บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดผลการศึกษา เป็นกระบวนการค้นหาปริมาณของคุณลักษณะ (trait) หรือความสามารถ (ability) ในตัวบุคคลซึ่งเป็นสิ่งที่ไม่สามารถสังเกตได้โดยตรง แต่สามารถศึกษาได้โดยใช้แบบวัดเป็นสิ่งเร้า ให้ผู้ถูกวัดแสดงพฤติกรรมการตอบสนองออกมา เพื่อนำพฤติกรรมเหล่านั้นมาแปลผลว่าบุคคลนั้นๆ มีคุณลักษณะหรือความสามารถที่ต้องการวัดมากน้อยเพียงใด

ไรท์และสโตน (Wright and Stone,1983:4) ได้กล่าวถึงลำดับขั้นในการวัดความสามารถของบุคคลว่าการที่จะวัดความสามารถในด้านใดด้านหนึ่งของบุคคล ผู้วัดจะต้องระบุคุณลักษณะที่ต้องการวัดให้ชัดเจน ดำเนินการสร้างข้อคำถามให้วัดได้ตรงและครอบคลุมคุณลักษณะนั้น โดยต้องสามารถพิสูจน์ให้เห็นว่าข้อคำถามนั้นให้ผลการวัดที่มีความสม่ำเสมอคงเส้นคงวา และก่อนที่จะนำผลการวัดไปใช้ต้องพิจารณาก่อนว่าแบบแผนการตอบของผู้สอบเป็นแบบแผนการตอบที่ตรงตามที่คาดหวังไว้หรือไม่ ซึ่งสอดคล้องกับ ฮาร์นิส์ช (Harnisch,1983) ที่กล่าวว่าการวิเคราะห์ผลการสอบโดยใช้คะแนนรวมหรือคะแนนการตอบข้อสอบที่ตอบถูกยังไม่สามารถบอกความสามารถที่แท้จริงของผู้สอบได้ เนื่องจากคะแนนรวมประกอบด้วยคะแนนความสามารถจริงรวมกับคะแนนความคลาดเคลื่อน ด้วยเหตุผลดังกล่าว นักวัดผลจึงได้พยายามพัฒนาวิธีการวิเคราะห์ผลการสอบ โดยใช้แบบแผนการตอบข้อสอบของผู้สอบเป็นรายบุคคล (item response pattern) มาร่วมพิจารณาด้วย โดยมีแนวความคิดว่า คะแนนรวมเพียงอย่างเดียวไม่เพียงพอที่จะประมาณค่าความสามารถที่แท้จริงของผู้สอบได้ เพราะผู้สอบที่มีคะแนนรวมเท่ากัน อาจมีความสามารถไม่เท่ากัน ถ้าแบบแผนการตอบข้อสอบแตกต่างกัน (Smith,1986) กล่าวได้ว่า การวิเคราะห์ผลการสอบเพื่อตรวจสอบความสามารถ หรือวินิจฉัยข้อบกพร่องของผู้สอบให้ชัดเจนและยุติธรรมนั้น ควรพิจารณาทั้งคะแนนรวมและแบบแผนการตอบของผู้สอบเป็นรายบุคคลโดยพิจารณาว่าการตอบข้อสอบของผู้สอบเหมาะสมหรือสอดคล้องกับความสามารถที่แท้จริงหรือไม่

ดัชนีบ่งชี้ความเหมาะสมของบุคคล (person-fit index) เป็นค่าสถิติที่บ่งบอกว่าบุคคลนั้นมีแบบแผนการตอบที่สอดคล้องกับแบบแผนการตอบข้อสอบ ตามแนวความคิดของกัตแมนที่เรียกว่า กัตแมนสมบูรณ์ (Perfect Guttman) มากน้อยเพียงใด (Levine and Rubin,1979; Wright and Stone,1979; and Van Der Flier,1982. cited in Meijer,1996) โดยกัตแมน มีความเชื่อว่า เมื่อผู้สอบตอบข้อสอบที่มีความยากระดับหนึ่งผิด ผู้สอบคนนั้นจะตอบข้อสอบทุกข้อที่ยากกว่าข้อดังกล่าวผิดด้วย และเมื่อผู้สอบตอบข้อสอบที่มีค่าความยากระดับหนึ่งได้ถูกต้อง ผู้สอบคนนั้นจะตอบข้อสอบทุกข้อที่ง่ายกว่าข้อนั้นได้ถูกต้องด้วย ดังนั้นความผิดปกติของแบบแผนการตอบจึงมี 2 ลักษณะ คือ ประการแรก เกิดจากการที่ผู้สอบตอบข้อสอบผิดในข้อที่มีความยากน้อยกว่าหรือเท่ากับความสามารถที่แท้จริง (true ability) ประการที่สอง เกิดจากการที่ผู้สอบตอบข้อสอบได้ถูกต้องในข้อที่มีความยากสูงกว่าความสามารถที่แท้จริง

ดัชนีบ่งชี้ความเหมาะสมของบุคคลสามารถประยุกต์เพื่อนำไปใช้ตรวจสอบบุคคล ที่มีแบบแผนการตอบที่ผิดปกติเนื่องจากมีมโนทัศน์ที่คลาดเคลื่อน (misconseption) (Birenbaum, Kelly & Tatsuoka,1992; Tatsuoka & Tatsuoka,1992; cited in Meijer,1996 ศิริเดช สุชีวะ,2537) ใช้ในการบ่งชี้แบบแผนการตอบที่ผิดปกติของตัวผู้สอบซึ่งเกิดจากพฤติกรรมการตอบข้อสอบ การบริหารการสอบ หรือความบกพร่องของข้อสอบ (Meijer,1996; Birenbaum,1986; Rudner, 1983; Hanisch & Linn,1981) หรือเพื่อบ่งชี้ความผิดพลาดในการแปลผลของคะแนนจากแบบสอบชุดนั้นๆ เนื่องจากความสามารถ (ability) ของผู้สอบ ที่ตอบข้อสอบแตกต่างไปจากที่คาดหวังไว้ในโมเดลการตอบข้อสอบ ได้จากการประมาณค่าที่คลาดเคลื่อน (Drasgow & Levine,1987; Drasgow,Levine & Williams,1985; Levine & Rubin,1979. cited in Drasgow,1996)

ดัชนีที่ใช้วิเคราะห์แบบแผนการตอบข้อสอบรายบุคคลนั้น มีผู้ศึกษาและพัฒนาไว้หลายตัวด้วยกัน ซึ่งสุนันท์ ศลโกสุม (2630) ได้เสนอเป็น 2 กลุ่ม ตามเทคนิคของการวิเคราะห์ดังนี้

กลุ่มแรก วิเคราะห์โดยการหาความสัมพันธ์ของแบบแผนการตอบถูกและผิดเป็นรายบุคคลกับคะแนนรวม ซึ่งเป็นทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory) แบบแผนของการตอบข้อสอบที่ถูกต้องและตรงตามลักษณะของข้อสอบที่มีคุณภาพดี ผู้สอบควรจะตอบข้อสอบที่ง่ายๆได้และตอบข้อสอบที่ยากไม่ได้ และคนที่มีคะแนนรวมเท่ากันควรตอบคำถามได้เหมือนๆ กัน นั่นคือต้องมีแบบแผนการตอบถูกและผิด เหมือนกันในกลุ่มที่มี

คะแนนเท่ากัน แต่ถ้าในกลุ่มผู้สอบที่ได้คะแนนเท่ากันมีการตอบถูกและผิดไม่เป็นรูปแบบเดียวกัน แสดงว่ามีความผิดปกติเกิดขึ้น อาจจะเนื่องจากข้อสอบหรือตัวของผู้สอบ

กลุ่มที่สอง ใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) ดัชนีกลุ่มนี้คำนวณโดยการประมาณค่าความน่าจะเป็นที่ผู้สอบแต่ละคนจะตอบข้อสอบแต่ละข้อได้ถูกต้อง เปรียบเทียบกับผลการตอบข้อสอบในข้อนั้นๆ เพื่อจะดูว่าจะผิดไปจากแบบแผนการตอบข้อสอบที่ปกติหรือไม่ ถ้าผิดปกติแสดงว่าผู้สอบคนนั้นไม่เหมาะสมกับโมเดลการตอบข้อสอบ (person misfit) หรือผู้สอบคนนั้นมีแบบแผนการตอบข้อสอบที่ผิดปกติ (aberrant response pattern)

รัดเนอร์ (Rudner,1983) ได้ศึกษาเปรียบเทียบดัชนีต่างๆ โดยใช้ข้อมูลมอนติคาโล (Monte Carlo Data) พบว่าดัชนีที่ใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) จะตรวจพบความผิดปกติของแบบแผนการตอบซึ่งเกิดจากบุคคลซึ่งไม่เหมาะสมกับโมเดล ได้ดีกว่าดัชนีที่ใช้หลักการพิจารณาแบบแผนการตอบถูกและผิดแล้วหาค่าสหสัมพันธ์กับคะแนนรวม ซึ่งศึกษาตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT)

วิธีการวิเคราะห์ความเหมาะสมของบุคคลกับโมเดลการตอบข้อสอบ (person-fit) ที่นักวัดผลการศึกษาพัฒนาขึ้น มีหลายวิธีที่สามารถใช้ตรวจค้นข้อสอบซึ่งเหมาะสมกับโมเดลการตอบ (item-fit) ได้ กล่าวคือ เมื่อวิเคราะห์ความเหมาะสมของบุคคล วิเคราะห์ผู้สอบ 1 คน ข้อสอบทั้งฉบับ ในทำนองเดียวกัน เมื่อวิเคราะห์ความเหมาะสมของข้อสอบ สามารถทำได้โดยวิเคราะห์ข้อสอบ 1 ข้อ ผู้สอบทุกคน บุคคลที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบ (person-misfit) ซึ่งวิเคราะห์ได้จากแบบแผนการตอบข้อสอบบ่งบอกว่าความสามารถ (ability) ที่ประมาณค่าได้ไม่ใช่ความสามารถที่แท้จริง แต่ได้จากการประมาณค่าที่คลาดเคลื่อนเนื่องจากบุคคลนั้นมีแบบแผนการตอบข้อสอบที่ผิดปกติ (aberrant response pattern) ด้วยพฤติกรรมการตอบข้อสอบที่แตกต่างกัน เช่นการสับสนในรูปแบบของแบบสอบ การเดาคำตอบ การทุจริตในการสอบ ความสะเพร่า การทำข้อสอบไม่ทันตามกำหนดเวลา หรือการมีมโนทัศน์ที่คลาดเคลื่อน ในเนื้อหาที่มุ่งวัด ส่วนข้อสอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบ (item-misfit) ซึ่งวิเคราะห์ได้จากแบบแผนการตอบข้อสอบบ่งบอกว่าข้อสอบข้อนั้นวัดไม่ได้ตรงกับคุณลักษณะหรือความสามารถที่มุ่งวัด (Meijer,1996; Nering,1995; Schmitt et al. ,1993; Reise & Due,1991; Reise,1990; Drasgow,1987)

ดัชนีที่ใช้วิเคราะห์ความเหมาะสมของบุคคล (person-fit) และข้อสอบซึ่งเหมาะสมกับโมเดลการตอบ (item-fit) ตามทฤษฎีการตอบสนองข้อสอบ มีวิธีประมาณค่าดัชนี 3 วิธี (Rost and Davier,1994) คือ

1. การทดสอบไค-สแควร์ของบุคคล (The Chi-Square Test) พัฒนาโดย ไรท์ และปัญจปาเกสาน (Wright & Panchapakesan,1969) บอค (Bock,1972) และ วอลเลนเบอร์ก (Wollenberg, 1979) โดยเปรียบเทียบสัดส่วนของความถี่ของการตอบที่สังเกตได้กับความถี่ของการตอบที่คาดหวังของกลุ่มผู้ตอบข้อสอบ ทดสอบความแตกต่างด้วยสถิติทดสอบไค-สแควร์ (Chi-square) χ^2- test แบบทางเดียว (one-tailed test) ถ้าดัชนีบ่งชี้ความเหมาะสมของบุคคลใดมีค่าแตกต่างจาก 0 อย่างมีนัยสำคัญที่ระดับ 0.05 คือ มีค่าสถิติ χ^2 มากกว่า 9.48 บุคคลนั้นมีแบบแผนการตอบไม่เหมาะสมกับโมเดลการตอบข้อสอบ (person misfit) หรือผู้สอบคนนั้นมีแบบแผนการตอบข้อสอบที่ผิดปกติ (aberrant response pattern) ซึ่งวิธีนี้มีข้อดีคือ จะไม่มีข้อจำกัดในการวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ มีความไวต่อความคลาดเคลื่อนในการตอบข้อสอบ แต่จะมีจุดด้อยเนื่องจากมีความลำเอียงในการปฏิเสธสมมติฐานเกี่ยวกับความเหมาะสมของบุคคลและข้อสอบในโมเดล (Reise,1990)

2. การประมาณค่าด้วยฟังก์ชันไลค์ลิฮูด (The Likelihood-based Approach) พัฒนาโดย เลวิน และ รูบิน (Levine & Rubin,1979) และ ดราสโกว์ (Drasgow,1985) คำนวณค่าดัชนีโดยการประมาณค่าความน่าจะเป็นที่ผู้สอบแต่ละคนจะตอบข้อสอบแต่ละข้อได้ถูกต้อง เปรียบเทียบกับผลการตอบข้อสอบ ในข้อนั้นๆ แปลงค่าสถิติความเหมาะสมของบุคคล เป็นค่ามาตรฐาน (Z- Value) ทดสอบความแตกต่างแบบสองทาง (two-tailed test) ถ้าดัชนีบ่งชี้ความเหมาะสมของบุคคลใดมีค่าแตกต่างจาก 0 อย่างมีนัยสำคัญที่ระดับ 0.05 คือ มีค่า $|L_z|$ มากกว่า 1.96 บุคคลนั้นมีแบบแผนการตอบ ไม่เหมาะสมกับโมเดลการตอบข้อสอบ (person misfit) หรือผู้สอบคนนั้นมีแบบแผนการตอบข้อสอบที่ผิดปกติ (aberrant response pattern) ดัชนีที่ประมาณค่าด้วยวิธีนี้ คือ ดัชนี L_z ซึ่ง ไรส์ (Reise,1990) ได้ศึกษาเปรียบเทียบประสิทธิภาพของดัชนี L_z และ χ^2 พบว่า ดัชนี L_z มีประสิทธิภาพในการ วิเคราะห์ความเหมาะสมของบุคคลและข้อสอบสูงกว่า χ^2 ดัชนีกลุ่มนี้มีข้อดีคือ คำนวณได้ง่าย ไม่มีความลำเอียง มีการแจกแจงเป็นโค้งปกติ อัตราการตรวจค้นขึ้นอยู่กับความยาวของแบบสอบและความสามารถของผู้สอบ เมื่อความยาวของแบบสอบเพิ่มขึ้นอัตราการตรวจค้นจะเพิ่มขึ้น

3. การประมาณค่าจากคะแนนส่วนที่เหลือ (The Score Residual Approach) พัฒนาโดย ไรท์ และ สโตน (Wright & Stone,1979) ไรท์ (Wright,1980) และ ราสช์ (Rasch,1980) วิธีนี้สามารถตรวจสอบแบบแผนการตอบข้อสอบได้ทั้งการให้คะแนนแบบ 2 ค่าและหลายค่า ดัชนีในกลุ่มนี้คือ W_1 (Wright,1979) และ W_3 (Rudner,1983) ประมาณค่าดัชนีโดยประมาณค่าสัดส่วนระหว่างผลรวมของกำลังสองของผลต่างของคะแนนที่สังเกตได้กับคะแนนที่คาดหวัง กับค่าที่คาดหวังจากโมเดล ซึ่งคะแนนสองชุดนี้เป็นสัดส่วนของความแปรปรวนของคะแนนส่วนที่เหลือซึ่งไม่เป็นอิสระจากกันมีการแจกแจงเป็นแบบ ที (t-distribution) ทดสอบความแตกต่างแบบสองทาง (two-tailed test) บุคคลซึ่งมีค่าสถิติบ่งชี้ความเหมาะสมของบุคคลแตกต่างจาก 0 อย่างมี นัยสำคัญที่ระดับ 0.05 คือ มีค่า $|W_1|$ มากกว่า 2.00 เป็นบุคคลซึ่งไม่เหมาะสมกับโมเดลการตอบ วิธีนี้มีข้อดี คือประมาณค่าโดยไม่มีความลำเอียง สามารถตรวจสอบแบบแผนการตอบได้ดีในทุกระดับความสามารถ รัดเนอร์ (Rudner,1996) วิเคราะห์ผลการสอบประเมินคุณภาพของนักเรียนระดับมัธยมศึกษาของสหรัฐอเมริกาโดยใช้ดัชนี W_3 เพื่อศึกษาความแตกต่างของผลสัมฤทธิ์ระหว่างรัฐ พบว่าข้อมูลจากการสอบมีความเหมาะสมกับโมเดลดีมาก และไม่พบความแตกต่างระหว่างรัฐ

ปัจจัยที่ส่งผลต่อประสิทธิผลในการตรวจค้นบุคคลและข้อสอบ ซึ่งไม่เหมาะสมกับโมเดลการตอบข้อสอบ คือ ความยาวของแบบสอบ ระดับความสามารถของผู้สอบ ช่วงความยากของแบบสอบ ค่าอำนาจจำแนก ค่าการเดา และมิติของแบบสอบ ไรส์ และ ดิวส์ (Reise & Due,1991) ได้ศึกษาถึงผลของปัจจัยเกี่ยวกับคุณลักษณะของแบบสอบ ต่อประสิทธิผลในการประมาณค่าของดัชนีบ่งชี้ความเหมาะสมของบุคคล ซึ่งประมาณค่าด้วยฟังก์ชันไลด์ลิฮูด (L_z index) โดยได้ศึกษาผลของ ช่วงความยากของแบบสอบ ค่าการเดาและความยาวของแบบสอบ ต่อประสิทธิผลในการตรวจค้นบุคคลซึ่งไม่เหมาะสมกับโมเดลการตอบข้อสอบ ณ ระดับความสามารถที่แตกต่างกันของผู้สอบ โดยศึกษาความยาวของแบบสอบ 5 ระดับ คือ 7, 14, 21, 35, 42 ข้อ ผลการศึกษาพบว่า อัตราการตรวจค้นจะเพิ่มขึ้นเมื่อความสามารถของผู้สอบสูงกว่า 1.5 หรือต่ำกว่า -1.5 หรือช่วงความยากของแบบสอบเพิ่มขึ้น และเมื่อค่าการเดาสูงขึ้นอัตราการตรวจค้นจะน้อยลง สำหรับผู้สอบที่มีความสามารถต่ำ ไรส์ และดิวส์ ให้ข้อสรุปว่า ความยาวของแบบสอบที่เหมาะสม จะต้องมีความยาวไม่น้อยกว่า 20 ข้อ

สมิธ (Smith,1994) ได้ศึกษาเปรียบเทียบประสิทธิผลในการตรวจค้นข้อสอบซึ่งไม่เหมาะสมกับโมเดลการตอบข้อสอบของดัชนีซึ่งประมาณค่าจากคะแนนส่วนที่เหลือ (W_1 index)

โดยศึกษาระดับความสามารถของผู้สอบ 4 ระดับ คือ ความสามารถเฉลี่ย 0.0, 0.5, 1.0, 1.5 และค่าความยากในระดับเดียวกับระดับความสามารถ แบบสอบมีความยาว 5 ระดับ คือ 10, 20, 30, 40, และ 50 ข้อ พบว่าระดับความสามารถของผู้สอบและค่าความยากของแบบสอบส่งผลอย่างไม่ดงที่ต่อการแจกแจงของค่าสถิติบ่งชี้ความเหมาะสมของข้อสอบ และความยาวของแบบสอบตั้งแต่ 20 ข้อขึ้นไป ส่งผลให้ประสิทธิผลในการตรวจค้นเพิ่มขึ้น

ชมิท และ คณะ (Schmitt et.al.,1993) ได้ศึกษาความตรงตามเกณฑ์สัมพันธ์ (Criterion-Related Validity) ของแบบวัดความสามารถทางเครื่องกล ซึ่งเป็นแบบสอบชุด (Battries Test) วิเคราะห์แบบแผนการตอบข้อสอบด้วยดัชนี L_{zm} ซึ่งเป็นดัชนี L_z ที่พัฒนาขึ้นให้สามารถใช้วิคราะห์แบบสอบชุด คัดเลือกบุคคลที่เหมาะกับโมเดล (person-fit) เมื่อวิเคราะห์หาคุณภาพของแบบสอบ แบบสอบซึ่งวิเคราะห์โดยใช้แบบแผนการตอบข้อสอบรายบุคคลมีความตรงตามเกณฑ์สัมพันธ์สูงกว่าการวิเคราะห์โดยไม่พิจารณาแบบแผนการตอบ การศึกษาครั้งนี้ของ ชมิท ได้ผลสรุปสอดคล้องกับ ดราสโกว์ (Drasgow,1987) กล่าวคือ เมื่อวิเคราะห์แบบแผนการตอบข้อสอบรายบุคคล เพื่อวิเคราะห์บุคคลที่เหมาะสมกับโมเดลการตอบข้อสอบ แล้วแยกบุคคลซึ่งไม่เหมาะสมกับโมเดลออกจากการวิเคราะห์ จะทำให้ค่าความตรงของแบบสอบที่วิเคราะห์เฉพาะบุคคลซึ่งมีแบบแผนการตอบปกติมีค่าเพิ่มขึ้น ดังนั้นการวิเคราะห์ความเหมาะสมของผู้ตอบกับโมเดล (analysis of person-fit) และความเหมาะสมของข้อสอบแต่ละข้อ (item-fit) จึงมีบทบาทสำคัญในการสร้างและศึกษาคุณภาพของแบบสอบและมาตรวัด ดัชนีบ่งชี้ความเหมาะสมของบุคคล (person-fit index) ที่มีคุณภาพจะต้องสามารถบ่งชี้บุคคลซึ่งมีแบบแผนการตอบผิดปกติได้อย่างถูกต้องแม่นยำ เป็นผลให้คุณภาพของแบบสอบที่ วิเคราะห์หลังจากนำบุคคลและข้อสอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบออกจากการวิเคราะห์เพิ่มขึ้น

จากรายงานการวิจัยที่กล่าวข้างต้น จะเห็นได้ว่ามีการศึกษาถึงความสามารถในการตรวจค้นผู้สอบและข้อสอบที่ไม่เหมาะสมกับกับโมเดลการตอบในเงื่อนไขต่างๆ เฉพาะวิธีการในการประมาณค่าดัชนีใดดัชนีหนึ่งยังไม่มีการศึกษาในเชิงเปรียบเทียบระหว่างวิธี ผู้วิจัยจึงสนใจที่จะศึกษาวิธีการในการวิเคราะห์แบบแผนการตอบข้อสอบโดยใช้ดัชนีบ่งชี้ความเหมาะสมของบุคคลด้วย วิธีการประมาณค่าด้วยฟังก์ชันไลค์ลิฮูด (The Likelihood-based Approach) และการประมาณค่าจากคะแนนส่วนที่เหลือ (The Score Residual Approach) โดยศึกษาผลการตรวจสอบบุคคลที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบ ของดัชนีแอลแซด (L_z) ซึ่งเป็นโมเดล

โลจิสติกแบบ 3 พารามิเตอร์ ประมาณค่าดัชนีด้วยฟังก์ชันไลด์ลิฮูด และดัชนีดับเบิ้ลยูวัน (W_1) ซึ่งเป็นโมเดลโลจิสติกแบบ 1 พารามิเตอร์หรือ ราส์ชโมเดล ประมาณค่าดัชนีจากคะแนนส่วนที่เหลือ เนื่องจากดัชนี 2 ตัวนี้มีวิธีประมาณค่าดัชนีที่แตกต่างกัน และมีรายงานการวิจัยสนับสนุนว่าดัชนีที่ประมาณค่าด้วยวิธีดังกล่าวมีประสิทธิภาพในการตรวจค้นบุคคลซึ่งมีแบบแผนการตอบที่ผิดปกติ สามารถนำไปประยุกต์ใช้ได้สถานการณ์การทดสอบโดยทั่วไป โดยเฉพาะดัชนี w_1 ได้มีการพัฒนาโปรแกรมการวิเคราะห์ใช้อย่างแพร่หลายในปัจจุบัน ในการศึกษาครั้งนี้ผู้วิจัยศึกษาระดับความสามารถของผู้สอบ 3 ระดับ คือ ผู้สอบที่มีความสามารถสูง ความสามารถปานกลาง และ ความสามารถต่ำ และความยาวของแบบสอบ 3 ขนาด คือ แบบสอบยาว 20 ข้อ 40 ข้อ และ 60 ข้อ เพื่อเปรียบเทียบว่าดัชนีที่ประมาณค่าด้วยวิธีการที่แตกต่างกันจะมีความสามารถในการตรวจค้นบุคคลซึ่งมีแบบแผนการตอบที่ผิดปกติ (aberrant response pattern) ได้เหมือนกันหรือแตกต่างกันอย่างไรตามเงื่อนไขที่กำหนด โดยเปรียบเทียบจำนวนบุคคลซึ่งไม่เหมาะกับโมเดลการตอบข้อสอบ (person misfit) ที่ตรวจค้นได้ด้วยดัชนีทั้ง 2 ตัว และเปรียบเทียบคุณภาพของแบบสอบที่ได้จากการวิคราะห์ความเหมาะสมของบุคคลเมื่อแยกบุคคลที่ไม่เหมาะสมกับโมเดลออกจากการวิเคราะห์แล้ว ในด้านค่าสารสนเทศของแบบสอบ ความเที่ยง และความตรงตามทฤษฎี เพื่อเป็นแนวทางในการวิเคราะห์ข้อสอบและแบบสอบโดยใช้แบบแผนการตอบข้อสอบรายบุคคล เพื่อตรวจสอบผู้สอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบต่อไป

วัตถุประสงค์ของการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์ เพื่อเปรียบเทียบผลการตรวจสอบความเหมาะสมของบุคคลระหว่างดัชนีแอลแชดและดัชนีดับเบิ้ลยูวัน โดยมีวัตถุประสงค์เฉพาะ ดังนี้

1. เพื่อเปรียบเทียบผลการตรวจสอบความเหมาะสมของบุคคลระหว่างดัชนีแอลแชดและดัชนีดับเบิ้ลยูวัน เมื่อผู้สอบมีระดับความสามารถต่างกัน

2. เพื่อเปรียบเทียบผลการตรวจสอบความเหมาะสมของบุคคลระหว่างดัชนีแอลแชดและดัชนีดับเบิ้ลยูวัน เมื่อแบบสอบมีความยาวต่างกัน

สมมติฐานการวิจัย

จากการศึกษาของ รัดเนอร์ (Rudner,1983) ซึ่งจำลองแบบแผนการตอบให้ผู้สอบมีคะแนนที่สูงกว่าความเป็นจริง (spuriosly high) และคะแนนที่ต่ำกว่าความเป็นจริง (spuriosly

low) พบว่า ดัชนี L_z ซึ่งประมาณค่าด้วยฟังก์ชันไลด์ลิฮูด สามารถตรวจค้นผู้สอบ ซึ่งไม่เหมาะสมกับโมเดลการตอบได้ดีกว่า W_1 ซึ่งประมาณค่าด้วยคะแนนส่วนที่เหลือ ดราสโกว์และคณะ (Drasgow et.al.,1987) ได้จำลองข้อมูลให้เป็นคะแนนที่ผู้สอบตอบได้สูงกว่าความเป็นจริง (Spuriously high) และต่ำกว่าความเป็นจริง (Spuriously low) ในระดับที่แตกต่างกัน คือ 15% และ 30% โดยใช้ดัชนีตามทฤษฎีการตอบสนองข้อสอบจำนวน 10 ดัชนี พบว่าดัชนี L_z สามารถตรวจค้นผู้สอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบได้ร้อยละ 76 และ 46 ส่วนดัชนี W_1 มีอัตราการตรวจค้น ร้อยละ 73 และ 21 เมื่อจำลองให้ผู้ตอบมีคะแนนสูงและต่ำกว่าความเป็นจริง ตามลำดับ จากการศึกษาเอกสารและรายงานการวิจัยที่เกี่ยวข้องดังกล่าว ผู้วิจัยตั้งสมมติฐานการวิจัย ดังนี้

1. ดัชนีแอลแชดและดัชนีดับเบิ้ลยูวันมีผลการตรวจสอบผู้สอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบต่างกัน เมื่อระดับความสามารถของผู้สอบต่างกัน

1.1 เมื่อผู้สอบมีความสามารถสูงดัชนีแอลแชด มีผลการตรวจสอบผู้สอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบสูงกว่าดัชนีดับเบิ้ลยูวัน

1.2 เมื่อผู้สอบมีความสามารถปานกลางดัชนีแอลแชดและดัชนีดับเบิ้ลยูวันมีผลการตรวจสอบผู้สอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบไม่แตกต่างกัน

1.3 เมื่อผู้สอบมีความสามารถต่ำดัชนีแอลแชด มีผลการตรวจสอบผู้สอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบสูงกว่าดัชนีดับเบิ้ลยูวัน

2. ดัชนีแอลแชด มีผลการตรวจสอบผู้สอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบสูงกว่าดัชนีดับเบิ้ลยูวัน ในทุกระดับความยาวของแบบสอบ

ขอบเขตของการวิจัย

1. วิธีการประมาณค่าดัชนีบ่งชี้ความเหมาะสมของบุคคลที่ใช้ในการวิจัย มี 2 วิธี คือ การประมาณค่าดัชนีจากฟังก์ชันไลด์ลิฮูด (The Likelihood-Based Approach) ซึ่งเป็นโมเดลโลจิสติกแบบ 3 พารามิเตอร์กับการประมาณค่าดัชนีจากคะแนนส่วนที่เหลือ (The Score Residual Approach) เป็นโมเดลโลจิสติกแบบ 1 พารามิเตอร์ หรือ ราส์ชโมเดล ภายใต้ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

2. ข้อมูลที่ใช้ในการวิจัยครั้งนี้ ใช้การจำลองข้อมูลด้วยโปรแกรม IRTDATA ซึ่งพัฒนาโดย โจแฮนสัน (Johanson,1992)

3. ตัวแปรในการวิจัย

3.1 ตัวแปรอิสระ (Independent Variable) มี 3 ตัว คือ

3.1.1 ดัชนีบ่งชี้ความเหมาะสมของบุคคล มี 2 วิธี คือ

1) ดัชนีแอลแชด
2) ดัชนีดับเบิ้ลยูวัน

3.1.2 ความยาวของแบบสอบ มี 3 ขนาด คือ 20,40 และ 60 ข้อ

3.1.3 ระดับความสามารถของผู้สอบ มี 3 ระดับ คือ

1) กลุ่มผู้สอบที่มีความสามารถสูง ($\overline{\theta}$ = 1.5)
2) กลุ่มผู้สอบที่มีความสามารถปานกลาง ($\overline{\theta}$ = 0.0)
3) กลุ่มผู้สอบที่มีความสามารถต่ำ ($\overline{\theta}$ = -1.5)

3.2 ตัวแปรตาม (Dependent Variable) เป็นผลการตรวจสอบความเหมาะสมของบุคคลกับโมเดลการตอบข้อสอบ วัดจากคุณภาพของแบบสอบหลังการวิเคราะห์ 3 ด้าน คือ

3.2.1 ค่าสารสนเทศของแบบสอบ (Test Information Function)

3.2.2 ความเที่ยงของแบบสอบ (Reliability)

3.2.3 ความตรงตามทฤษฎี (Construct Validity)

คำจำกัดความที่ใช้ในการวิจัย

แบบแผนการตอบข้อสอบ (Item Response Pattern) หมายถึง เมตริกซ์คำตอบของผู้สอบแต่ละคน ที่นำมาจัดเรียงลำดับตามแนวแถวจากข้อที่ง่ายที่สุดซึ่งเป็นข้อที่มีจำนวนผู้สอบภายในกลุ่มตอบได้ถูกต้องมากที่สุด ไปจนถึงข้อที่ยากที่สุดซึ่งเป็นข้อที่มีจำนวนผู้สอบภายในกลุ่มตอบได้ถูกต้องน้อยที่สุด

แบบแผนการตอบที่ปกติ (Nonaberant Response Pattern) หมายถึง แบบแผนการตอบข้อสอบของผู้สอบ ซึ่งตอบข้อสอบที่มีค่าความยากต่ำกว่า หรือเท่ากับความสามารถของตนได้ถูกต้อง และตอบข้อสอบผิดในข้อที่มีค่าความยากสูงกว่าระดับความสามารถ

แบบแผนการตอบที่ผิดปกติ (Aberrant Response Pattern) หมายถึง แบบแผนการตอบข้อสอบของผู้สอบ ซึ่งตอบข้อสอบผิดในข้อ ที่มีความยากน้อยกว่าหรือเท่ากับความสามารถที่แท้จริงหรือเกิดจากการที่ผู้สอบตอบข้อสอบที่มีความยากสูงกว่าความสามารถที่แท้จริงได้ถูกต้อง

โมเดลการตอบข้อสอบ (Item Response Model) หมายถึง โมเดลโลจิสติก แบบ 1 พารามิเตอร์ หรือราส์ชโมเดล และ โมเดลโลจิสติก แบบ 2 และ 3 พารามิเตอร์ ซึ่งเป็นแบบจำลองทางคณิตศาสตร์ ที่แสดงความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถของผู้สอบ

ค่าพารามิเตอร์ของข้อสอบ หมายถึง ค่าสถิติที่บ่งบอกคุณภาพของข้อสอบ ตามทฤษฎีการตอบสนองข้อสอบ มี 3 ค่า คือ ค่าอำนาจจำแนก (a) ค่าความยาก (b) และ ค่าการเดา (C)

ค่าอำนาจจำแนก (a) หมายถึง ค่าความชันของโค้งคุณลักษณะข้อสอบที่จุดเปลี่ยนโค้งในทางปฏิบัติจะใช้ข้อสอบซึ่งมีค่าระหว่าง0.5 ถึง 2.5

ค่าความยาก (b) หมายถึง ตำแหน่งของโค้งคุณลักษณะข้อสอบ ณ จุดที่โค้งมีความชันมากที่สุด มีพิสัยระหว่าง -∝ ถึง +∝ ในทางปฏิบัติจะใช้ข้อสอบที่มีค่าระหว่าง -3.0 ถึง +3.0

ค่าการเดา (C) หมายถึง ค่ากำกับต่ำสุดของโค้งคุณลักษณะของข้อสอบ มีพิสัยระหว่าง 0 ถึง 1 ในทางปฏิบัติจะใช้ข้อสอบที่มีค่าระหว่าง 0 ถึง 0.3

ระดับความสามารถ (θ) หมายถึง ความสามารถของผู้สอบที่ประมาณค่าได้จากโมเดลโลจิสติก แบบ 3 พารามิเตอร์ ตามทฤษฎีการตอบสนองข้อสอบ แบ่งเป็น 3 ระดับ คือ

ผู้สอบที่มีความสามารถสูง หมายถึง กลุ่มผู้สอบที่มีระดับความสามารถเฉลี่ยเท่ากับ 1.50

ผู้สอบที่มีความสามารถปานกลาง หมายถึง กลุ่มผู้สอบที่มีระดับความสามารถเฉลี่ยเท่ากับ 0.0

ผู้สอบที่มีความสามารถต่ำ หมายถึง กลุ่มผู้สอบที่มีระดับความสามารถเฉลี่ยเท่ากับ -1.50

ความยาวของแบบสอบ หมายถึง จำนวนข้อสอบซึ่งสุ่มมาจากเมตริกซ์คำตอบ มี 3 ชุด คือ ความยาว 20 ข้อ , 40 ข้อ และ 60 ข้อ

ดัชนีบ่งชี้ความเหมาะสมของบุคคล (Person-fit index) หมายถึง ค่าสถิติที่บ่งบอกว่าบุคคลนั้นมีแบบแผนการตอบสอดคล้องหรือแตกต่างจากค่าที่คาดหวังในโมเดลการตอบข้อสอบ ตามทฤษฎีการตอบสนองข้อสอบ สำหรับการวิจัยนี้ ได้แก่ ค่าดัชนีแอลแซด (L_Z) และ ดัชนีดับเบิ้ลยูวัน (W_1)

ดัชนีแอลแซด (L_Z index) หมายถึง ดัชนีบ่งชี้ความเหมาะสมของบุคคล ซึ่งประมาณค่าด้วยฟังก์ชันไลด์ลิฮูด (The Likelihood-based Approach) (Levine Rubin,1979)

ดัชนีดับเบิ้ลยูวัน (W_1 index) หมายถึง ดัชนีบ่งชี้ความเหมาะสมของบุคคลซึ่งประมาณค่าจากคะแนนส่วนที่เหลือ (The Score Residual Approach) ซึ่งมีค่าสถิติที่ใช้เป็นเกณฑ์ ในการพิจารณาความเหมาะสมของบุคคลกับโมเดลการตอบข้อสอบ 2 ตัว คือ ค่าสถิติ infit และ ค่าสถิติ outfit (Linacre & Wright,1994)

บุคคลที่เหมาะสมกับโมเดล (person fit) หมายถึง บุคคลที่มีแบบแผนการตอบข้อสอบสอดคล้องกับรูปแบบของการตอบของบุคคลที่ควรจะเป็นในการตอบตามที่คาดหวังไว้ในโมเดลการตอบข้อสอบตามทฤษฎีการตอบสนองข้อสอบ คือ มี $|L_Z|$ น้อยกว่า 1.96 และ $|W_1|$ น้อยกว่า 2.00

ความเที่ยง (reliability) หมายถึง ความสอดคล้องภายในของคะแนนการตอบแบบสอบตามเมตริกซ์คำตอบที่ได้จากการจำลองข้อมูล คำนวณโดยใช้สูตรของคูเดอร์ ริชาร์ดสันที่ 20 (KR-20)

คุณภาพของแบบสอบก่อนการวิเคราะห์ หมายถึง ค่าความเที่ยง และค่าสารสนเทศของแบบสอบก่อนการตรวจสอบความเหมาะสมของบุคคลกับโมเดลการตอบข้อสอบ

คุณภาพของแบบสอบหลังการวิเคราะห์ หมายถึง ค่าสารสนเทศ ค่าความเที่ยง และความตรงตามทฤษฎีของแบบสอบ ภายหลังการตรวจสอบความเหมาะสมของบุคคลของบุคคลกับโมเดลการตอบข้อสอบ

ค่าสารสนเทศของแบบสอบ (Test Information Function) หมายถึง ผลรวมของค่าสารสนเทศของข้อสอบ ก่อนและหลังการวิเคราะห์ผู้สอบที่เหมาะสมกับโมเดลการตอบข้อสอบด้วยโปรแกรม BILOG

ค่าประสิทธิภาพสัมพัทธ์ของแบบสอบ (Relative Efficiency) หมายถึง อัตราส่วนระหว่างค่าฟังก์ชันสารสนเทศของแบบสอบในแต่ละเงื่อนไข ก่อนและหลังการตรวจสอบผู้สอบที่เหมาะสมกับโมเดลการตอบข้อสอบ

ความตรงเชิงทฤษฎี (Construct validity) หมายถึง คุณสมบัติของแบบสอบ ซึ่งได้จากการจำลองข้อมูลที่ให้ผลการวัดสอดคล้องกับโมเดลตามทฤษฎีโดยพิจารณาจากดัชนีความเหมาะสมของข้อมูลกับโมเดลทางทฤษฎี จากการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis) ด้วยโปรแกรม LISREL

ผลการตรวจสอบความเหมาะสมของบุคคล หมายถึง ความสามารถของดัชนีบ่งชี้ความเหมาะสมของบุคคลในการบ่งชี้บุคคลที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบได้ถูกต้อง เมื่อขจัดผู้สอบที่ไม่เหมาะสมกับโมเดลออกไปแล้วเป็นผลให้แบบสอบที่ได้จากการวิเคราะห์มีค่าสารสนเทศ ความเที่ยง และความตรงเพิ่มขึ้น

ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้ทราบผลการตรวจสอบความเหมาะสมของบุคคลว่าระหว่างดัชนีแอลแซด (L_Z) และ ดัชนีดับเบิ้ลยูวัน (W_1) ดัชนีใดมีประสิทธิผลในการตรวจค้นบุคคลซึ่งมีแบบแผนการตอบข้อสอบที่ผิดปกติ ได้ดีกว่ากัน

2. เป็นแนวทางสำหรับเลือกวิธีการตรวจสอบความเหมาะสมของบุคคลกับโมเดลการตอบข้อสอบในเงื่อนไขที่แตกต่างกัน

3. เป็นพื้นฐานสำหรับการศึกษาเพื่อเสริมสร้างสารสนเทศ เกี่ยวกับการวิเคราะห์ความเหมาะสมของบุคคลกับโมเดลการตอบข้อสอบ (person-fit analysis)