัดส่วนการตอบข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่ม ในโมเดลล็อก-ลิเนียร์ มีขั้นตอนการดำเนินการดังนี้

ขั้นที่ 1 การแจกแจงผลการตอบข้อสอบแต่ละข้อ (ทั้งการตอบถูกและผิด) ลงในตารางการณ์จรแบบ 3 ทาง (ระดับความสามารถของผู้เข้าสอบ X กลุ่มผู้เข้าสอบ X ผลการตอบข้อสอบ) โดยระดับความสามารถของผู้เข้าสอบใช้การแบ่งคะแนนของแบบสอบทั้งฉบับออก เป็นช่วงๆ เช่นเดียวกับวิธีตารางการณ์จร

ขั้นที่ 2 กำหนดข้อมูลจำเพาะของโมเดล

ขั้นที่ 3 การคำนวณค่าสถิติวัดระดับความกลมกลืนระหว่างข้อมูลเชิงประจักษ์ กับ โมเดล เช่น การทดสอบอัตราส่วนไลค์ลิฮูดของค่าไค-สแควร์ (likelihood ratio of chi-square (G2)) เป็นต้น

ขั้นที่ 4 การทดสอบความแตกต่างอย่างมีนัยสำคัญระหว่างโมเดล (ค่า G2)

ในขั้นที่ 4 มีการทดสอบโมเดลทางคณิตศาสตร์ไม่อิมตัว (nonsatured models) 3 โมเดล ดังนี้ โมเดลที่หนึ่ง ทดสอบผลหลัก (main effect) ของระดับความสามารถของผู้เข้าสอบโมเดลที่สอง ทดสอบผลหลักของกลุ่มผู้เข้าสอบย่อย โมเดลที่สาม ทดสอบปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้เข้าสอบกับกลุ่มผู้เข้าสอบย่อย ถ้าโมเดลที่หนึ่งสอดคล้องกับข้อมูล แสดงว่า ข้อสอบทำหน้าที่ไม่ต่างกัน ถ้าโมเดลที่สองสอดคล้องกับข้อมูลดีกว่าโมเดลที่หนึ่งอย่างมีนัยสำคัญ แสดงว่า ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ (uniform DIF) และถ้าโมเดลที่สามสอดคล้องกับข้อมูล แสดงว่า ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (non- uniform DIF)

ข้อดีของวิธีนี้ ได้แก่ สามารถตรวจสอบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ และใช้กับกลุ่มตัวอย่างขนาดเล็กได้ ส่วนข้อด้อยได้แก่ การคำนวณค่าสถิติค่อนข้างยุ่งยากและเสียค่าใช้จ่ายสูง

วิธีนี้ได้รับการแนะนำให้ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในกรณีที่สนใจ การทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ (Scheuneman & Bleistein.1989)

6. วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel (MH) Method)

Mantel และ Haenszel (1959) ได้พัฒนาและ นำวิธีนี้ไปใช้ในงานวิจัยทางการแพทย์ตั้งแต่ปี ค.ศ. 1959 ต่อมา Holland และ Thayer (1986 ; 1988) ได้แนะนำให้นำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีแมนเทล-แฮนส์เซล เป็นวิธีตารางการณ์จรแบบไม่คำนวณซ้ำทวน ใช้หลักการประมาณและทดสอบค่าพารามิเตอร์ขององค์ประกอบร่วม 2 องค์ประกอบที่สัมพันธ์กัน ในตารางการณ์จร แบบ 3 ทาง (ผลการตอบข้อสอบ X กลุ่มผู้เข้าสอบ X ระดับความสามารถของผู้เข้าสอบ) ระดับความสามารถของผู้เข้าสอบใช้การแบ่งคะแนนของแบบสอบทั้งฉบับออกเป็นช่วงละหนึ่งหน่วยคะแนน (unit intervals) ทั้งนี้เพื่อหลีกเลี่ยงความคลาดเคลื่อนในการจัดกลุ่มช่วงคะแนน

หลักการสำคัญของวิธีนี้คือ การเปรียบเทียบผลการตอบข้อสอบ (ทั้งตอบถูกและตอบผิด) ของผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มสนใจ แล้วจำแนกข้อสอบที่แสดงว่าทำหน้าที่ต่างกันกับผู้เข้าสอบกลุ่มอ้างอิงหรือกลุ่มสนใจออกมา โดยวิเคราะห์ในทุกๆ ช่วงหนึ่งหน่วยคะแนน (unit intervals) ข้อสอบข้อใดที่ผู้เข้าสอบทั้งสองกลุ่มทำคะแนนได้เท่าๆ กัน แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน ในระหว่างผู้เข้าสอบทั้งสองกลุ่ม

ข้อดีของวิธีนี้ คือ คำนวณง่าย เสียค่าใช้จ่ายไม่แพง ใช้ได้กับกลุ่มตัวอย่างขนาดเล็กกว่ากลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ ส่วนข้อด้อยของวิธีนี้คือ ไม่ไวต่อการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ อย่างไรก็ตาม ขณะนี้อยู่ในระหว่างการพัฒนาเชิงวิธีการ เพื่อให้สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้ดียิ่งขึ้น

วิธีนี้ค่อนข้างใหม่ และได้รับความนิยมในทางปฏิบัติ หน่วยงานบริการทดสอบทางการศึกษาแห่งสหรัฐอเมริกา (Educational Testing Service: ETS) ได้แนะนำให้ใช้วิธีนี้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

บทสรุป

ดังได้กล่าวมาแล้วว่า วิธีการทางสถิติที่ใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบมีจุดเด่นจุดด้อยแตกต่างกันไป ดังนั้นการเลือกใช้วิธีการทางสถิติที่เหมาะสมเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จึงควรคำนึงถึงเรื่องต่อไปนี้

1. ขนาดของกลุ่มตัวอย่างและอุปกรณ์ที่ใช้ในการคำนวณ กลุ่มตัวอย่างเป็นประเด็นสำคัญ ในการเลือกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น ถ้ากลุ่มตัวอย่างต่ำกว่า 100 คน วิธีที่ได้ผลดีอาจได้แก่ วิธีแปลงค่าความยากของข้อสอบ หรือ วิธีตารางการณ์จร แต่ถ้ากลุ่มตัวอย่าง 200 คนขึ้นไป เราก็สามารถใช้วิธีแมนเทล-แฮนส์เซล หรือ วิธีการทำให้เป็นมาตรฐาน ได้ ส่วนกลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ อาจต้องใช้กลุ่มตัวอย่างขนาด 1,000 คน และต้องใช้คอมพิวเตอร์ที่มีประสิทธิภาพสูงในการคำนวณ

2. ความสำคัญของผลการตรวจสอบและความแม่นยำที่ต้องการ การตัดสินใจในเรื่องนี้ควรพิจารณาว่าต้องการผลการตรวจสอบไปใช้ทำอะไร เช่น ถ้าผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโครงการทดสอบนั้นๆ นำไปใช้ตัดสินชะตาชีวิตของผู้เข้าสอบ จำเป็นต้องใช้ผลการตรวจสอบ

ที่มีความแม่นยำสูง ก็อาจใช้วิธี IRT area แบบ 3 พารามิเตอร์ ในกรณีที่ข้อมูลสอดคล้องกับข้อตกลงเบื้องต้นของโมเดล IRT

3. ความสนใจของกลุ่มผู้ใช้ข้อมูล กลุ่มผู้ใช้ข้อมูลสนใจศึกษาข้อสอบทำหน้าที่ต่างกันประเภทใด เช่น ถ้าเชื่อว่ามีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้เข้าสอบกับความเป็นสมาชิกของกลุ่มผู้เข้าสอบย่อยๆ หรือ สนใจข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (non uniform DIF) เราก็อาจเลือกวิธี ล็อก-ลิเนียร์ หรือกลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ

อย่างไรก็ตาม วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทุกวิธีมีข้อจำกัดอยู่บ้าง และไม่มีวิธีการทางสถิติวิธีใดที่สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันในแบบสอบได้ทุกข้อ ดังนั้นในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโครงการสำคัญๆ จึงควรใช้หลายๆ วิธี จะเหมาะสมกว่าการใช้วิธีใดวิธีหนึ่งเพียงวิธีเดียว

บรรณานุกรม

กาญจนา วัธนสุนทร 2538 *การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ* วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต ภาควิชาวิจัยการศึกษา บัณฑิตวิทยาลัยจุฬาลงกรณ์มหาวิทยาลัย

ชัชชัย เผ่าพงศ์ 2527 *การวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์และภาษาระดับมัธยม ศึกษาตอนต้น* วิทยานิพนธ์ศึกษาศาสตรมหาบัณฑิต ขอนแก่น มหาวิทยาลัยขอนแก่น

ทัศนีย์ พรมนตรี 2530 *การวิเคราะห์ความลำเอียงของแบบสอบวิชาคณิตศาสตร์โครงการตรวจสอบคุณภาพการศึกษา ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526* วิทยานิพนธ์ครุศาสตรมหาบัณฑิต ภาควิชาวิจัยการศึกษา บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

นิรมล ชัยชวลิต 2537 *การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของแบบสอบความเข้าใจในการอ่านภาษาไทย ตามทฤษฎีคลาสลิคอลที่ใช้วิธีวิเคราะห์ต่างกัน* วิทยานิพนธ์การศึกษามหาบัณฑิต มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร

พัชรี ปิยภัณฑ์ 2531 *การวิเคราะห์ความลำเอียงของข้อสอบจากแบบสอบวัดผลสัมฤทธิ์ทางการเรียน ชั้นประถมศึกษาปีที่ 6* วิทยานิพนธ์การศึกษามหาบัณฑิต มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร

สุพัฒน์ สุกมลสันต์ 2534 *การวิเคราะห์ความลำเอียงของข้อทดสอบภาษาอังกฤษเข้ามหาวิทยาลัย ปี 2531 2533* กรุงเทพฯ สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย

สุรศักดิ์ อมรรัตนศักดิ์ 2531 *การศึกษาเปรียบเทียบผลของวิธีวิเคราะห์ความลำเอียงที่ต่างกัน 4 วิธี* วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต ภาควิชาวิจัยการศึกษา บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

Angoff. W H 1982. The use of difficulty and discrimination indices in the identification of biased test items. In R.A Berk (ed). *Handbook of methods for detect-*

ing test bias, pp. 96-116. Baltimore, MD: Johns Hopkins University Press.

Angoff, W.H. 1993. Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (eds.), *Differential item functioning*, pp. 3-23 Hillsdale, NJ: Lawrence Erlbaum Associates.

Camilli, G., & Shepard, L.A. 1994. *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Dorans, N.J., & Kulick, E. 1983. *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977. An application of the standardization approaches*. (Research Rep. No.83-9) Princeton, NJ: Educational Testin Service.

Dorans, N.J., & Kulick, E. 1986. Demonstrating the utilit of the standardization approachto assessing unexpected differential item performance on the Scholastic Aptitude Test *Journal of Educational Measurement* 23 (4) : 355-368

Dorans, N.J., Schmitt, A.P., & Bleistein, C.A. 1992. The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement 29(4)*: 309-319.

Hambleton, R.K., Clauser, B.E., Mazor, K.M., & Jones, R.W. 1993. *Advanced in the detection of differentially functioning test items*. (Research Rep. No 237) Amherst, MA: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluative.

Holland, P.W., & Thayer, D.T. 1986. *Differential item functioning and the Mantel-Haenszel procedure*. (Research Rep. No. 86-69) Prinction, NJ: Educational Testing Service

Holland, P.W. & Thayer, D.T. 1988. Differential item functioning and the Mantel-Haenszel procedure In P.W. Wainer & H T. Braun (eds.), *Test validity*, pp 129-145. Hillsdale, NJ Lawrence Erlbaum Associates.

Holland, P.W., & Wainer, H., eds. 1993 *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbuam Associates.

Ironson, G.H., & Subkoviak, M.J. 1979. A comparison of several methods of assessing item bias. *Journal of Educational Measurement* 18: 209-225.

Linn, R.L. 1993. The use of diferential item functioning statistics: A discussion of current practice and future implimentation In P.W. Holland & H. Wainer (eds.), *Differential item functioning*, pp. 349-364. Hillsdale, NJ: Lawrence Erlbuam Associates.

Lord, F.M. 1980. *Applications of item response theory to practical testing problems*. Hillsdale, NJ. Lawrence Erlbuam Associates.

Mantel, N., & Haenszel, W 1959. Statistical

aspects of the analysis of data from retrospective studies of disease *Journal of the National Cancer Institute* 22 (4) . 719-748

Mellenbergh. G.J 1982 Contingency table models for assessing item bias *Journal of Educational Statistics* 7(2): 105-118.

Millsap, R.E , & Everson H T 1993 Methodological review statistical approaches forassessing measurement bias. *Applied Psychological Measurement* 17(4) 297-334

Ramsay. J.O. 1993. Comments on the monte carlo study of Donoghue. Holland. and Thayer In P W Holland & H Wainer (eds.). *Differential item functioning*. pp 167-169 HillsdaleNJ Lawrence Erlbuam Associates

Scheuneman. J.D 1975 (April) *A new method of assessing bias in test items* Paper presented at the meeting of the American Educational Research Association Washington. DC

Scheuneman. J.D 1979 A method of assessing bias in test items. *Journal of Educational Measurement* 16 (3) 143-152

Scheuneman. J.D 1982. A posteriori analyses of biased items. In R.A. Berk (ed.). *Handbook of methods for detecting test bias*. pp. 180-198. Baltimore. MD: Johns Hopkins University Press.

Scheuneman. J D . & Bleistein. C.A. 1989. A consumer's guide to statistics foridentifying differential item functioning. *Applied Measurement in Education* 2 (3) 255-275.

Sticker. L.J 1982. Identifying test items that perform differentially in population subgroups A partial correlation index *Applied Psychological Measurement* 6 (3) 261-273

Swaminathan. H.. & Rogers. H.J. 1990. Detecting differential item functioning using logistic regression procedures *Journal of Educational Measurement* 27 (4) 361-370

Ziekey M 1993 Practical questions in the use of DIF statistics in test development. In P W Holland & H Wainer (eds.). *Differential item functioning*. pp. 337-347 Hillsdale. NJ. Lawrence Erlbaum Associates