บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์ 2 ประการ **ประการแรก** เพื่อพัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ ด้วยข้อมูลจำลองตามเงื่อนไขการปรับเทียบคะแนน โมเดลการตอบสนองข้อสอบ แบบแผนการเก็บรวบรวมข้อมูล และการหาคุณภาพของการปรับเทียบ **ประการที่สอง** เพื่อตรวจสอบคุณภาพของเกณฑ์ที่พัฒนาขึ้น ในด้านความตรงเชิงเกณฑ์สัมพันธ์ โดยตรวจสอบความสอดคล้องของผลการตัดสินด้วยเกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ของปิเตอร์เซนและคณะ และความสอดคล้องของผลการตัดสินด้วยเกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ความเสมอภาคของลอร์ด

ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลจำลองจากโปรแกรม IRTDATA เมื่อจำลองข้อมูลตามเงื่อนไขการปรับเทียบคะแนนที่อาจเกิดขึ้นได้เกือบทั้งหมดดังนี้คือ โมเดล 1 และ 3 พารามิเตอร์ แบบแผนการเก็บรวบรวมข้อมูลแบบใช้ข้อสอบร่วมและใช้กลุ่มสมมูล การหาคุณภาพการปรับเทียบคะแนนด้วยการปรับเทียบกลับสู่แบบสอบเดิมและการใช้กลุ่มสอบทานผล โดยกำหนดเงื่อนไขในการจำลองข้อมูลเป็นจำนวนข้อสอบ 10, 20, 30, ...,200 ข้อ และกลุ่มผู้สอบจำนวน 100, 200, 300, ...,3,000 คน

วิธีดำเนินการวิจัยมีขั้นตอนดังนี้ คือ 1) ตรวจสอบความถูกต้องโปรแกรม IRTDATA โดยพิจารณาจากการนำข้อมูลที่เป็นคะแนนตอบรายข้อที่จำลองจากโปรแกรม IRTDATA ไปวิเคราะห์ด้วยโปรแกรม BILOG แล้วนำค่าพารามิเตอร์ผู้ตอบและค่าพารามิเตอร์ข้อสอบไปเทียบกับผลจากโปรแกรม IRTDATA ว่าแตกต่างกันหรือไม่ และตรวจสอบความเป็นเอกมิติ (Unidimensionality) ของข้อมูลด้วยการวิเคราะห์ตัวประกอบ (Factor Analysis) ด้วยวิธีการวิเคราะห์ตัวประกอบสำคัญ (Princepal Component Analysis) และหมุนแกนด้วยวิธี Varimax 2) จำลองข้อมูลตามแบบแผนที่กำหนด 3) ปรับเทียบพารามิเตอร์ข้อสอบจากแบบสอบฉบับที่ 1 ผ่านแบบสอบฉบับที่ 2 และแบบสอบฉบับที่ 3 แล้วปรับเทียบกลับสู่แบบสอบฉบับที่ 1 เดิม สำหรับการปรับเทียบกลับสู่แบบสอบเดิม และปรับเทียบพารามิเตอร์ข้อสอบจากแบบสอบฉบับที่ 1 ไปสู่แบบสอบฉบับที่ 2 สำหรับกลุ่มสอบทานผล 4) คำนวณคะแนนจริง (True Score) จากทั้งพารามิเตอร์ข้อสอบจากแบบสอบฉบับที่ 1 ที่ไม่ได้ปรับเทียบกับพารามิเตอร์ข้อสอบที่ปรับเทียบแล้ว เมื่อใช้พารามิเตอร์ผู้สอบจากกลุ่มผู้สอบแบบสอบฉบับที่ 1 สำหรับการปรับเทียบกลับสู่แบบสอบเดิม และใช้พารามิเตอร์ผู้ตอบจากกลุ่มสอบทานผลสำหรับการใช้กลุ่มสอบทานผล 5) ตรวจสอบความแตกต่างระหว่างค่าเฉลี่ยคะแนนจริงที่ไม่ได้ปรับเทียบกับค่าเฉลี่ยของคะแนนจริงที่ปรับเทียบแล้ว ด้วยสถิติ t-test แบบ Two Dependent Samples เพื่อหาค่าวิกฤต t-test ที่นัยสำคัญ

ทางสถิติระดับ .01 6) กำหนดจุดตัดโดยคำนวณค่าดัชนีความแตกต่างเมื่อทดสอบความแตกต่างค่าเฉลี่ยของคะแนนจริงทั้งสองด้วยสถิติ t-test แล้วพบค่าวิกฤตที่นัยสำคัญทางสถิติระดับ .01 จุดตัดนี้จะแบ่งกลุ่มค่าดัชนีความแตกต่างออกเป็น 2 กลุ่ม คือ กลุ่มดัชนีที่แสดงคุณภาพการปรับเทียบระดับสูง และกลุ่มดัชนีที่แสดงคุณภาพการปรับเทียบระดับต่ำ จำแนกตามเงื่อนไขการปรับเทียบคะแนนเป็น 7 เงื่อนไข ดังนี้ คือ รวมทุกเงื่อนไข โมเดล 1 พารามิเตอร์ โมเดล 3 พารามิเตอร์ การใช้กลุ่มสมมูล การใช้ข้อสอบร่วม การปรับเทียบกลับสู่แบบสอบเดิม และการใช้กลุ่มสอบทานผล 7) กำหนดจุดตัดซึ่งได้จากการคำนวณค่าดัชนีความแตกต่าง เมื่อพบค่าวิกฤต t-test ที่นัยสำคัญทางสถิติ .02 จากการทดสอบความแตกต่างค่าเฉลี่ยของคะแนนจริงทั้งสองด้วยสถิติ t-test จุดตัดนี้จะแบ่งกลุ่มดัชนีที่แสดงคุณภาพการปรับเทียบระดับสูงออกเป็น 2 กลุ่มย่อย และจุดตัดที่คำนวณค่าดัชนีความแตกต่างเมื่อพบค่าวิกฤตที่นัยสำคัญทางสถิติระดับ .001 เป็นจุดตัดที่แบ่งกลุ่มดัชนีที่แสดงคุณภาพการปรับเทียบระดับต่ำออกเป็น 2 กลุ่มย่อย 8) ให้ความหมายแต่ละกลุ่มดัชนีความแตกต่างกำหนดเป็นเกณฑ์ 9) ตรวจสอบความสอดคล้องของผลการตัดสินคุณภาพการปรับเทียบคะแนนด้วยเกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ของปิเตอร์เซนและคณะ และความสอดคล้องของผลการตัดสินคุณภาพการปรับเทียบคะแนนด้วยเกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ความเสมอภาคของลอร์ด

สรุปผลการวิจัย

ผลการวิจัยในการศึกษาครั้งนี้ สรุปได้ดังนี้

1. ผลการตรวจสอบความถูกต้องโปรแกรม IRTDATA พบว่า

1.1 ค่าพารามิเตอร์ผู้ตอบ ค่าพารามิเตอร์ข้อสอบจากการจำลองข้อมูลด้วยโปรแกรม IRTDATA และจากการวิเคราะห์ด้วยโปรแกรม BILOG เมื่อใช้ข้อมูลคะแนนการตอบรายข้อชุดเดียวกันไม่แตกต่างกัน โดยมีการจำลองข้อมูลทั้งโมเดลการตอบสนองข้อสอบ 1 พารามิเตอร์ และโมเดล 3 พารามิเตอร์ แสดงให้เห็นว่าข้อมูลที่จำลองจากโปรแกรม IRTDATA มีความเชื่อถือได้

1.2 ค่าไอเกน (Eigenvalue) สูงสุดของแบบสอบ คือ 11.712 และคิดเป็นร้อยละ 29.28 ของความแปรปรวนทั้งหมดสำหรับข้อมูลจำลองจากโปรแกรม IRTDATA ตามโมเดลการตอบสนองข้อสอบ 1 พารามิเตอร์ และค่าไอเกนสูงสุดของแบบสอบ คือ 8.734 คิดเป็นร้อยละ 21.834 ของความแปรปรวนทั้งหมด สำหรับข้อมูลจำลองตามโมเดลการตอบสนองข้อสอบ 1 พารามิเตอร์ พิจารณาแล้วเห็นว่ามีค่ามากพอที่จะสรุปว่ามีความเป็นเอกมิติประเภทที่มีตัวประกอบหลักเด่นกว่าตัวประกอบอื่น (Essential Unidimensionality) สรุปได้ว่าข้อมูลที่ได้จากการใช้โปรแกรม IRTDATA มีความเป็นเอกมิติเป็นไปตามข้อตกลงของทฤษฎีการตอบสนองข้อสอบ

2. เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบที่พัฒนาขึ้น มี 7 เกณฑ์ดังนี้

2.1 เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ เกณฑ์รวมเงื่อนไขทั้งหมด

ระดับคุณภาพการปรับเทียบ	ดัชนีความแตกต่าง
น่าพอใจอย่างยิ่ง	AMD < 0.000183 MAD < 0.002391 RMS < 0.002907
น่าพอใจ	0.000183 ≤ AMD < 0.000244 0.002391 ≤ MAD < 0.002481 0.002907 ≤ RMS < 0.003032
ไม่น่าพอใจ	0.000244 ≤ AMD < 0.000352 0.002481 ≤ MAD < 0.002621 0.003032 ≤ RMS < 0.003181
ไม่น่าพอใจอย่างยิ่ง	0.000352 ≤ AMD 0.002621 ≤ MAD 0.003181 ≤ RMS

2.2 เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนสำหรับโมเดล 1 พารามิเตอร์

ระดับคุณภาพการปรับเทียบ	ดัชนีความแตกต่าง
น่าพอใจอย่างยิ่ง	AMD < 0.000182
	MAD < 0.002450
	RMS < 0.002924
น่าพอใจ	0.000182 ≤ AMD < 0.000261
	0.002450 ≤ MAD < 0.002508
	0.002924 ≤ RMS < 0.003045
ไม่น่าพอใจ	0.000261 ≤ AMD < 0.000355
	0.002508 ≤ MAD < 0.002574
	0.003045 ≤ RMS < 0.003073
ไม่น่าพอใจอย่างยิ่ง	0.000355 ≤ AMD
	0.002574 ≤ MAD
	0.003073 ≤ RMS

2.3 เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ เกณฑ์สำหรับโมเดล 3 พารามิเตอร์

ระดับคุณภาพการปรับเทียบ	ดัชนีความแตกต่าง
น่าพอใจอย่างยิ่ง	AMD < 0.000183
	MAD < 0.002402
	RMS < 0.002924
น่าพอใจ	0.000183 ≤ AMD < 0.000242
	0.002402 ≤ MAD < 0.002468
	0.002924 ≤ RMS < 0.003122
ไม่น่าพอใจ	0.000242 ≤ AMD < 0.000355
	0.002468 ≤ MAD < 0.002613
	0.003122 ≤ RMS < 0.003398
ไม่น่าพอใจอย่างยิ่ง	0.000355 ≤ AMD
	0.002613 ≤ MAD
	0.003398 ≤ RMS

2.4 เกณฑ์สำหรับแบบแผนการเก็บรวบรวมข้อมูลใช้กลุ่มสมมูล

ระดับคุณภาพการปรับเทียบ	ดัชนีความแตกต่าง
น่าพอใจอย่างยิ่ง	AMD < 0.000197
	MAD < 0.002353
	RMS < 0.002924
น่าพอใจ	0.000197 ≤ AMD < 0.000237
	0.002353 ≤ MAD < 0.002375
	0.002924 ≤ RMS < 0.002930
ไม่น่าพอใจ	0.000237 ≤ AMD < 0.000374
	0.002375 ≤ MAD < 0.002510
	0.002930 ≤ RMS < 0.003022
ไม่น่าพอใจอย่างยิ่ง	0.000374 ≤ AMD
	0.002510 ≤ MAD
	0.003022 ≤ RMS

2.5 เกณฑ์สำหรับแบบแผนการเก็บรวบรวมข้อมูลใช้ข้อสอบร่วม

ระดับคุณภาพการปรับเทียบ	ดัชนีความแตกต่าง
น่าพอใจอย่างยิ่ง	AMD < 0.000176
	MAD < 0.002410
	RMS < 0.002899
น่าพอใจ	0.000176 ≤ AMD < 0.000247
	0.002410 ≤ MAD < 0.002535
	0.002899 ≤ RMS < 0.003083
ไม่น่าพอใจ	0.000247 ≤ AMD < 0.000341
	0.002535 ≤ MAD < 0.002677
	0.003083 ≤ RMS < 0.003261
ไม่น่าพอใจอย่างยิ่ง	0.000341 ≤ AMD
	0.002677 ≤ MAD
	0.003261 ≤ RMS

2.6 เกณฑ์สำหรับการปรับเทียบกลับสู่แบบสอบเดิม

ระดับคุณภาพการปรับเทียบ	ดัชนีความแตกต่าง
น่าพอใจอย่างยิ่ง	AMD < 0.000169
	MAD < 0.002370
	RMS < 0.002875
น่าพอใจ	0.000169 ≤ AMD < 0.000261
	0.002370 ≤ MAD < 0.002508
	0.002875 ≤ RMS < 0.003045
ไม่น่าพอใจ	0.000261 ≤ AMD < 0.000336
	0.002508 ≤ MAD < 0.002637
	0.003045 ≤ RMS < 0.003124
ไม่น่าพอใจอย่างยิ่ง	0.000336 ≤ AMD
	0.002637 ≤ MAD
	0.003124 ≤ RMS

2.7 เกณฑ์สำหรับการปรับเทียบกลับสู่แบบสอบเดิม

ระดับคุณภาพการปรับเทียบ	ดัชนีความแตกต่าง
น่าพอใจอย่างยิ่ง	AMD < 0.000189
	MAD < 0.002402
	RMS < 0.002924
น่าพอใจ	0.000189 ≤ AMD < 0.000235
	0.002402 ≤ MAD < 0.002468
	0.002924 ≤ RMS < 0.003026
ไม่น่าพอใจ	0.000235 ≤ AMD < 0.000361
	0.002468 ≤ MAD < 0.002613
	0.003026 ≤ RMS < 0.003210
ไม่น่าพอใจอย่างยิ่ง	0.000361 ≤ AMD
	0.002613 ≤ MAD
	0.003210 ≤ RMS

3. ผลการตรวจสอบคุณภาพของเกณฑ์ที่พัฒนาขึ้น มีดังต่อไปนี้

3.1 ผลการตัดสินคุณภาพการปรับเทียบคะแนนเมื่อใช้เกณฑ์ที่พัฒนาขึ้นทั้ง 7 เกณฑ์ เปรียบเทียบกับผลการใช้ตามเกณฑ์ของปิเตอร์เซนและคณะ พบว่าไม่สอดคล้องกัน โดยผลการตัดสินคุณภาพการปรับเทียบทั้ง 4 ระดับจากเกณฑ์ที่พัฒนาขึ้น คือระดับคุณภาพการปรับเทียบน่าพอใจอย่างยิ่ง คุณภาพการปรับเทียบน่าพอใจ คุณภาพการปรับเทียบไม่น่าพอใจ และคุณภาพการปรับเทียบไม่น่าพอใจอย่างยิ่ง อยู่ในระดับน่าพอใจอย่างยิ่งเมื่อเทียบกับเกณฑ์ของปิเตอร์เซนและคณะ แสดงว่าเกณฑ์ที่พัฒนาขึ้นให้ความคลาดเคลื่อนของการปรับเทียบคะแนนน้อยกว่าเกณฑ์ของปิเตอร์เซนและคณะ

3.2 ผลการตัดสินคุณภาพการปรับเทียบคะแนนเมื่อใช้เกณฑ์ที่พัฒนาขึ้นทั้ง 7 เกณฑ์เปรียบเทียบกับผลการใช้เกณฑ์ความเสมอภาคของลอร์ด พบว่ามีความสอดคล้องกัน โดยผลการตัดสินคุณภาพการปรับเทียบระดับน่าพอใจอย่างยิ่งหรือระดับน่าพอใจและระดับไม่น่าพอใจหรือระดับไม่น่าพอใจอย่างยิ่งจากการใช้เกณฑ์ที่พัฒนาขึ้น อยู่ในระดับคุณภาพน่าพอใจและระดับคุณภาพไม่น่าพอใจตามลำดับ เมื่อใช้เกณฑ์ความเสมอภาคของลอร์ด แสดงว่าเกณฑ์ทั้งสองให้ผลการตัดสินคุณภาพการปรับเทียบคะแนนเป็นไปในทิศทางเดียวกัน

อภิปรายผลการวิจัย

จากผลการวิจัยที่กล่าวมาข้างต้นสามารถอภิปรายได้ใน 3 ประเด็น ได้แก่ **ประเด็นที่หนึ่ง** เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบที่พัฒนาขึ้น **ประเด็นที่สอง** ผลการตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบเมื่อใช้เกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ชองปิเตอร์เซนและคณะ และสุดท้าย**ประเด็นที่สาม** ผลการตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ เมื่อใช้เกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ความเสมอภาคของลอร์ด โดยมีรายละเอียดในแต่ละประเด็น ดังนี้

ประเด็นที่ 1 เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบที่พัฒนาขึ้น

เกณฑ์ที่พัฒนาขึ้นในครั้งนี้มี 7 เกณฑ์ มีความหลากหลายเพื่อให้ผู้ดำเนินการปรับเทียบคะแนนได้เลือกใช้ให้เหมาะสมกับสถานการณ์การปรับเทียบคะแนน ดังเช่น แฮร์ริส และ เคร้าส์ (Harris and Crouse, 1993) กล่าวว่าไม่มีเกณฑ์ที่สมบูรณ์ที่สุดที่จะใช้ตัดสินผลการปรับเทียบคะแนนได้ทุกเงื่อนไขการปรับเทียบคะแนน และเมื่อใช้เกณฑ์ที่ต่างกันแล้วจะทำให้ผลการตัดสิน

คุณภาพการปรับเทียบต่างกันด้วย (Skaggs, 1990b; Livingston and others cited in Harris and Crouse, 1993) ฉะนั้นผู้ใช้ผลการปรับเทียบคะแนนต้องพิจารณาเลือกใช้เกณฑ์ให้สอดคล้องกับกระบวนการปรับเทียบคะแนน ทำให้ผลตัดสินคุณภาพการปรับเทียบเป็นที่ยอมรับได้ (Harris and Crouse, 1993)

เกณฑ์ที่พัฒนาขึ้นทั้ง 7 เกณฑ์ เป็นกลุ่มค่าดัชนีความแตกต่างเรียงลำดับจากน้อยไปหามาก 4 กลุ่ม กลุ่มค่าดัชนีเหล่านี้แสดงถึงระดับคุณภาพการปรับเทียบคะแนน 4 ระดับ โดยกลุ่มแรกแสดงถึงระดับคุณภาพการปรับเทียบน่าพอใจอย่างยิ่ง ค่าดัชนีในกลุ่มที่ 2 แสดงถึงระดับคุณภาพการปรับเทียบน่าพอใจ ค่าดัชนีในกลุ่มที่ 3 แสดงถึงระดับคุณภาพการปรับเทียบไม่น่าพอใจ และกลุ่มดัชนีแสดงถึงระดับคุณภาพการปรับเทียบไม่น่าพอใจอย่างยิ่ง โดยจุดตัดที่ใช้แบ่งกลุ่มดัชนี คือ ค่าดัชนีความแตกต่างระหว่างคะแนนจริงที่ปรับเทียบแล้วกับคะแนนจริงที่ยังไม่ได้ปรับเทียบ เมื่อพบค่าวิกฤตจากการทดสอบความแตกต่างค่าเฉลี่ยของคะแนนจริงทั้งสองด้วยสถิติ t-test แบบTwo Dependent Sample Test ซึ่งจุดตัดมีความสำคัญกับการกำหนดช่วงของเกณฑ์ (Glaser and Nitko, 1971) การกำหนดจุดตัดที่มีหลักการและมีความเชื่อถือได้ จะทำให้เกณฑ์มีคุณภาพด้วยเพราะจุดตัดจะแบ่งช่วงของเกณฑ์ได้อย่างถูกต้องและเหมาะสม

เกณฑ์ที่พัฒนาขึ้นเป็นช่วงของค่าดัชนี AMD (Absolute Mean Difference), MAD (Mean Absolute Difference) และ RMS (Root Mean Square) ที่บอกระดับคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ ซึ่งเป็นดัชนีที่นิยมใช้ในกระบวนการปรับเทียบคะแนน (Harris and Crouse, 1993) เพราะเป็นดัชนีที่คำนวณได้ง่ายไม่ซับซ้อน แปลผลชัดเจน โดยค่าดัชนีที่มีค่าน้อยจะบอกถึงคุณภาพการปรับเทียบที่ดี เมื่อใช้ข้อมูลชุดเดียวกัน ดัชนี RMS จะให้ค่ามากที่สุด รองลงมาคือดัชนี MAD และดัชนี AMD ตามลำดับ

ค่าดัชนีความแตกต่างที่กำหนดเป็นเกณฑ์มีค่าน้อย เช่น ดัชนี AMDมีค่าตั้งแต่ 0.000056 ถึง 0.270909 ดัชนี MAD มีค่าตั้งแต่ 0.002223 ถึง 0.270909 และดัชนี RMS มีค่าตั้งแต่ 0.002704 ถึง 0.280773 เป็นเพราะค่าดัชนีเหล่านี้คำนวณจากความแตกต่างระหว่างคะแนนจริงที่ปรับเทียบแล้วกับคะแนนจริงที่ยังไม่ได้ปรับเทียบ จากผู้สอบกลุ่มเดียวกัน เนื่องจากดัชนีความแตกต่างมีค่าน้อยจะแสดงถึงการเกิดความคลาดเคลื่อนจากการปรับเทียบคะแนนมีน้อย สอดคล้องกับคำกล่าวของ แฮร์ริส และ เคร้าส์ (Harris and Crouse, 1993) ที่ว่าในการประเมินผลการปรับเทียบคะแนนเมื่อใช้ดัชนีความแตกต่าง เช่น ดัชนี RMS ผู้ใช้ผลการปรับเทียบต้องการที่จะให้ได้ค่าดัชนีความแตกต่างที่มีค่าน้อย ๆ

ประเด็นที่ 2 ผลการตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ เมื่อใช้เกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ของปิเตอร์เซนและคณะ

เกณฑ์ที่พัฒนาขึ้นเป็นช่วงของค่าดัชนีความแตกต่างที่บอกระดับคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ 4 ระดับ คือ ระดับคุณภาพปรับเทียบน่าพอใจอย่างยิ่ง คุณภาพการปรับเทียบน่าพอใจ คุณภาพการปรับเทียบไม่น่าพอใจ และคุณภาพปรับเทียบไม่น่าพอใจอย่างยิ่ง ส่วนเกณฑ์ของปิเตอร์เซนและคณะเป็นช่วงของกำลังสองของผลคูณระหว่างค่าคงที่กับความแปรปรวนของคะแนน มี 5 ช่วง แต่ละช่วงบอกระดับความพึงพอใจในการปรับเทียบคะแนน คือ ระดับน่าพอใจอย่างมาก ระดับน่าพอใจ ระดับปานกลาง ระดับไม่น่าพอใจ และระดับไม่น่าพอใจอย่างยิ่ง เมื่อต้องการทราบคุณภาพการปรับเทียบคะแนน สามารถคำนวณค่าดัชนีความแตกต่างแล้วพิจารณาว่าดัชนีตกอยู่ในช่วงใด ซึ่งจะบอกระดับคุณภาพการปรับเทียบคะแนนตามต้องการได้ จากการจำลองข้อมูลตามเงื่อนไขการปรับเทียบคะแนน แล้ววิเคราะห์หาคุณภาพการปรับเทียบคะแนนโดยใช้เกณฑ์ที่พัฒนาขึ้นทั้ง 7 เกณฑ์ พบว่าค่าดัชนีความแตกต่างที่แสดงถึงคุณภาพการปรับเทียบน่าพอใจอย่างยิ่ง คุณภาพการปรับเทียบน่าพอใจ คุณภาพการปรับเทียบไม่น่าพอใจ และคุณภาพการปรับเทียบไม่น่าพอใจอย่างยิ่ง ทั้ง 4 ระดับ จะตกอยู่ในระดับน่าพอใจอย่างยิ่งเมื่อใช้เกณฑ์ของปิเตอร์เซนและคณะ จากแนวคิดที่ แฮร์ริส และเคร้าส์ (Harris and Crouse, 1993) ได้เสนอแนะว่าในกระบวนการปรับเทียบคะแนนต้องการให้มีค่าดัชนีความแตกต่างที่น้อย แสดงว่าเกณฑ์ที่พัฒนาขึ้นให้ผลการตัดสินคุณภาพการปรับเทียบคะแนนได้ดีกว่าเกณฑ์ของปิเตอร์เซนและคณะ เพราะให้ค่าดัชนีความแตกต่างที่น้อยกว่าและค่าดัชนีความแตกต่างที่กำหนดเป็นจุดตัดแบ่งกลุ่มดัชนีที่แสดงถึงคุณภาพการปรับเทียบคะแนนสำหรับเกณฑ์ที่พัฒนาขึ้นมีค่าน้อย เพราะเป็นค่าที่ได้จากการทดสอบความแตกต่างค่าเฉลี่ยของคะแนนแล้วพบค่าวิกฤตที่มีนัยสำคัญทางสถิติ ส่วนจุดตัดในเกณฑ์ของปิเตอร์เซนและคณะ ได้จากการกำหนดของผู้เชี่ยวชาญ

เกณฑ์ของปิเตอร์เซนและคณะสามารถนำไปใช้ได้ทุกสถานการณ์การปรับเทียบคะแนน เมื่อนำไปใช้ในสถานการณ์เฉพาะ เช่น การปรับเทียบโมเดลการตอบสนองข้อสอบ 3 พารามิเตอร์ เกณฑ์ที่เหมาะสมกว่าควรเป็นเกณฑ์สำหรับโมเดล 3 พารามิเตอร์ ดังเช่น แฮร์ริส และ เคร้าส์ (Harris and Crouse, 1993) กล่าวว่าไม่มีเกณฑ์ที่สมบูรณ์ที่สุดที่จะใช้ตัดสินผลการปรับเทียบคะแนนได้ทุกเงื่อนไขการปรับเทียบคะแนน และเมื่อใช้เกณฑ์ที่ต่างกันแล้วจะทำให้ผลการตัดสินคุณภาพการปรับเทียบต่างกันด้วย (Skaggs, 1990b; Livingston and others cited in Harris and Crouse, 1993)

จากการใช้เกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ของปิเตอร์เซนและคณะ ทำให้ได้ระดับคุณภาพการปรับเทียบคะแนนแตกต่างกัน อาจเป็นเพราะเกณฑ์ที่พัฒนาขึ้นเป็นเกณฑ์สัมบูรณ์ ค่าดัชนีความแตกต่างที่บอกระดับคุณภาพการปรับเทียบคะแนนจะไม่แปรเปลี่ยนไปตามเงื่อนไขการปรับเทียบคะแนน และเป็นเกณฑ์ที่ได้จากเงื่อนไขที่อาจเป็นไปได้เกือบจะทั้งหมดของการปรับเทียบคะแนน ส่วนเกณฑ์ของปิเตอร์เซนและคณะจะแปรเปลี่ยนไปตามส่วนเบี่ยงเบนมาตรฐานของคะแนน ซึ่งการปรับเทียบคะแนนในแต่ละครั้งจะใช้เกณฑ์ที่แตกต่างกัน และในการวิจัยครั้งนี้ได้นำเกณฑ์ของปิเตอร์เซนและคณะ มาใช้กับข้อมูลจำลองหลากหลายสถานการณ์การปรับเทียบคะแนนและพบว่าการปรับเทียบคะแนนทุกสถานการณ์มีคุณภาพในระดับน่าพอใจอย่างยิ่ง แต่เมื่อใช้เกณฑ์ที่พัฒนาขึ้นและเกณฑ์ความเสมอภาคของลอร์ด ปรากฏว่าคุณภาพการปรับเทียบมีหลายระดับ แสดงว่าเกณฑ์ของปิเตอร์เซนและคณะให้ผลการตัดสินคุณภาพการปรับเทียบไม่คงที่ อาจเป็นเพราะการพัฒนาของปิเตอร์เซนและคณะมีการกำหนดช่วงของจุดตัดในเกณฑ์มีความกว้างเกินไป ทำให้ผลการตัดสินคุณภาพการปรับเทียบคะแนนมีความคลาดเคลื่อน

การใช้เกณฑ์ของปิเตอร์เซนและคณะในการตัดสินคุณภาพการปรับเทียบคะแนน มีขั้นตอนวิเคราะห์ค่อนข้างจะยุ่งยากซับซ้อน เพราะต้องนำข้อมูลไปคำนวณค่าจุดตัด 4 จุด ที่แบ่งระดับคุณภาพการปรับเทียบคะแนนออกเป็น 5 ระดับ และนำข้อมูลชุดเดิมไปคำนวณค่าดัชนีความแตกต่าง แล้วนำมาเทียบกับเกณฑ์ที่ทำไว้ว่าคุณภาพการปรับเทียบอยู่ในระดับใด ส่วนเกณฑ์ที่พัฒนาขึ้นจะคำนวณค่าดัชนีความแตกต่าง แล้วนำมาเทียบกับเกณฑ์ว่าคุณภาพการปรับเทียบอยู่ในระดับใด

ประเด็นที่ 3 ผลการตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ เมื่อใช้เกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ความเสมอภาคของลอร์ด

เกณฑ์ความเสมอภาคของลอร์ดเป็นเกณฑ์ที่กำหนดจากคุณสมบัติที่ดีของการปรับเทียบคะแนน คือ ความเสมอภาคของลอร์ด (Lord's Equity) ที่ว่าการปรับเทียบคะแนนระหว่างแบบสอบฉบับที่ 1 และฉบับที่ 2 จะถือว่ามีความเสมอภาคสำหรับผู้สอบก็ต่อเมื่อ ไม่ว่าผู้สอบมีความสามารถระดับใดก็ตาม เมื่อสอบแบบสอบฉบับที่ 1 หรือฉบับที่ 2 แล้วจะไม่ก่อให้เกิดผลที่แตกต่างกันกับผู้สอบ นั่นคือคะแนนที่ผ่านการปรับเทียบแล้วจะต้องมีการแจกแจงเหมือนกัน ดังนั้นเกณฑ์ความเสมอภาคของลอร์ดจึงแบ่งออกเป็น 2 ระดับ คือระดับคุณภาพการปรับเทียบน่าพอใจเมื่อมีการตรวจสอบการแจกแจงของคะแนนจากแบบสอบ 2 ฉบับที่ปรับเทียบแล้วปรากฏว่าไม่แตกต่างกัน ส่วนอีกระดับหนึ่งแสดงถึงคุณภาพการปรับเทียบไม่น่าพอใจเมื่อมีการตรวจสอบการแจกแจงของคะแนนจากแบบสอบทั้ง 2 ฉบับแล้วแตกต่างกัน โดยการตรวจสอบด้วยสถิติ Wilcoxon signed-rank test จากการจำลองตามเงื่อนไขการปรับเทียบคะแนนทั้งหมด แล้ว

ผลการตัดสินคุณภาพการปรับเทียบคะแนนเมื่อใช้เกณฑ์ที่พัฒนาขึ้นทั้ง 7 เกณฑ์เปรียบเทียบกับผลการใช้เกณฑ์ความเสมอภาคของลอร์ด พบว่ามีความสอดคล้องกัน โดยผลการตัดสินคุณภาพการปรับเทียบระดับน่าพอใจอย่างยิ่งหรือระดับน่าพอใจและระดับไม่น่าพอใจหรือระดับไม่น่าพอใจอย่างยิ่งจากการใช้เกณฑ์ที่พัฒนาขึ้น อยู่ในระดับคุณภาพน่าพอใจและระดับคุณภาพไม่น่าพอใจตามลำดับ เมื่อใช้เกณฑ์ความเสมอภาคของลอร์ด เป็นเพราะทั้งสองเกณฑ์ได้ใช้กระบวนการทางสถิติกำหนดจุดตัดแบ่งกลุ่มดัชนีที่แสดงคุณภาพการปรับเทียบคะแนน โดยเกณฑ์ที่พัฒนาขึ้นใช้สถิติ t-test แบบ Two Dependent Sample Test ส่วนเกณฑ์ความเสมอภาคของลอร์ดใช้สถิติ Wilcoxon signed-rank test ทดสอบ ทั้งสองสถิติมีข้อตกลงเบื้องต้นและกระบวนการวิเคราะห์ที่แตกต่างกัน แต่อยู่ภายใต้วัตถุประสงค์เดียวกันคือทดสอบค่าความคลาดเคลื่อนจากการปรับเทียบคะแนน

เมื่อเปรียบเทียบจำนวนระดับการตัดสินคุณภาพของการปรับเทียบคะแนนเมื่อใช้เกณฑ์ความเสมอภาคของลอร์ดและเกณฑ์ที่พัฒนาขึ้น จะเห็นว่าเกณฑ์ที่พัฒนาขึ้นให้ประโยชน์มากกว่าเพราะเกณฑ์ความเสมอภาคของลอร์ดให้ระดับคุณภาพปรับเทียบน้อยเพียง 2 ระดับเท่านั้น คือ ระดับน่าพอใจและระดับไม่น่าพอใจ ซึ่งให้ทางเลือกในการนำสารสนเทศเพื่อการตัดสินใจเกี่ยวกับการปรับเทียบคะแนนไปใช้ประโยชน์ได้น้อย ส่วนเกณฑ์ที่พัฒนาขึ้นมี 4 ระดับ คือ ระดับคุณภาพการปรับเทียบน่าพอใจอย่างยิ่ง คุณภาพการปรับเทียบน่าพอใจ คุณภาพการปรับเทียบไม่น่าพอใจและระดับคุณภาพการปรับเทียบไม่น่าพอใจอย่างยิ่ง

การใช้เกณฑ์ความเสมอภาคของลอร์ดกับเกณฑ์ที่พัฒนาขึ้นสำหรับการปรับเทียบคะแนนมีกระบวนการคล้อยคลึงกัน เพราะเป็นการคำนวณค่าดัชนีจากการปรับเทียบคะแนนแล้วนำไปเทียบกับดัชนีในเกณฑ์ว่าตกในช่วงใดของเกณฑ์ ซึ่งได้คำนวณค่าดัชนีความแตกต่างแล้วพิจารณาว่าค่าดัชนีตกอยู่ในระดับคุณภาพใดสำหรับเกณฑ์ที่พัฒนาขึ้น และทดสอบความแตกต่างการแจกแจงคะแนนที่ปรับเทียบแล้วกับคะแนนที่ยังไม่ได้ปรับเทียบ ถ้าการแจกแจงของคะแนนมีความแตกต่างกันอย่างมีนัยสำคัญแสดงว่าคุณภาพการปรับเทียบไม่น่าพอใจ ส่วนการแจกแจงคะแนนไม่มีความแตกต่างกันอย่างมีนัยสำคัญ แสดงว่ามีคุณภาพการปรับเทียบคะแนนในระดับน่าพอใจ

เกณฑ์ที่พัฒนาขึ้นมี 7 เกณฑ์ จำแนกตามเงื่อนไขการปรับเทียบคะแนน เพื่อให้ผู้ใช้ผลการปรับเทียบได้มีทางเลือกใช้เกณฑ์ได้เหมาะสมกับเงื่อนไข ส่วนเกณฑ์ความเสมอภาคของลอร์ดเป็นเกณฑ์รวมใช้ได้ทุกเงื่อนไข แต่ถ้าผู้ใช้ผลการปรับเทียบคะแนนต้องการให้ความสำคัญกับเงื่อนไขการปรับเทียบเป็นพิเศษ ก็ควรใช้เกณฑ์ที่พัฒนาขึ้นมี 7 เกณฑ์ คือ เกณฑ์รวมทุกเงื่อนไข เกณฑ์สำหรับโมเดล 1 พารามิเตอร์ เกณฑ์สำหรับโมเดล 3 พารามิเตอร์ เกณฑ์สำหรับกลุ่มสมมูล เกณฑ์สำหรับการใช้ข้อสอบร่วม เกณฑ์สำหรับการปรับเทียบกลับสู่แบบสอบเดิมและเกณฑ์สำหรับกลุ่มสอบทานผล

ข้อเสนอแนะในการนำเกณฑ์ที่พัฒนาขึ้นไปใช้

1. ในการนำเกณฑ์ไปใช้ตัดสินผลการปรับเทียบคะแนน สำหรับแบบสอบต่างฉบับ ซึ่งแบบสอบแต่ละฉบับต้องวัดคุณลักษณะเดียวกัน เป็นแบบสอบที่สร้างมาจากจุดประสงค์หรือตารางวิเคราะห์หลักสูตรเดียวกัน

2. เกณฑ์นี้จะเป็นประโยชน์อย่างยิ่งเมื่อมีการวางแผนและดำเนินการประเมินผลการปรับเทียบคะแนน และใช้เกณฑ์นี้ตัดสินคุณภาพการปรับเทียบไปพร้อมกับการปรับเทียบคะแนน

3. เกณฑ์เหล่านี้เป็นเกณฑ์ที่พัฒนาขึ้นเพื่อตัดสินคุณภาพการปรับเทียบคะแนนจริง (True Score) ตามทฤษฎีการตอบสนองข้อสอบ แต่สามารถใช้เป็นเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนดิบ (Raw Score) ได้ โดยใช้แบบแผนการเก็บรวบรวมข้อมูลแบบกลุ่มสมมูล ซึ่งเป็นการสุ่มผู้สอบเข้ากลุ่มแต่ละกลุ่ม หรือจัดผู้สอบต่างกลุ่มให้มีความสามารถใกล้เคียงกัน หาคุณภาพการปรับเทียบด้วยการใช้กลุ่มสอบทานผล เป็นกลุ่มผู้สอบอีกกลุ่มที่มีความสามารถคล้ายกับกลุ่มผู้สอบที่ทำแบบสอบฉบับที่ 1 และกลุ่มผู้สอบที่ทำแบบสอบฉบับที่ 2 กลุ่มสอบทานผลนี้จะทำแบบสอบทั้งสองฉบับ วิเคราะห์หาค่าพารามิเตอร์ผู้สอบ (θ) ของกลุ่มสอบทานผล วิเคราะห์แยกเป็น 2 ครั้ง โดยกำหนดให้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าพารามิเตอร์ผู้สอบ (θ) ให้เท่ากันก่อนการวิเคราะห์ จากนั้นนำคะแนนดิบจากแบบสอบฉบับที่ 1 และแบบสอบฉบับที่ 2 สำหรับกลุ่มสอบทานผลที่ระดับความสามารถเดียวกันไปคำนวณหาค่าดัชนีความแตกต่าง AMD, MAD และ RMS แล้วนำไปเทียบกับเกณฑ์ที่พัฒนาขึ้นว่าคุณภาพการปรับเทียบคะแนนอยู่ในระดับใด

4. เกณฑ์ที่พัฒนาขึ้นนี้ไม่เหมาะที่จะนำไปใช้ในกรณีที่มีการปรับเทียบคะแนนตามทฤษฎีการวัดแบบดั้งเดิม (Classical Test Thoery) แต่เมื่อต้องการทราบคุณภาพของการปรับเทียบคะแนนสามารถทำได้โดย นำคะแนนสอบรายข้อที่ได้จากการสอบไปตรวจสอบว่ามีการวัดมิติเดียว (Unidimensionality) หรือไม่ แล้วนำไปวิเคราะห์ปรับเทียบคะแนนและหาคุณภาพการปรับเทียบตามทฤษฎีการตอบสนองข้อสอบ แล้วจึงใช้เกณฑ์ที่พัฒนาขึ้นนี้ตัดสินคุณภาพการปรับเทียบ

5. เกณฑ์ที่พัฒนาขึ้นมีทั้งหมด 7 เกณฑ์ คือ เกณฑ์รวมทุกเงื่อนไข เกณฑ์สำหรับโมเดล 1 พารามิเตอร์ เกณฑ์สำหรับโมเดล 3 พารามิเตอร์ เกณฑ์สำหรับแบบแผนการเก็บรวบรวมข้อมูลแบบกลุ่มสมมูล เกณฑ์สำหรับแบบแผนการเก็บรวบรวมข้อมูลแบบใช้ข้อสอบร่วม เกณฑ์สำหรับการหาคุณภาพการปรับเทียบด้วยการปรับเทียบกลับสู่แบบสอบเดิม และเกณฑ์สำหรับการหาคุณภาพการปรับเทียบด้วยการใช้กลุ่มสอบทานผล ซึ่งจะอยู่ในดุลพินิจของผู้ใช้ผลการปรับเทียบคะแนนจะเลือกใช้ ตามความสำคัญของเงื่อนไขที่กำหนดไว้หรือตามจุดประสงค์ของการปรับเทียบคะแนนในแต่ละครั้ง

6. เกณฑ์ที่พัฒนาขึ้นมีทั้งหมด 7 เกณฑ์ แต่ละเกณฑ์เป็นช่วงของค่าดัชนีความแตกต่างระหว่างคะแนนจริงที่ไม่ได้ปรับกับคะแนนจริงที่ปรับเทียบแล้ว ได้แก่ดัชนี AMD (Absolute Mean

Diference) ดัชนี MAD (Mean Absolute Difference) และดัชนี RMS (Root Mean Square) แต่ดัชนี RMS เป็นค่าความคลาดเคลื่อนในการปรับเทียบคะแนน ซึ่งคำนวณได้จากรากที่สองของค่าเฉลี่ยของกำลังสองความแตกต่างระหว่างคะแนนจริงที่ปรับเทียบแล้วกับคะแนนจริงที่ไม่ได้ปรับเทียบ การได้มาสำหรับดัชนี RMS นี้มีความคงเส้นคงวาและมีความเชื่อถือได้มากกว่าดัชนี AMD และดัชนี MAD เพราะค่าดัชนี AMD และดัชนี MAD มีโอกาสเท่ากันในกรณีที่คะแนนจริงที่ปรับเทียบแล้วทุกค่ามากกว่าคะแนนจริงที่ไม่ได้ปรับเทียบ ค่าดัชนี RMS มีค่ามากกว่าดัชนี AMD และดัชนี MAD พิสัยในแต่ละช่วงของเกณฑ์มากกว่า ทำให้เห็นความแตกต่างได้ชัดเจนกว่าดัชนีทั้งสอง และในงานวิจัยโดยทั่วไปจะนิยมใช้ดัชนี RMS ฉะนั้นเมื่อใช้เกณฑ์ที่พัฒนาขึ้นและต้องการเลือกใช้ค่าดัชนีเพียงดัชนีเดียว ควรเลือกใช้ดัชนี RMS

ข้อเสนอแนะเพื่อทำวิจัยต่อไป

1. เกณฑ์ที่พัฒนาขึ้นนี้ใช้สำหรับการปรับเทียบคะแนนระหว่างแบบสอบที่มีการให้คะแนนแบบ 0-1 แต่ยังมีแบบสอบที่มีการให้คะแนนแบบพหุวิภาคหรือให้คะแนนหลายค่า เช่น 5, 4, 3, 2, 1 เพื่อขยายองค์ความรู้เรื่องเกณฑ์การปรับเทียบคะแนน และได้เกณฑ์ที่มีความเชื่อถือได้ จึงควรพัฒนาเกณฑ์สำหรับการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบสำหรับการปรับเทียบคะแนนที่มีการให้คะแนนแบบพหุวิภาค

2. เกณฑ์ที่พัฒนาขึ้นในครั้งนี้เป็นเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ ควรพัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการวัดแบบดั้งเดิม (Classical Test Theory) เพื่อให้ผู้นำผลการปรับเทียบคะแนนได้เลือกใช้เกณฑ์ให้เหมาะสมกับกระบวนการปรับเทียบคะแนนระหว่างแบบสอบ