# บทที่ 1

# บทนำ

## ความเป็นมาและความสำคัญของปัญหา

การทดสอบเป็นกระบวนการวัดเพื่อประมาณค่าคุณลักษณะหรือความสามารถของบุคคล ซึ่งเป็นสิ่งจำเป็นและเป็นประโยชน์กับการจัดการศึกษา ผลที่ได้จากการทดสอบเป็นข้อมูลในการตัดสินใจทั้งในระดับบุคคลและระดับสถาบันการศึกษา ในระดับบุคคลทำให้บุคคลได้เลือกเรียน เลือกลงทะเบียนเรียนหรือเลือกทำงานได้ตามความสามารถและความถนัดของตนเอง สำหรับระดับสถาบันการศึกษาช่วยในการคัดเลือกบุคคลเข้าเรียน สำเร็จการศึกษา เป็นข้อมูลในการพัฒนาและปรับปรุงการเรียนการสอน ช่วยสถานประกอบการในการคัดเลือกบุคคลเข้าทำงาน การจัดตำแหน่งหรือเลื่อนระดับให้สูงขึ้น นอกจากนี้ยังเป็นประโยชน์ในระดับนโยบายช่วยพัฒนาและปรับปรุงนโยบายจัดการศึกษา ยกเลิกหรือคงนโยบายไว้ใช้ในปีต่อไป ดังที่ ศิริชัย กาญจนวาสี (2545) ได้กล่าวไว้ว่า

> *" การทดสอบในอนาคต จะเป็นระบบที่บูรณาการของการใช้ศาสตร์แห่งทฤษฎีการวัด ( Measurement Theories ) กับความล้ำสมัยของเทคโนโลยีเพื่อสนองตอบความต้องการใช้สารสนเทศสำหรับการตัดสินใจ การติดตาม กำกับและพัฒนาทรัพยากรมนุษย์ได้อย่างเสมอภาคเที่ยงธรรม โครงสร้างและวิธีดำเนินการสอบจะมีความกระชับ รัดกุม ยืดหยุ่น สอดคล้องกับสภาพแวดล้อมและสนองตอบต่อความแตกต่างระหว่างบุคคลอย่างมีประสิทธิผลและมีประสิทธิภาพยิ่งขึ้น"*

การทดสอบได้ถูกนำมาใช้ให้เกิดประโยชน์กับกระบวนการจัดการศึกษากันอย่างกว้างขวาง เมื่อมีความจำเป็นที่ต้องใช้แบบสอบชุดเดียวกับคนจำนวนมากทดสอบพร้อมกันจะเกิดปัญหาในเรื่องสถานที่สอบไม่เพียงพอ คนคุมสอบไม่เพียงพอ สถานการณ์การสอบจัดได้ไม่เท่าเทียมกัน การรั่วของข้อสอบมีความเป็นไปได้จากการย้ายแบบสอบจากสถานที่หนึ่งไปอีกที่หนึ่ง หรือการสอบที่ใช้แบบสอบชุดเดียวกันทดสอบกับผู้สอบต่างกลุ่มต่างเวลา ปัญหาที่พบก็คือผู้สอบครั้งแรกจะนำข้อสอบไปเผยแพร่กับคนที่สอบครั้งหลังทราบและมีการศึกษาค้นคว้าเพิ่มเติม ทำให้ผลการสอบไม่ตรงกับความสามารถของผู้สอบ เกิดความไม่ยุติธรรมกับผู้สอบคนอื่น ๆ แนวทางแก้ไขคือการสร้างแบบสอบหลายชุดให้คู่ขนานกัน เพื่อให้ผลการสอบเป็นคะแนนที่มีความหมาย เปรียบเทียบกันได้โดยตรงจากการสอบแต่ละครั้งที่ใช้แบบสอบต่างชุด แต่ในทางปฏิบัติการสร้างแบบสอบคู่ขนาน ที่มีความเท่าเทียมกันทางด้านโครงสร้าง เนื้อหา ค่าสถิติของข้อสอบ หรือแบบสอบมีความเที่ยง ความ

ตรง ค่าเฉลี่ยและความแปรปรวนของคะแนนจากแบบสอบคู่ขนานให้เท่ากันนั้นมีโอกาสเป็นไปได้น้อยมาก แม้ว่าแบบสอบหลายฉบับสร้างจากโครงสร้างเนื้อหาเดียวกัน แต่ตัวคำถามที่ใช้ในแต่ละฉบับมีความแตกต่างกัน ตลอดจนการดำเนินการสอบในแต่ละครั้งไม่สามารถจัดสถานการณ์ให้เหมือนกันได้ เพราะอาจเปลี่ยนแปลงไปตามเวลาและสถานที่ จึงมีโอกาสน้อยที่จะมีความเท่าเทียมกันในระดับความยาก ทำให้คะแนนที่ได้จากการสอบไม่สามารถนำมาเปรียบเทียบกันได้ จึงมีผู้คิดวิธีการที่จะทำให้คะแนนที่วัดจากแบบสอบที่มีเนื้อหาเดียวกันแต่ต่างฉบับกัน สามารถนำผลออกมาใช้แทนกันได้เป็นกระบวนการใช้เทคนิคการสร้างแบบสอบหลาย ๆ ชุด ด้วยความรอบคอบระมัดระวังให้สามารถวัดคุณลักษณะเดียวกัน และเทคนิคการปรับทางสถิติเพื่อให้คะแนนจากแบบสอบต่างชุดกันปรับชดเชยกันสำหรับความแตกต่างในการสอบและคุณลักษณะของข้อสอบ จนสามารถเปรียบเทียบกันได้อย่างยุติธรรม (ศิริชัย กาญจนวาสี, 2545) กระบวนการนี้เรียกว่าการปรับเทียบคะแนนระหว่างแบบสอบ

สถานศึกษามีอิสระในการวัดและประเมินผลสัมฤทธิ์ทางการเรียน การจัดการศึกษาในสถานศึกษามีหลายหลักสูตรไม่ว่าจะเป็นการศึกษาระดับขั้นพื้นฐานหรือระดับอุดมศึกษา ที่มีเป้าหมายเดียวกัน เนื้อหาสาระในหลักสูตร จุดประสงค์การเรียนรู้ เนื้อหาในรายวิชาถูกระบุให้เหมือนกัน แต่กระบวนการเรียนการสอนและการวัดและประเมินผลสถานศึกษาดำเนินได้ ตามความเหมาะสม การวัดผลสัมฤทธิ์ทางการเรียนมีการใช้เครื่องมือที่หลากหลาย และแบบสอบยังเป็นเครื่องมือที่นิยมใช้กันมากในปัจจุบัน สถานศึกษาแต่ละแห่งได้ใช้แบบสอบต่างฉบับกันวัดในเนื้อหาวิชาเดียวกัน และต้องการใช้ผลการวัดร่วมกันเพื่อเป็นข้อมูลตัดสินให้ทุนการศึกษา การศึกษาต่อ เปรียบเทียบผลการสอบว่าโรงเรียนใดมีคะแนนสูงกว่า หรือการสอบในรายวิชาเดียวกันในแต่ละปีที่ใช้แบบสอบต่างกันเพื่อดูพัฒนาการการเรียนรู้ แม้ว่าแบบสอบต่างฉบับจะสร้างจากจุดประสงค์เดียวกัน เนื้อหาเดียวกัน แต่จะทำให้เท่าเทียมกันในเรื่องความยาก ค่าอำนาจจำแนกของข้อสอบ ค่าความเที่ยง ความตรงของแบบสอบ มีความเท่าเทียมกันนั้นเป็นไปได้น้อยมาก ทำให้ผลการสอบอยู่ต่างสเกลกัน เพื่อให้ผลการสอบมีประสิทธิภาพ มีความยุติธรรมกับผู้สอบ จึงต้องมีการปรับเทียบคะแนนก่อนนำผลไปใช้ร่วมกันหรือใช้ทดแทนซึ่งกันและกันได้

การสอบที่มีความจำเป็นต้องใช้เทคนิคการปรับเทียบคะแนนระหว่างแบบสอบ ตามสถานการณ์การสอบดังนี้ คือสถานการณ์ที่มีความจำเป็นต้องสร้างแบบสอบเนื้อหาเดียวกันขึ้นมาหลาย ๆ ฉบับ ด้วยเหตุผลของการนำไปใช้ในการทดสอบให้เกิดความยุติธรรมและป้องกันความลับของข้อสอบเมื่อใช้ต่างเวลากันสำหรับผู้สอบกลุ่มขนาดใหญ่ เพื่อปรับเทียบว่าคะแนนคะแนนที่ได้จากฉบับหนึ่งเทียบเป็นเท่าไรของอีกฉบับหนึ่ง ซึ่งวัดในระดับเดียวกัน จึงเป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างฉบับของวิชาเดียวกัน สำหรับกลุ่มผู้สอบระดับชั้นเดียวกัน ส่วนอีกสถานการณ์หนึ่งเป็นสถานการณ์เป็นสถานการณ์ที่มีความจำเป็นต้องสร้างแบบสอบเนื้อหาเดียวกัน แต่ต่างฉบับต่างมุ่งวัดความสามารถของผู้สอบที่ต่างระดับกัน เพื่อเปรียบเทียบว่าคะแนนที่สอบได้จากฉบับหนึ่ง

เทียบเป็นเท่าไรของฉบับอื่นที่วัดต่างระดับกัน จึงเป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างระดับของวิชาเดียวกัน สำหรับกลุ่มผู้สอบต่างระดับชั้นกัน การปรับเทียบคะแนนที่นำไปใช้ให้เกิดประโยชน์กับการจัดการศึกษามีขั้นตอนดังนี้ คือ (Kolen and Brennan, 1995) 1) กำหนดจุดมุ่งหมายการปรับเทียบคะแนน 2) สร้างแบบสอบหลายฉบับวัดเนื้อหาเดียวกัน 3) เลือกวิธีเก็บรวบรวมข้อมูล 4) เก็บรวบรวมข้อมูลตามรูปแบบที่กำหนด 5) เลือกนิยามเชิงปฏิบัติการของการปรับเทียบคะแนน 6) เลือกวิธีประมาณค่าสถิติที่ใช้วิเคราะห์ และ 7) ประเมินผลการปรับเทียบคะแนน

วัตถุประสงค์ของการปรับเทียบคะแนนเป็นตัวกำหนดแนวทางในการดำเนินการปรับเทียบคะแนน เพื่อให้เกิดประโยชน์ตามความต้องการจำแนกได้เป็น 2 กลุ่ม กลุ่มแรกมีวัตถุประสงค์เพื่อนำผลที่ได้จากการปรับเทียบคะแนนจากแบบสอบต่างฉบับไปเป็นข้อมูลไปใช้ในการตัดสินผลการเรียนหรือผลการศึกษาร่วมกัน กรณีนี้ใช้กระบวนการปรับเทียบในแนวนอน (Horizontal Equating) ส่วนอีกกลุ่มหนึ่งมีวัตถุประสงค์เพื่อนำผลการปรับเทียบไปใช้ในการพิจารณาพัฒนาการทางการเรียนหรือการเปลี่ยนแปลงทางการศึกษา ใช้การปรับเทียบคะแนนตามแนวตั้ง ( Vertical Equating ) แฮมเบิลตัน และ สวามินาธาน ( Hambleton and Swaminathan, 1985 ) ได้อธิบายถึงการปรับเทียบตามแนวนอนว่า เป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างฉบับ เมื่อแต่ละฉบับมุ่งวัดคุณลักษณะเดียวกัน มีระดับความยากใกล้เคียงกัน และกลุ่มผู้สอบมีการแจกแจงความสามารถอยู่ในประชากรกลุ่มเดียวกัน หรือมีความสามารถใกล้เคียงกัน ส่วนการปรับเทียบคะแนนตามแนวตั้งนั้น เป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างฉบับ แต่ละแบบสอบมุ่งวัดคุณลักษณะเดียวกัน แต่มีความระดับความยากแตกต่างกัน และกลุ่มผู้สอบมีการแจกแจงความสามารถอยู่ต่างกลุ่มประชากร หรือมีความสามารถแตกต่างกัน

การออกแบบวิธีการเก็บรวบรวมข้อมูล รูปแบบกลุ่มสุ่ม ( Random Group Design ) มีความยุ่งยากในการพัฒนาแบบสอบน้อยที่สุด ( Kolen and Brennan, 1995 ) เพราะว่าไม่มีการสร้างและพัฒนาข้อสอบร่วมที่เป็นตัวแทนเนื้อหาในแบบสอบที่นำมาปรับเทียบ และเนื่องจากมีการสุ่มแบบสอบให้กับกลุ่มผู้สอบที่มีจำนวนค่อนข้างมาก ทำให้มีปัญหากับข้อตกลงเบื้องต้นในการใช้สถิติน้อยที่สุด ถ้าขนาดกลุ่มตัวอย่างไม่เพียงพอที่จะใช้กับรูปแบบกลุ่มสุ่ม ควรใช้รูปแบบผู้สอบกลุ่มเดียวที่ดำเนินการจัดแบบสอบ 2 ฉบับ ให้กับผู้สอบแต่ละคน โดยจัดให้สมดุล (Single Group with Counterbalancing) อิทธิพลของกลุ่มผู้สอบที่ใช้ในการปรับเทียบคะแนนขึ้นอยู่กับรูปแบบการเก็บรวบรวมข้อมูล แบบสอบต่างฉบับที่สร้างขึ้นและนำไปทดสอบกับกลุ่มตัวอย่างขนาดใหญ่ตามรูปแบบกลุ่มสุ่ม ( Random Groups Design ) แล้วนำคะแนนที่ได้มาปรับเทียบ ผลที่ได้คือความสัมพันธ์ของการปรับเทียบไม่ขึ้นอยู่กับกลุ่มผู้สอบที่ใช้ในการปรับเทียบคะแนน ( Anggoff and Cowell, 1986; Harris and Kolen, 1986 ) สำหรับรูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันใช้แบบสอบร่วม

( Common-Item Nonequivalent Groups Design ) เมื่อขนาดกลุ่มผู้สอบกลุ่มเดิมและกลุ่มใหม่แตกต่างกันมาก ก่อให้เกิดปัญหาสำคัญในการประมาณค่าความสัมพันธ์ของการปรับเทียบ ทั้งวิธีการปรับเทียบแบบดั้งเดิมและการปรับเทียบตามทฤษฎีการตอบสนองข้อสอบ และขนาดของกลุ่มตัวอย่างที่แตกต่างกันมาก ทำให้ข้อตกลงเบื้องต้นทางสถิติใช้ไม่ได้กับวิธีการปรับเทียบคะแนนบางวิธี ( Cook and Petersen, 1987; Harris, 1993; Skaggs, 1990; Skaggs and Lissitz, 1986 cited in Kolen and Brennan, 1995 )

แบบสอบร่วมที่ใช้ในรูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้แบบสอบร่วม ควรสร้างจากกำหนดแบบแผนเดียวกันตามสัดส่วน มีจำนวนข้อสอบที่เป็นตัวแทนเนื้อหาสาระของข้อสอบอย่างเพียงพอ จำนวนข้อสอบร่วมที่ใช้ควรคำนึงถึงทั้งเนื้อหาและหลักสถิติ ตามหลักสถิติเมื่อข้อสอบร่วมมีจำนวนมาก ทำให้ความคลาดเคลื่อนเชิงสุ่มของการปรับเทียบลดลง ( Budescu, 1985 ; Wingersky, et al, 1987 ) แองกอฟ ( Angoff, 1971 ) แนะนำว่าควรใช้ข้อสอบร่วมจำนวน 20 % ของข้อสอบที่ใช้ในการปรับเทียบจำนวน 40 ข้อ หรือ มากกว่า หรือถ้าข้อสอบมีความยาวมาก ควรใช้ข้อสอบร่วมจำนวน 30 ข้อ และการใช้จำนวนข้อสอบร่วมควรคำนึงถึงองค์ประกอบของโครงการทดสอบที่มีความเฉพาะ หรือความแตกต่างด้านเนื้อหาวิชาด้วย เพราะข้อสอบร่วมเป็นตัวแทนของแบบสอบทั้งสองฉบับที่นำมาใช้ในการปรับเทียบคะแนน การจัดตำแหน่งของของข้อสอบร่วมในแบบสอบต่างฉบับที่มีความแตกต่างกัน เช่น ในแบบสอบชุดแรกอยู่ตอนต้น ในแบบสอบชุดที่สองจัดไว้ตอนปลาย จะส่งผลต่อความถูกต้องแม่นยำของการปรับเทียบคะแนน( Cook and Petersen, 1987; Eignor, 1985; Kolen and Harris, 1990 )

การเลือกวิธีประมาณค่าสถิติที่ใช้วิเคราะห์ เลือกให้สอดคล้องกับนิยามเชิงปฏิบัติการที่กำหนด มีวิธีการปรับจากค่าเฉลี่ย ( Mean Equating ) โดยพิจารณาคะแนนสมมูลกันเมื่อคะแนนจากแบบสอบต่างฉบับเบี่ยงเบนไปจากคะแนนเฉลี่ยเท่ากัน วิธีปรับเทียบเชิงเส้นตรง ( Linear Equating ) พิจารณาคะแนนสมมูลกันเมื่อคะแนนจากแบบสอบต่างฉบับมีคะแนนมาตรฐานเท่ากัน วิธีปรับเทียบอิควิเปอร์เซ็นต์ไทล์ ( Equipercentile Equating ) ที่คะแนนสมมูลกันเมื่อคะแนนจากแบบสอบต่างฉบับมีตำแหน่งเปอร์เซ็นต์ไทล์เท่ากัน และวิธีการปรับเทียบคะแนนโดยใช้สมการถดถอย (Regression Equating ) เป็นการสร้างสมการทำนายคะแนนคะแนนจากแบบสอบชุดหนึ่งไปยังอีกชุดหนึ่ง หรือได้จากคะแนนสมมูลกันเมื่อคะแนนของแบบสอบแต่ละฉบับทำนายคะแนนเกณฑ์ได้เท่ากัน ทั้ง 4 วิธีนี้ เป็นวิธีการปรับเทียบคะแนนตามทฤษฎีการทดสอบแบบดั้งเดิม ( Classical Test Theory ) ส่วนวิธีการปรับเทียบตามทฤษฎีการตอบสนองข้อสอบ ( Item Response Theory) เป็นการหาสัมประสิทธิ์การปรับเทียบ หรือค่าความชัน ( Slope ) และ ค่าคงที่ ของฟังก์ชันเชิงเส้นตรง

ที่เป็นความสัมพันธ์ของการปรับเทียบคะแนน วิธีการหาสัมประสิทธิ์การปรับเทียบ มีวิธีใช้ค่าเฉลี่ยของค่าอำนาจจำแนกและค่าเฉลี่ยของค่าความยากของข้อสอบ ( Mean and Mean Method ) วิธีใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าความยากของข้อสอบ ( Mean and $\sigma$ Method ) วิธีการทำให้ความแตกต่างระหว่างคะแนนจริงเดิมกับคะแนนจริงที่ปรับแล้วมีค่าน้อยที่สุด โดยใช้สถิติ F-test (Characteristic Curve Method ) และวิธีทำให้ความแตกต่างระหว่างพารามิเตอร์ข้อสอบเดิมกับที่ปรับแล้วมีค่าน้อยที่สุด โดยใช้สถิติ $\chi^2$- test ( Minimum $\chi^2$ )

การใช้สถิติในการปรับเทียบคะแนน วิธีการปรับค่าเฉลี่ย ( Mean Method ) และวิธีปรับเทียบคะแนนเชิงเส้นตรง ( Linear Equating ) เหมาะสำหรับการปรับเทียบที่มีกลุ่มตัวอย่างขนาดเล็ก แบบสอบต่างฉบับมีค่าความยากใกล้เคียงกัน มีขั้นตอนการปรับเทียบที่ดำเนินการได้ง่ายและชัดเจน วิเคราะห์ง่าย และเป็นวิธีที่อธิบายให้บุคคลโดยทั่วไปเข้าใจได้ง่าย วิธีการปรับเทียบอิควิเปอร์เซ็นต์ไทล์ (Equipercentile Equating) และวิธีการปรับเทียบตามทฤษฎีการตอบสนองข้อสอบ 3 พารามิเตอร์ นำไปใช้ในสถานการณ์ที่ความสัมพันธ์การปรับเทียบ (Equating Relationship) ไม่เป็นเชิงเส้นตรง ใช้กลุ่มตัวอย่างขนาดใหญ่ และวิธีการปรับเทียบตามทฤษฎีการตอบสนองข้อสอบมีข้อตกลงเบื้องต้นที่แข็งแกร่ง การตรวจสอบบริบทของการทดสอบมีความจำเป็นต้องทำ เพื่อให้มั่นใจว่าวิธีการแกร่งพอที่จะฝ่าฝืนข้อตกลงเบื้องต้นเหล่านี้ โดยเฉพาะวิธีของราส์ชที่ใช้กลุ่มตัวอย่างน้อยกว่าวิธีการปรับเทียบตามทฤษฎีการตอบสนองข้อสอบ 3 พารามิเตอร์ (Kolen and Brennan, 1995 ) สเก็กส์ ( Skaggs, 1986 ) มีความเห็นว่าการปรับเทียบคะแนนไม่มีวิธีการใดดีที่สุด ทั้งนี้ขึ้นอยู่กับสถานการณ์ของการปรับเทียบ และแฮมเบิลตันกับสวามินาธาน ( Hambleton and Swaminathan, 1985 ) แนะนำว่า เมื่อสถานการณ์การปรับเทียบมีให้เลือกหลายทาง การปรับเทียบตามทฤษฎีการตอบสนองข้อสอบดูเหมือนจะให้ผลน่าเชื่อถือได้ และเนื่องจากเงื่อนไขการปรับเทียบคะแนนจะต้องมีความเสมอภาค ( Equity ) มีความสมมาตร ( Symmetry ) และความไม่แปรเปลี่ยน (Invariance) ดังนั้นถ้าใช้วิธีการปรับเทียบตามแนวทฤษฎีการทดสอบแบบดั้งเดิมแล้วไม่สามารถเป็นไปตามเงื่อนไขเหล่านี้ได้ การปรับเทียบตามทฤษฎีการตอบสนองข้อสอบเท่านั้นที่เป็นไปตามเงื่อนไขเหล่านี้ได้

การประเมินผลการปรับเทียบคะแนน มีเกณฑ์ที่ใช้ดังนี้ ( Harris and Crouse, 1993) 1) ความเสมอภาค ( Weak Equity ) ของ Divgi และ Yen ที่พิจารณาจากความเท่าเทียมกันของการแจกแจงตามเงื่อนไขของคะแนนที่ได้จากแบบสอบต่างฉบับหลังจากการปรับเทียบแล้ว 2) ดัชนีสำหรับการเปลี่ยนแปลงคะแนน (Indices) ของ Angoff 3) ความคลาดเคลื่อนมาตรฐาน (Standard Error ) ของ Angoff เป็นการวิเคราะห์เพื่อประมาณความคลาดเคลื่อนของการปรับเทียบจากการ

สุ่มตัวอย่าง 4) ข้อมูลที่จำลองขึ้น ( Generated Data ) ของ Lord เพื่อใช้สำหรับการปรับเทียบคะแนน 5) การปรับเทียบคะแนนจากแบบสอบกลับสู่แบบสอบเดิม ( Equating a Test to Itself ) ของ Lord เป็นการปรับเทียบคะแนนกลับสู่แบบสอบเดิมโดยตรง หรือปรับผ่านแบบสอบอื่นก่อนปรับกลับสู่แบบสอบเดิม 6) กลุ่มตัวอย่างขนาดใหญ่ ( Large Sample ) ของ Angoff เป็นการใช้กลุ่มตัวอย่างขนาดใหญ่ในการปรับเทียบคะแนนซึ่งคล้ายกับการปรับเทียบคะแนนจากประชากร และใช้เปรียบเทียบกับผลการปรับเทียบคะแนนที่มีขนาดกลุ่มตัวอย่างน้อยกว่า 7) ความคงเส้นคงวา (Consistency) เป็นการประเมินผลการปรับเทียบคะแนนข้ามวิธีเพื่อหาความคงเส้นคงวา 8) ความคงที่ ( Stability ) ของ Angoff เป็นการปรับเทียบคะแนนซ้ำเพื่อตรวจความคงที่

ดิฟกิ ( Divgi, 1981) และ เยน (Yen, 1983 ) ได้แนะนำคุณสมบัติความเสมอภาค ( Equity ) ของ ลอร์ด ที่พิจารณาจากความเท่ากันของการแจกแจงอย่างมีเงื่อนไข (Conditional Distribution ) ของคะแนนที่ได้จากแบบสอบต่างฉบับหลังจากการปรับเทียบคะแนนแล้ว และได้นำไปใช้เป็นเกณฑ์เปรียบเทียบวิธีการปรับเทียบคะแนน ฮาร์ริสและครอส (Harris and Crouse,1995) มีความเห็นว่าเกณฑ์ความเสมอภาคนี้เหมาะที่จะนำไปใช้ในการปรับเทียบคะแนนตามนิยามความเสมอภาคของลอร์ดเท่านั้น และเป็นเกณฑ์ที่มีความยุ่งยากในการคำนวณและการแปลผล

ดัชนีความแตกต่างเป็นค่าที่แสดงให้เห็นว่า มีความคลาดเคลื่อนเกิดขึ้นจากการปรับเทียบคะแนนเพียงใด ซึ่งค่านี้ได้จากความแตกต่างระหว่างค่าจริงหรือค่าที่ใช้เป็นเกณฑ์ (True or Criterion Value) กับคะแนนที่ปรับแล้ว ถ้าความแตกต่างมีค่าน้อยหมายความว่า ความคลาดเคลื่อนที่เกิดจากการปรับเทียบคะแนนในครั้งนั้นน้อย การปรับเทียบคะแนนดังกล่าวย่อมมีความเหมาะสมกับสถานการณ์ที่ใช้ ดัชนีความแตกต่างได้ถูกใช้เป็นเกณฑ์เปรียบเทียบคุณภาพของการใช้วิธีการปรับเทียบคะแนนที่ต่างกัน หรือเปรียบเทียบคุณภาพการใช้วิธีการปรับเทียบคะแนนวิธีเดียวต่างเงื่อนไข ดัชนีความแตกต่างที่นักวัดผลนำมาใช้เป็นเกณฑ์ในกระบวนการปรับเทียบกันมากได้แก่ ดัชนี RMS (Root Mean Square) ดัชนี MAD (Mean Absolute Difference) และดัชนี MSD (Mean Sighed Difference) (Skaggs and Lissitz,1986b; Harris, 1987; Bejar and Wingersky,1982; Harris and Kolen, 1990; Livingston, Dorans and Wright, 1990; Eignor and Cook, 1991; Harris, 1991c cited in Harris and Crouse, 1993 ) เป็นเพราะดัชนีเหล่านี้คำนวณง่าย ซึ่งค่าดัชนี RMS ได้จากรากที่สองของค่าเฉลี่ยกำลังสองของผลต่างของค่าจริงหรือค่าที่ใช้เป็นเกณฑ์กับคะแนนที่ได้จากการปรับเทียบ ดัชนี MAD ได้จากค่าสัมบูรณ์เฉลี่ยของผลต่างระหว่างค่าจริงหรือค่าที่ใช้เป็นเกณฑ์กับคะแนนที่ได้จากการปรับเทียบ และดัชนี MSD ได้จากค่าเฉลี่ยของผลต่างระหว่างค่าจริงหรือค่าที่ใช้เป็นเกณฑ์กับคะแนนที่ได้จากการปรับเทียบ

ฮาร์ริส และ เคร้าส์ ( Harris and Crouse, 1993 ) ได้ศึกษาเกณฑ์เพื่อใช้ในการปรับเทียบคะแนน การปรับเทียบคะแนนกลับสู่แบบกลับสู่แบบสอบชุดเดิม ( Equating a Test to Itself ) เป็นเกณฑ์ที่ใช้กับสถานการณ์ปรับเทียบคะแนนระหว่างแบบสอบต่างฉบับ 3 ชุด คือ ชุด X, Y, Z ไปทดสอบกับกลุ่มผู้สอบ 3 กลุ่ม แต่ละกลุ่มทำแบบสอบคนละฉบับ การปรับคะแนนเริ่มจากการปรับคะแนนจากแบบสอบชุด X ไปยังชุด Y ปรับคะแนนจากแบบสอบชุด Y ต่อไปยังชุด Z และปรับคะแนนจากแบบสอบชุด Z กลับไปยังแบบสอบชุด X เดิม ถ้าการปรับเทียบมีคุณภาพเพียงพอ คะแนน 5 คะแนนจากแบบสอบชุด X เดิม กับชุด X ที่ปรับจะมีค่าเท่ากัน เกณฑ์นี้สามารถนำไปใช้กับการเก็บรวบรวมข้อมูลรูปแบบกลุ่มสุ่ม และรูปแบบกลุ่มไม่เท่าเทียมกันใช้แบบสอบร่วม เกณฑ์นี้ไม่เหมาะที่จะนำไปใช้ในการประเมินเพื่อเปรียบเทียบวิธีการปรับเทียบคะแนน ความยุ่งยากในการใช้เกณฑ์นี้คือการกำหนดแบบสอบชุดเริ่มต้นหรือชุดสุดท้าย (Harris and Crouse, 1993) แองกอฟ (Angoff, 1987) กล่าวว่าเกณฑ์การปรับเทียบคะแนนกลับสู่แบบสอบชุดเดิมเป็นเกณฑ์ที่มีประโยชน์เพราะทำให้รู้ผลการลดความคลาดเคลื่อน ใช้ในการประเมินความคลาดเคลื่อนในวงจรลูกโซ่การเปรียบได้อย่างแท้จริง เบรนแนน และ โคเลน ( Brennan and Kolen, 1987a,b ) แสดงให้เห็นว่าการปรับเทียบคะแนนกลับสู่แบบสอบชุดเดิม ที่ใช้การเก็บรวบรวมข้อมูลรูปแบบกลุ่มไม่เท่าเทียมกันใช้แบบสอบร่วม เมื่อใช้แบบสอบชุด X ปรับกลับสู่ชุด X เดิม โดยผ่านชุด Y และ ชุด Z ให้ผลแตกต่างกันกับการปรับชุด Z กลับสู่ชุด Z เดิม โดยผ่านชุด Y และ ชุด X

ฮาร์ริส และ เคร้าส์ ( Harris and Crouse, 1993 ) ได้รายงานเกี่ยวกับการใช้การปรับเทียบคะแนนที่จำลองขึ้น ( Simulated Equating ) โดยกำหนดโมเดลทางจิตมิติเพื่อใช้ในการปรับเทียบจริง และสร้างข้อมูลให้เหมาะสม ( Fit )กับโมเดล การปรับเทียบจริงที่เกิดขึ้นสามารถใช้เป็นเกณฑ์ได้

ปิเตอร์เซน มาร์โค และ สตีเวอร์ท (Pertersen, Marco and Stewart, 1982) ได้เสนอเกณฑ์ในการตัดสินคุณภาพของการปรับเทียบคะแนน โดยคำนวณค่าดัชนี MSE (Mean Square Error) ซึ่งเป็นค่าเฉลี่ยกำลังสองของผลต่างของคะแนนฐานกับคะแนนที่ปรับแล้ว ถ่วงน้ำหนักด้วยความแปรปรวนคะแนนฐาน และได้กำหนดเกณฑ์ตามความเห็นของคณะผู้วิจัย เป็นดังนี้ คือ

ระดับ น่าพอใจอย่างมาก เมื่อ $MSE < (.05\ S_X)^2$

ระดับน่าพอใจ เมื่อ $(.05\ S_X)^2 \leq MSE < (.10\ S_X)^2$

ระดับปานกลาง เมื่อ $(.10\ S_X)^2 \leq MSE < (.15\ S_X)^2$

ระดับไม่น่าพอใจ เมื่อ $(.15\ S_X)^2 \leq MSE < (.20\ S_X)^2$ และ

ระดับไม่น่าพอใจอย่างมาก เมื่อ $(.20\ S_X)^2 \leq MSE$

โดยที่ $S_x$ คือ ส่วนเบี่ยงเบนมาตรฐานของคะแนนฐาน

จากที่กล่าวมาการประเมินผลการปรับเทียบคะแนนเป็นขั้นตอนที่สำคัญยิ่ง เพราะจะทำให้ทราบว่าการปรับเทียบครั้งนี้มีความถูกต้องแม่นยำหรือไม่ มีคุณภาพระดับใด ผลการประเมินมีความมีความเชื่อถือได้หรือไม่ ข้อมูลเหล่านี้จะเป็นประโยชน์กับผู้ที่จะตัดสินใจนำผลการปรับเทียบคะแนนไปใช้ ซึ่งจะต้องมีเกณฑ์เป็นตัวตัดสิน เกณฑ์ที่ใช้เพื่อตัดสินคุณภาพการปรับเทียบคะแนนที่นิยมใช้กันมากคือดัชนีความแตกต่างระหว่างคะแนนจริงหรือคะแนนที่ใช้เป็นเกณฑ์ กับคะแนนที่ได้จากการปรับเทียบคะแนน เป็นเพราะค่าดัชนีนี้มีความชัดเจนในที่มาและการแปลผลที่ง่ายของเกณฑ์ (Harris and Crouse, 1993) และดัชนีความแตกต่างมีหลายตัวจำแนกตามหลักการที่ได้มาของค่าดัชนี ดังเช่น ดัชนี RMS, MAD และดัชนี MSD ค่าดัชนีเหล่านี้แสดงให้เห็นถึงความคลาดเคลื่อนของการปรับเทียบคะแนนซึ่งพิจารณาจากผลที่ได้จากการปรับเทียบคะแนน จะนำไปใช้เพื่อตัดสินผลการใช้วิธีการปรับเทียบที่แตกต่างกันว่าวิธีใดมีคุณภาพดีกว่า หรือตัดสินผลการใช้เงื่อนไขการปรับเทียบที่แตกต่างกันเมื่อใช้วิธีการปรับเทียบวิธีเดียวกัน มีทั้งการใช้ดัชนีเพียงตัวเดียวหรือหลายตัว แต่เมื่อต้องการตัดสินผลการปรับเทียบคะแนนที่เกิดขึ้นเพียงครั้งเดียว ไม่มีกลุ่มเปรียบเทียบ จะต้องอาศัยเกณฑ์ซึ่งเป็นค่าดัชนีที่แสดงถึงคุณภาพที่ยอมรับได้ของการปรับเทียบคะแนน จากการศึกษาพบว่ามีเกณฑ์ของปิเตอร์เซนและคณะที่กำหนดเป็นช่วงของค่าดัชนีมี 5 ระดับ ซึ่งเกณฑ์นี้ได้จากการใช้ดัชนี MSE แล้วถ่วงน้ำหนักด้วยส่วนเบี่ยงเบนมาตรฐานของคะแนนจริง กำหนดช่วงของค่าดัชนีโดยคณะผู้วิจัย เกณฑ์นี้ได้นำไปใช้ตัดสินคุณภาพการปรับเทียบคะแนนกันอย่างแพร่หลาย ( Petersen and Others, 1982; ภาวิณี ศรีสุขวัฒนานันท์, 2528; พรพิมล นาคเวช, 2537; พิชัย ละแมนชัย, 2538 และ วรรณ์ดี แสงประทีปทอง, 2538) เนื่องจากจุดตัดที่แบ่งระดับคุณภาพการปรับเทียบคะแนนสำหรับเกณฑ์ของปิเตอร์เซนและคณะ ได้กำหนดขึ้นโดยคณะผู้วิจัยเป็นการกำหนดเกณฑ์โดยผู้เชี่ยวชาญ โดยไม่ได้หาค่าดัชนีต่ำสุดที่แสดงถึงระดับคุณภาพที่ยอมรับได้ ซึ่งเป็นคุณสมบัติที่ดีในการพัฒนาเกณฑ์ ทำให้เกิดปัญหาว่าเกณฑ์ของปิเตอร์เซนและคณะมีความเชื่อถือได้มากน้อยเพียงใด ทำให้ผู้วิจัยสนใจที่จะพัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนที่กำหนดจุดตัดแตกต่างจากเกณฑ์ของปิเตอร์เซนและคณะ

การปรับเทียบคะแนนจำแนกตามทฤษฎีการวัด ได้แก่ การปรับเทียบตามทฤษฎีการวัดแบบดั้งเดิมที่มีวิธีการปรับเทียบอิควิเปอร์เซ็นไทล์ วิธีเปรียบเทียบค่าเฉลี่ย วิธีปรับเทียบเชิงเส้นตรง การใช้สมการถดถอย และการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ แต่เนื่องจากการปรับเทียบคะแนนตามทฤษฎีการวัดแบบดั้งเดิมนั้นไม่สามารถแก้ปัญหาบางจุดได้ เช่น ความเสมอภาค ความสมมาตรและความไม่แปรเปลี่ยน ด้วยเหตุนี้ผู้วิจัยจึงเลือกศึกษาตามแนวทฤษฎีการตอบสนองข้อสอบซึ่งสามารถแก้ปัญหาเหล่านี้ได้ และโมเดลที่นิยมใช้กันในทฤษฎีการตอบสนองข้อสอบคือ โมเดล 1 และ 3 พารามิเตอร์ การศึกษาครั้งนี้จึงเลือกศึกษาโมเดล 1 และ 3 พารามิเตอร์ ส่วน

วิธีการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบที่นิยมใช้กันมี 5 วิธี ดังนี้ คือ 1) วิธีกำหนดค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความสามารถผู้สอบทั้งสองกลุ่มให้เท่ากัน 2) วิธี Mean and Mean 3) วิธี Mean and Sigma 4) วิธี Characteristic Curve และ 5) วิธี Minimum $\chi^2$ การศึกษาครั้งนี้เลือก 4 วิธีแรก เนื่องจาก วิธี Characteristic Curve และวิธี Minimum $\chi^2$ ให้ผลการปรับเทียบคะแนนใกล้เคียงกัน (Kim and Cohen, 1995)

ในการวิจัยครั้งนี้ต้องการพัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ เป็นเกณฑ์ที่กำหนดจากค่าดัชนีความแตกต่างระหว่างคะแนนจริงที่ไม่ได้ปรับเทียบ ($\tau$) กับคะแนนจริงที่ปรับเทียบแล้ว ($\tau'$) และได้วิเคราะห์หาจุดตัดซึ่งเป็นจุดที่แบ่งดัชนีความแตกต่างออกเป็นกลุ่มดัชนีที่แสดงถึงคุณภาพการปรับเทียบระดับต่ำ และกลุ่มดัชนีที่แสดงถึงคุณภาพการปรับเทียบคะแนนในระดับสูง จุดตัดนี้ชี้ให้เห็นถึงความไม่แตกต่างกันระหว่างค่าเฉลี่ยของคะแนนจริงที่ไม่ได้ปรับกับค่าเฉลี่ยของคะแนนจริงที่ปรับเทียบแล้ว สำหรับกลุ่มผู้สอบกลุ่มเดียวกัน เป็นกลุ่มผู้สอบที่ได้มาจากการสุ่ม และค่า $\tau$ กับ $\tau'$ ระหว่างคู่เป็นอิสระจากกัน เนื่องจากการได้มาของข้อมูลเป็นไปตามข้อตกลงเบื้องต้นของสถิติ t-test ดังนั้นผู้วิจัยจึงเลือกใช้สถิติ t-test แบบ Two Dependent Sample Test เป็นสถิติทดสอบความแตกต่างระหว่างค่าเฉลี่ยของคะแนนจริง $\tau$ กับค่าเฉลี่ยคะแนนจริง $\tau'$ เพื่อกำหนดดัชนีความแตกต่างที่เป็นจุดตัดเมื่อค่า t จากการคำนวณมีค่าใกล้ค่าวิกฤตมากที่สุด และจัดเป็นกลุ่มดัชนีที่แสดงคุณภาพการปรับเทียบในระดับต่ำในกรณีที่ผลการทดสอบว่าค่าเฉลี่ยของคะแนนจริงทั้งสองแตกต่างกันอย่างมีนัยสำคัญ ส่วนในกรณีที่ผลการทดสอบแล้วว่าค่าเฉลี่ยของคะแนนจริงทั้งสองไม่แตกต่างกันจัดเป็นกลุ่มดัชนีที่แสดงคุณภาพการปรับเทียบคะแนนในระดับสูง

การหาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนที่นิยมใช้กันมาก ได้แก่การปรับเทียบกลับสู่แบบสอบเดิม เป็นการปรับเทียบคะแนนผ่านแบบสอบฉบับอื่นแล้วปรับกลับสู่แบบสอบเดิม จากนั้นคำนวณค่าดัชนีความแตกต่างระหว่างคะแนนจากแบบสอบเดิมกับคะแนนที่ปรับเทียบผ่านแบบสอบอื่นแล้วปรับกลับสู่แบบสอบเดิมเป็นเกณฑ์ และการใช้กลุ่มสอบทานผลเป็นกลุ่มผู้สอบที่ต้องทำแบบสอบทั้งสองฉบับ คำนวณค่าดัชนีความแตกต่างระหว่างคะแนนที่ไม่ได้ปรับเทียบกับคะแนนที่ปรับเทียบแล้วสำหรับกลุ่มสอบทานผลเป็นเกณฑ์ การหาเกณฑ์ทั้งสองวิธีนี้มีกระบวนการที่เชื่อถือได้ การแปลผลมีความชัดเจน และนิยมใช้กันโดยทั่วไป ในการวิจัยครั้งนี้ผู้วิจัยจึงใช้วิธีการหาเกณฑ์ทั้งสองรูปแบบนี้ และค่าดัชนีความแตกต่างที่ใช้เป็นเกณฑ์ได้แก่ ดัชนี RMS (Root Mean Square) ดัชนี MAD (Mean Absolute Difference) และดัชนี MSD (Mean Sighed Difference) ผู้วิจัยได้ใช้ทั้งสามดัชนี

ในเรื่องแบบแผนการเก็บรวบรวมข้อมูลในการปรับเทียบคะแนน ถึงแม้ว่าจะมี 3 รูปแบบ คือ รูปแบบกลุ่มเดียวจัดให้สมดุล รูปแบบกลุ่มสมมูล และรูปแบบใช้แบบสอบร่วม แต่การวิจัยที่ผ่านมามักจะไม่ค่อยพบเห็นการศึกษาตามรูปแบบกลุ่มเดี่ยวจัดให้สมดุล เพราะในทางปฏิบัติรูปแบบนี้นำมาใช้น้อยมาก ดังนั้นการวิจัยครั้งนี้จึงเลือกศึกษารูปแบบกลุ่มสมมูลและรูปแบบใช้แบบสอบร่วม โดยเฉพาะรูปแบบใช้แบบสอบร่วม ได้ใช้ข้อสอบร่วมจำนวน 20 % ของจำนวนข้อสอบที่ใช้ในการปรับเทียบคะแนนตามคำแนะนำของแองกอฟ (Angoff, 1984)

จากการศึกษางานวิจัยที่เกี่ยวข้องกับการปรับเทียบคะแนนพบว่าความยาวแบบสอบและจำนวนผู้สอบเป็นองค์ประกอบที่มีผลต่อการปรับเทียบคะแนน ผู้วิจัยจึงกำหนดความยาวแบบสอบและจำนวนผู้สอบเป็นเงื่อนไขในการจำลองข้อมูลเพื่อให้ได้ข้อมูลเกือบทุกเงื่อนไข โดยกำหนดความยาวแบบสอบและจำนวนผู้สอบตามที่ได้พบเห็นในการสอบทั่ว ๆ ไป ซึ่งความยาวแบบสอบเริ่มจาก 10, 20, 30 เพิ่มขึ้นครั้งละ 10 จนถึง 200 ข้อ เป็นความยาวแบบสอบ 20 ค่า ส่วนจำนวนผู้สอบเริ่มจาก 100, 200, 300 เพิ่มขึ้นครั้งละ 100 จนถึง 3,000 คน เป็นจำนวนผู้สอบ 30 ค่า

การจำลองข้อมูลเป็นเทคนิคที่ใช้กันมากในการศึกษาเกี่ยวกับการปรับเทียบคะแนน เป็นเพราะสามารถใช้ข้อมูลจำลองขึ้นในการปรับเทียบคะแนนได้หลากหลายแนวทาง ทำให้นักวิจัยเสียค่าใช้จ่ายน้อยในการทำวิจัย (Harris and Crouse, 1993) เมื่อเปรียบเทียบข้อดีและข้อเสียของการใช้ข้อมูลที่จำลองขึ้นและข้อมูลจริงพบว่า วิธีการเก็บข้อมูลจริงมักมีความคลาดเคลื่อนเกิดขึ้นเสมอเพราะมีตัวแปรแทรกซ้อนมากผู้เก็บข้อมูลไม่สามารถควบคุมตัวแปรเหล่านี้ได้ (ภาวิณี ศรีสุขวัฒนานันท์, 2528) ศึกษาตัวแปรได้น้อย และยังสิ้นเปลืองทรัพยากรที่มากกว่าอีกด้วย คุ๊กและปีเตอร์เซน (Cook and Petersen, 1987) แนะนำว่าควรได้มีการศึกษาตัวแปรที่เกี่ยวข้องกับการปรับเทียบคะแนนโดยใช้วิธีการจำลองข้อมูลเพราะจะทำให้ได้ผลที่ชัดเจนกว่า ดังนั้นในการวิจัยครั้งนี้ผู้วิจัยจึงเลือกศึกษาจากข้อมูลที่จำลองขึ้น

เพื่อให้ได้เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนที่มีความเชื่อถือได้ สามารถนำไปใช้ในสถานการณ์ที่หลากหลายสำหรับการปรับเทียบคะแนน ผู้วิจัยจึงได้พัฒนาเกณฑ์หรือจุดตัดที่แบ่งระดับคุณภาพจากดัชนี RMS, MAD และ AMD ที่ได้จากการปรับดัชนี MSD ( Mean Sign Difference) โดยใช้ค่าสัมบูรณ์ทำให้ค่าดัชนี MSD เป็นบวก และเปลี่ยนชื่อดัชนีเป็นดัชนี AMD (Absolute Mean Difference) ซึ่งได้จากข้อมูลการปรับเทียบคะแนนตามเงื่อนไขต่าง ๆ ดังนี้ คือ วิธีการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบโมเดล 1 และ 3 พารามิเตอร์ แต่ละโมเดลใช้วิธีการปรับเทียบ 4 วิธี คือ วิธีกำหนดให้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์ผู้สอบทั้งสองกลุ่มให้เท่ากันก่อนการวิเคราะห์ วิธี Mean & Mean วิธี Mean & Sigma และวิธีโค้งคุณลักษณะข้อสอบ แบบแผนการเก็บรวบรวมข้อมูลแบบกลุ่มสุ่มและแบบกลุ่มไม่เท่าเทียม

กันใช้แบบสอบร่วม รูปแบบการหาเกณฑ์การปรับเทียบคะแนนกลับสู่แบบสอบเดิมและการใช้กลุ่มสอบทานผล ใช้ขนาดกลุ่มตัวอย่าง 30 ขนาด คือ 100, 200,...,3,000 คน และจำนวนข้อสอบในแต่ละฉบับจำนวน 20 ขนาด คือ 10, 20, 30,...,200 ข้อ เป็นเงื่อนไขในการจำลองข้อมูล ใช้ t-test แบบ Two Dependent Sample Test วิเคราะห์เพื่อกำหนดค่าดัชนีที่เป็นจุดตัดแบ่งกลุ่มดัชนีที่แสดงระดับคุณภาพการปรับเทียบคะแนน จุดตัดนี้โดยวิเคราะห์หาค่า t จากการคำนวณที่มีค่าเท่ากับค่าวิกฤต ซึ่งค่าวิกฤตนี้เป็นค่าที่บอกความแตกต่างกันอย่างมีนัยสำคัญระหว่างค่าเฉลี่ยของคะแนนจริงที่ไม่ได้ปรับเทียบ ($\tau$) กับค่าเฉลี่ยคะแนนจริงที่ปรับเทียบแล้ว ($\tau'$) สำหรับผู้สอบกลุ่มเดียวกัน แล้วจึงแบ่งค่าดัชนีออกเป็นช่วง ๆ จากจุดที่เริ่มต้นที่มีคุณภาพยอมรับได้ อันจะทำให้ได้เกณฑ์ที่มีความเชื่อถือได้ทางวิชาการ เพื่อใช้ในการตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ และเป็นการขยายองค์ความรู้ด้านเกณฑ์ตัดสินผลการปรับเทียบคะแนนให้พัฒนายิ่งขึ้นไป โดยมีคำถามวิจัยดังนี้

1. เกณฑ์ที่ได้จากการศึกษาครั้งนี้มีลักษณะเป็นอย่างไร
2. เกณฑ์ที่ได้จากการศึกษาครั้งนี้มีคุณภาพเป็นอย่างไร

## วัตถุประสงค์ของการวิจัย

การวิจัยครั้งนี้ มีวัตถุประสงค์หลักเพื่อพัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ โดยมีวัตถุประสงค์เฉพาะดังนี้

1. เพื่อพัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ ด้วยข้อมูลจำลอง จากรูปแบบการหาเกณฑ์ 2 แบบ คือ การปรับเทียบคะแนนกลับสู่แบบสอบเดิม และการใช้กลุ่มสอบทานผล
2. เพื่อตรวจสอบคุณภาพของเกณฑ์ที่พัฒนาขึ้น ในด้านความตรงเชิงเกณฑ์สัมพันธ์ โดยตรวจสอบความสอดคล้องของผลการตัดสินด้วยเกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ของปิเตอร์เซนและคณะ และความสอดคล้องของผลการตัดสินด้วยเกณฑ์ที่พัฒนาขึ้นกับเกณฑ์ความเสมอภาคของลอร์ด

## ขอบเขตของการวิจัย

1. เกณฑ์ที่พัฒนาขึ้นในแต่ละแบบแผน ใช้การจำลองข้อมูลแต่ละครั้งจากกลุ่มตัวอย่างต่างกลุ่ม
2. ตัวแปรในการวิจัยประกอบด้วย
   2.1 ตัวแปรอิสระ มีดังต่อไปนี้
      2.1.1 แบบแผนการเก็บรวบรวมข้อมูล ได้แก่ การใช้กลุ่มสมมูลและการใช้ข้อสอบร่วม

2.1.2 กระบวนการวิเคราะห์หาคุณภาพการปรับเทียบคะแนน ได้แก่ การปรับเทียบกลับสู่แบบสอบเดิมและการใช้กลุ่มสอบทานผล

2.1.3 โมเดลการตอบสนองข้อสอบ ได้แก่ โมเดล 1 และ 3 พารามิเตอร์

2.1.4 วิธีการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ ได้แก่ วิธี Same Scaling Convention, วิธี Mean and Mean, วิธี Mean and Sigma, และ วิธี Characteristic Curve

2.2 ตัวแปรตาม ได้แก่ เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการการตอบสนองข้อสอบ

**คำนิยามเชิงปฏิบัติการที่ใช้ในการวิจัย**

**การปรับเทียบคะแนน หมายถึง** การแปลงระบบคะแนนของแบบสอบต่างฉบับที่วัดเนื้อหาเดียวกัน โดยแปลงคะแนนของแบบสอบฉบับหนึ่งให้อยู่บนมาตรหรือสเกลของแบบสอบอีกฉบับหนึ่ง ที่ถือเป็นคะแนนที่สมมูลกัน เพื่อทำให้คะแนนจากแบบสอบต่างฉบับกันสามารถเปรียบเทียบกันได้โดยตรง สำหรับการวิจัยในครั้งนี้เป็นการแปลงคะแนนของแบบสอบต่างฉบับที่วัดเนื้อหาเดียวกัน ด้วยวิธีการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ

**วิธีการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ หมายถึง** การนำหลักการของทฤษฎีการตอบสนองข้อสอบ (tem Response Theory) โดยการประมาณค่าพารามิเตอร์ผู้สอบและพารามิเตอร์ข้อสอบมาใช้ในการปรับเทียบคะแนน โดยยึดหลักการว่า ถ้าพารามิเตอร์ผู้สอบอยู่บนสเกลเดียวกันแล้ว สามารถนำคะแนนจริงจากแบบสอบสองฉบับมาเทียบกันได้ สำหรับการวิจัยครั้งนี้หมายถึงการนำคะแนนจริงจากแบบสอบฉบับที่ 1 (X) ไปเทียบกับคะแนนจริงของแบบสอบอีกฉบับหนึ่ง (Y) ที่ระดับความสามารถ $\theta$ เดียวกัน

**แบบแผนการหาเกณฑ์จากการปรับเทียบคะแนนกลับสู่แบบสอบเดิม หมายถึง** กระบวนการหาดัชนีความแตกต่างที่ใช้เป็นเกณฑ์ตัดสินคุณภาพของการปรับเทียบคะแนน โดยใช้แบบสอบอย่างน้อย 3 ชุด เช่น ชุด X, Y และ Z แต่ละชุดใช้ทดสอบกับกลุ่มผู้สอบ 3 กลุ่ม นำคะแนนจากแบบสอบชุด X ปรับเทียบไปยังชุด Y ปรับคะแนนจากชุด Y ต่อไปยังชุด Z และปรับคะแนนจากชุด Z กลับไปยังชุด X เดิม สำหรับการปรับเทียบที่มีคุณภาพคะแนนจากแบบสอบ X เดิม กับคะแนนจากแบบสอบ X ที่ปรับผ่านแบบสอบ Y และ Z จะแตกต่างกันเพียงเล็กน้อย โดยเกณฑ์ตัดสินคุณภาพของการปรับเทียบคะแนนพัฒนาจากดัชนีความแตกต่าง AMD (Absolute Mean Square) ดัชนี MAD ( Mean Absolute Difference) หรือดัชนี RMS ( Root Mean Square) ระหว่างคะแนนจาก

แบบสอบชุด X ที่ไม่ได้ปรับกับคะแนนจากแบบสอบชุด X ปรับผ่านแบบสอบชุด Y และชุด Z แล้วปรับเข้าสู่แบบสอบชุด X เดิม

**แบบแผนการหาเกณฑ์โดยใช้กลุ่มสอบทานผล หมายถึง** กระบวนการหาดัชนีความแตกต่างที่ใช้เป็นเกณฑ์ โดยใช้แบบสอบต่างฉบับที่วัดในเนื้อหาเดียวกัน ไปทดสอบกับกลุ่มผู้สอบคนละกลุ่มกัน ทำตารางปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ กำหนดกลุ่มสอบทานผลเป็นกลุ่มผู้สอบที่สุ่มมาจากประชากรเดียวกัน หรือเป็นกลุ่มที่มีความสามารถใกล้เคียงกับกลุ่มผู้สอบสองกลุ่มข้างต้น โดยกลุ่มนี้ทำแบบสอบทั้งสองฉบับ ปรับคะแนนที่ได้จากกลุ่มสอบทานผลจากแบบสอบชุด X สู่ชุด Y โดยใช้ตารางปรับเทียบเดิม หาค่าดัชนีความแตกต่างจากคะแนนจากแบบสอบชุด Y ที่ไม่ได้ปรับเทียบ กับคะแนนจากแบบสอบชุด X ที่ปรับเทียบแล้วจากกลุ่มสอบทานผล

**คุณภาพของการปรับเทียบคะแนน หมายถึง** ความถูกต้องแม่นยำของการปรับเทียบคะแนน ซึ่งพิจารณาจากเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนน ประกอบด้วย ค่าดัชนี AMD ดัชนี MAD และ ดัชนี RMS ที่กำหนดเป็นเกณฑ์

**เกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนน หมายถึง** ค่าที่แสดงถึงระดับคุณภาพที่ยอมรับได้ของผลการปรับเทียบคะแนน เป็นช่วงของค่าดัชนี AMD ดัชนี MAD และดัชนี RMS ที่บอกระดับคุณภาพของการปรับเทียบ

**ค่าดัชนี** AMD (Absolute Mean Difference) **หมายถึง** ค่าความคลาดเคลื่อนในการปรับเทียบคะแนน ซึ่งคำนวณได้จากค่าสัมบูรณ์ของค่าเฉลี่ยของความแตกต่างระหว่างคะแนนจริงที่ได้จากแบบสอบชุดที่ 1 ที่ไม่ได้ปรับเทียบกับคะแนนจริงที่ได้จากการปรับจากแบบสอบชุดที่ 1 ผ่านแบบสอบชุดที่ 2 และชุดที่ 3 แล้วปรับกลับไปยังแบบสอบชุดที่ 1 เดิม เมื่อหาเกณฑ์จากการปรับเทียบคะแนนกลับสู่แบบสอบเดิม และความแตกต่างระหว่างคะแนนจริงที่ไม่ได้ปรับกับคะแนนจริงที่ปรับมาจากอีกแบบสอบหนึ่ง เมื่อการหาเกณฑ์โดยใช้กลุ่มสอบทานผล

**ค่าดัชนี** MAD (Mean Absolute Difference) **หมายถึง** ค่าความคลาดเคลื่อนในการปรับเทียบคะแนน ซึ่งคำนวณได้จากค่าเฉลี่ยของค่าสัมบูรณ์ของความแตกต่างระหว่างคะแนนจริงที่ได้จากแบบสอบชุดที่ 1 ที่ไม่ได้ปรับเทียบกับคะแนนจริงที่ได้จากการปรับจากแบบสอบชุดที่ 1 ผ่านแบบสอบชุดที่ 2 และชุดที่ 3 แล้วปรับกลับไปยังแบบสอบชุดที่ 1 เดิม เมื่อหาเกณฑ์จากการปรับเทียบคะแนนกลับสู่แบบสอบเดิม และความแตกต่างระหว่างคะแนนจริงที่ไม่ได้ปรับกับคะแนนจริงที่ปรับมาจากอีกแบบสอบหนึ่ง เมื่อการหาเกณฑ์โดยใช้กลุ่มสอบทานผล

**ค่าดัชนี** RMS (Root Mean Square) **หมายถึง** ค่าความคลาดเคลื่อนในการปรับเทียบคะแนน ซึ่งคำนวณได้จากรากที่สองของค่าเฉลี่ยของกำลังสองความแตกต่างระหว่างคะแนนจริงที่ได้จากแบบสอบชุดที่ 1 ที่ไม่ได้ปรับเทียบกับคะแนนจริงที่ได้จากการปรับจากแบบสอบชุดที่ 1 ผ่านแบบสอบชุดที่ 2 และชุดที่ 3 แล้วปรับกลับไปยังแบบสอบชุดที่ 1 เดิม เมื่อหาเกณฑ์จากการปรับเทียบคะแนนกลับสู่แบบสอบเดิม และความแตกต่างระหว่างคะแนนจริงที่ไม่ได้ปรับกับคะแนนจริงที่ปรับมาจากอีกแบบสอบหนึ่ง เมื่อการหาเกณฑ์โดยใช้กลุ่มสอบทานผล

**ประโยชน์คาดว่าจะได้รับ**

1. ได้เกณฑ์ที่มีความเชื่อถือได้ทางวิชาการ เพื่อใช้ในการตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ
2. ได้เกณฑ์ที่เหมาะสมกับสภาพการปรับเทียบที่หลากหลาย เพื่อใช้ในการตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ
3. ได้ขยายองค์ความรู้ด้านเกณฑ์ตัดสินผลการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ