

ผลการเปรียบเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT ภายใต้เงื่อนไขที่แตกต่างกัน

The Comparison of Test Scores Derived Through Kernel Equating and IRT Equating Methods Under Varied Conditions

ศศิธร ชุตินันท์กุล¹ ไพรัตน์ วงษ์นาม² สมพงษ์ ปันหูน³

Sasithorn Chutinuntakul¹ Pirat Wongnam² Sompong Panhoon³

¹ นิติศาสตรบัณฑิต สาขาวิชานิติศาสตร์ วิทยาลัยนิติศาสตร์ มหาวิทยาลัยบูรพา

² รองศาสตราจารย์ ดร. ประจำคณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

³ อาจารย์ ดร. ประจำคณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาวิธีการปรับเทียบคะแนนที่มีคุณภาพภายใต้เงื่อนไขรูปแบบข้อสอบร่วม ขนาดตัวอย่างและรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกัน และ 2) เปรียบเทียบความสอดคล้องของผลการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขที่กำหนด ผลการวิจัยพบว่า การปรับเทียบคะแนนโดยวิธีเคอเนล มีคุณภาพในการปรับเทียบคะแนนสูงกว่าวิธี IRT 2 พารามิเตอร์ โดยที่การวิเคราะห์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสูงกับตัวอย่างขนาด 700 คน และเงื่อนไขข้อสอบร่วมมีความยากอย่างสูงและตัดข้อสอบที่ไม่มีคุณภาพทั้งกับตัวอย่างขนาด 500 คน มีคุณภาพในการปรับเทียบคะแนนสูงที่สุด การตัดเกรดจากคะแนนก่อนการปรับเทียบคะแนนและคะแนนหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ ตามเงื่อนไขต่าง ๆ ภายใต้การตัดเกรด 3 ระดับ และ 8 ระดับ ส่วนใหญ่พบว่าไม่สอดคล้องกัน จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนที่จะตัดเกรด การขยายระดับการตัดเกรดเป็น 8 ระดับ การใช้ตัวอย่างขนาดใหญ่ขึ้น จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากขึ้น ส่วนการตัดข้อสอบที่ไม่มีคุณภาพทั้งก่อนการปรับเทียบคะแนนด้วยวิธีเคอเนล จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนกว่าวิธี IRT 2 พารามิเตอร์ ขณะที่การใช้ข้อสอบร่วมที่มีความยากอยู่ในช่วง .4 -.6 ทั้งสองวิธีจะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสูง การตัดเกรดหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ ของทุกเงื่อนไขมีความสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 โดยที่การตัดเกรด 3 ระดับ หลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสูง กับทุกขนาดตัวอย่างมีความสัมพันธ์กันในระดับดีมาก

คำสำคัญ: การปรับเทียบคะแนน เคอเนล IRT

Abstract

The objectives of this research were 1) to examine the appropriate test equating methods under varied anchor test patterns, sample sizes, and data formats, and 2) to compare the consistency of the grading results of scores before and after test equating under the specified conditions. The findings were as follows: Kernel equating method with the condition of randomized anchor items under samples size 700 and the condition of randomized anchor items with non-quality questions removed when using 500 samples were slightly best quality. Grading of the scores received before and after equating using Kernel method and 2-parameter IRT method based on varied conditions under 3-level and 8-level grading mainly resulted in inconsistency. The scores needed to be equated before grading. Expanding grading into 8 levels and bigger sample size resulted in more obvious inconsistency. Removing non-quality questions before equating with Kernel method resulted in more obvious inconsistency than 2-parameter IRT method. However, using anchor items with a difficulty of .4-.6 in both methods resulted in more obvious inconsistency in grading than with randomized difficulty levels. Besides this study found that the relationship of grading after equating by using Kernel method and 2-parameter IRT method under any conditions were statistically significant at .05 level. The two methods under 3-level grading by using randomized anchor test all the same sample size were the best relationship.

Keywords: Equating, Kernel, IRT

ความเป็นมาและความสำคัญของปัญหา

การวัดและการประเมินผลทางการศึกษา เป็นองค์ประกอบสำคัญของกระบวนการจัดการศึกษา ที่จะทำให้ทราบว่าผลของการจัดการศึกษามัธยฐานตามวัตถุประสงค์ของการจัดการศึกษามากน้อยเพียงใด โดยอาศัยเครื่องมือวัดก็คือ แบบสอบโดยเฉพาะแบบสอบแบบเลือกตอบซึ่งเชื่อว่าผลการสอบจะมีความเที่ยง ความตรง ความยุติธรรมสูง และมีความสะดวกในการตรวจให้คะแนน แบบสอบจึงเป็นเครื่องมือที่สำคัญ ทำอย่างไรจึงจะทำให้การใช้แบบสอบเกิดความเป็นธรรมกับผู้เข้าสอบทุกคน โดยเฉพาะในสถานการณ์ที่จำเป็นจะต้องใช้แบบสอบหลายฉบับในการวัดผลคราวเดียวกัน ซึ่งมักจะพบในสถานการณ์ที่มีผู้เข้าสอบจำนวนมาก ถ้ามาสอบพร้อมกัน จะมีปัญหาเรื่องสถานที่สอบไม่เพียงพอ ไม่มีคนคุมสอบ จำเป็นต้องจัดสอบหลายครั้งต่างเวลา ต่างสถานที่สอบ ปัญหาที่พบคือ ผู้สอบครั้งแรกนำข้อสอบไปเผยแพร่ เกิดการรั่วไหลของข้อสอบ ผู้สอบครั้งหลังรู้ข้อสอบ มีการศึกษาค้นคว้าเพิ่มเติม เพื่อแก้ไขปัญหาดังกล่าวจึงใช้แบบสอบหลายฉบับ ซึ่งโดยทั่วไปจะทำเป็นแบบสอบคู่ขนานโดยคาดหวังว่าแบบสอบจะมีความเท่าเทียมกัน แต่ในทางปฏิบัติการสร้างแบบสอบให้เป็นคู่ขนานที่มีความเท่าเทียมกันทางด้านโครงสร้าง เนื้อหา ค่าสถิติของข้อสอบ มีความเที่ยง ความตรง ค่าเฉลี่ย ความแปรปรวนและความยากง่าย มีโอกาสเป็นไปได้น้อยมาก ซึ่งถ้าจะสร้างให้คู่ขนานกันจริง ๆ จะต้องใช้เวลาและงบประมาณจำนวนมาก และอาจจะไม่คุ้มค่ากับการลงทุนเพราะถ้าหากมีการเปลี่ยนแปลงเนื้อหาหรือเปลี่ยนจุดเน้นที่ต้องการวัด หรือเพิ่มวัตถุประสงค์

บางประการ จะทำให้แบบสอบชุดใหม่มีโครงสร้างแตกต่างไปจากชุดเดิม การสร้างแบบสอบคู่ขนานจึงมีข้อจำกัด และทำได้ยากมาก ในทางปฏิบัติจึงเป็นเพียงการทำให้แบบสอบมีความใกล้เคียงกันทางด้านเนื้อหา แต่มีความแตกต่างกันในเรื่องความยากและค่าสถิติอื่นๆของแบบสอบ เป็นเหตุให้โครงสร้างของแบบสอบทั้งสองชุดนั้น ยังคงแตกต่างกัน (Holland & Rubin, 1982) จึงมีผู้คิดค้นวิธีการใหม่ โดยอาศัยวิธีการทางสถิติมาช่วยทำให้คะแนนที่วัดจากแบบสอบต่างชุดกันนี้ ให้สามารถเทียบแทนกันได้ กระบวนการนี้เรียกว่า “การปรับเทียบคะแนน (Test Equating)” ซึ่งเป็นกระบวนการใช้เทคนิคการปรับทางสถิติทำให้คะแนนจากแบบสอบต่างชุดกัน สามารถปรับชุดเซชสำหรับความแตกต่างในการสอบและคุณลักษณะของข้อสอบ ให้สามารถเทียบแทนกันได้อย่างยุติธรรม (ศิริชัย กาญจนวาสี, 2555, น.157)

นอกจากเรื่องของความยุติธรรมกับผู้สอบแล้ว ยังมีมุมมองในเรื่องของสิทธิมนุษยชนที่ทุกคนควรจะได้รับ ความคุ้มครอง ในสิทธิเบื้องต้นที่จะได้รับความเท่าเทียมกันในการสอบ จากแบบสอบที่มีความยากง่ายเท่าเทียมกัน เพราะผลของการสอบจากแบบสอบที่มีความยากแตกต่างกันดังกล่าวจะนำไปสู่การสอบได้สอบตก การได้เกรดสูง-ต่ำ อันจะส่งผลต่อเกรดเฉลี่ย ต่อการสำเร็จการศึกษา ต่อการสมัครเข้าทำงาน ต่อการศึกษาในระดับที่สูงขึ้น ต่อการเลื่อนระดับหรือการจัดตำแหน่ง ฯลฯ จึงจำเป็นจะต้องดำเนินการในเรื่องของการวัดและประเมินด้วยความถูกต้องและเป็นธรรม

การปรับเทียบคะแนนจึงเป็นเทคนิควิธีที่มีบทบาทสำคัญ ในการนำมาใช้แก้ปัญหาในสถานการณ์ต่าง ๆ ดังกล่าวข้างต้น แต่จากการศึกษางานวิจัยพบว่ยังไม่มีวิธีการปรับเทียบคะแนนวิธีการใดที่ดีที่สุด ขึ้นอยู่กับเงื่อนไขของการปรับเทียบคะแนนที่จะนำไปใช้ในสถานการณ์ต่างๆ ที่แตกต่างกันในแต่ละบริบทที่ทำการศึกษา (พัชรจิณทร์เพ็ง, 2551, น.5) แบบสอบที่มีความยากแตกต่างกันการปรับเทียบคะแนนด้วยทฤษฎีตอบข้อสอบจะมีความเหมาะสมมากที่สุด รองลงมา คือ การเทียบคะแนนแบบเชิงเส้นตรง (Kolen, 1981, p. 9) ส่วนงานวิจัยที่ศึกษาการเปรียบเทียบคะแนนในแนวตั้งของ Patience โดยใช้ วิธีอิกวิเปอร์เซ็นไทล์ วิธีการของเชอร์สโตน วิธีการใช้รูปแบบโลจิสติก 1 2 และ 3 พารามิเตอร์ พบว่าให้ผลการปรับเทียบคล้ายคลึงกันแบบสอบฉบับง่ายและยากปานกลาง แต่ถ้าเป็นแบบสอบฉบับยากวิธีการของเชอร์สโตนและวิธี 2 พารามิเตอร์จะให้ผลการปรับเทียบที่ดีกว่า (Patience, 1990 อ้างอิงถึงใน ภัทรพร เกษสังข์, 2546, น.5) สำหรับขนาดกลุ่มตัวอย่างที่มีขนาดใหญ่วิธีการปรับเทียบคะแนนที่เหมาะสมได้แก่ วิธีอิกวิเปอร์เซ็นไทล์ วิธี IRT 3 พารามิเตอร์ ถ้ากลุ่มตัวอย่างมีขนาดเล็กวิธีการปรับเทียบคะแนนที่เหมาะสมได้แก่วิธีการปรับเทียบเชิงเส้นตรง (Linear equating) วิธีของราสช์ (Rasch) วิธีคอนเนล (KE : Kernel Equating) แต่ถ้ากลุ่มตัวอย่างมีขนาดเล็กมากๆ วิธีการปรับเทียบคะแนนที่เหมาะสมได้แก่วิธี Identity วิธี Mean (Kolen & Brennan, 2004, pp. 293 – 294; Dvorak, 2009) นอกจากนี้พบว่าวิธี KE จะใช้กลุ่มตัวอย่างที่น้อยกว่าวิธีอื่น ๆ และให้ค่าความคลาดเคลื่อนในการปรับเทียบที่น้อยกว่าวิธีอื่น ๆ (Dvorak, 2009, p. 76) และยังพบว่าวิธี KE ให้ผลการปรับเทียบที่ดีกว่าวิธีเชิงเส้นตรงและอิกวิเปอร์เซ็นไทล์ที่เป็นวิธีการปรับเทียบแบบดั้งเดิม (Mao et al., 2006; von Davier et al., 2006 cited in Meng, 2012, p. 40) นอกจากนี้พบว่าวิธีตามทฤษฎีตอบข้อสอบแบบโลจิสติก 2 พารามิเตอร์ ที่ใช้แบบสอบร่วมภายใน ที่มีความยาว 15 ข้อ มีความยากเฉลี่ยระดับยากมาก จะให้ค่าความคลาดเคลื่อนมาตรฐานของการเชื่อมโยงคะแนนต่ำสุด รวมทั้งมีความเพียงพอของการเชื่อมโยงคะแนนดีที่สุดเมื่อเทียบกับวิธีตามทฤษฎีตอบข้อสอบแบบโลจิสติก 3 พารามิเตอร์ (ภัทรพร เกษสังข์, 2546, น.

130) ดังนั้นในบริบทที่ชุดวิชาต่าง มีผู้เรียนตั้งแต่ร้อยละหนึ่งจนกระทั่งมีจำนวนมากจึงจำเป็นต้องเลือกใช้วิธีการเปรียบเทียบที่เหมาะสม เพื่อความเหมาะสมกับกลุ่มขนาดเล็กผู้วิจัยจึงเลือกศึกษาวิธีการเปรียบเทียบโดยใช้วิธีคอนเนล ส่วนกลุ่มตัวอย่างที่มีขนาดปานกลางและใหญ่เลือกใช้วิธีตามทฤษฎีตอบสนองข้อสอบแบบโลจิสติก 2 พารามิเตอร์ ที่มีความเหมาะสมกับขนาดตัวอย่างมากกว่าวิธี 1 พารามิเตอร์ซึ่งเหมาะกับกลุ่มขนาดเล็กส่วนวิธี 3 พารามิเตอร์ เหมาะกับกลุ่มที่มีขนาดใหญ่ที่มีขนาดมากกว่า 1,000 คน (Kolen & Brennan, 2004, p. 288)

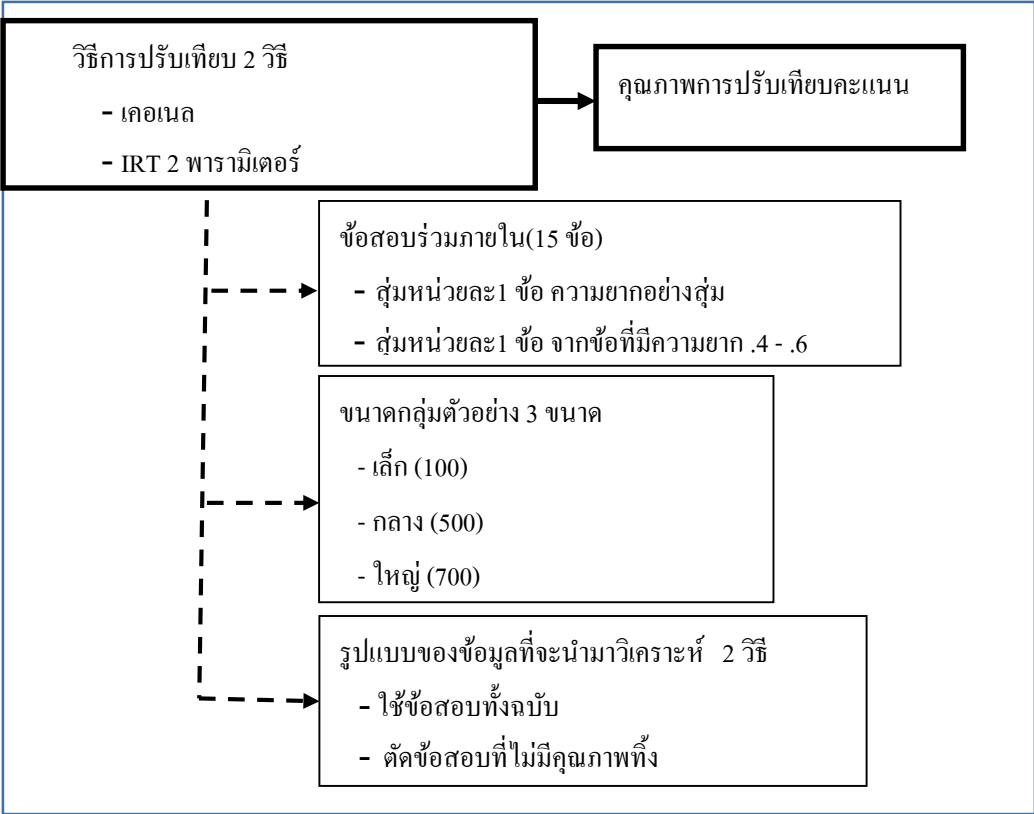
สำหรับการใช้ข้อสอบรวม ซึ่งเป็นอีกองค์ประกอบที่สำคัญมากต่อการเปรียบเทียบคะแนน โดยเฉพาะใน สถานการณ์ที่กลุ่มผู้สอบไม่ได้มาจากกลุ่มประชากรเดียวกัน รูปแบบการเปรียบเทียบโดยใช้ข้อสอบรวมจึงมีความ เหมาะสม ข้อสอบรวมที่ดีจะต้องเป็นตัวแทนทั้งเนื้อหาและค่าสถิติจากแบบสอบชุดเก่าที่จะนำไปใช้ในการ เปรียบเทียบ ข้อสอบรวมจึงต้องกระจายทั่วทั้งฉบับ และไม่มีการเปลี่ยนแปลงคำพูดหรือภาษาใหม่ในแบบสอบทั้ง ใหม่และเก่า (Kolen & Brennan, 2004, p.19) รวมทั้งไม่มีการเปลี่ยนแปลงตำแหน่งของข้อสอบรวมทั้ง 2 ฟอรัม เช่นในฉบับแรกข้อสอบรวมอยู่ตอนต้นของแบบสอบ แต่ฉบับหลังข้อสอบรวมอยู่ตอนท้าย เพราะจะมีผลต่อการ เปรียบเทียบ (Kolen & Brennan, 2004, p.19; Cook & Petersen, 1987; Eignor, 1985; Kolen & Harris, 1990) และควร หลีกเลี่ยงการปรับปรุงแก้ไขหรือเรียงตัวเลือกใหม่ (Cizek, 1994) นอกจากนี้หากข้อสอบรวมมีความยากปานกลาง จะช่วยลดความคลาดเคลื่อนในการเปรียบเทียบคะแนนได้ดี (Holland & Sinharay, 2007) ซึ่งสอดคล้องกับการศึกษา ของคาลด์เวลล์ (Caldwell, 1984) การศึกษาประสิทธิภาพของโมเดลการเปรียบเทียบเชิงเส้นตรงกับราชส์โมเดล โดยใช้ ข้อสอบรวม 2 รูปแบบ คือข้อสอบรวมที่ยากมาก กับยากปานกลาง ผลปรากฏว่าข้อสอบรวมที่มีความยากปานกลาง ให้ประสิทธิภาพที่ดีกว่า (Caldwell, 1984 อ้างถึงในพิชัย ละแมนชัย, 2538, น.49) ส่วนกรณีความยาวของข้อสอบรวม แองกอฟฟ์ (Angoff 1984, p.107) กล่าวว่าความยาวของข้อสอบรวมควรมีจำนวนไม่น้อยกว่าร้อยละ 20 ของจำนวน ข้อสอบในแบบสอบหรือควรมีจำนวนไม่น้อยกว่า 20 ข้อ แล้วแต่จำนวนไหนจะมากกว่าให้ใช้จำนวนนั้น ส่วนไรท์ และสโตน (Wright & Stone, 1979, p.98) กล่าวว่าข้อสอบรวมที่วัดเรื่องเดียวกันกับแบบสอบทั้งสองชุด มีจำนวน เพียง 10 ข้อก็เพียงพอแล้ว นอกจากนี้จากการศึกษาของแมคคินเลย์และเล็กเคส (Mckinley & Rackase 1981 cited in Cook & Petersen, 1987) ในการใช้ข้อสอบรวม 3 ขนาด คือ 5 15 และ 25 ข้อ โดยใช้ข้อมูลจริงในการเปรียบเทียบ รูปแบบ IRT 3 พารามิเตอร์ พบว่าข้อสอบรวมตั้งแต่ 15 ข้อ เป็นจำนวนที่เพียงพอสำหรับการเปรียบเทียบ (สุนิสา จุ้ยม่วงศรี, 2537, น.4) การวิจัยครั้งนี้จึงเลือกข้อสอบรวมที่ 15 ข้อ เพื่อใช้เป็นตัวแทนของเนื้อหาวิชาในชุดวิชาที่ ประกอบด้วย 15 หน่วย หน่วยละ 1 ข้อ

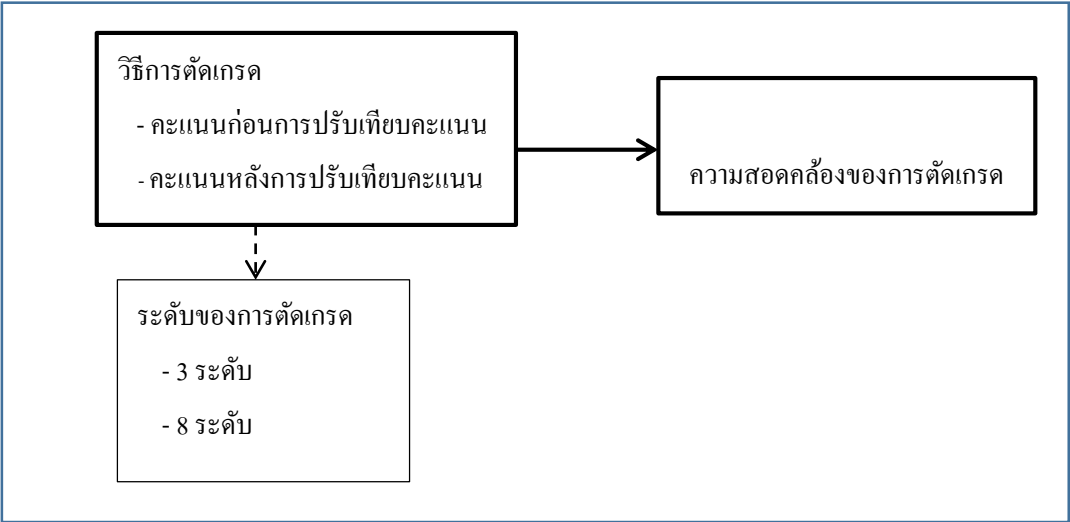
ตัวแปรต่างๆ ไม่ว่าจะเป็นวิธีการเปรียบเทียบ ลักษณะและขนาดของข้อสอบรวมที่มีผลต่อคุณภาพของการ เปรียบเทียบคะแนน ยังมีขนาดของกลุ่มตัวอย่างที่มีผลต่อการเปรียบเทียบคะแนน โดยที่ขนาดของกลุ่มตัวอย่างมี ผลกระทบโดยตรงต่อความคลาดเคลื่อนอย่างสุ่มของการเปรียบเทียบ ขนาดของกลุ่มตัวอย่างยิ่งมากยิ่งดี (Kolen & Brennan, 2004, p. 288) จึงควรเลือกใช้ขนาดตัวอย่างให้เหมาะสมในการเปรียบเทียบคะแนน ซึ่งยูเม็ง (Meng, 2012, p.45) ได้กำหนดว่าถ้าใช้กลุ่มตัวอย่างขนาด 200 คนต่อฉบับ ถือเป็นกลุ่มตัวอย่างขนาดเล็ก ขนาด 500 คนต่อฉบับ ถือเป็นกลุ่มตัวอย่างขนาดกลาง และขนาด 2,000 คนต่อฉบับ ถือเป็นกลุ่มตัวอย่างขนาดใหญ่ อย่างไรก็ตามไรท์ และสโตน (Wright & Stone, 1979, p. 98) กล่าวว่าในทางปฏิบัติถ้าเป็นการใช้แบบสอบจากคลังข้อสอบที่มี

โครงสร้างแบบเดียวกัน สามารถใช้กลุ่มตัวอย่างขนาดเล็กขนาด 100 คนได้ ดังนั้นในการวิจัยครั้งนี้จึงกำหนดขนาดกลุ่มตัวอย่างเป็น 3 ขนาด คือ ขนาดเล็ก (100 คน) ขนาดกลาง (500 คน) และขนาดใหญ่ (700 คน)

จากเหตุผลดังกล่าวข้างต้น ผู้วิจัยจึงเห็นความสำคัญในการที่จะสร้างคามยุติธรรมในการสอบให้กับนักศึกษาทุกคน เมื่อจำเป็นจะต้องใช้แบบสอบต่างฉบับในการสอบคราวเดียวกัน เพื่อให้สามารถนำผลคะแนนมาเปรียบเทียบกันได้อย่างยุติธรรม จึงเป็นที่มาของการทำวิจัยเรื่องนี้ ด้วยการเลือกวิธีการเปรียบเทียบคะแนน และการออกแบบการเปรียบเทียบให้สอดคล้องกับสถานการณ์ คือเลือกใช้วิธีการเปรียบเทียบคะแนน 2 วิธี ได้แก่ วิธีเคอร์เนล (Kernel) และวิธีตามแนวทฤษฎีการตอบข้อสอบ (Item Response Theory) 2 พารามิเตอร์ ที่ใช้ได้ดีกับกลุ่มตัวอย่างขนาดเล็กและกลุ่มตัวอย่างขนาดใหญ่ สอดคล้องกับบริบทของสถานศึกษาที่ผู้วิจัยทำการศึกษา ที่มีนักศึกษาลงทะเบียนเรียนที่สามารถแบ่งเป็นขนาดต่าง ๆ ได้ ทั้งขนาดเล็ก กลาง และขนาดใหญ่ ด้วยการออกแบบวิธีการเปรียบเทียบคะแนนโดยใช้รูปแบบของข้อสอบรวมภายในที่มีความยากปานกลาง และความยากอย่างสุมตามที่มีเอกสารงานวิจัยที่เกี่ยวข้องยืนยันข้อมูลว่าจะทำให้การเปรียบเทียบคะแนนมีคุณภาพและลดความคลาดเคลื่อนในการเปรียบเทียบคะแนนได้ดี หลังจากนั้นจึงศึกษาความสอดคล้องของการตัดเกรดจากคะแนนก่อนการปรับเทียบคะแนนกับการตัดเกรดจากคะแนนหลังการปรับเทียบคะแนนให้อยู่บนสเกลเดียวกัน เพื่อหาวิธีการปรับเทียบภายใต้รูปแบบและเงื่อนไขที่ดีที่สุดที่จะนำไปสู่การตัดเกรดที่ยุติธรรม ไม่ให้นักศึกษาเกิดการได้เปรียบเสียเปรียบกันจากการสอบด้วยแบบสอบต่างฉบับ

ผู้วิจัยศึกษาเอกสารแนวคิิตทฤษฎีต่าง ๆ ที่เกี่ยวข้องนำมากำหนดเป็นตัวแปรในการวิจัยและสร้างเป็นกรอบแนวคิดการวิจัย ได้ดังนี้





ภาพที่ 1 กรอบแนวคิดการวิจัย

วัตถุประสงค์การวิจัย

1. เพื่อศึกษาคุณภาพของวิธีการปรับเทียบคะแนนภายใต้เงื่อนไขรูปแบบข้อสอบร่วม ขนาดตัวอย่าง และรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกัน
2. เพื่อเปรียบเทียบความสอดคล้องของผลการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขที่กำหนด

วิธีการดำเนินการวิจัย

ผู้วิจัยแบ่งวิธีการดำเนินการวิจัย เป็น 2 ขั้นตอน คือ ขั้นตอนของวิธีการปรับเทียบคะแนน กับขั้นตอนของการตัดเกรดและตรวจสอบความสอดคล้องของการตัดเกรด รายละเอียดดังนี้

1. ขั้นตอนของวิธีการปรับเทียบคะแนน เป็นการปรับเทียบคะแนนโดยใช้แบบสอบปลายภาค ของภาคการศึกษา 1/2556 ภาค 1/2557 และภาค1/2558 ที่มีข้อสอบร่วมภายใน จำนวน 15 ข้อ ได้มาจากการสุ่มจากทุกหน่วยการเรียน ๆ ละ 1 ข้อ มีการจัดเรียงลำดับข้อสอบร่วมในตำแหน่งที่ตรงกัน และใช้แบบสอบของภาคการศึกษาที่ 1/2556 เป็นฐานสำหรับให้แบบสอบฉบับอื่น ๆ ปรับสู่แบบสอบฉบับนี้ วิธีการในการปรับเทียบคะแนนมีด้วยกัน 2 วิธี คือ วิธีคอนและ (Kernel) และวิธี IRT 2 พารามิเตอร์ ดำเนินการภายใต้เงื่อนไขข้อสอบร่วมที่มีความยากอยู่ในช่วง .4 - .6 เงื่อนไขข้อสอบร่วมมีความยากอย่างสูง เงื่อนไขข้อสอบร่วมที่มีความยากอยู่ในช่วง .4 - .6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และเงื่อนไขข้อสอบร่วมมีความยากอย่างสูงและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง วิเคราะห์กับตัวอย่างขนาด 100 คน 500 คน และ 700 คน ตามลำดับ หลังจากนั้นตรวจสอบคุณภาพของการปรับเทียบคะแนนของทั้งสองวิธี ด้วยค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน (Standard error of equating: SEE)
2. ขั้นตอนของการตัดเกรดและตรวจสอบความสอดคล้องของการตัดเกรด เป็นการนำคะแนนก่อนการปรับเทียบคะแนนและคะแนนหลังจากการปรับเทียบคะแนนมาตัดเกรด แบ่งการตัดเกรด 2 วิธี คือ

- วิธีที่ 1 ตัดเกรด 3 ระดับ คือ U และ H
- วิธีที่ 2 ตัดเกรด 8 ระดับ คือ $A^- B^+ B^- C^+ C^- D^+ D^-$ และ F

หลังจากนั้นหาความสัมพันธ์และระดับของความสัมพันธ์ของการตัดเกรดหลังการเปรียบเทียบคะแนนด้วยวิธีคอนและ IRT 2 พารามิเตอร์ของทุกเงื่อนไข

ประชากรและตัวอย่าง

ประชากรเป็นนักศึกษาระดับปริญญาตรีที่สอบในภาคการศึกษาที่ 1/2556 จำนวน 1210 คน ภาคการศึกษาที่ 1/2557 จำนวน 927 คน และภาคการศึกษาที่1/2558 จำนวน 857 คน ของชุดวิชาหนึ่งที่ผู้วิจัยทำการศึกษา

ตัวอย่างที่ใช้ในการวิเคราะห์ แบ่งเป็น 3 ขนาด คือ 100 คน 500 คน และ 700 คน ได้มาจากวิธีการสุ่มอย่างง่ายจากประชากรในแต่ละภาคโดยไม่มีการนับซ้ำ ยกเว้นตัวอย่างขนาด 700 คน

เครื่องมือที่ใช้ในการวิจัย

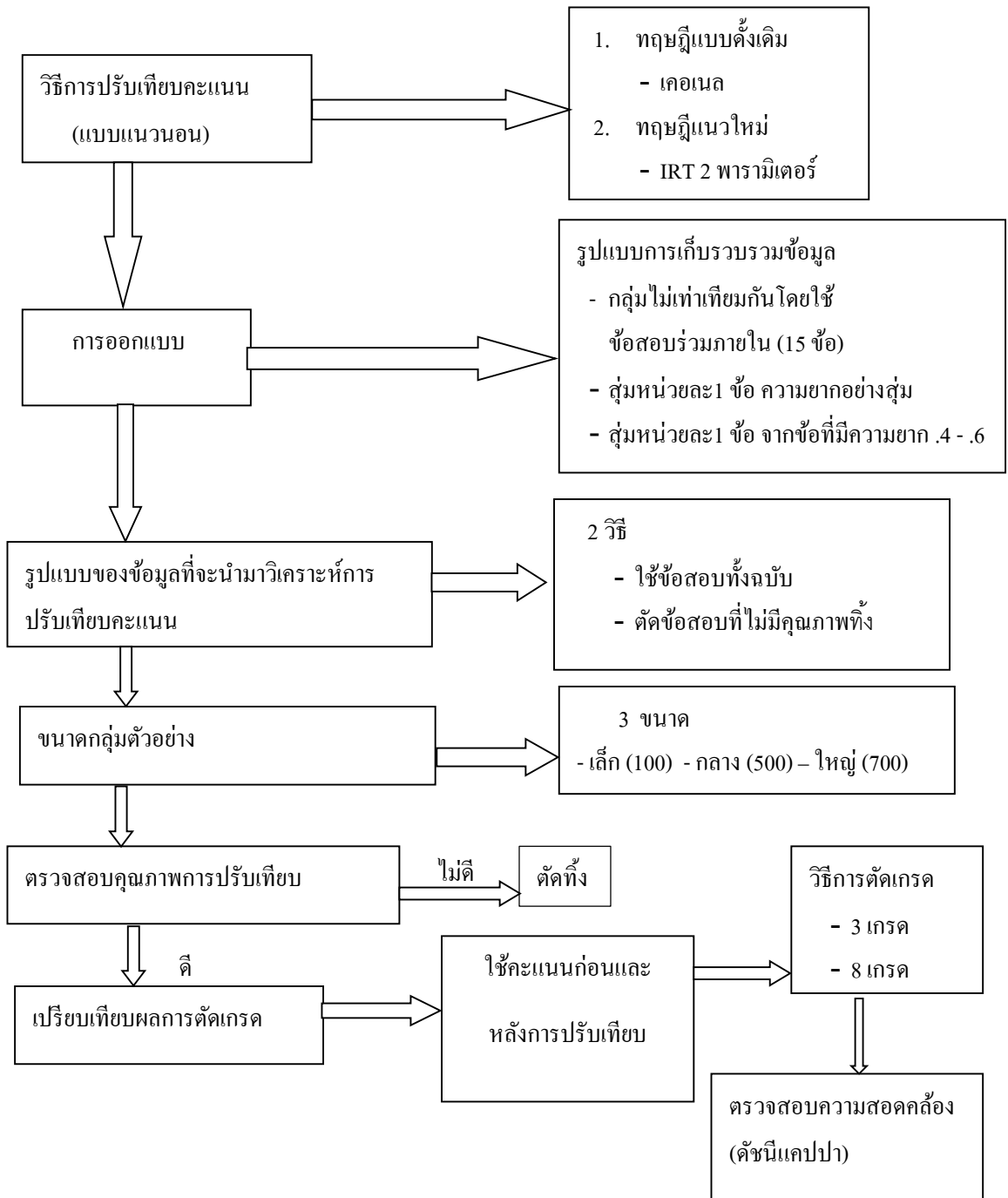
เครื่องมือที่ใช้ในการวิจัย เป็นแบบสอบปลายภาค ของภาคการศึกษา 1/2556 ภาค 1/2557 และภาค 1/2558 จำนวน 3 ฉบับ ที่มีความเป็นคู่ขนานด้านเนื้อหา โครงสร้าง รูปแบบของข้อสอบ และเวลาในการสอบที่เหมือนกันทั้ง 3 ฉบับโดยสุ่มจากระบบคลังข้อสอบที่มีโครงสร้างหลักเกณฑ์การสุ่มแบบเดียวกัน เป็นแบบสอบแบบเลือกตอบ 5 ตัวเลือก จำนวน 120 ข้อ ที่มีข้อรวมภายในจำนวน 15 ข้อใช้เวลาในการสอบ 3 ชั่วโมง แบบสอบทุกฉบับจะถูกปรับให้อยู่บนสเกลเดียวกันกับแบบสอบของภาคการศึกษา 1/2556

การเก็บรวบรวมข้อมูล

การเปรียบเทียบคะแนนครั้งนี้ได้ออกแบบแผนการเก็บรวบรวมข้อมูล โดยใช้แบบแผนกลุ่มไม่เทียบกันโดยใช้ข้อสอบร่วม (Common-item Nonequivalent Groups Design) ในสถานการณ์จริงของการสอบ 3 ภาคการศึกษา มีรูปแบบ ดังนี้

ภาคการศึกษา	ข้อสอบร่วม (15ข้อ)	ข้อสอบที่เหลือ	รวม
1/ 2556	$V_1 V_2$	X	120 ข้อ
1/2557	V_1	Y	120 ข้อ
1/2558	V_2	Z	120 ข้อ

โดยผ่านขั้นตอนการดำเนินการ ดังภาพที่ 2



ภาพที่ 2 ขั้นตอนการดำเนินการวิจัย

การวิเคราะห์ข้อมูล

1. วิเคราะห์คุณภาพข้อสอบรายข้อโดยใช้ทฤษฎีการทดสอบแบบดั้งเดิม เพื่อใช้ในการกำหนดข้อสอบรวมที่มีค่าความยากอยู่ระหว่าง .4 - .6 และใช้ในการตัดข้อที่ไม่มีคุณภาพทิ้งก่อนที่จะทำการเปรียบเทียบคะแนนตามเงื่อนไขหนึ่งของการเปรียบเทียบคะแนนที่กำหนด
2. ตรวจสอบข้อตกลงเบื้องต้นของการเปรียบเทียบคะแนนตามแนวทฤษฎีการตอบข้อสอบ (Item response theory)
3. วิเคราะห์ข้อมูล โดยใช้โปรแกรมในการเปรียบเทียบคะแนนวิธีเคอเนล (Kernel) และวิธีตามแนวทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์โดยใช้วิธีการปรับค่าพารามิเตอร์พร้อมกัน ภายใต้เงื่อนไขที่กำหนดตามรูปแบบข้อสอบรวม วิธีการวิเคราะห์ และขนาดของกลุ่มตัวอย่าง
4. หากคุณภาพของการเปรียบเทียบคะแนน ด้วยการหาความคลาดเคลื่อนของการเปรียบเทียบคะแนนของแต่ละวิธีเพื่อดูว่าเงื่อนไขมีคุณภาพในการเปรียบเทียบคะแนนได้ดีกว่ากัน
5. ตัดเกรดตามรูปแบบการตัดเกรดที่กำหนดเป็น 3 ระดับและ 8 ระดับ จากคะแนนก่อนการเปรียบเทียบคะแนน และคะแนนหลังจากการเปรียบเทียบคะแนน
6. ตรวจสอบความสอดคล้องของการตัดเกรดทั้ง 2 แบบ จากคะแนนก่อนการเปรียบเทียบคะแนน และหลังจากการเปรียบเทียบคะแนน ตามเงื่อนไขต่าง ๆ ของการเปรียบเทียบคะแนน โดยใช้เกณฑ์ค่าสัมประสิทธิ์แคปปา (Coefficient Kappa) ตั้งแต่ .8 ขึ้นไปจึงจะถือว่าการตัดเกรดมีความสอดคล้องกัน ไม่จำเป็นจะต้องทำการเปรียบเทียบคะแนน แต่ถ้าได้ค่าต่ำกว่านี้จะต้องทำการเปรียบเทียบคะแนนก่อนที่จะตัดเกรด
7. หาความสัมพันธ์และระดับของความสัมพันธ์ของการตัดเกรดหลังการเปรียบเทียบคะแนนด้วยวิธีเคอเนล และ IRT 2 พารามิเตอร์

ผลการวิจัยพบว่า

1. การเปรียบเทียบคะแนนด้วยวิธีเคอเนล ภายใต้ตัวอย่างขนาด 500 คน และ 700 คน ให้ค่า SEE (Standard error of equate) ต่ำใกล้เคียงกันทุกเงื่อนไข ยกเว้นขนาดตัวอย่าง 100 คนที่ให้ค่า SEE ค่อนข้างสูง (ค่า SEE ต่ำแสดงถึงคุณภาพของการเปรียบเทียบคะแนน) โดยเงื่อนไขข้อสอบรวมมีความยากอยู่ในช่วง .4 - .6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อวิเคราะห์กับขนาดตัวอย่าง 700 คน มีคุณภาพมากที่สุด การตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนเปรียบเทียบคะแนนจะมีคุณภาพของการเปรียบเทียบคะแนนมากกว่าไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และการเพิ่มขนาดตัวอย่างมีแนวโน้มทำให้ค่า SEE เพิ่มขึ้น
2. การเปรียบเทียบคุณภาพของการเปรียบเทียบคะแนนระหว่างวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ให้ผลว่าวิธีเคอเนลมีคุณภาพสูงกว่าโดยที่การวิเคราะห์ภายใต้เงื่อนไขข้อสอบรวมมีความยากอย่างสม่ำเสมอกับตัวอย่างขนาด 700 คน และเงื่อนไขข้อสอบรวมมีความยากอย่างสม่ำเสมอและตัดข้อสอบที่ไม่มีคุณภาพทิ้งกับตัวอย่างขนาด 500 คน มีคุณภาพในการเปรียบเทียบคะแนนสูงที่สุด
3. การตัดเกรดจากคะแนนก่อนการเปรียบเทียบคะแนนและคะแนนหลังการเปรียบเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ ตามเงื่อนไขต่าง ๆ ภายใต้การตัดเกรด 3 ระดับ และ 8 ระดับ ส่วนใหญ่พบว่าไม่

สอดคล้องกัน จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนที่จะตัดเกรด การขยายระดับการตัดเกรดเป็น 8 ระดับ การใช้ตัวอย่างขนาดใหญ่ขึ้น จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากขึ้น ส่วนการตัดข้อสอบที่ไม่มีคุณภาพทั้งก่อนการปรับเทียบคะแนนด้วยวิธีเคอเนล จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนกว่าวิธี IRT 2 พารามิเตอร์ ขณะที่การใช้ข้อสอบร่วมที่มีความยากอยู่ในช่วง .4 - .6 ทั้งสองวิธีจะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม

4. การใช้ตัวอย่างขนาด 500 คน 700 คน ในการปรับเทียบคะแนน จะเห็นความไม่สอดคล้องของการตัดเกรดได้อย่างชัดเจนไม่ว่าจะวิเคราะห์ด้วยวิธีเคอเนลหรือวิธี IRT 2 พารามิเตอร์ด้วยเงื่อนไขใดก็ตาม แต่ถ้าใช้ตัวอย่างขนาด 100 คนส่วนใหญ่ไม่พบความแตกต่างของการตัดเกรด จึงไม่จำเป็นต้องทำการปรับเทียบคะแนน

5. การหาความสัมพันธ์ของการตัดเกรดหลังการปรับเทียบคะแนน ทั้ง 3 ระดับ และ 8 ระดับ ด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ พบว่ามีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติทุกเงื่อนไข โดยที่การตัดเกรด 3 ระดับ หลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม กับทุกขนาดตัวอย่างมีความสัมพันธ์กันในระดับดีมาก ส่วนเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทั้ง มีความสัมพันธ์อยู่ในระดับดีที่เหลือสัมพันธ์กันในระดับต่ำ ส่วนการตัดเกรด 8 ระดับ หลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ มีเพียงเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม เพียงเงื่อนไขเดียว ที่การตัดเกรดหลังการปรับเทียบคะแนนของทั้งสองวิธีมีความสัมพันธ์กันอยู่ในระดับดี ส่วนเงื่อนไขที่เหลือสัมพันธ์กันในระดับต่ำ

อภิปรายผลการวิจัย

ผลการวิจัยดังกล่าวข้างต้น สามารถนำมาอภิปรายผลได้ ในประเด็นดังต่อไปนี้

1. วิธีการปรับเทียบคะแนน ด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ เมื่อวิเคราะห์กับตัวอย่างขนาดเล็ก 100 คน วิธีที่มีคุณภาพมากที่สุดคือวิธีเคอเนลโดยใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4 - .6 ซึ่งสอดคล้องกับตำราของโคเลนและแบรนแนน และรีเบคคา (Kolen & Brennan, 2004, pp. 293 – 294; Rebecca, Rebecca, Rebecca & Dvorak, 2009) แต่เมื่อตัวอย่างมีขนาดใหญ่ในที่นี่ใช้ 700 คน การวิเคราะห์ด้วยวิธี IRT 2 พารามิเตอร์ไม่ได้มีคุณภาพสูงกว่าการวิเคราะห์ด้วยวิธีเคอเนล ซึ่งไม่ได้เป็นไปตามสมมติฐานข้อที่ 2 ทั้งนี้อาจจะมีสาเหตุจากการใช้ตัวอย่างขนาด 700 คน ที่ไม่ใหญ่เพียงพอ ซึ่งแฮร์ริสกล่าวว่าความคลาดเคลื่อนของการปรับเทียบคะแนนที่แสดงถึงคุณภาพของการปรับเทียบคะแนน ขึ้นอยู่กับขนาดของตัวอย่างที่ (Harris, 1993 cited in Kolen & Brennan, 2004, p. 288)

2. วิธีการปรับเทียบคะแนน ด้วยวิธีเคอเนลส่วนใหญ่พบว่าคะแนนในช่วงต้นและช่วงปลายค่อนข้างกว้างไม่ว่าจะปรับเทียบโดยใช้เงื่อนไขใดก็ตาม ซึ่งทำให้ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนมีแนวโน้มสูงกว่าช่วงกลาง ๆ ของข้อมูล ซึ่งเป็นไปตามคุณลักษณะของการวิเคราะห์ของวิธีนี้ ที่ทำการวิเคราะห์ด้วยวิธี Gaussian สเตลของคะแนนจะเริ่มตั้งแต่ $-\infty$ ถึง $+\infty$ ทำให้คะแนนในช่วงต้นและปลายมีความคลาดเคลื่อนสูง (Mao et al., 2006 cited in Meng, 2012, p. 74) ส่วนความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน (SEE) จะลดลงเมื่อตัวอย่างมีขนาดใหญ่สอดคล้องกับข้อค้นพบของ เม็ง ที่ว่ากลุ่มตัวอย่างในการวิเคราะห์มีขนาด

ใหญ่ขึ้นย่อมเป็นตัวแทนประชากรที่ดีกว่า และทำให้การประมาณค่าได้ดีกว่า (Meng, 2012, p. 71) ซึ่งสอดคล้องกับที่เฮนสันและเบ็กกัน และลีได้กล่าวไว้ (Hanson & Beguin, 2002; Lee, 2007, cited in Meng, 2012, p. 27)

3. การตัดข้อสอบที่ไม่มีคุณภาพทั้งก่อนที่จะนำไปปรับเทียบคะแนนของทั้งเงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างลุ่ม ของทั้งวิธีเคนเนลและวิธี IRT 2 พารามิเตอร์ ส่วนใหญ่ไม่ได้ทำให้คุณภาพของการปรับเทียบคะแนนดีขึ้น ทั้งนี้อาจจะเป็นผลมาจากการตัดข้อสอบที่ทำให้ความยาวของแบบสอบสั้นลง ซึ่งความยาวของแบบสอบมีผลต่อการประมาณค่าพารามิเตอร์ และมีผลต่อคุณภาพของการปรับเทียบคะแนนซึ่งความยาวของแบบสอบที่เพิ่มขึ้น สามารถลดขนาดตัวอย่างในการวิเคราะห์ให้น้อยลงได้โดยยังคงไว้ซึ่งคุณภาพที่ไม่แปรเปลี่ยน (พิชัย ละแมนชัย, 2538, น.135)

4. การใช้ข้อสอบร่วมที่มีความยากปานกลาง (.4-.6) ในการปรับเทียบคะแนนทั้งวิธีเคนเนล และ IRT 2 พารามิเตอร์ ส่วนใหญ่มีคุณภาพในการปรับเทียบคะแนนต่ำกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างลุ่ม ทั้งนี้ อาจจะเป็นเนื่องมาจากการใช้ข้อสอบร่วมที่มียากปานกลาง (.4-.6) มีช่วงที่แคบกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างลุ่มที่มีความเป็นตัวแทนของแบบสอบทั้งฉบับได้ดีกว่า จึงทำให้ความคลาดเคลื่อนต่ำกว่า

5. การตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขต่าง ๆ ภายใต้การตัดเกรด 3 ระดับ ด้วยวิธีเคนเนลและวิธี IRT 2 พารามิเตอร์ ส่วนใหญ่ให้ผลตรงกันว่าไม่มีความสอดคล้องกัน แสดงว่าแบบสอบทั้ง 2 ฉบับที่นำมาปรับเทียบคะแนนไม่มีความเท่าเทียมกันทำให้ผู้สอบเกิดการได้เปรียบเสียเปรียบกัน เช่นเดียวกันกับการตัดเกรด 8 ระดับ ซึ่งช่วงการตัดเกรดยิ่งแคบจะเห็นความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากยิ่งขึ้น

ข้อเสนอแนะในการนำไปใช้

1. ควรนำวิธีการปรับเทียบคะแนนวิธีเคนเนล ไปใช้กับแบบสอบที่มีนักศึกษาลงทะเบียนเรียนเป็นจำนวนมาก ซึ่งถ้าใช้กับนักศึกษาตั้งแต่ 500 คนขึ้นไป ตามผลการวิจัยนี้จะเป็นจำนวนที่ให้ผลการปรับเทียบคะแนนมีความคงที่ โดยใช้เงื่อนไขข้อสอบร่วมมีความยากอย่างลุ่มทั้งที่ตัดและไม่ตัดข้อสอบที่ไม่มีคุณภาพทั้งในการวิเคราะห์ข้อมูล แต่ถ้านักศึกษามีเพียง 100 คน ไม่จำเป็นจะต้องทำการปรับเทียบคะแนน

2. ความสอดคล้องของการตัดเกรดหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ให้ค่าที่ใกล้เคียงกันกันระหว่างการตัดหรือไม่ตัดข้อสอบที่ไม่มีคุณภาพทั้งไม่ว่าจะใช้เงื่อนไขใดในการปรับเทียบคะแนนมากกว่าวิธีเคนเนล ส่วนกระบวนการของการปรับเทียบคะแนนในทางปฏิบัติ การปรับเทียบด้วยวิธี IRT 2 พารามิเตอร์ จะมีความยุ่งยากมากกว่าวิธีเคนเนล ดังนั้นวิธีเคนเนล จึงมีความสะดวกในการนำไปใช้มากกว่า

3. ควรทำการตรวจสอบชุดวิชาอื่น ๆ ว่าแบบสอบต่างฉบับมีความเท่าเทียมกันหรือไม่ เพื่อการยืนยันหาแบบสอบต่างฉบับก่อให้เกิดการได้เปรียบหรือเสียเปรียบกัน ก็ควรทำการปรับเทียบคะแนนก่อนการตัดเกรดเพื่อสร้างความเป็นมาตรฐานเดียวกันของแบบสอบและการตัดเกรด เป็นการให้ความยุติธรรมกับผู้สอบทุกคน

4. ผลจากการวิจัย การตัดเกรด 3 ระดับที่มีช่วงคะแนนกว้างจากแบบสอบต่างฉบับยังก่อให้เกิดความได้เปรียบกันในการตัดเกรด การขยายระดับของการตัดเกรดดีขึ้น เป็น 8 ระดับจะยิ่งเพิ่มความได้เปรียบเสียเปรียบ

ของการตัดเกรดชัดเจนมากขึ้น ดังนั้นถ้าจะขยายการตัดเกรดเป็น 8 ระดับ จำเป็นอย่างยิ่งที่จะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด

บรรณานุกรม

- ประสพชัย พสุนนท์. (2558). การประเมินความเชื่อมั่นระหว่างผู้ประเมินโดยใช้สถิติแคลปป์ วารสารวิชาการศิลป/ ศึกษาศาสตร์ประยุกต์. 8(1), 2-20.
- พัชร จันทร์เพ็ง. (2550). การเปรียบเทียบคุณภาพของวิธีการเชื่อมโยงคะแนนตามทฤษฎีการตอบข้อสอบแบบพหุ มิติภายใต้การหมุนแกน โครงสร้างเชิงมิติและระดับความสัมพันธ์ที่แตกต่างกัน (วิทยานิพนธ์ครุศาสตร ดุษฎีบัณฑิต ไม่ได้ตีพิมพ์). จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร.
- พิชัย ละแมนชัย. (2538). ขนาดกลุ่มตัวอย่างขั้นต่ำสำหรับการปรับเทียบคะแนน ในแนวระดับ. (วิทยานิพนธ์ ครุศาสตรดุษฎีบัณฑิต ไม่ได้ตีพิมพ์). จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร.
- ภัทรพร เกษสังข์. (2546). การศึกษาพัฒนาการความสามารถทางคณิตศาสตร์ของนักเรียนชั้นมัธยมศึกษาตอนต้น โดยการปรับเทียบคะแนนในแนวตั้งที่ใช้วิธีการที่เหมาะสม. (ปริญญานิพนธ์ดุษฎีบัณฑิต ไม่ได้ตีพิมพ์). มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, กรุงเทพมหานคร.
- ศิริชัย กาญจนวาสี. (2555). ทฤษฎีการทดสอบแนวใหม่. (ฉบับปรับปรุง). กรุงเทพมหานคร ศูนย์ตำรา เอกสารทาง วิชาการ คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- _____. (2551). การพัฒนาวิธีการปรับเทียบผลการเรียนเฉลี่ยสะสมตามกลุ่มสาระการเรียนรู้โดยใช้คะแนน ONET ของนักเรียนมัธยมศึกษาตอนปลาย. สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ.
- สำนักงานคณะกรรมการข้อมูลข่าวสารของทางราชการ. (2558). พระราชบัญญัติข้อมูลข่าวสารทางราชการ พ.ศ. 2540. สืบค้นจาก http://www.oic.go.th/web2014/ACTOfficial_Information.ht
- สุนิสา จั้วม่วงศรี. (2537). ผลของความยาวของแบบสอบร่วมที่มีต่อคุณภาพของวิธีการเทียบมาตราเชิงเส้นตรง (วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต ไม่ได้ตีพิมพ์). จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร.
- อุทัยวรรณ พงศ์อร่าม. (2545). การศึกษาขนาดกลุ่มตัวอย่างที่เหมาะสมสำหรับการปรับเทียบคะแนนด้วยวิธีอิกวิ เปรอร์เซ็นไทล์ และวิธีเชิงเส้นตรงตามแบบจำลองคะแนนจริงสัมพันธ์ ที่มีแบบแผนการปรับเทียบและ ความยาวของแบบสอบแตกต่างกัน. (ปริญญานิพนธ์ดุษฎีบัณฑิต ไม่ได้ตีพิมพ์). มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, กรุงเทพมหานคร.
- Angoff, W.H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.
- Caldwell, L.J. (1984). A comparison of equating error in linear and rasch model test equating methods. Doctoral Dissertation, The Florida State University, *Dissertation Abstracts International*, 49: 2847
- Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, 54, 8-20.
- Cook, L.,&Petersen,N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, 11, 225-244.

- Dvorak., K. L. N. (2009). *A comparison of Kernel equating to the test characteristic curve method* . Lincoln, Nebraska. University of Nebraska.
- Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT verbal and mathematical sections*. Princeton,NJ: Educational Testing Service.
- Godfrey, K. E. (2007). *A comparison of Kernel equating and IRT true score equating methods*. University of North Carolina at Greensboro.
- Holland, P.W. and Donald B Rubin. (1982). *Test equating*. New Jersey: Education Testing Service, Princeton, Academic Press.
- Holland, P.W. and Sinharay, S. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed. *Journal of Educational Measurement*, 44(3), 249 – 275.
- Kolen, M. J. and Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*. 27, 27-39.
- Kolen, M.J. and Brennan, R.L. (2004). *Test equating, scaling, and linking*, (2nd ed.). New York: Springer
- Lord, F.M. (1980). *Application of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates inc.
- Meng. (2012). *Comparison of Kernel equating and item response theory equating methods*. United States, University of Massachusetts.
- Petersen, N.S., M.J. Kolen and H.D. Hoover. (1989). Scaling, norming, and equating, in R.L. Linn (Ed.) *Educational Measurement*, 221- 262. (3rd ed.). New York : Macmillan.
- Stocking, M.L. and Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201 – 210.
- Tiscari, R.S. (1990). *A comparison of classical and item response theory equating methods using a composite score of direct and indirect writing assessment*. The University of Texas at Austin.
- Wendy Lam. (2009). *Linking current and future score scales for the AICPA Uniform CPA Exam*. University of Massachusetts Amherst.
- Wright, B.D. and Stone, M.H. (1979). *Best test design: Rasch Measurement*. Chicago Mesa Press.