ผลการปรับเทียบคะแนนด้วยวิธีเคอเนล และวิธี IRT ภายใต้เงื่อนไขที่แตกต่างกัน

ศศิธร ชุตินันทกุล

ดุษฎีนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาวิจัย วัดผลและสถิติการศึกษา
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
กรกฎาคม 2560

คณะกรรมการควบคุมดุษฎีนิพนธ์และคณะกรรมการสอบดุษฎีนิพนธ์ ได้พิจารณาดุษฎีนิพนธ์ของ ศศิธร ชุตินันทกุล ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมดุษฎีนิพนธ์

……..อาจารย์ที่ปรึกษาหลัก
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

…….. อาจารย์ที่ปรึกษาร่วม
(ดร.สมพงษ์ ปั้นหุ่น)

คณะกรรมการสอบดุษฎีนิพนธ์

……..ประธาน
(ผู้ช่วยศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

……..กรรมการ
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

…….. กรรมการ
(ดร.สมพงษ์ ปั้นหุ่น)

…….. กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์)

คณะศึกษาศาสตร์อนุมัติให้รับดุษฎีนิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพา

…….. คณบดีคณะศึกษาศาสตร์
(รองศาสตราจารย์ ดร.วิชิต สุรัตน์เรืองชัย)
วันที่ 26 เดือน กรกฎาคม พ.ศ. 2560

กิตติกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี ด้วยความกรุณาอย่างยิ่ง จากอาจารย์ที่ปรึกษาหลัก รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม ที่ได้ให้ความรู้ทั้งในเชิงทฤษฎีและการปฏิบัติเรื่องการใช้โปรแกรมสถิติด้านการวัดและประเมินผลการศึกษา ตลอดจนให้คำแนะนำและชี้แนวทางที่เป็นประโยชน์ จนกระทั่งผู้วิจัยสามารถพัฒนาเป็นดุษฎีนิพนธ์และทำจนสำเร็จ ผู้วิจัยขอขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้ และขอขอบพระคุณอาจารย์ ดร.สมพงษ์ ปั้นหุ่น อาจารย์ที่ปรึกษาร่วมที่ได้ให้คำปรึกษา แนะนำด้วยดีมาตลอด และขอขอบพระคุณประธานและกรรมการสอบปากเปล่าดุษฎีนิพนธ์ที่ได้ให้คำแนะนำทำให้งานวิจัยชิ้นนี้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก ที่ได้อนุเคราะห์โปรแกรมสำเร็จรูปเคอเนล และสอนวิธีการใช้โปรแกรมพร้อมทั้งคำแนะนำที่เป็นประโยชน์ต่องานวิจัย

ขอขอบพระคุณ Won-Chan Lee and Michael J. Kolen จาก University of Iowa ที่พัฒนาโปรแกรม IRT-CLASS ซึ่งผู้วิจัยนำมาประยุกต์ใช้ในงานวิจัยชิ้นนี้จนกระทั่งสำเร็จ

ขอขอบพระคุณคุณแม่ พี่ ๆ น้อง ๆ ตลอดจนเพื่อน ๆ ที่คอยห่วงใย เป็นกำลังใจ และช่วยเหลือด้วยดีเสมอมา

ท้ายที่สุดขอกราบขอบพระคุณครู อาจารย์ทุกท่านทั้งในอดีตและปัจจุบัน ที่ได้ประสาทความรู้ ทำให้ผู้วิจัยประสบความสำเร็จทั้งด้านการงานและการศึกษา

ศศิธร ชุตินันทกุล

55810102: สาขาวิชา: วิจัย วัดผลและสถิติการศึกษา; ปร.ด. (วิจัย วัดผลและสถิติการศึกษา)
คำสำคัญ: การปรับเทียบคะแนน/ เคอเนล/ IRT

ศศิธรชุตินันทกุล: ผลการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธีIRTภายใต้เงื่อนไขที่แตกต่างกัน (THE COMPARISON OF TEST SCORES DERIVED THROUGH KERNEL EQUATING AND IRT EQUATING METHODS UNDER VARIED CONDITIONS) คณะกรรมการควบคุมดุษฎีนิพนธ์: ไพรัตน์ วงษ์นาม, ค.ด., สมพงษ์ ปั้นหุ่น, ค.ด., 198 หน้า. ปี พ.ศ. 2560.

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) เพื่อศึกษาคุณภาพของวิธีการปรับเทียบคะแนนภายใต้เงื่อนไขรูปแบบข้อสอบร่วม ขนาดตัวอย่างและรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกัน 2) เพื่อเปรียบเทียบความสอดคล้องของผลการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขที่กำหนด

ข้อมูลที่ใช้ในการวิจัย เป็นผลการตอบข้อสอบปลายภาคของนักศึกษาระดับปริญญาตรีของชุดวิชาหนึ่งที่สอบในภาคการศึกษา 1/ 2556 ภาค 1/ 2557 และภาค 1/ 2558 แบบสอบเป็นแบบเลือกตอบ 5 ตัวเลือก จำนวน 120 ข้อ ที่มีข้อสอบร่วมภายใน จำนวน 15 ข้อ แบบสอบทุกฉบับจะถูกปรับให้อยู่บนสเกลเดียวกันกับแบบสอบของภาคการศึกษา 1/ 2556

ผลการวิจัยพบว่า

1. การปรับเทียบคะแนนโดยวิธีเคอเนล ภายใต้ตัวอย่างขนาด 500 คน และ 700 คน ให้ค่า SEE ต่ำใกล้เคียงกันทุกเงื่อนไข ยกเว้นขนาดตัวอย่าง 100 คน ที่ให้ค่า SEE ค่อนข้างสูง โดยเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4 -.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อวิเคราะห์กับขนาดตัวอย่าง 700 คน มีคุณภาพมากที่สุด การตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนปรับเทียบคะแนนจะมีคุณภาพของการปรับเทียบคะแนนมากกว่าไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และการเพิ่มขนาดตัวอย่างมีแนวโน้มทำให้ค่า SEE ลดลง

2. การเปรียบเทียบความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนระหว่างวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ วิธีเคอเนลให้ค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด เมื่อใช้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มกับกลุ่มตัวอย่างขนาด 700 คน และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

3. การตัดเกรดจากคะแนนก่อนการปรับเทียบคะแนนและคะแนนหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ ตามเงื่อนไขต่าง ๆ ภายใต้การตัดเกรด 3 ระดับ และ 8 ระดับ ส่วนใหญ่พบว่า ไม่สอดคล้องกัน จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนที่จะตัด

เกรด การขยายระดับการตัดเกรดเป็น 8 ระดับ การใช้ตัวอย่างขนาดใหญ่ขึ้น จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากขึ้น ส่วนการตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนการปรับเทียบคะแนนด้วยวิธีเคอเนล จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนกว่าวิธีวิธี IRT 2 พารามิเตอร์ ขณะที่การใช้ข้อสอบร่วมที่มีความยากอยู่ในช่วง .4 -.6 ทั้งสองวิธีจะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม

4. การใช้ตัวอย่างขนาด 500 คน 700 คน ในการปรับเทียบคะแนน จะเห็นความไม่สอดคล้องของการตัดเกรดได้อย่างชัดเจนไม่ว่าจะวิเคราะห์ด้วยวิธีเคอเนลหรือวิธี IRT 2 พารามิเตอร์ ไม่ว่าจะใช้เงื่อนไขใดก็ตาม ส่วนตัวอย่างขนาด 100 คน ส่วนใหญ่แล้วการปรับเทียบคะแนนจะมีความสอดคล้องกัน

5. การตัดเกรดหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ของทุกเงื่อนไขมีความสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 โดยที่การตัดเกรด 3 ระดับหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม กับทุกขนาดตัวอย่างมีความสัมพันธ์กันในระดับดีมาก

55810102 MAJOR: EDUCATIONAL RESEARCH, MEASUREMENT AND STATISTICS;
Ph.D. (EDUCATIONAL RESEARCH, MEASUREMENTAND STATISTICS)
KEYWORDS: EQUATING/ KERNEL/ IRT

SASITHORN CHUTINUNTAKUL: THE COMPARISON OF TEST SCORES DERIVED THROUGH KERNEL EQUATING AND IRT EQUATING METHODS UNDER VARIED CONDITIONS. DISSERTATION ADVISORS: PAIRATTANA WONGNAM, Ph.D., SOMPONG PANHOON, Ph.D. 198 P. 2017.

The objectives of this research were; 1) to examine the appropriate test equating methods under varied anchor test patterns, sample sizes, and data formats, 2) to compare the consistency of the grading results of scores before and after test equating under the specified conditions.

Data used in the research was the final test results of undergraduates in a subject taken in semester 1/ 2013, 1/ 2014, and 1/ 2015. The test was a five-choice test containing 120 questions with 15 anchor items. Each test was equated to be on the same scale as that of semester 1/ 2013.

The findings were as follows:

1. Kernel equating method under a sample size of 500 and 700 students resulted in similarly low SEE values in every condition, while the size of 100 students resulted in a relatively high SEE value. The anchor items condition had a difficulty level of .4-.6 with non-quality questions removed. Considering the sample sizes, 700 students showed the best quality. Removing non-quality questions before test equating resulted in better quality of test equating compared to reserving them. It simplied that the bigger the size, the lower SEE value.

2. The comparison of Kernel and IRT 2 parameters equating method found that Kernel equating method with a condition of randomized anchor items gave the lowest SEE under samples size 700, as well as the condition of randomized anchor items with non-quality questions removed when using 500 samples size.

3. Grading of the scores received before equating 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.

4. Using a sample size of 500 and 700 students in test equating resulted in obviously inconsistency in grading, whether analyzed by Kernel method or 2-parameter IRT method under any conditions. Otherwise, the same sample size of 100 students was consistency in grading.

5. The relationship of grading after equating by using Kernel method and 2-parameter IRT method under any conditions were statistically significant at .o5 level. The two methods under 3-level grading by using randomized anchor test all the same sample size were best relationship.

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดและการประเมินผลทางการศึกษา เป็นองค์ประกอบสำคัญของกระบวนการจัดการศึกษา ที่จะทำให้ทราบว่าผลของการจัดการศึกษาบรรลุตามวัตถุประสงค์ของการจัดการศึกษามากน้อยเพียงใดด้วยการใช้เครื่องมือวัดผลที่หลากหลาย ในการวัดผลการเรียนรู้ของผู้เรียนยังไม่มีเครื่องมือใดที่ดีที่สุดที่จะสามารถวัดได้โดยตรง ต้องอาศัยการวัดโดยอ้อมจึงมักมีความคลาดเคลื่อนรวมอยู่ด้วยเสมอ ในการวัดผลจึงจำเป็นต้องเลือกใช้เครื่องมือวัดให้เหมาะสมกับสิ่งที่ต้องการวัด เพื่อให้ผลการวัดถูกต้องและเชื่อถือได้ ซึ่งเครื่องมือวัดผลที่ใช้กันอยู่โดยทั่วไปมีความหลากหลายขึ้นอยู่กับวัตถุประสงค์ เนื้อหาที่จะวัด ตลอดจนลักษณะของผู้ที่ต้องการวัด แต่เครื่องมือวัดผลทุกชนิดมีลักษณะที่สอดคล้องกันอยู่ประการหนึ่งก็คือ ต่างพยายามวัดคุณลักษณะของผู้สอบเพื่อจะบอกถึงระดับความสามารถของผู้สอบ ซึ่งเครื่องมือวัดผลที่ได้รับความนิยมและใช้กันอยู่มาก คือ แบบสอบ โดยเฉพาะแบบสอบปรนัย เพราะเชื่อว่าผลการสอบจะมีความเที่ยง ความตรง ความยุติธรรมสูง และมีความสะดวกในการตรวจให้คะแนน

แบบสอบจึงเป็นเครื่องมือที่สำคัญ ทำอย่างไรจึงจะทำให้การใช้แบบสอบเกิดความเป็นธรรมกับผู้เข้าสอบทุกคน โดยเฉพาะในสถานการณ์ที่จำเป็นจะต้องใช้แบบสอบหลายฉบับในการวัดผลคราวเดียวกัน ซึ่งมักจะพบในสถานการณ์ที่มีผู้เข้าสอบจำนวนมาก ถ้ามาสอบพร้อมกันจะมีปัญหาเรื่องสถานที่สอบไม่เพียงพอ ไม่มีคนคุมสอบ จำเป็นต้องจัดสอบหลายครั้งต่างเวลาต่างสถานที่สอบ ปัญหาที่พบคือ ผู้สอบครั้งแรกนำข้อสอบไปเผยแพร่ เกิดการรั่วไหลของข้อสอบ ผู้สอบครั้งหลังรู้ข้อสอบ มีการศึกษาค้นคว้าเพิ่มเติม หรือในกรณีที่ในภาคการศึกษาเดียวกันเปิดให้มีการสอบหลายครั้งตามความพร้อมของนักศึกษา (Walk–in exam) หรือในสถานการณ์ของการสอบแข่งขันเพื่อรับเข้าทำงานที่มีผู้เข้าสอบจำนวนมาก มีการแข่งขันกันสูงเพื่อป้องกันการทุจริตในการสอบจำเป็นจะต้องใช้แบบสอบหลายฉบับ ในสถานการณ์เช่นนี้จึงมีการใช้แบบสอบต่างฉบับเพื่อต้องการรักษาความลับของแบบสอบ เพื่อป้องกันไม่ให้เกิดการทุจริตในการสอบ ซึ่งในการทำแบบสอบหลาย ๆ ฉบับในการสอบคราวเดียวกัน โดยทั่วไปจะจัดทำเป็นแบบสอบคู่ขนาน โดยคาดหวังว่าแบบสอบที่สร้างจะมีความเท่าเทียมกัน เพื่อไม่ให้เกิดความได้เปรียบเสียเปรียบระหว่างผู้สอบ แต่ในทางปฏิบัติการสร้างแบบสอบให้เป็นคู่ขนานให้มีความเท่าเทียมกันทางด้านโครงสร้างเนื้อหา ค่าสถิติของข้อสอบ มีความเที่ยง ความตรง ค่าเฉลี่ย ความแปรปรวนและความยากง่ายให้เท่ากัน

นั้นมีโอกาสเป็นไปได้น้อยมาก ซึ่งถ้าจะสร้างให้คู่ขนานกันจริง ๆ จะต้องใช้เวลา และงบประมาณจำนวนมาก และอาจจะไม่คุ้มค่ากับการลงทุน เพราะถ้าหากมีการเปลี่ยนแปลงเนื้อหาหรือเปลี่ยนจุดเน้นที่ต้องการวัด หรือเพิ่มวัตถุประสงค์บางประการ จะทำให้แบบสอบชุดใหม่มีโครงสร้างแตกต่างไปจากชุดเดิม ดังนั้นการสร้างแบบสอบให้คู่ขนานกันอย่างแท้จริง จึงมีข้อจำกัด และทำได้ยากมาก ในทางปฏิบัติการสร้างแบบสอบคู่ขนานจึงเป็นเพียงการทำให้แบบสอบมีความใกล้เคียงกันทางด้านเนื้อหา แต่มีความแตกต่างกันในเรื่องความยากและค่าสถิติอื่น ๆ ของแบบสอบ เป็นเหตุให้โครงสร้างของแบบสอบทั้งสองชุดนั้นยังคงแตกต่างกัน (Holland & Rubin, 1982) จึงมีผู้คิดค้นวิธีการใหม่โดยอาศัยวิธีการทางสถิติมาช่วยทำให้คะแนนที่วัดจากแบบสอบต่างชุดกันที่วัดเนื้อหาเดียวกัน แต่มีค่าสถิติของข้อสอบที่แตกต่างกันให้สามารถนำผลคะแนนมาใช้เทียบแทนกันได้ กระบวนการนี้เรียกว่า “การปรับเทียบคะแนน (Test equating) ซึ่งเป็นกระบวนการใช้เทคนิคการสร้างแบบสอบหลาย ๆ ชุด ด้วยความรอบคอบระมัดระวังให้สามารถวัดคุณลักษณะเดียวกัน และใช้เทคนิคการปรับทางสถิติให้คะแนนจากแบบสอบต่างชุดกันปรับชดเชยกัน สำหรับความแตกต่างในการสอบ และคุณลักษณะของข้อสอบให้สามารถเทียบแทนกันได้อย่างยุติธรรม (ศิริชัย กาญจนวาสี, 2555, หน้า 157) วิธีการนี้จะทำการแปลงคะแนนจากแบบสอบชุดหนึ่งไปสู่สเกลคะแนนของแบบสอบอีกชุดหนึ่ง ดังนั้นคะแนนที่นำมาเปรียบเทียบจึงอยู่บนสเกลการวัดเดียวกัน ทำให้สามารถหาคะแนนสมมูลระหว่างแบบสอบต่างชุดนำมาเปรียบเทียบกันได้ สามารถแปลความอย่างมีความหมาย ไม่ว่าแบบสอบต่างชุดนั้นจะคู่ขนานกันหรือไม่

นอกจากเรื่องของความยุติธรรมกับผู้สอบแล้ว ยังมีมุมมองในเรื่องของสิทธิมนุษยชนที่ทุกคนควรจะได้รับความคุ้มครอง ในสิทธิเบื้องต้นที่จะได้รับความเท่าเทียมกันในการสอบจากแบบสอบที่มีความยากง่ายเท่าเทียมกัน เพราะผลจากการวัดและประเมินมีผลต่อการตัดสินผลการเรียนรู้ คือ เกรดที่ผู้เรียนจะได้รับ การสอบจากแบบสอบต่างฉบับในการสอบคราวเดียวกัน ผลของการสอบจากแบบสอบที่มีความยากแตกต่างกัน อาจจะส่งผลให้ผู้เรียนบางคนแทนที่จะได้เกรด A กลับได้เกรด B เพราะทำแบบสอบฉบับที่ยากกว่า ในทางกลับกันผู้เรียนบางคนได้เกรด A แทนที่ได้เกรด B เพราะทำแบบสอบฉบับที่ง่ายกว่าทำให้ได้คะแนนที่สูงกว่า หากไม่ทำการปรับเทียบคะแนนก่อนที่นำไปตัดเกรดย่อมจะทำให้เกิดความได้เปรียบและเสียเปรียบกันในระหว่างผู้สอบที่ได้แบบสอบต่างฉบับกัน ซึ่งผลจากการใช้แบบสอบต่างฉบับดังกล่าวจะนำไปสู่การสอบได้สอบตก การได้เกรดสูง-ต่ำ อันจะส่งผลต่อเกรดเฉลี่ย ต่อการสำเร็จการศึกษา ต่อการสมัครเข้าทำงาน ต่อการศึกษาในระดับที่สูงขึ้น ต่อการเลื่อนระดับหรือการจัดตำแหน่ง ฯลฯ ซึ่งขึ้นอยู่กับจุดมุ่งหมายในการใช้ผลที่แตกต่างกัน จึงจำเป็นจะต้องดำเนินการในเรื่องของการวัดและประเมินด้วยความถูกต้องและเป็นธรรม

มุมมองอื่น ๆ ในเรื่องของการตรวจสอบข้อมูลข่าวสาร กระบวนการการวัดผลประเมินผล เรื่องของการตรวจสอบผลคะแนนจากการสอบไม่เป็นเอกสารลับที่ไม่สามารถเปิดเผยได้อีกต่อไป สืบเนื่องจากพระราชบัญญัติข้อมูลข่าวสารทางราชการ พ.ศ. 2540 มาตรา 11 ได้กำหนดให้บุคคลสามารถขอข้อมูลข่าวสารอื่นใดของราชการ และคำขอของผู้นั้นระบุข้อมูลข่าวสารที่ต้องการในลักษณะที่เข้าใจได้ตามควร ให้หน่วยงานของรัฐผู้รับผิดชอบจัดหาข้อมูลข่าวสารนั้น แก่ผู้ขอภายในเวลาอันสมควร ยกเว้นกรณีข้อมูลข่าวสารของราชการที่มีคำสั่งมิให้เปิดเผยตามมาตรา 15 ซึ่งเปิดเผยจะก่อให้เกิดความเสียหายต่อความมั่นคงของประเทศ เกิดอันตรายต่อชีวิตความปลอดภัยของบุคคลหนึ่งบุคคลใด ฯลฯ (พระราชบัญญัติข้อมูลข่าวสารทางราชการ พ.ศ. 2540, 2540, หน้า 4) ดังนั้นเรื่องของเกณฑ์การวัดผล การตัดเกรด คะแนนผลการสอบโดยเฉพาะการสอบแข่งขันที่มีผลต่อการสอบเข้าทำงานหรือการเลื่อนตำแหน่งเลื่อนระดับ ผู้สอบสามารถยื่นคำร้องขอตรวจสอบและขอความเป็นธรรมได้ ซึ่งถ้าหากแบบสอบมีความยากไม่ทัดเทียมกัน แต่นำมาตัดสินผลด้วยเกณฑ์เดียวกัน ผู้ที่ทำข้อสอบฉบับที่ยากย่อมเสียเปรียบสามารถนำไปฟ้องร้องต่อศาลปกครองได้ ดังนั้นเรื่องของความยุติธรรมในการวัดประเมินจึงมีความสำคัญยิ่ง

ดังนั้นเมื่อจำเป็นจะต้องใช้แบบสอบหลาย ๆ ฉบับในการสอบคราวเดียวกัน หรือใช้แบบสอบต่างฉบับในการสอบต่างเวลากัน โดยที่แบบสอบสร้างมาจากเนื้อหาเดียวกัน ข้อสอบประเภทเดียวกัน และมีพิสัยความยากอยู่ในระดับเดียวกัน แต่ข้อคำถามที่ใช้ในแต่ละฉบับมีความแตกต่างกัน มีโอกาสน้อยที่แบบสอบจะมีความเท่าเทียมกันในระดับความยาก คะแนนดิบที่ได้จึงไม่สามารถนำมาเปรียบเทียบกันได้โดยตรง ทางออกที่ดีที่สุด คือ จะต้องทำการปรับเทียบคะแนนเพื่อให้การแปลผลคะแนนการสอบ สามารถเทียบเคียงกันได้อย่างยุติธรรมกับผู้เข้าสอบทุกคน ไม่ว่าจะเป็นการสอบแข่งขันเพื่อบรรจุคนเข้าทำงานหรือสอบวัดผลสัมฤทธิ์ทางการเรียน

การปรับเทียบคะแนนจึงเป็นเทคนิควิธีที่มีบทบาทสำคัญ ที่ได้รับการยอมรับว่าสามารถแปลงระบบคะแนนจากแบบสอบต่างฉบับให้อยู่บนสเกลเดียวกัน ทำให้คะแนนจากแบบสอบต่างฉบับนั้น สามารถนำมาเปรียบเทียบกันได้อย่างยุติธรรมกับผู้เข้าสอบทุกคน และตรงกับความสามารถที่แท้จริงของแต่ละบุคคล จากการศึกษางานวิจัย พบว่า ไม่มีวิธีการปรับเทียบคะแนนวิธีการใดที่ดีที่สุดขึ้นอยู่กับเงื่อนไขของการปรับเทียบคะแนนที่จะนำไปใช้ในสถานการณ์ต่าง ๆ ที่แตกต่างกันในแต่ละบริบทที่ทำการศึกษา (พัชรี จันทร์เพ็ง, 2550, หน้า 5) แบบสอบที่มีความยากแตกต่างกันการปรับเทียบคะแนนด้วยทฤษฎีตอบข้อสอบจะมีความเหมาะสมมากที่สุด รองลงมา คือ การเทียบคะแนนแบบเชิงเส้นตรง (Kolen, 1981, p. 9) ส่วนงานวิจัยที่ศึกษาการเปรียบเทียบคะแนนในแนวตั้งของ Patience โดยใช้ วิธีอิควิเปอร์เซ็นไทล์ วิธีการของเธอร์สโตน วิธีการใช้รูปแบบโลจิสติก 1 2 และ 3 พารามิเตอร์ พบว่า ให้ผลการปรับเทียบคล้ายคลึงกันในแบบสอบฉบับง่ายและยากปานกลาง แต่ถ้าเป็นแบบสอบ

ฉบับยากวิธีการของเธอร์สโตนและวิธี 2 พารามิเตอร์จะให้ผลการปรับเทียบที่ดีกว่า (Patience, 1990 อ้างอิงถึงใน ภัทราพร เกษสังข์, 2546, หน้า 5) สำหรับขนาดกลุ่มตัวอย่าง ที่มีขนาดใหญ่วิธีการปรับเทียบคะแนนที่เหมาะสม ได้แก่ วิธีอิควิเปอร์เซ็นไทล์ วิธี IRT 3 พารามิเตอร์ ถ้ากลุ่มตัวอย่างมีขนาดเล็กวิธีการปรับเทียบคะแนนที่เหมาะสม ได้แก่ วิธีการปรับเทียบเชิงเส้นตรง (Linear equating) วิธีของราสซ์ (Rasch) วิธีเคอเนล (KE: Kernel equating) แต่ถ้ากลุ่มตัวอย่างมีขนาดเล็กมาก ๆ วิธีการปรับเทียบคะแนนที่เหมาะสม ได้แก่ วิธี Identity วิธี Mean (Kolen & Brennan, 2004, pp. 293-294; Rebecca & Dvorak, 2009) นอกจากนี้พบว่า วิธี KE จะใช้กลุ่มตัวอย่างที่น้อยกว่าวิธีอื่น ๆ และให้ค่าความคลาดเคลื่อนในการปรับเทียบที่น้อยกว่าวิธีอื่น ๆ (Rebecca & Dvorak, 2009, p. 76) และยังพบว่าวิธี KE ให้ผลการปรับเทียบที่ดีกว่าวิธีเชิงเส้นตรงและอิควิเปอร์เซ็นต์ไทล์ที่เป็นวิธีการปรับเทียบแบบดั้งเดิม (Mao et al., 2006; Von Davier et al., 2006 cited in Meng, 2012, p. 40) ส่วนวิธี IRT เป็นวิธีที่อยู่บนข้อตกลงที่สมเหตุสมผล อย่างเช่นกรณีการปรับเทียบด้วยราสซ์โมเดล เมื่อใช้รูปแบบกลุ่มตัวอย่างแบบกลุ่มสุ่ม หรือแบบผู้สอบต่างกลุ่มโดยใช้แบบสอบร่วม ควรอยู่บนเงื่อนไข แบบทดสอบต่างฉบับต้องมีโครงสร้างอย่างเดียวกัน มีความยากใกล้เคียงกัน ใช้กับกลุ่มตัวอย่างขนาดเล็ก ผลลัพธ์ที่ได้จึงจะมีความแกร่งในเรื่องการวิเคราะห์และประมาณค่า รวมทั้งจะมีความถูกต้องเมื่อค่าของข้อมูลอยู่บริเวณใกล้ ๆ กับค่าเฉลี่ย ส่วนวิธีอื่น ๆ ก็จะมีการกำหนดเงื่อนไขเฉพาะที่แตกต่างกันไปเหล่านี้เป็นต้น จึงถือว่าวิธี IRT เป็นวิธีที่อยู่บนข้อตกลงที่สมเหตุสมผล (Kolen & Brennan, 2004, p. 294) และยังพบว่า วิธีตามทฤษฎีตอบข้อสอบแบบโลจิสติก 2 พารามิเตอร์ ที่ใช้แบบสอบร่วมภายใน ที่มีความยาว 15 ข้อ มีความยากเฉลี่ยระดับยากมาก จะให้ค่าความคลาดเคลื่อนมาตรฐานของการเชื่อมโยงคะแนนต่ำสุด รวมทั้งมีความเพียงพอของการเชื่อมโยงคะแนนดีที่สุดเมื่อเทียบกับวิธีตามทฤษฎีตอบข้อสอบแบบโลจิสติก 3 พารามิเตอร์ (ภัทราพร เกษสังข์, 2546, หน้า 130) ดังนั้นในบริบทที่ชุดวิชาต่าง มีผู้เรียนตั้งแต่น้อยจนกระทั่งมีจำนวนมากจึงจำเป็นจะต้องเลือกใช้วิธีการปรับเทียบที่เหมาะสม เพื่อความเหมาะสมกับกลุ่มขนาดเล็ก ผู้วิจัยจึงเลือกศึกษาวิธีการปรับเทียบโดยใช้วิธีเคอเนล ส่วนกลุ่มตัวอย่างที่มีขนาดปานกลางและใหญ่เลือกใช้วิธีตามทฤษฎีตอบข้อสอบแบบโลจิสติก 2 พารามิเตอร์ที่มีความเหมาะสมกับขนาดตัวอย่างมากกว่าวิธี 1 พารามิเตอร์ซึ่งเหมาะกับกลุ่มขนาดเล็กส่วนวิธี 3 พารามิเตอร์เหมาะกับกลุ่มที่มีขนาดใหญ่ที่มีขนาดมากกว่า 1,000 คน (Kolen & Brennan, 2004, p. 288)

ส่วนการออกแบบการปรับเทียบโดยใช้ข้อสอบร่วม ซึ่งเป็นการนำข้อสอบจำนวนหนึ่งที่เหมือนกันไปใช้ร่วมกันระหว่างแบบสอบ 2 ฉบับ ที่จะนำมาปรับเทียบ ด้วยการนำคะแนนจากแบบสอบร่วมมาใช้ในกระบวนการวิเคราะห์สำหรับการปรับเทียบคะแนน ข้อสอบร่วมจึงเป็นอีกองค์ประกอบที่สำคัญมากต่อการปรับเทียบ โดยเฉพาะในสถานการณ์ที่เป็นการสอบต่างเวลากลุ่ม

ผู้สอบไม่ได้มาจากกลุ่มประชากรเดียวกัน รูปแบบการปรับเทียบโดยใช้ข้อสอบร่วมจึงมีความเหมาะสม ซึ่งการใช้แบบแผนข้อสอบร่วมในกลุ่มไม่เท่าเทียมกันในสถานการณ์ดังกล่าว ข้อสอบร่วมที่ดีจะต้องเป็นตัวแทนทั้งเนื้อหาและค่าสถิติจากแบบสอบชุดเก่าที่จะนำไปใช้ในการปรับเทียบ ข้อสอบร่วมจึงต้องกระจายทั่วทั้งฉบับ และไม่มีการเปลี่ยนแปลงคำพูดหรือภาษาใหม่ในแบบสอบทั้งใหม่และเก่า (Kolen & Brennan, 2004, p. 19) รวมทั้งไม่มีการเปลี่ยนแปลงตำแหน่งของข้อสอบร่วมทั้ง 2 ฟอร์ม เช่น ในฉบับแรกข้อสอบร่วมอยู่ตอนต้นของแบบสอบ แต่ฉบับหลังข้อสอบร่วมอยู่ตอนท้าย เพราะจะมีผลต่อการปรับเทียบ (Kolen & Brennan, 2004, p. 19; Cook & Petersen, 1987; Eignor, 1985; Kolen & Harris, 1990) และควรหลีกเลี่ยงการปรับปรุงแก้ไขหรือเรียงตัวเลือกใหม่ (Cizek, 1994) นอกจากนี้หากข้อสอบร่วมมีความยากปานกลางจะช่วยลดความคลาดเคลื่อนในการปรับเทียบคะแนนได้ดี (Holland & Sinharay, 2007) ซึ่งสอดคล้องกับการศึกษาของคาลด์เวลล์ (Caldwell, 1984) การศึกษาประสิทธิผลของโมเดลการปรับเทียบเชิงเส้นตรงกับราสช์โมเดล โดยใช้ข้อสอบร่วม 2 รูปแบบ คือ ข้อสอบร่วมที่ยากมากกับยากปานกลาง ผลปรากฏว่าข้อสอบร่วมที่มีความยากปานกลางให้ประสิทธิผลที่ดีกว่า (Caldwell, 1984 อ้างถึงใน พิชัย ละแมนชัย, 2538, หน้า 49)

นอกจากนี้คลีนและโคเลน กล่าวว่า หากผู้สอบทั้งสองกลุ่มมีความสามารถใกล้เคียงกัน ความยาวของข้อสอบร่วมไม่ได้มีผลต่อคุณภาพของการเชื่อมโยงคะแนนหรือการปรับเทียบคะแนน (Klein & Kolen, 1985 cited in Wendy, 2009, p. 12) กรณีความยาวของข้อสอบร่วม Angoff (1984, p. 107) กล่าวว่าความยาวของข้อสอบร่วมควรมีจำนวนไม่น้อยกว่าร้อยละ 20 ของจำนวนข้อสอบในแบบสอบหรือควรมีจำนวนไม่น้อยกว่า 20 ข้อ แล้วแต่จำนวนไหนจะมากกว่าให้ใช้จำนวนนั้น ส่วนไรท์และสโตน (Wright & Stone, 1979, p. 98) กล่าวว่า ข้อสอบร่วมที่วัดเรื่องเดียวกันกับแบบสอบทั้งสองชุด มีจำนวนเพียง 10 ข้อก็เพียงพอแล้ว นอกจากนี้จากการศึกษาของ Mckinley and Rackase (1981 cited in Cook & Petersen, 1987) ในการใช้ข้อสอบร่วม 3 ขนาด คือ 5 15 และ 25 ข้อ โดยใช้ข้อมูลจริงในการปรับเทียบรูปแบบ IRT 3 พารามิเตอร์ พบว่า ข้อสอบร่วมตั้งแต่ 15 ข้อ เป็นจำนวนที่เพียงพอสำหรับการปรับเทียบ (สุนิสา จุ้ยม่วงศรี, 2537, หน้า 4) การวิจัยครั้งนี้จึงเลือกข้อสอบร่วมที่ 15 ข้อ เพื่อใช้เป็นตัวแทนของเนื้อหาวิชาในชุดวิชาที่นำมาใช้ในการวิจัยซึ่งมีเนื้อหาประกอบด้วย 15 หน่วย สุ่มข้อสอบร่วมมาหน่วยละ 1 ข้อ เพราะถ้าหากเลือกมา 2 ข้อต่อหน่วย ก็จะมีจำนวนข้อสอบซ้ำของเดิมมากเกินไป เกิดความไม่เป็นธรรมกับผู้ที่เคยสอบไม่ผ่านมาก่อนและมาสอบซ้ำและจำข้อสอบได้

ตัวแปรต่าง ๆ ไม่ว่าจะเป็นวิธีการปรับเทียบ ลักษณะและขนาดของข้อสอบร่วมที่มีผลต่อคุณภาพของการปรับเทียบคะแนน ยังมีขนาดของกลุ่มตัวอย่างที่มีผลต่อการปรับเทียบคะแนน โดยที่

ขนาดของกลุ่มตัวอย่างมีผลกระทบโดยตรงต่อความคลาดเคลื่อนอย่างสุ่มของการปรับเทียบ ขนาดของกลุ่มตัวอย่างยิ่งมากยิ่งดี (Kolen & Brennan, 2004, p. 288) กรณีที่กลุ่มตัวอย่างมีขนาดไม่ใหญ่ รูปแบบการปรับเทียบคะแนนโดยใช้ทฤษฎีการตอบข้อสอบ 1 พารามิเตอร์จะมีความเหมาะสม แต่ถ้ากลุ่มตัวอย่างมีขนาดใหญ่ควรเลือกใช้รูปแบบ 2 หรือ 3 พารามิเตอร์จะเหมาะสมกว่า (Slinda & Linn, 1978, pp. 23-35) ซึ่ง Meng (2012, p. 45) ได้กำหนดว่าถ้าใช้กลุ่มตัวอย่างขนาด 200 คนต่อฉบับ ถือเป็นกลุ่มตัวอย่างขนาดเล็ก ขนาด 500 คนต่อฉบับ ถือเป็นกลุ่มตัวอย่างขนาดกลาง และขนาด 2,000 คนต่อฉบับ ถือเป็นกลุ่มตัวอย่างขนาดใหญ่ อย่างไรก็ตามWright and Stone (1979, p. 98) กล่าวว่าในทางปฏิบัติถ้าเป็นการใช้แบบสอบจากคลังข้อสอบที่มีโครงสร้างแบบเดียวกัน สามารถใช้กลุ่มตัวอย่างขนาดเล็ก ขนาด 100 คนได้ ดังนั้นในการวิจัยครั้งนี้จึงกำหนดขนาดกลุ่มตัวอย่างเป็น 3 ขนาด คือ ขนาดเล็ก (100 คน) ขนาดกลาง (500 คน) และขนาดใหญ่ (700 คน)

ตัวแปรต่าง ๆ ที่เกี่ยวข้องกับการปรับเทียบไม่ว่าจะเป็นขนาดของกลุ่มตัวอย่าง ความยาวของแบบสอบ ขนาดของข้อสอบร่วมและความแตกต่างของความสามารถของกลุ่มตัวอย่าง ล้วนมีผลต่อการปรับเทียบคะแนน (Ricker & Von Davier, 2007; Wang, Haiming, Donghong & Youming, 2008) ข้อสอบร่วมและขนาดกลุ่มตัวอย่างที่เพิ่มขึ้นจะช่วยลดความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบด้วยวิธี IRT (Hanson & Beguin, 2002) รวมทั้งการเพิ่มขนาดกลุ่มตัวอย่างและเพิ่มความยาวของข้อสอบจะทำให้การปรับเทียบด้วยวิธีของ Kernel มีความเหมาะสมมากขึ้น (Lee, 2007) โดยทั่วไปแล้วความยาวของข้อสอบร่วม ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่างมีความสัมพันธ์เชิงบวกกับการปรับเทียบคะแนน ในขณะที่ความแตกต่างของสามารถและความยากของข้อสอบร่วมมีความสัมพันธ์เชิงลบต่อการปรับเทียบ (Huang, 2009; Sinharay & Holland, 2007) ในการออกแบบการปรับเทียบจึงควรคำนึงถึงตัวแปรเหล่านี้ (Meng, 2012, p. 27)

จากเหตุผลดังกล่าวข้างต้น ผู้วิจัยจึงเห็นความสำคัญในการที่จะสร้างความยุติธรรมในการสอบให้กับนักศึกษาทุกคน จากการสอบที่ต่างเวลา จากแบบสอบที่ต่างฉบับ ให้สามารถนำผลคะแนนมาเปรียบเทียบกันได้อย่างยุติธรรม ได้รับความยุติธรรมในการตัดสินผล หรือได้รับความยุติธรรมในเรื่องของเกรดที่ได้รับ และสามารถใช้เป็นหลักประกันในเรื่องของมาตรฐานการศึกษาว่าจะสามารถคงไว้ซึ่งเกณฑ์และมาตรฐานเดียวกันไม่ว่าจะเป็นการสอบในปีการศึกษาใด และจัดให้กับนักศึกษากลุ่มใดก็ตามทั้งในอดีต ปัจจุบัน และในอนาคต ยิ่งไปกว่านั้นผู้วิจัยเห็นว่า การวัดผลการศึกษานอกจากการคำนึงถึงความยุติธรรมแล้ว ควรตระหนักในเรื่องของมนุษยชนที่นักศึกษาพึงได้รับการคุ้มครองว่าได้รับการปฏิบัติที่เท่าเทียมกัน การเรียกร้องสิทธิ การเปิดเผยข้อมูลข่าวสาร ซึ่งนับวันเรื่องเหล่านี้จะมีบทบาทมากขึ้น และเป็นเรื่องใกล้ตัวของนักวัดผลการศึกษาที่จะต้องได้รับการตรวจสอบ

ดังนั้นเมื่อจำเป็นจะต้องใช้แบบสอบต่างฉบับในการวัดผลคราวเดียวกัน เพื่อการตัดสินผลให้เกิดความเป็นธรรมกับนักศึกษาทุกคนจำเป็นจะต้องใช้วิธีการการปรับเทียบคะแนนเข้ามาช่วย โดยเฉพาะในสถานการณ์ที่มีผู้สอบต่างกลุ่มที่ไม่เท่าเทียมกัน และกลุ่มผู้สอบมีขนาดแตกต่างกันไป ตั้งแต่ขนาดเล็กจนถึงขนาดใหญ่ ผู้วิจัยจึงสนใจศึกษาวิธีการปรับเทียบคะแนน และการออกแบบการปรับเทียบให้สอดคล้องกับสถานการณ์และบริบทดังกล่าว ด้วยการเลือกใช้วิธีการปรับเทียบคะแนน 2 วิธี คือวิธีเคอเนล (Kernel) และวิธีตามแนวทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์ ที่ใช้ได้ดีกับกลุ่มตัวอย่างขนาดเล็กและกลุ่มตัวอย่างขนาดใหญ่ สอดคล้องกับบริบทของสถานศึกษาที่ผู้วิจัยทำการศึกษา โดยเลือกศึกษากับชุดวิชาที่มีนักศึกษาลงทะเบียนเรียนที่สามารถแบ่งเป็นขนาดต่าง ๆ ได้ ทั้งขนาดเล็ก กลาง และขนาดใหญ่ ด้วยการออกแบบวิธีการปรับเทียบคะแนนโดยใช้รูปแบบของข้อสอบร่วมภายในที่เป็นตัวแทนของเนื้อหาทั้งหมด เปรียบเทียบระหว่างข้อสอบร่วมที่มีความยากปานกลาง และความยากอย่างสุ่ม ตามที่มีเอกสารงานวิจัยที่เกี่ยวข้องยืนยันข้อมูลว่าจะทำให้การปรับเทียบคะแนนมีคุณภาพและลดความคลาดเคลื่อนในการปรับเทียบคะแนนได้ดี สำหรับรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ ในการปรับเทียบจะศึกษาจากคะแนนทั้งฉบับกับคะแนนที่ได้จากการตัดข้อสอบบางข้อที่ไม่มีคุณภาพทิ้ง หลังจากนั้นจึงศึกษาความสอดคล้องของการตัดเกรดจากคะแนนก่อนการปรับเทียบคะแนน กับการตัดเกรดจากคะแนนหลังการปรับเทียบคะแนนให้อยู่บนสเกลเดียวกัน เพื่อหาวิธีการปรับเทียบภายใต้รูปแบบและเงื่อนไขที่ดีที่สุด ที่จะนำไปสู่การตัดเกรดที่ยุติธรรม ไม่ให้นักศึกษาเกิดการได้เปรียบเสียเปรียบกันจากการสอบด้วยแบบสอบที่ต่างฉบับกัน ซึ่งงานวิจัยที่ศึกษาความสอดคล้องของการตัดเกรดหลังการปรับเทียบคะแนนในประเทศไทยมีน้อยมากมีเพียงเยาวดี รางชัยกุล เคยศึกษากระบวนการปรับเทียบคะแนนกับการสอบไล่ชั้นประโยคมัธยมศึกษาตอนปลาย (มศ. 5) เมื่อนานมาแล้ว พบว่า สัดส่วนของนักเรียนที่น่าจะสอบได้แต่ถูกตัดสินให้สอบตกมีสัดส่วนที่น่าสนใจ กล่าวคือ มีนักเรียนร้อยละ 37.6 ที่สอบตกใน พ.ศ. 2516 และร้อยละ 9 ที่ตกใน พ.ศ. 2517 ควรจะเป็นผู้ที่สอบได้หลังการปรับเทียบคะแนนให้เทียบเท่ากับแบบสอบใน พ.ศ. 2516 (เยาวดี รางชัยกุล, 2518 อ้างถึงใน ภาวิณี ศรีสุขวัฒนานันท์, 2528, หน้า 2) ผู้วิจัยจึงสนใจนำมาศึกษาซ้ำในบริบทใหม่กับสถานการณ์การสอบจริง ที่ไม่ใช่การจำลองข้อสอบดังเช่นที่งานวิจัยส่วนใหญ่ที่ศึกษาเกี่ยวกับเรื่องของการปรับเทียบคะแนนนิยมศึกษา

คำถามการวิจัย

1. วิธีการปรับเทียบภายใต้เงื่อนไขรูปแบบของข้อสอบร่วม ขนาดกลุ่มตัวอย่าง และรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกันวิธีการใดมีคุณภาพดีกว่ากัน

2. ผลการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขต่าง ๆ วิธีการใดให้ผลสอดคล้องมากกว่ากัน

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาคุณภาพของวิธีการปรับเทียบคะแนนภายใต้เงื่อนไขรูปแบบข้อสอบร่วม ขนาดตัวอย่างและรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกัน

2. เพื่อเปรียบเทียบความสอดคล้องของผลการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขที่กำหนด

สมมติฐานของการวิจัย

การเปรียบเทียบผลการตัดเกรดจากคะแนนก่อนและหลังการปรับเทียบคะแนน ด้วยวิธีปรับเทียบที่มีรูปแบบแตกต่างกันภายใต้เงื่อนไขต่าง ๆ ที่กำหนด จากการศึกษาหลักการเชิงทฤษฎีและงานวิจัยที่เกี่ยวข้อง พบว่า วิธีการปรับเทียบ ขนาดของกลุ่มตัวอย่าง ลักษณะของข้อสอบร่วม ข้อมูลที่จะนำมาวิเคราะห์ ล้วนมีผลต่อคุณภาพของการปรับเทียบคะแนน โดยที่กลุ่มตัวอย่างมีขนาดใหญ่วิธีการปรับเทียบคะแนนที่เหมาะ ได้แก่ วิธีอิควิเปอร์เซ็นไทล์ วิธี IRT 3 พารามิเตอร์ ถ้ากลุ่มตัวอย่างมีขนาดเล็กวิธีการปรับเทียบคะแนนที่เหมาะ ได้แก่ วิธีการปรับเทียบเชิงเส้นตรง (Linear equating) วิธีของราสซ์ (Rasch) วิธีเคอเนล (KE: Kernel equating) (Kolen & Brennan, 2004, pp. 293-294; Rebecca & Dvorak, 2009) นอกจากนี้พบว่า วิธีอิควิเปอร์เซ็นไทล์ วิธีการของเธอร์สโตน วิธีการใช้รูปแบบโลจิสติก 1 2 และ 3 พารามิเตอร์ ให้ผลการปรับเทียบคล้ายคลึงกันในแบบสอบฉบับง่าย และยากปานกลาง แต่ถ้าเป็นแบบสอบฉบับยากวิธีการของเธอร์สโตนและวิธี 2 พารามิเตอร์จะให้ผลการปรับเทียบที่ดีกว่า (Patience, 1990 อ้างถึงใน ภัทราพร เกษสังข์, 2546, หน้า 5) ซึ่งขนาดของกลุ่มตัวอย่างมีผลกระทบโดยตรงต่อความคลาดเคลื่อนอย่างสุ่มของการปรับเทียบ โดยขนาดของกลุ่มตัวอย่างยิ่งมากยิ่งดี (Kolen & Brennan, 2004, p. 288)

ส่วนประเด็นของข้อสอบร่วม ข้อสอบร่วมที่มีความยากปานกลางให้ประสิทธิผลที่ดีกว่า (Caldwell, 1984) รวมทั้งช่วยลดความคลาดเคลื่อนในการปรับเทียบคะแนนได้ดี (Holland & Sinharay, 2007) จากข้อค้นพบที่ได้ผู้วิจัยจึงนำมาเป็นแนวทางในการกำหนดสมมติฐานของการวิจัยดังนี้

1. วิธีการปรับเทียบคะแนน วิธีเคอเนล (Kernel) น่าจะเหมาะและใช้ได้ดีกับกลุ่มตัวอย่างขนาดเล็ก ส่วนวิธีตามแนวทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์ น่าจะเหมาะและใช้ได้ดีกับกลุ่มตัวอย่างขนาดใหญ่

2. การปรับเทียบคะแนนโดยใช้ข้อสอบร่วมที่มีความยากง่ายปานกลาง (ค่า p อยู่ระหว่าง .4-.6) และการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่มมีคุณภาพในการปรับเทียบคะแนนแตกต่างกัน

3. การนำข้อมูลข้อสอบทั้งหมดมาวิเคราะห์การปรับเทียบคะแนน กับตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนที่จะทำการปรับเทียบคะแนนมีคุณภาพแตกต่างกันโดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน

4. การเปรียบเทียบความสอดคล้องของผลการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขที่กำหนด จะให้ผลแตกต่างกัน

ขอบเขตของการวิจัย

1. การวิจัยครั้งนี้ใช้วิธีการศึกษาจากข้อมูลในสถานการณ์จริง โดยใช้วิธีการปรับเทียบคะแนนในแนวนอน (Horizontal equating) 2 วิธี คือ วิธีเคอเนล (Kernel equating) ที่วิเคราะห์ตามแนว Chain equating (CE) กับวิธีตามแนวทฤษฎีการตอบข้อสอบ (Item sesponse theory) 2 พารามิเตอร์ โดยใช้ข้อสอบร่วมภายใน และใช้รูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกัน เพื่อให้ได้คะแนนที่สามารถเทียบแทนกันได้ จากนั้นจึงทำการเปรียบเทียบผลการตัดเกรดระหว่างการใช้คะแนนก่อนการปรับเทียบกับคะแนนหลังการปรับเทียบ ที่แบ่งการตัดเกรดเป็น 3 ระดับ และ 8 ระดับ

2. ประชากรและการเลือกกลุ่มตัวอย่าง

2.1 ประชากรที่ใช้ในการวิจัยครั้งนี้ เป็นนักศึกษาระดับปริญญาตรีของมหาวิทยาลัยแห่งหนึ่ง ที่สอบในชุดวิชาที่ผู้วิจัยทำการศึกษา

2.2 กลุ่มตัวอย่าง เป็นนักศึกษาที่เข้าสอบในชุดวิชาที่ผู้วิจัยเลือก 1 ชุดวิชา ที่สอบในภาคการศึกษาที่ 1/ 2556 ภาคการศึกษา 1/ 2557 และภาคการศึกษาที่ 1/ 2558 ตามเงื่อนไขของการปรับเทียบที่ใช้รูปแบบข้อร่วม 2 รูปแบบ โดยใช้แบบสอบที่สอบในภาคการศึกษาที่ 1/ 2556 เป็นฐานในการปรับเทียบ

3. ตัวแปรที่ใช้ในการวิจัย

การวิจัยครั้งนี้ มุ่งศึกษาใน 2 ประเด็นหลัก คือ

3.1 การปรับเทียบคะแนน ประกอบด้วย ตัวแปร ดังนี้

3.1.1 ตัวแปรอิสระ ได้แก่

3.1.1.1 วิธีการปรับเทียบ 2 วิธี คือ วิธีตามทฤษฎี IRT 2 พารามิเตอร์ กับวิธีเคอเนลที่วิเคราะห์ตามแนว Chain equating

3.1.1.2 รูปแบบของข้อสอบร่วม มี 2 รูปแบบ คือ สุ่มข้อสอบจากทุกหน่วยการเรียน จำนวน 15 ข้อ โดยกำหนดค่าความยาก (ค่า p) ระหว่าง .4-.6 ซึ่งมีความยากปานกลาง กับสุ่มข้อสอบร่วมอย่างง่ายโดยไม่กำหนดค่าความยาก

3.1.1.3 ขนาดกลุ่มตัวอย่างในการวิเคราะห์ มี 3 ขนาด คือ กลุ่มตัวอย่างขนาดเล็ก (100 คน) กลุ่มตัวอย่างขนาดกลาง (500 คน) และกลุ่มตัวอย่างขนาดใหญ่ (700 คน)

3.1.1.4 รูปแบบข้อมูลที่จะนำมาวิเคราะห์ มี 2 รูปแบบ คือ ใช้ข้อสอบทั้งฉบับมาวิเคราะห์ กับตัดข้อสอบบางข้อที่ไม่มีคุณภาพทิ้ง

3.1.2 ตัวแปรตาม คุณภาพของการปรับเทียบคะแนน

3.2 การตัดเกรด ประกอบด้วย ตัวแปร ดังนี้

3.2.1 ตัวแปรอิสระ ได้แก่

3.2.1.1 วิธีการตัดเกรด 2 วิธี คือ แบ่งการตัดเกรดเป็น 3 เกรด กับแบ่งการตัดเกรดเป็น 8 เกรด

3.2.1.2 คะแนนที่ใช้ในการตัดเกรด 2 แบบ คือ คะแนนก่อนการปรับเทียบกับคะแนนหลังจากการปรับเทียบแล้ว

3.2.2 ตัวแปรตาม ได้แก่ ความสอดคล้องของการตัดเกรด

4. ข้อมูลที่ใช้ในการศึกษาครั้งนี้ เป็นการศึกษาจากสถานการณ์จริงของการสอบจากแบบสอบของชุดวิชาหนึ่ง ของมหาวิทยาลัยของรัฐแห่งหนึ่ง จากแบบสอบแบบเลือกตอบที่มีความยาว 120 ข้อ ที่ใช้สอบในภาคการศึกษาที่ 1/ 2556 ภาคการศึกษาที่ 1/ 2557 และภาคการศึกษาที่ 1/ 2558 โดยใช้ข้อสอบของภาคการศึกษาที่1/ 2556 เป็นแบบสอบฉบับฐานในการปรับเทียบคะแนน โดยใช้ข้อสอบร่วมจำนวน 15 ข้อ

5. การหาคุณภาพของการปรับเทียบคะแนน จะพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของปรับเทียบคะแนน (Standard error of equating) ของแต่ละวิธีเปรียบเทียบกันตามเงื่อนไขต่าง ๆ ของการปรับเทียบที่กำหนด

6. การหาความสอดคล้องของการตัดเกรด พิจารณาระดับความสอดคล้องของการตัดเกรดจากการใช้คะแนนก่อนที่จะทำการปรับเทียบคะแนนกับคะแนนหลังจากปรับเทียบคะแนนแล้วมาใช้ในการตัดเกรด ว่ามีความสอดคล้องกันเพียงใด โดยดูจากค่าดัชนีของแคปปา

นิยามศัพท์เฉพาะ

1. การปรับเทียบคะแนน (Test equating) หมายถึง กระบวนการนำคะแนนที่ได้จากแบบสอบ 2 ฉบับ ที่วัดในเนื้อหาวิชาเดียวกันนำมาปรับให้อยู่บนสเกลเดียวกันเพื่อให้คะแนนที่ได้

จากแบบสอบทั้งสองสามารถเทียบแทนกันได้ ในการวิจัยครั้งนี้จะใช้วิธีการปรับเทียบ 2 วิธี คือวิธีเคอเนล (Kernel equating) กับวิธีทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์ โดยใช้แบบแผนการเก็บรวบรวมข้อมูลแบบกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบร่วม

2. การปรับเทียบวิธีเคอเนล (Kernel equating) หมายถึง วิธีการปรับเทียบตามทฤษฎีดั้งเดิมมีกระบวนการในการปรับเทียบ 5 ขั้นตอน คือ ขั้นการปรับเรียบ การประมาณคะแนนความน่าจะเป็น การทำคะแนนให้ต่อเนื่อง การปรับเทียบ และการหาความคลาดเคลื่อนของการปรับเทียบ

3. การปรับเทียบวิธีทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์ หมายถึง การนำหลักการของทฤษฎีการตอบข้อสอบ ที่ใช้พารามิเตอร์ของข้อสอบ 2 ตัว คือ ค่าความยากและค่าอำนาจจำแนกมาใช้ในการปรับเทียบคะแนน โดยยึดหลักว่าเมื่อความสามารถของผู้สอบอยู่บนสเกลเดียวกันหรือเท่ากัน สามารถนำคะแนนจากแบบสอบสองฉบับมาปรับเทียบกันได้

4. แบบแผนในการปรับเทียบคะแนน (Equating design) หมายถึง วิธีการต่าง ๆ ในการออกแบบเลือกกลุ่มตัวอย่าง และเลือกจัดรูปแบบของแบบสอบในการปรับเทียบคะแนน สำหรับการวิจัยครั้งนี้ กำหนดเป็นแบบแผนกลุ่มไม่เท่าเทียมกันโดยใช้แบบสอบร่วม ผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้แบบสอบร่วม (Non-equivalent groups anchor test design หรือ Common-item nonequivalent groups design)

5. ข้อสอบร่วม (Anchor-test) หมายถึง ข้อสอบกลุ่มหนึ่ง ที่ผู้สอบทั้ง 2 กลุ่ม ทำเหมือนกัน ซึ่งอาจจะจัดผนวกเข้าเป็นแบบสอบฉบับเดียวกัน หรือจะจัดแยกฉบับต่างหากจากแบบสอบที่ใช้ในการปรับเทียบคะแนนก็ได้ สำหรับการวิจัยครั้งนี้จะใช้ข้อสอบร่วมภายใน (Internal anchor test)

6. ข้อสอบร่วมภายใน (Internal anchor test) หมายถึง ข้อสอบที่เหมือนกันจำนวนหนึ่งที่ใช้ร่วมกันระหว่างแบบสอบ 2 ฉบับใด ๆ ที่จะนำมาปรับเทียบคะแนน โดยที่นำข้อสอบร่วมนี้ไปจัดฉบับรวมอยู่ในฉบับเดียวกันของทั้งสองฉบับ ไม่ได้จัดแยกเป็นคนละฉบับต่างหาก งานวิจัยเรื่องนี้กำหนดเงื่อนไขที่เกี่ยวข้องกับข้อสอบร่วมแบ่งออกเป็น 4 ลักษณะ คือ

6.1 ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 เป็นข้อสอบร่วมที่มีค่าความยาก (p) ตามทฤษฎีการทดสอบแบบดั้งเดิมที่มีค่า p ตั้งแต่ .4-.6 ที่สุ่มมาจากข้อสอบในแต่ละหน่วย ๆ ละ 1 ข้อ

6.2 ข้อสอบร่วมมีความยากอย่างสุ่ม เป็นข้อสอบร่วมที่สุ่มมาจากข้อสอบในแต่ละหน่วย หน่วยละ 1 ข้อ โดยไม่ระบุค่าความยากของข้อสอบร่วม

6.3 ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เป็นเงื่อนไขที่เกี่ยวข้องกับรูปแบบของข้อมูลที่นำมาวิเคราะห์ในการปรับเทียบคะแนน ซึ่งข้อสอบร่วมจะมีลักษณะเช่นเดียวกันกับของข้อ 6.1 แต่ในการวิเคราะห์ข้อมูลจะตัดข้อสอบที่มีค่าความยาก ค่า p ต่ำกว่า 0.2 และค่า p ที่มีค่ามากกว่า 0.8 ทิ้ง

6.4 ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เป็นเงื่อนไขที่เกี่ยวข้องกับรูปแบบของข้อมูลที่นำมาวิเคราะห์ในการปรับเทียบคะแนน ซึ่งข้อสอบร่วมจะมีลักษณะเช่นเดียวกันกับของข้อ 6.2 แต่ในการวิเคราะห์ข้อมูลจะตัดข้อสอบที่มีค่าความยาก ค่า p ต่ำกว่า 0.2 และค่า p ที่มีค่ามากกว่า 0.8 ทิ้ง

7. แบบสอบฉบับฐาน หมายถึง แบบสอบที่ใช้เป็นหลักให้แบบสอบฉบับอื่น ๆ ปรับเทียบคะแนนเข้าสู่แบบสอบฉบับนี้

8. คุณภาพของการปรับทียบ หมายถึง ความถูกต้องของการปรับเทียบคะแนน ซึ่งพิจารณาของแต่ละวิธีแยกกันตามค่าที่ได้จากโปรแกรมการวิเคราะห์ของแต่ละวิธี

9. ผลการตัดเกรด หมายถึง การนำคะแนนที่ได้จากการสอบมาใช้ในการกำหนดระดับผลการเรียน ในการวิจัยครั้งนี้จะใช้คะแนนก่อนการปรับเทียบ และคะแนนหลังการปรับเทียบมาใช้ในการตัดเกรด โดยกำหนดระดับผลการตัดเกรดเป็น 2 รูปแบบ คือ

9.1 รูปแบบที่ 1 เป็นการตัดเกรดตามเกณฑ์ที่มหาวิทยาลัยกำหนด แบ่งเป็น 3 ระดับ คือ

U ได้คะแนนต่ำกว่าร้อยละ 60

S ได้คะแนนร้อยละ 60-75

H ได้คะแนนร้อยละ 76 ขึ้นไป

9.2 รูปแบบ 2 กำหนดเป็น 8 ระดับ เป็นการตัดเกรดตามเกณฑ์งานวิจัยที่ศึกษาเพื่อนำมาใช้ในการปรับระบบการตัดเกรดใหม่ คือ

A ได้คะแนนร้อยละ 76-100

B^+ ได้คะแนนร้อยละ 70-75

B ได้คะแนนร้อยละ 65-69

C^+ ได้คะแนนร้อยละ 60-64

C ได้คะแนนร้อยละ 55-59

D^+ ได้คะแนนร้อยละ 50-54

D ได้คะแนนร้อยละ 45-49

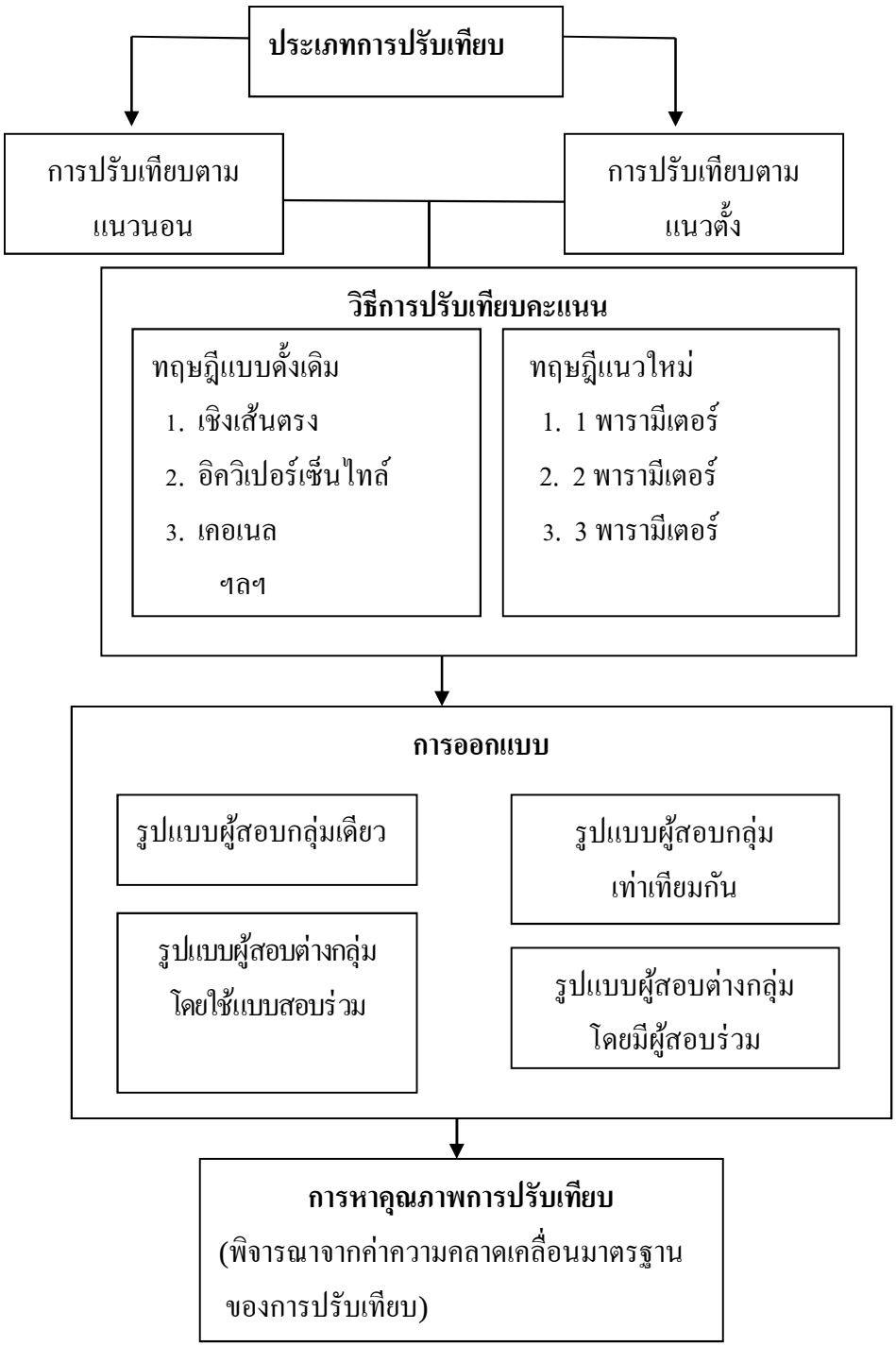
F ได้คะแนนต่ำกว่าร้อยละ 45

10. ความสอดคล้องของการตัดเกรด หมายถึง ระดับความสัมพันธ์ของการตัดเกรดจากคะแนนก่อนการปรับเทียบคะแนนกับการตัดเกรดจากคะแนนหลังการปรับเทียบคะแนน ที่มีค่าตั้งแต่ .81 ขึ้นไป ซึ่งถ้าการตัดเกรดมีความความสอดคล้องกัน ก็ไม่จำเป็นจะต้องทำการปรับเทียบคะแนน

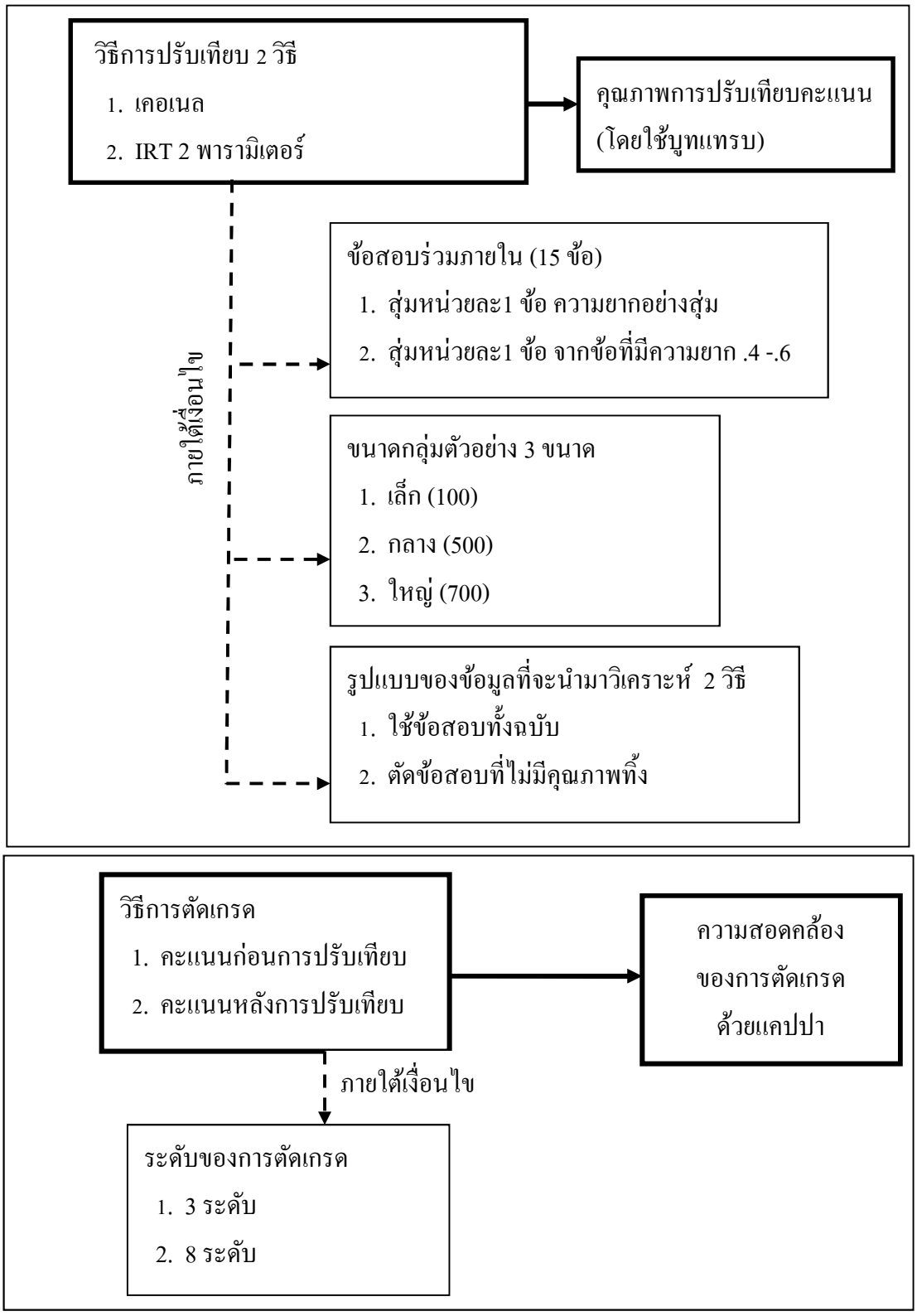
11. สัมประสิทธิ์แคปปา (Coefficient kappa) หมายถึง ตัวบ่งชี้ถึงระดับความสอดคล้องในการตัดสินใจของผู้ตัดสินตั้งแต่ 2 คนขึ้นไป ในการวิจัยครั้งนี้จะนำมาประยุกต์ใช้ในการตัดสินความสอดคล้องของการตัดเกรด เมื่อใช้คะแนนก่อนการปรับเทียบกับคะแนนหลังการปรับเทียบในการตัดเกรดว่ามีความสอดคล้องกันเพียงใด ตามเงื่อนไขต่าง ๆ ที่กำหนดของวิธีการปรับเทียบและวิธีการตัดเกรดที่ต่างกัน

กรอบแนวคิดการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์ เพื่อศึกษาวิธีการปรับเทียบคะแนน 2 วิธี คือ วิธีเคอเนลกับวิธีทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์ โดยมีรูปแบบของข้อสอบร่วม ขนาดกลุ่มตัวอย่าง และรูปแบบของข้อมูลที่จะนำมาวิเคราะห์การปรับเทียบคะแนนที่แตกต่างกัน หลังจากนั้นจึงนำคะแนนที่ได้จากการปรับเทียบ และคะแนนที่ไม่ผ่านการปรับเทียบไปใช้ในการตัดเกรด 2 รูปแบบ คือ การตัดเกรด 3 ระดับ กับการตัดเกรด 8 ระดับ เปรียบเทียบผลการตัดเกรดว่าสอดคล้องกันเพียงใด โดยนำกรอบแนวคิดเชิงทฤษฎี มาใช้เป็นแนวทางในการกำหนดกรอบแนวคิดในการวิจัย ดังภาพที่ 1-1 ถึง 1-2



ภาพที่ 1-1 กรอบแนวคิดเชิงทฤษฎีของการปรับเทียบคะแนน



ภาพที่ 1-2 กรอบแนวคิดในการวิจัย

ประโยชน์ที่ได้รับ

1. ผลจากการศึกษาสามารถใช้เป็นแนวทางในการตัดสินผล ให้กับสถาบันการศึกษาที่มีการสอบหลายครั้งในชุดวิชาที่มีเนื้อหาเดียวกัน และใช้แบบสอบหลายชุดในการตัดสินผลการเรียนในภาคการศึกษาเดียวกัน หรือต่างภาคการศึกษาเพื่อให้เกิดความยุติธรรมต่อผู้เข้าสอบทุกคน

2. สถาบันการศึกษาสามารถนำวิธีการจากการศึกษาไปประยุกต์และพัฒนาระบบการปรับเทียบคะแนนที่เหมาะสมให้กับชุดวิชาที่เปิดสอน เพื่อใช้เป็นหลักประกันมาตรฐานคุณภาพการศึกษาว่าจะผลิตบัณฑิตออกมาแต่ละรุ่นที่มีมาตรฐานด้านวิชาการที่เป็นมาตรฐานเดียวกัน

3. แบบสอบที่ใช้ทดสอบแต่ละปีแม้จะมีความแตกต่างในเรื่องของความยาก แต่สามารถใช้วิธีการทางสถิติในการปรับเทียบคะแนน เพื่อไม่ให้เกิดความลำเอียงกับนักศึกษาต่างปี/ ภาคการศึกษา สร้างความยุติธรรมให้กับผู้เข้าสอบทุกคน

4. ได้ขยายองค์ความรู้ในเรื่องของวิธีการปรับเทียบคะแนนโดยใช้วิธีการ รูปแบบของข้อสอบร่วม ขนาดกลุ่มตัวอย่างและรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกันให้กว้างขวางมากขึ้น

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการปรับเทียบคะแนนระหว่างแบบสอบ (Test equating) ในช่วงแรกจะมีการศึกษากันเฉพาะกลุ่มนักจิตมิติเท่านั้น ต่อมานักวัดผลเริ่มเห็นความสำคัญและนำการปรับเทียบคะแนนมาใช้ในเรื่องของการวัดและประเมินในประเด็นของความยุติธรรมเกี่ยวกับแบบสอบ จึงเป็นจุดเริ่มต้นของศาสตร์ด้านการวัดผลที่นำวิธีการของการปรับเทียบคะแนนมาศึกษาและพัฒนาขึ้นตามลำดับ ผู้วิจัยจึงขอเสนอผลการศึกษาเอกสารที่เกี่ยวข้องกับการปรับเทียบคะแนนตามลำดับ ดังนี้

ตอนที่ 1 แนวคิดการวัดและประเมินผลการศึกษา

ตอนที่ 2 แนวคิดทฤษฎีของการปรับเทียบคะแนนระหว่างแบบสอบ

ตอนที่ 3 การหาคุณภาพของการปรับเทียบคะแนน

ตอนที่ 4 งานวิจัยที่เกี่ยวข้องกับการปรับเทียบคะแนน

ตอนที่ 1 แนวคิดการวัดและประเมินผลการศึกษา

1. การวัดและประเมินผลการศึกษา

พจน์ สะเพียรชัย (2549, หน้า 123-128) กล่าวว่า การประเมินผลการศึกษา จะต้องทำการวิเคราะห์ใน 2 เรื่องใหญ่ ๆ คือ ทำการวิเคราะห์ระบบการศึกษา ซึ่งนักการศึกษาจะมองและวิเคราะห์เกี่ยวกับหลักสูตร วิเคราะห์ปรัชญา ความเชื่อที่ใช้ในการจัดการศึกษา วิเคราะห์ความมุ่งหมาย และนโยบายการจัดการศึกษา วิเคราะห์เนื้อหาวิชา ระบบการเรียนการสอน การวัดประมินผล ตลอดจนการบริหารจัดการที่เกี่ยวข้องกับการจัดการเรียนการสอน กับการวิเคราะห์ระบบการประเมินผล ซึ่งเกี่ยวข้องกับการหาเทคนิควิธี เพื่อให้ได้มาซึ่งข้อมูล หลักฐานเพื่อประกอบการตัดสินใจว่าการจัดการศึกษานั้นเหมาะสมหรือไม่เพียงใด ในการประเมินผลผู้ประเมินจะต้องทำการศึกษาสิ่งที่มุ่งประเมินให้เข้าใจอย่างถ่องแท้ก่อนว่ากำลังจะประเมินอะไร และจะต้องกำหนดจุดมุ่งหมายของการประเมินให้ชัดเจน จากนั้นจึงกำหนดเทคนิควิธีการประเมิน สร้างเครื่องมือในการเก็บรวบรวมข้อมูลให้มีความคลาดเคลื่อนน้อยที่สุด และใช้เทคนิคการวิเคราะห์ข้อมูลนำผลการวัดประเมินเปรียบเทียบกับมาตรฐานบางอย่างตามจุดมุ่งหมายของเรื่องนั้น ๆ เป็นเกณฑ์การประเมินผลการศึกษา

สำหรับการประเมินผลการศึกษาของไทย ส่วนใหญ่แล้วจะประเมินใน 2 ด้าน คือ การประเมินผลแผนพัฒนาการศึกษา กับการประเมินผลการศึกษาเกี่ยวกับการสอน เช่นการสอบวัดผลสัมฤทธิ์ การสอบเลื่อนชั้น การสอบเทียบวุฒิ การสอบคัดเลือก เป็นต้น ในที่นี้จะกล่าวเฉพาะการประเมินผลการศึกษาที่เกี่ยวข้องกับการเรียนการสอน

2. รูปแบบการประเมินผลทางการศึกษา

การประเมินผลทางการศึกษา ถ้าแบ่งตามการใช้ผลการประเมินสามารถแบ่งออกได้เป็น 4 ประเภท (ชยุตม์ ภิรมย์สมบัติ, 2556, หน้า 28-29) ได้แก่

2.1 การประเมินระหว่างเรียน (Formative assessment) เป็นการประเมินที่ทำในระหว่างการเรียนการสอน เพื่อตรวจสอบว่าผู้เรียนมีความรู้ความเข้าใจหรือไม่ จะทำการวัดและประเมินขณะที่การเรียนการสอนกำลังดำเนินการอยู่ หากไม่เข้าใจจะทำการปรับปรุงแก้ไขได้ทันที

2.2 การประเมินระหว่างทาง (Interim assessment) เป็นการประเมินภายหลังจากการเรียนการสอนดำเนินไปแล้วระยะเวลาหนึ่ง เช่นอาจจะประเมินหลังจากเรียนจบเป็นบท ๆ ไป หรือจะประเมินครั้งละ 3 บท หรือจะประเมินหลังจากเรียนไปแล้วทุก ๆ 1 เดือน เป็นต้น

2.3 การประเมินสรุปผล (Summative assessment) เป็นการประเมินหลังจากเสร็จสิ้นการเรียนการสอนปลายภาคเรียน หรือจบการเรียนในรายวิชานั้น ๆ

2.4 การประเมินระดับชาติ (National assessment) เป็นการประเมินการจัดการศึกษาในภาพรวมทั้งระบบของประเทศ ที่ดำเนินการโดยหน่วยงานกลางภายนอก

สำหรับการวิจัยครั้งนี้จะใช้ผลการประเมินสรุปผล (Summative assessment) ปลายภาคเรียนในการปรับเทียบคะแนน

3. การตัดสินผล เป็นการตัดสินค่าที่ได้จากการวัดประเมินซึ่งผู้ตัดสินจะต้องเลือกเกณฑ์ที่เหมาะสมเป้าประสงค์หรือบริบทของแต่ละสถาบัน โดยทั่วไปเกณฑ์ในการประเมินผล แบ่งออกเป็น 3 ประเภท คือ การประเมินแบบอิงเกณฑ์ (Criterion-referenced evaluation) ซึ่งเป็นการตัดสินผลโดยใช้เกณฑ์ตามที่กำหนดไว้ล่วงหน้าตามวัตถุประสงค์การเรียนรู้ที่กำหนด ผลการเรียนรู้ที่ประเมินของแต่ละคนจะเทียบกับเกณฑ์ที่กำหนดไม่ขึ้นกับบุคคลอื่น ส่วนการประเมินแบบอิงกลุ่ม (Norm-referenced evaluation) เป็นการตัดสินผลการเรียนรู้โดยการเปรียบเทียบกับบุคคลอื่นในกลุ่มเดียวกัน ผลการตัดสินจึงขึ้นอยู่กับระดับความสามารถของบุคคลอื่นในกลุ่มเดียวกัน เป็นการจัดจำแนกความสามารถบุคคลเปรียบเทียบกับกลุ่มว่าอยู่ระดับใดเมื่อเทียบกับกลุ่ม สุดท้ายเป็นการประเมินแบบอิงตนเอง (Self-referenced evaluation) เป็นการตัดสินโดยการเปรียบเทียบกับตนเอง ว่ามีความรอบรู้เป็นอย่างไรจากระดับเดิม เป็นการตัดสินเพื่อดูพัฒนาการของแต่ละบุคคล ไม่เน้นการเปรียบเทียบกับบุคคลอื่น (สุวิมล ว่องวาณิช, 2550, หน้า 56-57)

การตัดสินผลการเรียนรู้ เป็นการประกันว่า ผู้เรียนมีผลการเรียนรู้ หรือมีศักยภาพเป็นอย่างไร บรรลุผลตามจุดมุ่งหมายของการวัดหรือไม่ ซึ่งในการตัดสินผลนิยมให้เป็นเกรด มีทั้งการกำหนดเป็นตัวอักษร เช่น A, B, C, D, E หรือกำหนดเป็นตัวเลข เช่น 4, 3, 2, 1, 0 หรือจัดแยกเป็นประเภท เช่น ดีเยี่ยม ดีมาก ดี ปานกลาง พอใช้ ควรปรับปรุง

สำหรับการวิจัยครั้งนี้จะใช้การตัดสินผลแบบอิงเกณฑ์ โดยให้เป็นเกรด 2 รูปแบบ ตามที่สถานศึกษากำหนด และตามงานวิจัยที่คาดว่าจะนำไปใช้ในการปรับปรุงระบบการตัดเกรดแบบใหม่ ดังนี้

การตัดเกรด 3 เกรด ตามที่สถานศึกษากำหนด กำหนดระดับและเกณฑ์คะแนน ดังนี้

เกรด	ระดับ	คะแนน
H	4.00	มากกว่า 75%
S	2.33	60-75%
U	0	ต่ำกว่า 60%

การตัดเกรด 8 เกรด ตามงานวิจัยที่คาดว่าจะนำไปใช้ในการปรับปรุงระบบการตัดเกรดแบบใหม่กำหนดระดับ และเกณฑ์คะแนน ดังนี้

เกรด	ระดับ	คะแนน
A	4.00	76-100%
B^+	3.50	70-75%
B	3.00	65-69%
C^+	2.50	60-64%
C	2.00	55-59%
D^+	1.50	50-54%
D	1.00	45-49%
F	0.00	ต่ำกว่า 45%

4. การนำการตัดสินผลไปใช้ประโยชน์ การนำผลการตัดสินผลการเรียนรู้ของผู้เรียนมีประโยชน์ทั้งต่อตัวผู้เรียน ผู้สอน ผู้ปกครอง สถาบันการศึกษา และต่อประเทศชาติในประเด็นต่าง ๆ ดังนี้

4.1 ด้านผู้เรียน นำผลการตัดสินมาเป็นข้อมูลในการปรับปรุง ซ่อมเสริม หรือส่งเสริมให้ประสบความสำเร็จตามศักยภาพที่มีอยู่ นำไปใช้ในการวางแผนการเรียน การศึกษาต่อ หรือการประกอบอาชีพ

4.2 ด้านผู้สอน ผลการเรียนรู้ของผู้เรียนจะเป็นข้อมูลส่วนหนึ่งในการสะท้อนด้านการสอนของตนเอง ซึ่งผู้สอนสามารถนำ ๆ ไปปรับปรุงและพัฒนาการสอนให้มีประสิทธิภาพยิ่งขึ้น นำส่งให้ฝ่ายแนะแนวใช้ในการให้คำปรึกษา

4.3 ด้านผู้ปกครอง ผลการเรียนรู้ของบุตร หลานในความปกครองเป็นตัวสะท้อนระดับความสามารถของผู้ที่อยู่ในความปกครอง ซึ่งผู้ปกคองสามารถนำไปใช้ในการให้ความช่วยเหลือสนับสนุนส่งเสริมให้ประสบความสำเร็จสูงสุดตามศักยภาพ

4.4 ด้านสถาบันการศึกษา และต่อประเทศชาติ นำผลการเรียนรู้ในภาพรวมไปใช้ในการกำหนดเป็นนโยบายด้านการศึกษา วางแผนกำหนดแนวทางเพื่อพัฒนาด้านการศึกษา

ตอนที่ 2 แนวคิดทฤษฎีของการปรับเทียบคะแนนระหว่างแบบสอบ

ความหมายของการปรับเทียบคะแนน

ลอร์ด (Lord, 1980, p. 195) ให้ความหมายของการปรับเทียบคะแนนระหว่างแบบสอบว่า เป็นการแปลงคะแนนจากแบบสอบต่างฉบับ ให้มีความหมายสับเปลี่ยนกันได้เพื่อความเสมอภาคของบุคคล

แองกอฟ (Angoff, 1984, p. 85) ให้ความหมายของการปรับเทียบคะแนนระหว่างคะแนนของแบบสอบว่า หมายถึง กระบวนการแปลงระบบคะแนนของแบบสอบฉบับหนึ่งไปสู่ระบบคะแนนของแบบสอบอีกฉบับหนึ่งซึ่งวัดสิ่งเดียวกัน เพื่อให้คะแนนแปลงจากแบบสอบทั้ง 2 ฉบับนั้น สามารถเทียบเท่ากันและใช้แทนกันได้

ปีเตอร์เซน, โคเลน และฮูเว่อร์ (Petersen, Kolen & Hoover, 1989, p. 242) ให้ความหมายของการปรับเทียบคะแนนระหว่างแบบสอบว่า เป็นกระบวนการเชิงประจักษ์ ที่ใช้ในการแปลงคะแนนจากแบบสอบฉบับหนึ่งไปยังแบบสอบอีกฉบับหนึ่ง โดยที่แบบสอบทั้ง 2 ฉบับนั้นต้องวัดคุณลักษณะเดียวกัน

โคเลน และเบรนเนน (Kolen & Brennan, 2004, p. 2) ให้ความหมายของการปรับเทียบคะแนนว่าเป็นกระบวนการทางสถิติที่ใช้ในการปรับคะแนนจากแบบสอบต่างฉบับเพื่อให้คะแนนจากแบบสอบต่างฉบับสามารถเทียบแทนกันได้

จากความหมายดังกล่าว สรุปได้ว่า การปรับเทียบคะแนนเป็นกระบวนการในการแปลงคะแนนจากแบบสอบฉบับหนึ่งไปยังแบบสอบอีกฉบับหนึ่ง ที่วัดในคุณลักษณะเดียวกันให้เป็นคะแนนที่สมมูล เพื่อทำให้คะแนนนั้นสามารถนำมาเปรียบเทียบกันได้โดยตรง

เงื่อนไขของการปรับเทียบคะแนน

เนื่องจากการปรับเทียบคะแนนระหว่างแบบสอบ มีวัตถุประสงค์เพื่อที่จะทำให้คะแนนจากแบบสอบต่างฉบับมีความเท่าเทียมกันมากที่สุด และวิธีการปรับเทียบคะแนนจะต้องมี

การวางแผนในการเก็บรวบรวมข้อมูล มีเกณฑ์ในการแปลงคะแนนจากแบบสอบฉบับหนึ่งไปยังแบบสอบอีกฉบับหนึ่ง จึงต้องกำหนดเงื่อนไขบางประการในการปรับเทียบซึ่งเลวิน (Lavin, 1955 cited in Holland & Lubin, 1982) กล่าวว่า แบบสอบที่จะนำมาปรับเทียบคะแนนนอกจากจะวัดคุณลักษณะเดียวกันแล้วจะต้องมีความเที่ยงสูง และเป็นแบบสอบที่คู่ขนานกันในด้านโครงสร้าง (Structure) เวลาที่ใช้สอบ (Timing) ชนิดของข้อสอบ (Item types) รูปแบบ (Formats) และเนื้อหา (Subject matter) ซึ่งข้อสอบแต่ละฉบับอาจมีความยากแตกต่างกันแต่สามารถใช้เทคนิคการปรับเทียบคะแนนระหว่างแบบสอบ ปรับคะแนนจากแบบสอบต่างฉบับให้มีความเท่าเทียมกัน

นอกจากนี้ ลอร์ด (Load, 1980 อ้างถึงใน พิชัย ละแมนชัย, 2538, หน้า 16) ได้กล่าวถึงเงื่อนไขในการปรับเทียบคะแนนว่าแบบสอบทั้งสองจะต้องมีคุณสมบัติ 4 ประการ คือ

1. แบบสอบทั้งสองฉบับจะต้องวัดความสามารถเดียวกัน (Same ability) คือแบบสอบทั้งสองวัดในคุณลักษณะเดียวกัน คุณลักษณะนี้อาจเป็นคุณลักษณะแฝงหรือความสามารถหรือทักษะอย่างใดอย่างหนึ่งก็ได้

2. มีความเสมอภาค (Equity) คือ เมื่อทุกกลุ่มมีความสามารถเดียวกันการแจกแจงคะแนนของแบบสอบที่ได้หลังจากที่มีการแปลงคะแนนแล้ว จะมีการแจกแจงเหมือนกับการแจกแจงของคะแนนจากแบบสอบที่ใช้เป็นฉบับเทียบคะแนน

3. ความไม่แปรเปลี่ยนตามกลุ่ม (Invariance across groups) คือ คะแนนที่ได้จากการแปลงคะแนนไม่ว่าจะมาจากกลุ่มใด ๆ ก็ตามจะมีค่าเท่าเทียมกันหรือมีค่าไม่เปลี่ยนแปลงไปตามกลุ่มของผู้เข้าสอบ

4. มีความสมมาตร (Symmetry) คือ ผลของการปรับเทียบคะแนนจะต้องเหมือนกันไม่ว่าจะเทียบจากแบบสอบฉบับ X ไปยังฉบับ Y หรือเทียบจากแบบสอบฉบับ Y ไปยังฉบับ X

ประเภทของการปรับเทียบคะแนน

การปรับเทียบคะแนนระหว่างแบบสอบ แบ่งออกได้เป็น 2 ประเภท ดังนี้ (ศิริชัย กาญจนวาสี, 2555, หน้า 160)

1. การปรับเทียบคะแนนตามแนวนอน (Horizontal equating) เป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างฉบับ ที่วัดคุณลักษณะเดียวกัน มีระดับความยากใกล้เคียงกัน เป็นเทคนิคที่เหมาะสมในสถานการณ์ที่มีความจำเป็นจะต้องสร้างแบบสอบที่มีเนื้อหาเดียวกันหลาย ๆ ฉบับ เพื่อนำไปใช้ในการสอบเพื่อให้เกิดความยุติธรรม และป้องกันความลับของข้อสอบ เมื่อใช้ต่างเวลากันเพื่อปรับเทียบคะแนนที่ได้จากฉบับหนึ่งเทียบเป็นเท่าไรของอีกฉบับหนึ่ง ที่วัดในระดับเดียวกัน จึงเป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างฉบับของวิชาเดียวกัน สำหรับกลุ่มผู้สอบระดับชั้นเดียวกัน

2. การปรับเทียบคะแนนตามแนวตั้ง (Vertical equating) เป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างฉบับ เมื่อแต่ละฉบับมุ่งวัดลักษณะเดียวกัน แต่มีระดับความยากแตกต่างกัน และกลุ่มผู้สอบมีการแจกแจงความสามารถอยู่ต่างประชากรกัน หรือมีความสามารถแตกต่างกัน เป็นเทคนิคที่เหมาะสมในสถานการณ์ที่มีความจำเป็นต้องสร้างแบบสอบเนื้อหาเดียวกัน แต่ต่างฉบับก็มุ่งวัดความสามารถของผู้สอบที่ต่างระดับกัน เพื่อปรับเทียบว่าคะแนนที่สอบได้จากฉบับหนึ่งเทียบเป็นเท่าไรของฉบับอื่นที่วัดต่างระดับกัน จึงเป็นการปรับเทียบคะแนนระหว่างแบบสอบต่างระดับของวิชาเดียวกัน สำหรับกลุ่มผู้สอบต่างระดับชั้นกัน

สำหรับการวิจัยครั้งนี้จะเลือกใช้วิธีการปรับเทียบคะแนนตามแนวนอน เนื่องจากเป็นการปรับเทียบของชุดวิชาเดียวกันและอยู่ในระดับชั้นเดียวกัน

รูปแบบวิธีการปรับเทียบคะแนน (Designs for equating method) การปรับเทียบคะแนนจำแนกตามแนวคิดพื้นฐานได้ 3 รูปแบบ คือ

1. การปรับเทียบคะแนนตามแนวทฤษฎีการวัดแบบมาตรฐานเดิม (Classical test theory)

2. การปรับเทียบคะแนนแนวทฤษฎีการตอบข้อสอบ (Item response theory)

3. การปรับเทียบคะแนนวิธีเชิงเส้นตรงตามแบบจำลองคะแนนจริงสัมพันธ์ ซึ่งวิธีการปรับเทียบคะแนนแต่ละวิธีมีรายละเอียด ดังนี้

4. วิธีการปรับเทียบคะแนนตามแนวทฤษฎีการวัดแบบดั้งเดิม (Classical test theory) ที่นิยมใช้กันมี 4 วิธี คือ

4.1 วิธีการปรับเทียบคะแนนแบบอิควิเปอร์เซ็นไทล์ (Equipercentile equating) เป็นวิธีการปรับเทียบที่ยึดหลักการว่า การแจกแจงของคะแนนจากแบบสอบ X และแบบสอบ Y มีลักษณะคล้ายกัน การเทียบหาคะแนนสมมูลทำได้โดยใช้คะแนน ณ ตำแหน่งเปอร์เซ็นไทล์เดียวกันของคะแนน 2 ชุดนั้น ผลการปรับเทียบคะแนนแสดงได้ด้วยกราฟ ถ้าหากความยากของแบบสอบทั้ง 2 ชุดนั้นใกล้เคียงกัน เส้นกราฟจะใกล้เคียงกับเส้นตรง แต่ถ้าหากแบบสอบ 2 ชุดนั้นมีความยากแตกต่างกัน เส้นกราฟจะเป็นเส้นโค้ง การนำวิธีการปรับเทียบวิธีนี้ไปใช้มีข้อแนะนำ ดังนี้

4.1.1 ควรใช้กลุ่มผู้สอบขนาดใหญ่ ควรเป็นกลุ่มที่มีความสามารถค่อนข้างกระจายและกระจายพอ ๆ กัน ถ้าใช้กลุ่มตัวอย่างขนาดเล็กจะทำให้คะแนนมีความไวต่อความแปรปรวนเชิงสุ่ม ผลการปรับเทียบคะแนนสามารถแปรผันไปตามกลุ่มผู้สอบได้

4.1.2 การสร้างกราฟเปอร์เซ็นไทล์เพื่อปรับเทียบคะแนนระหว่างแบบสอบควรสร้างด้วยความระมัดระวังและไม่ลำเอียง

4.1.3 แบบสอบที่นำมาปรับเทียบคะแนนกัน ควรมีความเที่ยงใกล้เคียงกัน ถ้าต่างกันมากจะทำให้การปรับเทียบขาดความคงที่

4.1.4 การปรับเทียบคะแนนควรอยู่ในช่วงพิสัยของคะแนนสังเกต การปรับเทียบคะแนนที่อยู่นอกพิสัยของคะแนนสังเกตจะมีความคลาดเคลื่อนสูง (Angoff, 1984)

4.2 วิธีปรับเทียบคะแนนเชิงเส้น (Linear equating) เป็นวิธีการใช้สมการเส้นตรงในการแปลงคะแนน โดยมีหลักการว่าคะแนนจากแบบสอบทั้ง 2 ฉบับ จะมีความเท่าเทียมกันก็ต่อเมื่อคะแนนของแต่ละฉบับมีคะแนนมาตรฐาน (Z-score) เท่ากัน วิธีปรับเทียบคะแนนเชิงเส้นตรง อาจกล่าวได้ว่าเป็นกรณีเฉพาะของวิธีปรับเทียบแบบอิควิเปอร์เซ็นไทล์ เนื่องจากว่าถ้าการแจกแจงคะแนนจากแบบสอบ X และ Y เหมือนกัน วิธีอิควิเปอร์เซ็นไทล์จะให้ผลการปรับเทียบคะแนนเหมือนกันกับวิธีปรับเทียบคะแนนเชิงเส้นตรง การนำวิธีปรับเทียบคะแนนวิธีนี้ไปใช้มีข้อแนะนำ ดังนี้

4.2.1 ควรตรวจสอบความสอดคล้องของการแจกแจงก่อนปรับเทียบคะแนนและหลังการปรับเทียบคะแนนว่าเป็นความสัมพันธ์เชิงเส้นตรง

4.2.2 ถ้าความยากของแบบสอบที่นำมาปรับเทียบคะแนนกันมีความแตกต่างกัน การแจกแจงคะแนนอาจมีความสัมพันธ์เชิงเส้นโค้ง การใช้วิธีการปรับเทียบคะแนนเชิงเส้นตรงจะไม่เหมาะสม ควรเลือกใช้วิธีปรับเทียบอิควิเปอร์เซ็นไทล์จะได้ผลดีกว่า

4.2.3 วิธีปรับเทียบคะแนนโดยใช้สมการถดถอย (Regression equating) เป็นการใช้สมการเส้นตรงในการทำนายตัวแปรตาม จากตัวแปรอิสระซึ่งมีลักษณะไม่สมมาตร กล่าวคือ สมการทำนาย Y จาก X หรือการทำนาย X จาก Y เป็นสมการที่ไม่สมมาตร หรือให้ผลการทำนายไปในทิศทางเดียวกัน นอกจากนี้คะแนนจากแบบสอบที่ใช้เป็นตัวทำนายยังมีข้อตกลงเบื้องต้นว่าจะต้องมีค่าความเที่ยงเป็น 1 การนำวิธีปรับเทียบคะแนนวิธีนี้ไปใช้มีข้อแนะนำ ดังนี้

4.2.3.1 แบบสอบที่จะนำมาปรับเทียบต้องเป็นแบบสอบที่วัดคุณลักษณะเดียวกัน มีความเป็นคู่ขนานและมีความเที่ยงสูง

4.2.3.2 แบบสอบที่นำมาปรับเทียบคะแนนกันต้องมีความสัมพันธ์กับคะแนนเกณฑ์เท่าเทียมกัน มิเช่นนั้นแล้วจะทำให้คะแนนจากแบบสอบฉบับหนึ่งจะสามารถทำนายคะแนนเกณฑ์ได้แม่นยำกว่าอีกฉบับหนึ่ง ทำให้การปรับเทียบคะแนนมีความผันแปรตามกลุ่มที่ศึกษา

4.3 วิธีเคอเนล (KE: Kernel equating) เป็นวิธีการปรับเทียบที่นำเสนอโดย Holland and Thayer ใน ค.ศ. 1989 ต่อมา Von Davier, Holland and Thayer ได้พัฒนาใหม่ใน ค.ศ. 2004 และให้ข้อสังเกตว่า KE จะใช้กลุ่มตัวอย่างที่น้อยกว่าวิธีอื่น ๆ และให้ค่าความคลาดเคลื่อนในการปรับเทียบน้อย นอกจากนี้ยังให้ค่าฟังก์ชันการปรับเทียบ และความคลาดเคลื่อนที่สอดคล้องกันในทุกค่าของการออกแบบเก็บรวบรวมข้อมูล การปรับเทียบด้วยอิควิเปอร์เซ็นต์ไทล์จะใช้วิธีการ

เชิงเส้นในการปรับการแจกแจงให้ต่อเนื่อง แต่เคอเนลจะใช้วิธี Gaussian ซึ่งมีเงื่อนไขต่ำกว่าแทน กระบวนการในการปรับเทียบด้วยเคอเนล ประกอบด้วย 5 ขั้นตอน คือ

4.3.1 ขั้นการปรับเรียบ (Pre-smoothing) โดยใช้ loglinear model ในการปรับเรียบเพื่อดูการกระจายของคะแนนและสร้างเมทริกซ์หลังจากคำนวณความคลาดเคลื่อนมาตรฐานในการปรับเทียบ

4.3.2 ประมาณคะแนนความน่าจะเป็น (Estimating the score probabilities) จากคะแนนการแจกแจงที่ปรับเรียบของแบบสอบ X และ Y

4.3.3 ทำคะแนนให้ต่อเนื่อง (Continuization) จากคะแนนการแจกแจงที่ไม่ต่อเนื่อง

4.3.4 ทำการปรับเทียบ (Equating) สร้างฟังก์ชันการปรับเทียบจาก X ไป Y

4.3.5 หาความคลาดเคลื่อนของการปรับเทียบ (Standard error of equating: SEE)

สำหรับเทคนิควิธีที่ใช้ในการปรับเทียบเคอเนล กรณีออกแบบใช้กลุ่มที่ไม่เท่าเทียมกัน โดยใช้ข้อสอบร่วม (NEAT: Non-equivalent groups anchor test) มี 2 วิธี คือ วิธี Chain equating (CE) และวิธี Post-stratification equating (PSE) ซึ่งวิธีการทั้งสองให้ผลไม่แตกต่างกัน (Rebecca & Dvorak, 2009, p. 41)

1. การปรับเทียบคะแนนตามแนวทฤษฎีการตอบข้อสอบ (Item response theory)

การปรับเทียบคะแนนตามแนวทฤษฎีการตอบข้อสอบ Lord (1980) ได้กล่าวไว้ว่า แบบสอบ 2 ฉบับใด ๆ ที่เทียบคะแนนกันนั้น ต้องเป็นแบบสอบที่มีมิติเดียวกัน คือ มีการวัดคุณลักษณะเดียวหรือวัดความสามารถเดียว ข้อตกลงเกี่ยวกับความเป็นมิติเดียวเป็นเรื่องที่ซับซ้อนและยุ่งยากมาก เนื่องจากมีปัจจัยที่มีผลต่อคะแนนสอบ เช่น ปัจจัยด้านความรู้ความเข้าใจ บุคลิกภาพ ปัจจัยด้านการจัดการสอบ ความวิตกกังวล ความสามารถในการทำงานได้รวดเร็ว ฯลฯ เมื่อเป็นเช่นนี้การที่จะทำให้แบบสอบเป็นไปตามข้อตกลงนี้จึงทำได้ยาก การทดสอบความเป็นมิติเดียวจะใช้วิธีการวิเคราะห์ตัวประกอบวัดจากค่าไอเกน (Eigen value) การปรับเทียบวิธีนี้มีข้อตกลงที่สำคัญ 3 ประการ คือ

1.1 ความเสมอภาค (Equity) คือ ถ้าพิจารณาที่ระดับความสามารถ (θ) ใด ๆ การแจกแจงความถี่อย่างมีเงื่อนไขของคะแนนแปลงX(Y) หรือ Y (คะแนนจากแบบสอบฉบับ X ที่แปลงมาสู่สเกลเดียวกับแบบสอบ Y) ที่ θ ที่กำหนดให้ด้วยการเทียบคะแนนต้องเหมือนกับการแจกแจงความถี่อย่างมีเงื่อนไขของคะแนนจากแบบสอบที่ต้องการปรับเทียบ (X)

1.2 ความไม่ผันแปรตามกลุ่ม (Invariance across of groups) คือคะแนนแปลงจะคงที่โดยไม่แปรเปลี่ยนไปตามประชากรที่นำมาสร้างสมการเทียบคะแนน

1.3 ความสมมาตร (Symmetry) คือ คะแนนจากการปรับเทียบคะแนนนั้นต้องเหมือนกันไม่ว่าการเทียบนั้นจะเทียบจากแบบสอบฉบับ X ไปยังแบบสอบฉบับ Y หรือจากแบบสอบฉบับ Y ไปยังแบบสอบฉบับ X

การปรับเทียบคะแนนโดยใช้ทฤษฎีการตอบข้อสอบ มีการเทียบคะแนนอยู่ 2 รูปแบบ คือ การใช้คะแนนจริงและคะแนนสังเกต การเทียบด้วยคะแนนจริงไม่สามารถอธิบายคะแนนที่อยู่ต่ำกว่าระดับการเดาได้ โดยวิธีนี้จะอธิบายคะแนนสมมูลเฉพาะคะแนนที่อยู่เหนือค่าเฉลี่ยของการเดา ถึงแม้ว่าวิธีนี้จะเป็นการเทียบโดยใช้คะแนนจริง แต่ก็ยังคงเป็นคะแนนจริงที่ได้จากการคำนวณความคลาดเคลื่อนยังมีอยู่ ส่วนการเทียบโดยใช้คะแนนสังเกตเป็นการเทียบคะแนนโดยประมาณที่สามารถอธิบายคะแนนสมมูลจาก X และ Y ทั้งสองวิธีนี้มีความสอดคล้องกันมาก แต่การใช้คะแนนสังเกตมีความยุ่งยากซับซ้อนมากกว่าการใช้คะแนนจริง (Lord & Winggerskey, 1984; พรพิมล นาคเวช, 2535 อ้างถึงใน อุทัยวรรณ พงศ์อร่าม, 2545, หน้า 30)

ทฤษฎีการตอบข้อสอบ (Item response theory) เป็นทฤษฎีทางการวัดผลที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในบุคคลกับพฤติกรรมการตอบสนองข้อสอบแต่ละข้อว่ามีความน่าจะเป็นในการตอบถูกเพียงใด ทฤษฎีนี้ตั้งอยู่บนหลักพื้นฐานที่สำคัญ 2 ประการ (Hambletor, Swarminathan & Roger, 1991 อ้างอิงใน อดิศร ศรีบุญวงษ์, 2545 หน้า 28-29) คือ

1. ความสามารถของบุคคลในการตอบข้อสอบได้ถูกหรือผิด สามารถอธิบายได้ด้วยคุณลักษณะภายในหรือความสามารถ (Latent trait or ability) ของบุคคลนั้น ๆ

2. ความสัมพันธ์ระหว่างความสามารถในการตอบข้อสอบได้ถูกต้องกับความสามารถของผู้สอบที่วัดจากแบบสอบ สามารถอธิบายได้ด้วยโค้งลักษณะข้อสอบ (Item characteristic curve: ICC)

พัฒนาการของทฤษฎีการตอบข้อสอบ เริ่มจากราส์ซ (Rasch) นักคณิตศาสตร์ชาวเดนมาร์ก ได้เสนอโมเดลราส์ซ (Rasch model แบบ 1 พารามิเตอร์ โดยมีแนวคิดว่าค่าความยากของข้อสอบเป็นสิ่งเดียวที่มีอิทธิพลต่อการตอบสนองข้อสอบ โดยกำหนดโมเดลเป็นฟังก์ชันทางคณิตศาสตร์ ดังสมการ

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

ต่อมาปี ค.ศ. 1952 Lord ได้เสนอฟังก์ชันนอร์มัลโอไจฟ์ แบบ 2 พารามิเตอร์ โดยเพิ่มพารามิเตอร์ อำนาจจำแนก ว่าเป็นสิ่งที่มีอิทธิพลต่อการตอบสนองข้อสอบ โดยกำหนดโมเดลเป็นฟังก์ชันทางคณิตศาสตร์ ดังสมการ

$$P_i(\theta) = \frac{e^{Da\ (\theta - b_i)}}{1 + e^{Da\ (\theta - b_i)}}$$

และในปี ค.ศ. 1974 Lord ได้เสนอโมเดลโลจิสติก 3 พารามิเตอร์ โดยเพิ่มพารามิเตอร์โอกาสการเดาข้อสอบ โดยกำหนดโมเดลเป็นฟังก์ชันทางคณิตศาสตร์ ดังสมการ

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da(\theta-b_i)}}{1+e^{Da(\theta-b_i)}}$$

ปัจจุบันทฤษฎีการตอบข้อสอบได้พัฒนาเพิ่มมากขึ้น โดยบาร์ตัน และลอร์ด (Barton and Lord) ได้เสนอโมเดลโลจิสติกแบบ 4 พารามิเตอร์ขึ้น โดยเขาเชื่อว่านักเรียนที่มีความสามารถสูงมากไม่จำเป็นที่จะต้องตอบข้อสอบถูกเสมอไป ขึ้นอยู่กับความรอบคอบหรือความระมัดระวังในการทำข้อสอบ (อดิศร บุญวงศ์, 2545, หน้า 29)

การปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ ลอร์ดได้แบ่งเป็น 2 วิธีใหญ่ ๆ คือ การปรับเทียบคะแนนโดยใช้คะแนนจริง (True-score equating) กับใช้คะแนนสังเกต (Observed score equating) โดยที่การปรับเทียบคะแนนโดยใช้คะแนนจริงหาได้จากสมการ

$$\xi = \xi(\theta) = \sum_{i=1}^{n_x} P_i(\theta) \text{ และ}$$

$$\eta = \eta(\theta) = \sum_{j=1}^{n_y} P_j(\theta) \text{ เมื่อ}$$

ξ, η คือ คะแนนจริงจากแบบสอบฟอร์ม X และฟอร์ม Y ตามลำดับ

ส่วนการปรับเทียบคะแนนโดยใช้คะแนนสังเกต จะเริ่มด้วยการประมาณความสามารถของกลุ่มรวม ($r(\theta)$) ของผู้เข้าสอบทั้งหมดที่ทำแบบสอบร่วม การประมาณการกระจายของคะแนนสังเกต X หาได้จากสมการ

$$\hat{\phi}_X(X) = \frac{1}{N}\sum_{a=1}^{N}\hat{\phi}_X\left(X\middle|\hat{\phi}\right) \text{ เมื่อ a = 1, 2, 3, …, N เป็นผู้สอบในกลุ่มรวม}$$

โดยที่ $\hat{\phi}_x(X) = \int_{-\alpha}^{\alpha}\hat{\phi}_x\left(X\middle|\theta\right)\hat{r}(\theta)d\theta$

เนื่องจากแบบสอบ X และ Y เป็นอิสระจากกัน เมื่อกำหนดความสามารถ (θ) คงที่จะสามารถประมาณการกระจายร่วมของความสามารถจากคะแนนแบบสอบ X และ Y ได้จากสมการ

$$\hat{\phi}(X,Y) = \frac{1}{N}\sum_{a=1}^{N}\hat{\phi}_x\left(X\middle|\hat{\theta}_a\right)\hat{\phi}_y\left(Y\middle|\hat{\theta}_a\right) \text{ หรือ}$$

$$\hat{\phi}(X,Y) = \int_{-\alpha}^{\alpha}\hat{\theta}_x(X|\theta)\hat{\phi}_y(Y|\theta)\hat{r}(\theta)d\theta$$

ลอร์ด กล่าวว่า การปรับเทียบคะแนนทั้ง 2 วิธีนี้มีความสอดคล้องกันมาก แต่การสรุปผลเพื่อนำไปอ้างอิงจะต้องทำอย่างพิถีพิถัน การปรับเทียบโดยใช้คะแนนสังเกตจะมีสูตรที่ยุ่งยากซับซ้อนงานวิจัยในการปรับเทียบโดยใช้คะแนนสังเกตจึงมีน้อยมาก (Lord, 1980, pp. 202-203)

ขั้นตอนการปรับเทียบคะแนนโดยใช้ทฤษฎี IRT

แบบสอบที่จะนำมาปรับเทียบถ้าวิเคราะห์และคำนวณค่าพารามิเตอร์ของข้อสอบแล้ว สามารถนำผลการตอบข้อสอบไปใช้วิธีการ Maximum likelihood ประมาณค่าความสามารถของผู้สอบ (θ) ซึ่งค่า θ จะเข้าสู่ค่าที่แท้จริงเมื่อมีจำนวนข้อสอบเพิ่มมากขึ้น และค่า θ อยู่บนสเกลร่วมกันอยู่แล้ว จึงปรับเทียบกันได้โดยตรงไม่จำเป็นต้องปรับเทียบคะแนน แต่กรณีที่แบบสอบต่างฉบับกันนั้นไม่ทราบค่าพารามิเตอร์ จะต้องใช้วิธีประมาณค่าร่วมกันไปทั้งค่าพารามิเตอร์ของข้อสอบและค่าความสามารถของผู้สอบ θ การประมาณค่าพารามิเตอร์ของข้อสอบแตกต่างกันได้ตามการกำหนดค่าเมตริกซ์ θ ของผู้สอบ ดังนั้น จึงมีความจำเป็นต้องใช้วิธีการปรับเทียบคะแนนระหว่างแบบสอบ โดยมีขั้นตอนดังนี้

1. เลือกรูปแบบการปรับเทียบ ซึ่งขึ้นอยู่กับธรรมชาติของแบบสอบและกลุ่มผู้สอบ
2. กำหนดโมเดลการตอบสนองข้อสอบ ซึ่งรูปแบบที่นิยมมี 3 โมเดล คือ โมเดล 1 2 และ 3 พารามิเตอร์ กรณีกลุ่มตัวอย่างขนาดไม่ใหญ่นัก ควรเลือกใช้โมเดล 1 พารามิเตอร์ แต่ถ้ากลุ่มตัวอย่างขนาดใหญ่ควรใช้โมเดล 2 หรือ 3 พารามิเตอร์
3. สร้างเมตริกซ์สเกลร่วมของความสามารถและพารามิเตอร์ข้อสอบ
4. กำหนดสเกลสำหรับรายงานคะแนนสอบที่ปรับเทียบแล้ว ดังนี้

ถ้าต้องการรายงานผลปรับเทียบด้วยคะแนนความสามารถ θ จากสมการการปรับเทียบ θ เมื่อคำนวณผลตามสมการก็สามารถสร้างตารางรายงานผลเปรียบเทียบได้ทันที

ถ้าต้องการรายงานผลการปรับเทียบด้วยคะแนนจริง สามารถคำนวณคะแนนจริงที่สมมูลกันตรงตำแหน่ง θ ต่าง ๆ ได้ เพื่อนำมาสร้างตารางปรับเทียบ หรือพล๊อตเป็นกราฟก็จะได้ค่าคะแนนที่สมมูลกัน

ถ้าต้องการรายงานผลการปรับเทียบด้วยคะแนนดิบ หรือคะแนนที่สังเกตได้จะต้องคำนวณความถี่ของการแจกแจงคะแนนดิบอย่างมีเงื่อนไขตามทฤษฎีจากกลุ่มตัวอย่าง คำนวณความถี่รวมของการแจกแจงคะแนนดิบตามทฤษฎีของแต่ละแบบสอบ คำนวณค่าเปอร์เซ็นไทล์ของคะแนนจากแต่ละแบบสอบ และปรับเทียบโดยหลักของอิควิเปอร์เซ็นไทล์ สุดท้ายสร้างตารางปรับเทียบหรือพล๊อตกราฟที่ปรับเทียบค่าคะแนนดิบที่สมมูลกันจากแบบสอบที่ต่างกัน

วิธีการปรับเทียบคะแนนโดยใช้ทฤษฎี IRT สามารถจำแนกตามวิธีการปรับค่าพารามิเตอร์ได้ 2 แนวทาง คือ วิธีปรับค่าพารามิเตอร์พร้อมกัน กับการปรับค่าพารามิเตอร์แยกกัน (Kolen & Brennan, 2004, pp. 166-168) รายละเอียด ดังนี้

แนวทางที่ 1 ใช้วิธีปรับค่าพารามิเตอร์พร้อมกัน (Simultaneous or concurrent calibration) ใช้หลักการปรับค่าพารามิเตอร์ผู้สอบจากแบบสอบต่างฉบับ ให้อยู่ในมาตรวัดความสามารถเดียวกัน ด้วยการนำข้อมูลจากแบบสอบทั้ง 2 ชุด มาต่อกันแล้ววิเคราะห์พร้อมกัน สำหรับข้อสอบอีกชุดที่ผู้สอบต่างกลุ่มต่างไม่ได้ทำ ในการวิเคราะห์ข้อมูลจะใช้คำว่า “Not reach”ในการประมาณค่าสัมประสิทธิ์การเทียบคะแนน สามารถใช้โปรแกรม BILOG-MG และ ICL ในการวิเคราะห์ด้วยการใช้โค้งลักษณะข้อสอบ(Characteristic curve) ในการประมาณค่าซึ่งมีทั้งวิธีที่นำเสนอโดยเฮบารา (Haebara, 1980) และวิธีของสตอกกิงและลอร์ด (Stocking & Lord, 1983) การปรับเทียบคะแนนเมื่อข้อสอบร่วมมีขนาดเล็ก วิธีโค้งลักษณะข้อสอบจะให้ผลดีกว่าวิธีการปรับค่าพารามิเตอร์พร้อมกันและให้ผลที่คล้ายคลึงกันเมื่อใช้จำนวนข้อสอบร่วมที่มากขึ้น แต่วิธีการปรับค่าพารามิเตอร์พร้อมกันจะให้ผลไม่ดี ถ้าผู้สอบมีความสามารถแตกต่างกันมาก (Kim & Cohen, 1998, pp. 141-142)

แนวทางที่ 2 ใช้วิธีปรับค่าพารามิเตอร์แยกกัน (Separate calibration) เป็นการนำข้อมูลที่ได้มาวิเคราะห์แยกกัน การเชื่อมโยงมาตรวัดจะใช้สารสนเทศที่ได้จากข้อสอบร่วมแปลงค่าพารามิเตอร์ความสามารถเชิงเส้น ด้วยสัมประสิทธิ์การเทียบคะแนนหาความชันและจุดตัดเพื่อแปลงมาตรวัดจากแบบสอบฉบับหนึ่งไปสู่แบบสอบอีกฉบับหนึ่ง ซึ่งวิธีการประมาณค่าสัมประสิทธิ์การเทียบคะแนนมีด้วยกันหลายวิธี ดังนี้

1. ถ้าใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์ความยากมาประมาณค่าสัมประสิทธิ์ความชัน และจุดตัดเรียกว่าวิธี Mean and sigma วิธีการปรับคะแนนจากสเกล I ให้อยู่ในสเกล J จะต้องหาค่าความชัน (A) และค่าคงที่ (B) จากสมการ

$$A = \frac{\sigma(b_j)}{\sigma(b_i)}, \qquad B = \mu(b_j) - (A\mu(b_i)$$

เมื่อ $\sigma(b_j)$ เป็นส่วนเบี่ยงเบนมาตรฐานของค่าความยากจากแบบสอบชุด J

$\sigma(b_i)$ เป็นส่วนเบี่ยงเบนมาตรฐานของค่าความยากจากแบบสอบชุด I

$\mu(b_j)$ เป็นค่าเฉลี่ยของค่าความยากจากแบบสอบชุด J

$\mu(b_i)$ เป็นค่าเฉลี่ยของค่าความยากจากแบบสอบชุด I

2. ถ้าใช้ค่าเฉลี่ยของพารามิเตอร์อำนาจจำแนกและพารามิเตอร์ความยากมาประมาณค่าสัมประสิทธิ์ความชัน และจุดตัดเรียกว่าวิธี Mean and mean วิธีการปรับคะแนนจากสเกล I ให้อยู่ในสเกล J จะต้องหาค่าความชัน (A) และค่าคงที่ (B) จากสมการ

$$A = \frac{\mu(a_j)}{\mu(a_i)}, \qquad B = \mu(b_j) - A\mu(b_i)$$

เมื่อ $\mu(a_j)$ เป็นค่าเฉลี่ยของค่าอำนาจจำแนกจากแบบสอบชุด J

$\mu(a_i)$ เป็นค่าเฉลี่ยของค่าอำนาจจำแนกจากแบบสอบชุด I

$\mu(b_j)$ เป็นค่าเฉลี่ยของค่าความยากจากแบบสอบชุด J

$\mu(b_i)$ เป็นค่าเฉลี่ยของค่าความยากจากแบบสอบชุด I

ทั้งสองวิธีนี้ใช้ได้กับการออกแบบการปรับเทียบกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบร่วม (Commom item nonequivalent groups design)

3. วิธีโค้งลักษณะข้อสอบ (Characteristic curve or response function transformation methods) เป็นวิธีการที่ช่วยแก้ปัญหากรณีที่ข้อสอบมีความยากแตกต่างกันมาก แต่กลับให้โค้งลักษณะข้อสอบที่เหมือนกัน ซึ่งจะส่งผลต่อวิธี Mean and mean ที่ในการประมาณค่าพารามิเตอร์ไม่ได้ประมาณค่าทุกค่าพร้อมกัน Haebara (1980) จึงได้เสนอวิธีการประมาณค่าพารามิเตอร์ข้อสอบทุกค่าพร้อมกัน ต่อมา Stocking and Lord (1983) เสนอวิธีการคล้ายกับของเฮบารา และเรียกวิธีการทั้งคู่ว่า วิธีโค้งลักษณะข้อสอบ อาศัยแนวคิดการแปลงความสามรถ I และ J จากสมการ ใช้ได้ทั้งการปรับเทียบคะแนนตามแนวตั้งและแนวนอน การหาค่าสัมประสิทธิ์การปรับเทียบคะแนน A และ B จากสมการ

$$\text{Pij}(\theta_{Ji}; a_{Jj}, b_{Jj}, ¢_{Jj}) = p_{ij}(A\theta_{Ii} + B; \frac{a_{Ij}}{A}, Ab_{Ij} + B, c_{Ij})$$

4. วิธีไคสแควร์น้อยที่สุด (Minimum χ^2 method) เป็นวิธีการปรับเทียบคะแนนที่ใช้ได้ทั้งการปรับเทียบตามแนวนอนและตามแนวตั้ง ที่เสนอโดย ดิฟกิ ใช้วิธีการประมาณค่าพารามิเตอร์ข้อสอบแบบแยกกัน ในการประมาณค่าอำนาจจำแนก และความยากของข้อสอบหาได้จากสูตร

$$a^*_{i2} = a_{i2} / A$$

$$b^*_{i2} = Ab_{i2} + B_i$$

เมื่อ A คือ ความชัน

B คือ ค่าคงที่

คะแนนความสามารถ (θ) ของคนที่ j หาได้จาก

$$\theta^*_{j2} = A\theta_{j2} + B$$

การหาสัมประสิทธิ์การปรับเทียบที่เหมาะสม ใช้วิธีประมาณค่าไคสแควร์ต่ำสุด จากสมการ

$$\chi^2 = \sum_{i=1}^{n} (a_{i1} - a^*_{i2}, b_{i1} - b^*_{i2})(\sum i1 + \sum {}^*_{i2})^{-1}(a_{i1} - a^*_{i2}, b_{i1} - b^*_{i2})'$$

จากการศึกษางานวิจัยที่เกี่ยวข้องเกี่ยวกับวิธีการประมาณค่าสัมประสิทธิ์การปรับเทียบ สรุปได้ว่า วิธีการปรับเทียบคะแนนด้วยวิธีโค้งคุณลักษณะแบบสอบกับวิธี robust mean and sigma จะพบว่า วิธีโค้งคุณลักษณะแบบสอบให้ผลดีกว่า (Stocking & Lord, 1983, p. 207) และเมื่อเปรียบเทียบวิธีโค้งคุณลักษณะแบบสอบกับวิธี Mean and mean วิธีโค้งคุณลักษณะแบบสอบก็ให้ผลที่แม่นยำกว่าเช่นเดียวกัน (Baker & Al-karni, 1991, p. 161) แต่เมื่อเปรียบเทียบวิธีการปรับค่าพารามิเตอร์พร้อมกันกับวิธีโค้งคุณลักษณะ แบบสอบกับผลปรากฏว่าวิธีการปรับค่าพารามิเตอร์พร้อมกันเป็นวิธีที่ดีกว่า (Pertersen, Cook & Stocking, 1983 cited in Bastari, 2000, p. 22) นอกจากนี้พบว่า เมื่อขนาดของข้อสอบร่วมมีขนาดเล็ก วิธีโค้งคุณลักษณะแบบสอบให้ผลดีกว่าวิธีการปรับค่าพารามิเตอร์พร้อมกัน (Kim & Cohen, 1998, pp. 141-142)

วิธีปรับเทียบวิธีนี้แบบสอบแต่ละชุดจะวัดคุณลักษณะเดียวกัน หรือแต่ละคู่ต้องมีคะแนนจริงสัมพันธ์กันเป็นเส้นตรง ซึ่งแมคแคน (MacCann, 1989 อ้างถึงใน อุทัยวรรณ พงศ์อร่าม, 2545, หน้า 40-41) ได้ศึกษาวิธีการปรับเทียบคะแนนสอบ โดยใช้แบบสอบที่เป็นคะแนนจริงสัมพันธ์ โดยอาศัยข้อตกลงของการปรับเทียบคะแนนที่ว่า

1. แบบสอบทั้ง 2 ต้องวัดคุณลักษณะเดียวกัน และการแจกแจงของคะแนนสอบจากแบบสอบทั้ง 2 ต้องสมมูลกัน ถ้าแบบสอบมีความเที่ยงสูง จะทำให้รูปทรงของการแจกแจงเป็นรูปแบบเดียวกัน คือ มีการกระจายของคะแนนจริงที่สมมูลกัน คะแนนจริงของแบบสอบจึงจะมีความสัมพันธ์กันอย่างสมบูรณ์หรือสัมพันธ์กันเป็นเส้นตรง

2. แบบสอบทั้ง 2 ฉบับที่นำมาเทียบคะแนนสอบจะต้องมีความเที่ยงเท่าเทียมกัน การทำให้แบบสอบมีความเที่ยงเท่าเทียมกันจะต้องใช้กลุ่มผู้เข้าสอบที่มีความสามารถเท่าเทียมกัน

ขั้นตอนดำเนินการปรับเทียบคะแนน

การปรับเทียบคะแนนมีขั้นตอนดำเนินการ ดังต่อไปนี้ (Kolen & Brennan, 1995 อ้างถึงใน อดิศร ศรีบุญวงษ์, 2545, หน้า 18-21)

1. กำหนดจุดมุ่งหมายในการปรับเทียบคะแนนว่าจะนำผลจากการปรับเทียบเพื่อดูพัฒนาการเปลี่ยนแปลงทางการศึกษา หรือเพื่อเทียบความสามารถผู้สอบเป็นข้อมูลใช้ตัดสินผลร่วมกัน ใช้แทนกันได้

2. สร้างแบบสอบหลายฉบับ แต่ละฉบับวัดเนื้อหาเดียวกัน และสร้างตามแบบแผนการออกข้อสอบเดียวกัน (Item specification) เป็นแบบสอบที่มีลักษณะของความเป็นคู่ขนานในด้านเนื้อหา โครงสร้าง รูปแบบ ชนิดของข้อสอบ และเวลาที่ใช้สอบ

3. เลือกวิธีการเก็บรวบรวมข้อมูล ว่าจะเลือกใช้รูปแบบใด จะเป็นรูปแบบกลุ่มสุ่ม (Random groups design) โดยสุ่มกลุ่มตัวอย่างจากประชากรเดียวกัน แต่ละกลุ่มทำแบบสอบ

คนละชุด รูปแบบผู้สอบกลุ่มเดียว (Single-groups design) ผู้สอบกลุ่มเดียวทำแบบสอบทั้งสองชุด รูปแบบกลุ่มเดียวที่ได้รับการจัดให้สมดุล (Single-groups design with counterbalancing) แบ่งผู้สอบแต่ละกลุ่ม แบบสอบแต่ละชุดเป็น 3 ส่วน ให้ผู้สอบกลุ่มย่อยแรกสอบแบบสอบชุดที่ 1 ตอนแรก ตามด้วยแบบสอบชุดที่ 2 ตอนหลัง และผู้สอบกลุ่มย่อยที่ 2 ทำแบบสอบชุดที่ 2 ตอนแรก ตามด้วยแบบสอบชุดที่ 1 ตอนหลัง และรูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้แบบสอบร่วม (Common item nonequivalent groups design) ผู้สอบต่างกลุ่มประชากรทำแบบสอบคนละชุด และผู้สอบทุกคนทำแบบสอบร่วมที่อาจจะแทรกภายในแบบสอบทั้งสองฉบับ (Internal common item) ผู้สอบต่างกลุ่มประชากรทำแบบสอบคนละชุด และผู้สอบทุกคนทำแบบสอบร่วมที่อาจจะแทรกภายในแบบสอบทั้งสองฉบับ (Internal common item) หรือแยกออกจากแบบสอบ (External common item) หรือใช้รูปแบบผู้สอบต่างกลุ่มโดยมีผู้สอบร่วม (Common-person design)

4. เก็บรวบรวมข้อมูลตามรูปแบบที่กำหนดไว้

5. เลือกนิยามเชิงปฏิบัติการของการปรับเทียบคะแนน เพื่อตัดสินใจว่าจะใช้วิธีการปรับเทียบคะแนนเชิงเส้นตรง (Linear equating methods) หรือวิธีการปรับเทียบที่ไม่เป็นเชิงเส้นตรง (Nonlinear equating methods)

6. เลือกวิธีประมาณค่าสถิติที่ใช้วิเคราะห์ เลือกให้สอดคล้องกับนิยามเชิงปฏิบัติการที่กำหนด มีวิธีการปรับจากค่าเฉลี่ย (Mean equating) โดยพิจารณาคะแนนสมมูลกันเมื่อคะแนนจากแบบสอบต่างฉบับเบี่ยงเบนไปจากคะแนนเฉลี่ยเท่ากัน วิธีปรับเทียบเชิงเส้นตรง ((Linear equating) พิจารณาคะแนนสมมูลกันเมื่อคะแนนจากแบบสอบต่างฉบับมีคะแนนมาตรฐานเท่ากัน วิธีการปรับเทียบอิควิเปอร์เซ็นต์ไทล์ (Equipercentile equating) ที่คะแนนสมมูลกันเมื่อคะแนนจากแบบสอบต่างฉบับมีตำแหน่งเปอร์เซ็นต์ไทล์เท่ากัน และวิธีการปรับเทียบคะแนนโดยใช้สมการถดถอย (Regression equating) เป็นการสร้างสมการทำนายคะแนน คะแนนจากแบบสอบชุดหนึ่งไปยังอีกชุดหนึ่ง หรือได้จากคะแนนสมมูลกันเมื่อคะแนนของแบบสอบแต่ละฉบับทำนายคะแนนเกณฑ์ได้เท่ากัน ทั้ง 4 วิธีนี้ เป็นวิธีการปรับเทียบคะแนนตามทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory) ส่วนวิธีการปรับเทียบตามทฤษฎีการตอบข้อสอบ (Item respons theory) เป็นการหาสัมประสิทธิ์การปรับเทียบ หรือค่าความชัน (Slope) และค่าคงที่ ของฟังก์ชันเชิงเส้นตรงที่เป็นความสัมพันธ์ของการปรับเทียบคะแนน วิธีการหาสัมประสิทธิ์การปรับเทียบ มีวิธีใช้ค่าเฉลี่ยของค่าอำนาจจำแนกและค่าเฉลี่ยของค่าความยากของข้อสอบ (Mean and mean method) วิธีใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าความยากของข้อสอบ (Mean and σ method) วิธีการทำให้ความแตกต่างระหว่างคะแนนจริงเดิมกับคะแนนจริงที่ปรับแล้วมีค่าน้อยที่สุด โดยใช้สถิติ F-test (Characteristic curve method) และวิธีทำให้ความแตกต่างระหว่างพารามิเตอร์ข้อสอบเดิม

กับที่ปรับแล้วมีค่าน้อยที่สุด โดยใช้สถิติ F-test (Characteristic curve method) และวิธีทำให้ความแตกต่างระหว่างพารามิเตอร์ข้อสอบเดิมกับที่ปรับแล้วมีค่าน้อยที่สุด โดยใช้สถิติ χ^2-test (Minimum χ^2)

7. ประเมินผลการปรับเทียบคะแนน มีเกณฑ์ที่ใช้ดังนี้ (Harris & Croude, 1993 อ้างถึงใน พัชรี จันทร์เพ็ง, 2550, หน้า 95) ความเสมอภาค (Weak equity) ของ Divgi and Yen ที่พิจารณาจากความเท่าเทียมกันของการแจกแจงตามเงื่อนไขของคะแนนที่ได้จากแบบสอบต่างฉบับหลังจากการปรับเทียบแล้ว ดัชนีสำหรับการเปลี่ยนแปลงคะแนน (Indices) ของ Angoff ความคลาดเคลื่อนมาตรฐาน (Standard error) ของ Angoff เป็นการวิเคราะห์เพื่อประมาณความคลาดเคลื่อนของการปรับเทียบจากการสุ่มตัวอย่าง ข้อมูลที่จำลองขึ้น (Generated data) ของ Load เพื่อใช้สำหรับการปรับเทียบคะแนน การปรับเทียบคะแนนจากแบบสอบกลับสู่แบบสอบเดิม (Equating a test to itself) ของ Load เป็นการปรับเทียบคะแนนกลับสู่แบบสอบเดิมโดยตรง หรือปรับผ่านแบบสอบอื่นก่อนปรับกลับสู่แบบสอบเดิม การใช้สอบกับกลุ่มตัวอย่างขนาดใหญ่ (Large sample) ของ Angoff เป็นการใช้กลุ่มตัวอย่างขนาดใหญ่ในการปรับเทียบคะแนนซึ่งคล้ายกับการปรับเทียบคะแนนจากประชากร และใช้เปรียบเทียบกับผลการปรับเทียบคะแนนที่มีขนาดกลุ่มตัวอย่างน้อยกว่า ความคงเส้นคงวา (Consistency) เป็นการประเมินผลการปรับเทียบคะแนนข้ามวิธี เพื่อหาความคงเส้นคงวา ความคงที่ (Stability) ของ Angoff เป็นการปรับเทียบคะแนนซ้ำเพื่อตรวจทานความคงที่ของผลการปรับเทียบคะแนน

เงื่อนไขที่ส่งเสริมให้การปรับเทียบคะแนนเป็นที่น่าพอใจมีรายละเอียดดังนี้ (Kolen & Brennan, 1995)

1. ลักษณะทั่วไป

1.1 เป้าหมายของการที่นำคะแนนมาปรับเทียบมีการระบุอย่างชัดเจน

1.2 การออกแบบเก็บรวบรวมข้อมูล การวางแผนเชื่อมโยงการปรับเทียบคะแนน วิธีการทางสถิติที่ใช้ และการเลือกผลที่ได้จากการปรับเทียบหลายวิธี มีความเหมาะสมกับความสำเร็จของการปรับเทียบคะแนน

1.3 มีการติดตามและควบคุมคุณภาพของกระบวนการปรับเทียบคะแนนอย่างเพียงพอ

2. การพัฒนาแบบสอบสำหรับการเก็บรวบรวมข้อมูลทุกรูปแบบ

2.1 กำหนดเนื้อหาแบบสอบและกำหนดแบบแผนทางสถิติข้อสอบเป็นอย่างดี มีความคงที่

2.2 การสร้างแบบสอบเพื่อนำมาปรับเทียบพิจารณาจากสถิติข้อสอบ เช่น ค่าความยาก ค่าอำนาจจำแนก จากการสอบครั้งก่อน

2.3 ข้อสอบที่ใช้มีความยาวพอเหมาะ อย่างน้อยมีจำนวน 30 ข้อ หรือมากกว่า

3. การพัฒนาแบบสอบสำหรับรูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันใช้แบบสอบร่วม

3.1 ข้อสอบร่วมต้องเป็นตัวแทนของแบบสอบที่ใช้ในการปรับเทียบ ทั้งคุณลักษณะของเนื้อหาและค่าสถิติของข้อสอบ

3.2 ข้อสอบร่วมมีจำนวนพอเหมาะ อย่างน้อย 30 ข้อหรือ 20 % ของแบบสอบที่มีความยาว 40 ข้อ หรือมากกว่า (Kolen & Brennan, 2004, p. 271) สำหรับการปรับเทียบด้วย IRT กรณีใช้การปรับค่าพารามิเตอร์พร้อมกัน สามารถใช้ข้อสอบร่วม 5-6 ข้อที่เลือกมาอย่างดีได้ (Wingersky & Lord, 1984 cited in Yu Meng, 2012, p. 33) ขณะที่ไรท์และสโตน (Wright & Stone, 1979, p. 98) กล่าวว่าแบบสอบร่วมควรเป็นข้อสอบที่วัดเรื่องเดียวกันกับแบบสอบที่ใช้ในการปรับเทียบและมีจำนวนเพียง 10 ข้อก็เพียงพอ

3.3 ข้อสอบร่วมแต่ละข้อสร้างให้มีลักษณะที่ใกล้เคียงกับแบบสอบที่นำมาปรับเทียบทั้งตัวคำถามและตัวเลือก

4. กลุ่มตัวอย่าง

4.1 กลุ่มตัวอย่างมีลักษณะเป็นตัวแทน

4.2 กลุ่มตัวอย่างมีความคงที่

4.3 กลุ่มตัวอย่างมีขนาดใหญ่เพียงพอ

4.4 กลุ่มตัวอย่าง 2 กลุ่ม สำหรับกลุ่มผู้สอบไม่เท่าเทียมกันโดยใช้แบบสอบร่วม กลุ่มผู้สอบควรมีความสามารถไม่แตกต่างกันมากนัก

5. การบริหารการสอบ

5.1 ดำเนินการทดสอบอย่างปลอดภัย

5.2 การสอบในแต่ละครั้งดำเนินการอย่างระมัดระวัง และจัดสภาพการสอบให้เหมือนกัน

6. หลักสูตร เนื้อหาการปฏิบัติ หรือสาระที่นำมาใช้ในการปรับเทียบคะแนนมีความคงที่

ผลของการปรับเทียบคะแนนมีความเชื่อถือได้ หรือมีความถูกต้องแม่นยำขึ้นอยู่กับการเลือกสถิติที่ใช้ในการปรับคะแนนให้มีความเหมาะสมกับแบบแผนการออกข้อสอบ การเก็บรวบรวมข้อมูล และการดูแลการสอบที่เป็นไปอย่างมีมาตรฐาน

การออกแบบเพื่อรวบรวมข้อมูลและการจัดกระทำทางสถิติ

การปรับเทียบคะแนน อาศัยวิธีการทางสถิติ และการออกแบบรวบรวมข้อมูลซึ่งผลของการปรับเทียบคะแนนจะมีความน่าเชื่อถือ ขึ้นอยู่กับการเลือกวิธีการปรับเทียบ การออกแบบแผนการออกข้อสอบ การเก็บรวบรวมข้อมูล ที่สอดคล้องเหมาะสมและสามารถจัดกระทำข้อมูล

ภายใต้สภาพการณ์ของความจำกัด ที่ไม่สามารถวางเงื่อนไขให้มีความเพียงพอตามนิยามได้ทุกประการ ซึ่งแองกอฟ (Angoff, 1984, pp. 94-121) ได้จำแนกการออกแบบรูปแบบการปรับเทียบออกเป็น 2 กลุ่มใหญ่ คือ กลุ่มปรับเทียบโดยใช้แบบสอบร่วม (Anchor test) กับกลุ่มการปรับเทียบโดยไม่ใช้แบบสอบร่วม ซึ่งในกรณีที่ไม่ใช้แบบสอบร่วมนี้เป็นรูปแบบที่ต้องการเงื่อนไขที่เข้มงวดหลายประการ คือ 1) ความเป็นตัวอย่าง 2) สุ่มสมมูลของผู้รับการทดสอบ ความเป็นคู่ขนานของแบบสอบ โดยเน้นความสมมูลของค่าความเที่ยงของแบบสอบทั้ง 2 ชุด เมื่อเงื่อนไขเข้มงวดมากการปฏิบัติก็จะมีข้อขัดข้องมาก เพราะในสภาพการณ์จริงโอกาสการสุ่ม และการตรวจความสมมูลของความเที่ยงย่อมเป็นไปได้ยากมาก ดังนั้นการออกแบบที่ใช้แบบสอบร่วมจึงเป็นทางออกที่เหมาะสมกว่า ในที่นี้จะขอกล่าวรายละเอียดเฉพาะการออกแบบการปรับเทียบคะแนนโดยใช้แบบสอบร่วม (Anchor test) หรือการใช้ข้อสอบร่วม (Common item)

แบบสอบร่วม เป็นกลุ่มของข้อสอบที่กำหนดให้ผู้สอบ 2 กลุ่ม ทำแบบสอบคนละชุด โดยที่แบบสอบแต่ละชุดนั้นมีข้อสอบร่วมกันจำนวนหนึ่ง ซึ่งเรียกว่าแบบสอบร่วม แบบสอบร่วมจะทำหน้าที่ลดความแตกต่างของกลุ่มผู้สอบทั้ง 2 กลุ่ม อันเนื่องมาจากความแปรปรวนของการสุ่ม (Angoff, 1984, p. 106) คะแนนจากแบบสอบร่วมนำมาใช้เป็นฐานเมื่อประมาณความแตกต่างของกลุ่มคน 2 กลุ่ม และสารสนเทศนี้นำไปใช้ในการประมาณความแตกต่างระหว่างแบบสอบ 2 ชุด ที่ใช้สอบเพื่อหาคะแนนแปลงที่เทียบเท่ากันต่อไป แบบสอบร่วมที่ดีจะต้องวัดความรู้ความสามารถเดียวกับแบบสอบที่ต้องการปรับเทียบและจะต้องมีค่าสัมประสิทธิ์สหสัมพันธ์กับแบบสอบที่ต้องการปรับเทียบสูง หรือกล่าวอีกนัยได้ว่าถ้าสามารถทำแบบสอบร่วมให้เป็นแบบสอบที่มีลักษณะเป็นฉบับย่อส่วนจะทำให้การปรับเทียบคะแนนมีความถูกต้องแม่นยำขึ้น (Angoff, 1984, p. 107)

แบบสอบร่วม แบ่งออกเป็น 2 ชนิด คือ แบบสอบร่วมภายใน (Internal anchor test) เป็นการผนวกแบบสอบร่วมเข้ากับแบบสอบที่ใช้สอบในการปรับเทียบคะแนน ถ้าแบบสอบร่วมมีความยาวมากเท่าไร ผู้สอบที่เคยรับการทดสอบมาก่อนย่อมได้เปรียบ เพราะเกิดการเรียนรู้หรือจำข้อสอบในส่วนของข้อสอบร่วมได้ ส่วนแบบสอบร่วมภายนอก (External anchor test) เป็นการใช้แบบสอบร่วมเป็นส่วนหนึ่งของการปรับเทียบคะแนน โดยไม่ได้รวมข้อสอบร่วมไว้ในฉบับเดียวกันกับแบบสอบที่ใช้ในการปรับเทียบคะแนน

ความยาวของแบบสอบร่วม (Angoff, 1984, p. 107) แนะนำว่าควรใช้ข้อสอบร่วมที่มีความยาวมากพอที่จะก่อให้เกิดสารสนเทศที่จำเป็น ควรมีจำนวนไม่น้อยกว่า 20 ข้อ หรือไม่น้อยกว่า 20% ของจำนวนข้อ แล้วแต่จำนวนไหนจะมากกว่า ไวท์ และสโตน์ (Wright & Stone, 1979, p. 98) แนะนำว่าถ้าเป็นแบบสอบที่วัดเรื่องเดียวกันกับเรื่องที่ทำการวัดในแบบสอบทั้ง 2 ชุด ข้อสอบร่วมที่มีความยาก .3 ขึ้นไป ใช้เพียง 10 ข้อ ก็เพียงพอแล้ว

การออกแบบเพื่อเก็บรวบรวมข้อมูล ลอร์ด ได้เสนอไว้ 4 รูปแบบ (Lord, 1975 cited in Kolen & Brennen, 2004, pp. 13-23) ดังนี้

1. รูปแบบผู้สอบกลุ่มเดียว (Single-group design) แบ่งเป็น 2 แบบย่อย คือ

1.1 ผู้สอบกลุ่มเดียวที่ไม่ได้รับการจัดให้สมดุล (Uncounterbalanced design) ใช้ผู้สอบกลุ่มเดียว แต่ละคนทำข้อสอบทั้ง 2 ฉบับ จากนั้นนำคะแนนมาปรับเทียบกันบนพื้นฐานความสามารถที่เท่ากัน แต่ในทางปฏิบัติการสอบฉบับหลัง อาจได้รับปัจจัยที่ผลกระทบจากการเรียนรู้การสอบฉบับแรก ความเมื่อยล้า ฯลฯ ซึ่งอาจมีอิทธิพลต่อการปรับเทียบ

1.2 ผู้สอบกลุ่มเดียวได้รับการจัดให้สมดุล (Counterbalanced design) รูปแบบนี้ขจัดผลของลำดับก่อน-หลัง โดยสุ่มผู้สอบแยกเป็น 2 กลุ่ม แต่ละกลุ่มได้รับการสอบ 2 ฉบับ โดยกลุ่มแรกสอบแบบสอบฉบับที่ 1 ตามด้วยฉบับ 2 ส่วนกลุ่มหลังให้สอบแบบสอบฉบับที่ 2 ตามด้วยฉบับที่ 1 เพื่อให้เกิดความสมดุล

2. รูปแบบผู้สอบกลุ่มเท่าเทียมกัน (Equivalent-group design) โดยจัดกลุ่มผู้สอบให้มีความคล้ายคลึงกันมากที่สุด แล้วให้แต่ละกลุ่มทำแบบสอบเพียงฉบับเดียว เพื่อหลีกเลี่ยงปัญหาการเรียนรู้ ความเมื่อยล้า แต่อาจจะมีปัญหากลุ่มที่ใช้อาจมีการแจกแจงความสามารถที่แตกต่างกัน วิธีการที่จะลดความแตกต่าง ควรใช้กลุ่มตัวอย่างให้มีขนาดใหญ่

3. รูปแบบผู้สอบต่างกลุ่มโดยใช้แบบสอบร่วม (Anchor-test design) แบ่งเป็น 2 แบบย่อย คือ

3.1 ผู้สอบกลุ่มสุ่มโดยใช้แบบสอบร่วม (Anchor-test random group design) เริ่มจากการสุ่มกลุ่มตัวอย่าง มา 2 กลุ่ม แต่ละกลุ่มทำแบบสอบฉบับเดียว แบบสอบแต่ละฉบับมีข้อสอบร่วมจำนวนหนึ่ง ซึ่งอาจจะจัดรวมไว้ในแบบสอบเดียวกัน (Internal anchor test) หรือจัดแยกต่างหาก (External anchor test) รูปแบบนี้ช่วยลดความลำเอียงที่มีสาเหตุจากการเรียนรู้ ความเมื่อยล้า และความแตกต่างเกี่ยวกับระดับความสามารถของกลุ่มผู้สอบ แต่ประสิทธิภาพของการปรับเทียบคะแนนยังขึ้นอยู่กับคุณภาพของแบบสอบร่วม จึงต้องคำนึงถึงเนื้อหา ระดับความยากที่จะต้องมีความคล้ายกับแบบสอบที่ต้องการปรับทียบคะแนนและความยาวของแบบสอบร่วมก็เป็นปัจจัยสำคัญที่ส่งผลต่อคุณภาพของการปรับเทียบ

3.2 ผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้แบบสอบร่วม (Non-equivalent groups anchor-test design) ใช้ผู้สอบกลุ่มไม่เทียมกันแต่ละกลุ่มทำข้อสอบฉบับเดียวที่มีข้อสอบร่วมภายในหรือแบบสอบร่วมภายนอกก็ได้ รูปแบบนี้จะใช้ในสถานการณ์ที่ผู้สอบต่างโปรแกรม ต่างเวลา หรือต่างระดับ

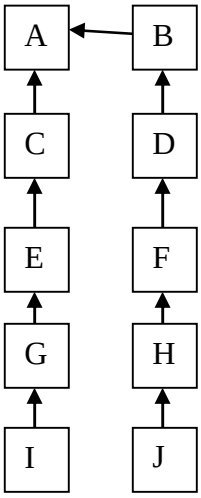
4. รูปแบบผู้สอบต่างกลุ่มโดยมีผู้สอบร่วม (Common-person design) เป็นการใช้ผู้สอบต่างกลุ่มกันทำแบบสอบต่างฉบับ แต่มีจำนวนผู้สอบจำนวนหนึ่งจากทั้ง 2 กลุ่ม ทำแบบสอบทั้ง 2 ฉบับ การออกแบบทั้ง 4 รูปแบบดังกล่าว นำมาเขียนเป็นตารางได้ ดังนี้

ตารางที่ 2-1 แบบแผนการเก็บรวบรวมข้อมูลเพื่อปรับเทียบคะแนน

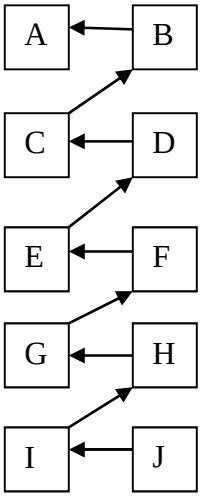
แบบแผนการเก็บรวบรวมข้อมูล	กลุ่มตัวอย่าง	แบบสอบ		
		ฉบับที่ 1	ฉบับที่ 2	แบบสอบร่วม
1. รูปแบบผู้สอบกลุ่มเดียว				
1.1 รูปแบบผู้สอบกลุ่มเดียว ที่ไม่ได้รับการจัดให้สมดุล	p_1	X	Y	
1.2 รูปแบบผู้สอบกลุ่มเดียวที่ได้รับ	p_1	X	Y	
การจัดให้สมดุล	p_2	Y	X	
2. รูปแบบผู้สอบกลุ่มเท่าเทียมกัน	p_1	X		
	p_2		Y	
3. รูปแบบผู้สอบต่างกลุ่มโดยใช้ แบบสอบร่วม				
3.1 รูปแบบผู้สอบกลุ่มสุ่มโดยใช้				
แบบสอบร่วม	p_1	XV		
3.1.1 แบบสอบร่วมภายใน	p_2		YV	
	p_1	X		V
3.1.2 แบบสอบร่วมภายนอก	p_2		Y	V
3.2 รูปแบบผู้สอบกลุ่มไม่เท่าเทียม กันโดยใช้แบบสอบร่วม				
3.2.1 แบบสอบร่วมภายใน	p_1	XV		
	Q_2		YV	
3.2.2 แบบสอบร่วมภายนอก	p_1	X		V
	Q_2		Y	V
4. รูปแบบผู้สอบต่างกลุ่มโดยมี ผู้สอบร่วม	p_1	X		
	p			
	p_{10}	X	Y	
	Q_1		Y	
	Q			
	Q_{10}	X	Y	

สำหรับการออกแบบข้อสอบร่วมของรูปแบบกลุ่มไม่เท่าเทียมกัน(Common-item nonequivalent groups design) โคเรน และเบรนเนน (Kolen & Brennan, 2004, p. 281-284) ได้นำเสนอการออกแบบการปรับเทียบคะแนนในกลุ่มไม่เท่าเทียมกันด้วยข้อสอบร่วม ของการสอบในฤดูใบไม้ผลิและฤดูใบไม้ร่วงที่สอบในระยะเวลาติดต่อกัน 5 ปี ไว้ 4 รูปแบบ ดังนี้

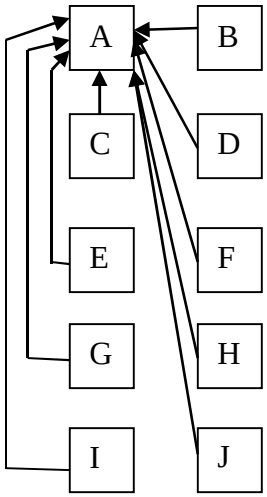
1. แบบแผนที่ 1



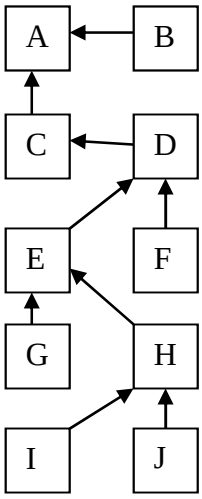
2. แบบแผนที่ 2



3. แบบแผนที่ 3



4. แบบแผนที่ 4



ลูกศรของแต่ละแบบแผนแทนการเชื่อมโยงคะแนนระหว่างแบบสอบด้วยข้อสอบร่วม เช่นการปรับเทียบคะแนนของฟอร์ม J จะใช้ข้อสอบร่วมของฟอร์ม H ในการออกแบบการปรับเทียบคะแนนกลุ่มสุ่มรูปแบบกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบร่วม มีกฎ 4 ข้อที่สามารถนำไปใช้เกี่ยวกับข้อสอบร่วมภายใน ดังนี้

1. หลีกเลี่ยงการฝืนการปรับเทียบด้วยการทำให้จำนวนการเชื่อมโยงคะแนนเข้าสู่ฟอร์มเดิมให้เหลือน้อยที่สุด ซึ่งตามแบบแผนที่ 1 ไม่ได้เป็นไปตามกฎข้อนี้ เช่นกว่าจะเชื่อมโยงคะแนนจากฟอร์ม J สู่ฟอร์ม I จะต้องผ่านการเชื่อมโยงคะแนนจาก J to H, H to F, F to D, …, กว่าจะถึง G to I จะต้องผ่านถึง 9 ครั้ง เป็นต้น

2. ถ้าเป็นไปได้ให้เชื่อมโยงคะแนนในช่วงเวลาเดียวกัน แบบแผนที่ 1 ออกแบบเป็นไปตามกฎข้อนี้

3. ลดจำนวนการเชื่อมโยงสู่ฟอร์มเดิมให้เหลือน้อยที่สุด แทนที่จะไปผ่านฟอร์มอื่น ๆ ก่อนที่เชื่อมโยงสู่ฟอร์มเดิม แบบแผนที่ 3 ออกแบบเป็นไปตามกฎข้อนี้และกฎข้อที่ 1

4. หลีกเลี่ยงการเชื่อมโยงคะแนนกลับสู่ฟอร์มเดิมบ่อยมากเกินไป

แบบแผนที่ 2 และ 4 ออกแบบเป็นไปตามกฎข้อนี้และกฎข้อที่ 1 ในการออกแบบการเชื่อมโยงคะแนนครั้งหนึ่ง ๆ เป็นไปไม่ได้ที่จะทำให้เป็นไปตามกฎได้ทุกกฎในคราวเดียวกัน จะต้องเลือกพิจารณาให้สอดคล้องกับเนื้อหา โครงสร้างในแบบสอบ

ขนาดกลุ่มตัวอย่าง กลุ่มตัวอย่างที่ใช้ในแบบแผนการเชื่อมโยงคะแนนระหว่างแบบสอบของกลุ่มที่ไม่เท่าเทียมกันโดยใช้ข้อสอบร่วม ขนาดของกลุ่มตัวอย่างจะขึ้นอยู่กับระดับความสัมพันธ์ระหว่างคะแนนรวม (Total score) กับคะแนนของข้อสอบร่วม (Common-item score) (Budescu, 1985) และรูปร่างการแจกแจงที่มีผลต่อความคลาดเคลื่อนในการปรับเทียบ สำหรับการปรับเทียบโดยวิธี IRT ขนาดของกลุ่มตัวอย่างขึ้นค่าพารามิเตอร์ ถ้าเป็น 1 พารามิเตอร์ จะใช้กลุ่มตัวอย่างขนาด 400 คน/ ฟอร์ม เหมือนกันกับวิธีเชิงเส้นตรง แต่ถ้าเป็น 3 พารามิเตอร์ จะใช้กลุ่มตัวอย่างที่ใหญ่กว่าขนาด 1,000 คน/ ฟอร์ม เหมือนกันกับวิธีอิควิเปอร์เซ็นไทล์ ซึ่งขนาดของกลุ่มตัวอย่างมีผลกระทบโดยตรงต่อความคลาดเคลื่อนอย่างสุ่มของการปรับเทียบ โดยขนาดของกลุ่มตัวอย่างยิ่งมากยิ่งดี (Kolen & Brennan, 2004, p. 288) และกลุ่มตัวอย่างยิ่งมีความเหมือนกันมากเท่าใด จะทำให้การปรับเทียบคะแนนมีความเหมาะสมมากเท่านั้น (Kolen & Brennan, 2004, p. 281) นอกจากนี้ ลีวิงสตัน (Livingston, 1993 cited in Kolen & Brennan, 2004, p. 288) กล่าวว่า ขนาดของกลุ่มตัวอย่างมีผลโดยตรงต่อความคลาดเคลื่อนอย่างสุ่มของการปรับเทียบคะแนน

สำหรับการออกแบบการปรับเทียบคะแนน แบบแผนกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบร่วม (Common-item nonequivalent groups design) หากกลุ่มตัวอย่างมีความคล้ายคลึงกันควรเลือกใช้วิธีอิควิเปอร์เซ็นไทล์ในการปรับเทียบ แต่ถ้ากลุ่มตัวอย่างมีความแตกต่างกันวิธีปรับเทียบที่เหมาะสมกว่าคือ วิธี Levine observed score หรือวิธี IRT (Kolen & Brennan, 2004, p. 298)

สัมประสิทธิ์แคปปา (Coefficient kappa): แนวคิดและการประยุกต์ใช้

สัมประสิทธิ์แคปปา (สุรชัย มีชาญ, 2547, หน้า 15-25) เป็นตัวบ่งชี้ถึงระดับความสอดคล้องในการตัดสินใจของผู้ตัดสินตั้งแต่ 2 คนขึ้นไป ซึ่งค่าสัมประสิทธิ์แคปปามีค่าได้ตั้งแต่ -1 ถึง 1 เมื่อใดที่มีค่าใกล้ 1 แสดงว่าผู้ตัดสินทั้งสองสามารถตัดสินได้อย่างสอดคล้องกันสูงมาก แต่ถ้ามีค่าใกล้ -1 แสดงว่ามีความขัดแย้งกันอย่างมาก ดังนั้นหากผู้ตัดสินทั้งสองคือผู้ที่ได้รับการฝึกฝนและมีความรู้ความเข้าใจเกี่ยวกับเกณฑ์และวิธีการตัดสินอย่างดีแล้ว สัมประสิทธิ์นี้ก็จะ

ช่วยบ่งชี้ว่าเกณฑ์การตัดสินที่ได้กำหนดไว้นั้นมีความชัดเจนหรือมีมาตรฐานเพียงพอที่จะนำไปใช้ในวงกว้างต่อไปหรือไม่ และในกรณีที่ผู้ตัดสินอีกคนหนึ่งมีความสามารถในการตัดสินมากน้อยเพียงใด สมบูรณ์พร้อมแล้วหรือยังต้องพัฒนาเพิ่มอีก นอกจากนี้ยังนิยมนำไปประยุกต์ใช้ในการตรวจสอบความเชื่อมั่น (Reliability) ของแบบสอบอิงเกณฑ์ (Criterion-referenced test) ได้อีกด้วย โดยอาจใช้ตรวจสอบว่าแบบสอบ 2 ฉบับที่คู่ขนานกัน (Parallel form) จะสามารถจำแนกผู้สอบเป็นผู้รอบรู้และไม่รอบรู้ได้อย่างสอดคล้องกันหรือไม่ (ซึ่งควรสอดคล้องกันเพราะทั้งสองฉบับถูกสร้างขึ้นมาเพื่อให้ใช้แทนกันได้ ตามเจตนาของการสร้างแบบสอบคู่ขนาน) รวมทั้งการตรวจสอบว่าแบบสอบฉบับหนึ่งที่ถูกนำไปทดลองใช้กับผู้สอบกลุ่มเดิมสองครั้งจะสามารถจำแนกผู้สอบเป็นผู้รอบรู้และไม่รอบรู้ได้อย่างคงที่มากน้อยเพียงใด เพราะโดยความคาดหวังนั้น หากครั้งที่ 1 นักเรียนคนใดที่สอบผ่านเกณฑ์ (ได้รับการตัดสินว่าเป็นผู้รอบรู้) ในครั้งที่ 2 เขาก็ควรจะสอบผ่านเกณฑ์เช่นเดิม ส่วนผู้ที่สอบไม่ผ่านเกณฑ์ก็ควรสอบไม่ผ่านทั้งสองครั้งด้วย เราจึงจะเชื่อถือในแบบสอบฉบับนั้นมากขึ้น ซึ่งคุณสมบัติเหล่านี้สำคัญอย่างมากสำหรับแบบสอบอิงเกณฑ์

เรายังสามารถใช้สัมประสิทธิ์แคปปาเพื่อบ่งชี้ถึงระดับความสอดคล้องในการตัดสินได้อย่างหลากหลาย ทั้งในกรณีที่ต้องการวัดความสอดคล้องในการตัดสิน จาก 2 แหล่ง หรือการจำแนกผลงานประดิษฐ์ของนักเรียนออกเป็น 3 กลุ่ม คือ “ดี-พอใช้-ต่ำกว่ามาตรฐาน” หรือกรณีอื่น ๆ ในทำนองเดียวกันนี้

ทั้งนี้เราสามารถนำสัมประสิทธิ์แคปปาไปประยุกต์ใช้ เพื่อประโยชน์ทางด้านการเรียนการสอนในสถานศึกษาได้อย่างหลากหลาย ในที่นี้จะขอยกตัวอย่างดังต่อไปนี้

ตัวอย่าง

จากการทดสอบความสามารถของนักศึกษาคณะศึกษาศาสตร์ในการจำแนกจุดประสงค์การเรียนรู้ออกเป็น 3 ด้านตามพฤติกรรมที่มุ่งหวัง ได้แก่ ด้านพุทธิพิสัย (Cognitive domain) ด้านจิตพิสัย (Affective domain) และด้านทักษะพิสัย (Psychomotor domain) โดยกำหนดจุดประสงค์การเรียนรู้ที่ประมวลมาจากรายวิชาต่าง ๆ จำนวน 30 ข้อ จากนั้นให้นักศึกษาจำแนกจุดประสงค์การเรียนรู้ออกเป็นด้าน ๆ เทียบกับผลการตัดสินของผู้เชี่ยวชาญ ถ้าผลการตัดสินของนักศึกษาคนหนึ่งเมื่อเทียบกับผลการตัดสินของผู้เชี่ยวชาญ (ซึ่งถือเป็น “เฉลย”) ได้ผลดังตารางที่ 2-2

ตารางที่ 2-2 ผลการตัดสินของนักศึกษาเทียบกับผลการตัดสินของผู้เชี่ยวชาญ

		ผลการจำแนกโดยนักศึกษา			
		ด้านพุทธิพิสัย	**ด้านจิตพิสัย**	**ด้านทักษะพิสัย**	**รวม**
ผลการจำแนก	ด้านพุทธิพิสัย	10 ข้อ	2 ข้อ	3 ข้อ	15 ข้อ
โดยผู้เชี่ยวชาญ	ด้านจิตพิสัย	4 ข้อ	12 ข้อ	-	16 ข้อ
(เฉลย)	ด้านทักษะพิสัย	1 ข้อ	3 ข้อ	5 ข้อ	9 ข้อ
	รวม	15 ข้อ	17 ข้อ	8 ข้อ	40 ข้อ

จงคำนวณหาระดับความสอดคล้องในการตัดสิน ซึ่งบ่งชี้ถึงระดับความรู้ความสามารถของนักศึกษาคนดังกล่าว โดยใช้สัมประสิทธิ์แคปปา ($K_{2/1}$)

จากตารางที่ 2-2 แสดงว่า มีจุดประสงค์การเรียนรู้ที่กำหนดมาให้ทั้งหมด จำนวน 40 ข้อ เป็นจุดประสงค์ทางด้านพุทธพิสัย จำนวน 15 ข้อ ด้านจิตพิสัย จำนวน 17 ข้อ และด้านทักษะพิสัย จำนวน 8 ข้อ ซึ่งนักศึกษาคนนี้สามารถจำแนกได้ถูกต้อง(สอดคล้องกับ “มาตรฐาน”) ไม่มากนัก โดยในด้านพุทธิพิสัยสามารถจำแนกได้ถูกต้อง 10 ข้อ (จำแนกผิด 5 ข้อ โดยระบุว่าเป็นด้านจิตพิสัย 2 ข้อ ส่วนอีก 3 ข้อ ระบุว่าเป็นด้านทักษะพิสัย) สำหรับด้านอื่น ๆ ก็แปลความหมายได้ในทำนองเดียวกัน

จากข้อมูลทั้งหมดที่กล่าวมานี้ เราสามารถใช้ $K_{2/1}$ เพื่อตรวจสอบว่านักศึกษาคนนี้มีความสามารถมากน้อยเพียงใด และยังสามารถใช้ K_1 เพื่อบ่งชี้เพิ่มเติมได้อีกว่ามีความสามารถเด่นหรือด้อยในด้านใดบ้าง ทั้งนี้เพื่อที่ครูและผู้เกี่ยวข้องจะได้หาแนวทางปรับปรุงแก้ไขได้อย่างเหมาะสมในโอกาสต่อไป

ดังนั้น ด้วยกระบวนการเดียวกับที่กล่าวมาแล้วในตัวอย่างที่ 1 และ 2 เราสามารถคำนวณหาสัดส่วนของแต่ละช่อง (Cell) ได้ดังตารางที่ 2-3

ตารางที่ 2-3 สัดส่วนความสอดคล้องในการตัดสินของนักศึกษาและผู้เชี่ยวชาญ

		ผลการจำแนกโดยนักศึกษา			
		ด้านพุทธิพิสัย	**ด้านจิตพิสัย**	**ด้านทักษะพิสัย**	**รวม**
ผลการจำแนก	ด้านพุทธิพิสัย	.250	.050	.075	.375
โดยผู้เชี่ยวชาญ	ด้านจิตพิสัย	.100	.300	.000	.400
(เฉลย)	ด้านทักษะพิสัย	.025	.075	.125	.225
	รวม	.375	.425	.200	1.000

จากตารางที่ 2-3 จะได้

$P_{1.} = 0.375$

นั่นคือ มีจุดประสงค์การเรียนรู้ที่ได้รับการตัดสินจากผู้เชี่ยวชาญ (เฉลย) ว่าอยู่ในด้านที่ 1 หรือด้านพุทธิพิสัย ร้อยละ 37.50 ของจุดประสงค์ที่กำหนดมาให้ทั้งหมด หรือคิดเป็นสัดส่วนเท่ากับ 0.375

ในทำนองเดียวกัน จะได้

$P_{2.} = 0.400$

$P_{3.} = 0.225$

สามารถคำนวณหาสัมประสิทธิ์แคปปา ในแต่ละด้าน (ซึ่งก็คือความสอดคล้องในการตัดสินหรือระดับความรู้ความสามารถในแต่ละด้าน) ได้ดังต่อไปนี้

$$K_1 = \frac{P_{11} - P_{1.}P_{.1}}{P_{1.} - P_{1.}P_{.1}}$$

$$= \frac{0.25 - (0.375)(0.375)}{0.375 - (0.375)(0.375)}$$

$$= 0.4667$$

$$K_2 = \frac{P_{22} - P_{2.}P_{.2}}{P_{2.} - P_{2.}P_{.2}}$$

$$= \frac{0.30 - (0.400)(0.425)}{0.400 - (0.400)(0.425)}$$

$$= 0.5652$$

$$K_3 = \frac{P_{33} - P_{3.}P_{.3}}{P_{3.} - P_{3.}P_{.3}}$$

$$= \frac{0.125 - (0.225)(0.200)}{0.225 - (0.225)(0.200)}$$

$$= 0.4444$$

จากข้อมูลทั้งหมด แทนค่าต่าง ๆ ลงในสูตรคำนวณ จะได้

$$K_{2/1} = \sum p_{i.} K_i$$

$$= (0.375)(0.4667) + (0.400)(0.5652) + (0.225)(0.4444)$$

$$= 0.5011$$

สัมประสิทธิ์แคปปา มีค่าเท่ากับ 0.5011 แสดงว่า นักศึกษาคนนี้มีความสามารถในการจำแนกจุดประสงค์การเรียนรู้ไม่มากนัก โดยจำแนกได้ถูกต้อง (สอดคล้องกับผู้เชี่ยวชาญ) มากที่สุดในด้านจิตพิสัย ($K_2 = 0.5652$) รองลงมาคือ ด้านพุทธิพิสัยและด้านทักษะพิสัย ตามลำดับ ($K_1 = 0.4667$ และ $K_3 = 0.5011$) จึงจำเป็นต้องพัฒนาความรู้ความสามารถในประเด็นเหล่านี้ได้มากยิ่งขึ้น

จากทั้งหมดที่กล่าวมาจะเห็นได้ว่า เราสามารถนำสัมประสิทธิ์แคปปาไปประยุกต์ใช้ในกระบวนการเรียนการสอนได้อย่างหลากหลาย อีกทั้งขั้นตอนการคำนวณก็ไม่ได้ยุ่งยากสลับซับซ้อนมากนัก จึงคาดหวังว่าผู้เกี่ยวข้องจะนำไปปรับใช้ให้เกิดประโยชน์ต่อการศึกษาและสังคมวงกว้างต่อไป

สำหรับเกณฑ์ที่ใช้ในการตัดสินความสอดคล้องของสัมประสิทธิ์แคปปา แลนดิส และโคช (Landis & Koch, 1977, pp. 159-174) กำหนดระดับความสอดคล้อง ไว้ดังนี้

0.0-0.20	สอดคล้องกันน้อยมาก
0.21-0.40	สอดคล้องกันพอสมควร
0.41-0.60	สอดคล้องกันปานกลาง
0.61-0.80	สอดคล้องกันมาก
0.81-1.00	สอดคล้องกันมากที่สุด

ซึ่งคริปเพนดอป (Krippendorff, 1980 อ้างถึงใน ประสพชัย พสุนนท์, 2558, หน้า 8) กล่าวว่า ค่าความสอดคล้องควรมีค่ามากกว่า 0.8 จึงจะสามารถยอมรับได้ในทางปฏิบัติ

ตอนที่ 3 การหาคุณภาพของการปรับเทียบคะแนน

การหาคุณภาพของการปรับเทียบคะแนน นิยมหาจากความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน ซึ่งหามาจากความคลาดเคลื่อนในการประมาณค่าของการปรับเทียบคะแนน โดยที่แหล่งของความคลาดเคลื่อนมีด้วยกัน 2 แหล่ง คือ แหล่งแรกมาจากความคลาดเคลื่อนอย่าง

สุ่มของการปรับเทียบ (Random equating error) มาจากการเลือกตัวอย่างจากประชากรมาใช้ในการประมาณค่าพารามิเตอร์ (เช่น ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน การจัดอันดับเปอร์เซ็นไทล์ เป็นต้น) ซึ่งความคลาดเคลื่อนของการปรับเทียบคะแนน (Standard error of equating: SEE) เป็นดัชนีชี้ความคลาดเคลื่อนอย่างสุ่มของการปรับเทียบ (Random equating error) เมื่อขนาดของกลุ่มตัวอย่างใหญ่ขึ้นความคลาดเคลื่อนของการปรับเทียบคะแนนก็จะน้อยลง จึงสามารถลดความคลาดเคลื่อนอย่างสุ่มของการปรับเทียบด้วยการกำหนดขนาดตัวอย่างให้ใหญ่ขึ้น และด้วยการใช้วิธีการออกแบบการปรับเทียบที่เหมาะสม แหล่งที่สองมาจากความคลาดเคลื่อนอย่างเป็นระบบของการปรับเทียบ (Systematic equating error) เป็นผลมาจากการหลีกเลี่ยงข้อตกลงเบื้องต้นหรือเงื่อนไขของการปรับเทียบ เช่นในกรณีการออกแบบการปรับเทียบใช้กลุ่มสุ่ม (Random groups design) แต่ปรากฏว่ากลุ่มตัวอย่างที่ได้มีความสามารถไม่เท่าเทียมกัน หรือการใช้ข้อสอบร่วมที่ไม่ได้เป็นตัวแทนของแบบสอบทั้งฉบับ เป็นต้น (Kolen & Brennan, 2004, pp. 23-24)

การศึกษาคุณภาพของการปรับเทียบคะแนน จึงใช้การประเมินคุณภาพของการปรับเทียบโดยการหาค่าความคลาดเคลื่อนของการปรับเทียบคะแนน (Standard error of equating: SEE) หรือใช้การตรวจสอบความเพียงพอของการปรับเทียบคะแนน วิธีการตรวจสอบคุณภาพสามารถทำได้หลายวิธี ดังนี้

1. ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนเชิงเส้นตรง เมื่อกลุ่ม 2 กลุ่มทำแบบสอบกลุ่มละชุด ชุด X หรือชุด Y และทำข้อสอบร่วมเหมือนกัน ค่าประมาณคลาดเคลื่อนหาได้จากสูตร

$$SEE_x^2 = 2\sigma_X^2 \ (1-\hat{r}^2) \ \frac{(1-\hat{r}^2)z_Y^2 + 2}{N_t}$$

โดยมีข้อตกลงว่า

$$\hat{r} = \frac{b_{xv\alpha}\hat{\sigma}_v}{\hat{\sigma}_x} = \frac{b_{xv\beta}\hat{\sigma}_v}{\hat{\sigma}_y}$$

เมื่อ $b_{xv\alpha}$ และ $b_{yv\beta}$ คือ สัมประสิทธิ์การถดถอย

และ $\hat{\sigma}_v$ $\hat{\sigma}_x$ และ $\hat{\sigma}_y$ คือ ค่าประมาณส่วนเบี่ยงเบนมาตรฐานของข้อสอบร่วม (v) ข้อสอบฉบับ x และข้อสอบฉบับ y ตามลำดับ (Angoff, 1984, p. 106)

2. ความคลาดเคลื่อนมาตรฐาน ของรูปแบบอิควิเปอร์เซ็นไทล์

การประมาณความคลาดเคลื่อนมาตรฐานของการเทียบมาตราตามวิธีนี้ คำนวณจากสูตรความแปรปรวน ดังนี้

$$\text{Var } X^* = \frac{PQ}{g^2_{\sigma_Y}} \left(\frac{1}{n_\sigma} + \frac{1)}{n_\beta}\right.$$

เมื่อ SEE = $\sqrt{VarX^*}$ (Lord, 1982, cited in Davier, Holland & Thayer, 2003, p. 68)

Var X* คือ ความแปรปรวนของคะแนนแปลง

Q คือ สัดส่วนประชากรที่อยู่ต่ำกว่าคะแนนที่กำหนด สำหรับการประมาณค่า y_0 นี้ แทนด้วยค่าของตัวอย่าง

P คือ 1-Q แทนด้วยค่าของตัวอย่าง

g_0 คือ ความน่าจะเป็นของการแจกแจงของคะแนน X เมื่อกำหนดคะแนน y_0 หาค่าจากข้อมูลเชิงประจักษ์ด้วยสูตร g_0 = Ordinate/ $S.D_x$

เมื่อ Ordinate คือ ค่าที่อ่านจากตารางแจกแจงโค้งปกติ ตามค่า P ที่กำหนด

$S.D_x$ คือ ส่วนเบี่ยงเบนมาตรฐานของคะแนนใน X ที่ศึกษา

3. หาค่าประสิทธิภาพสัมพัทธ์ (Relative efficiency) ของความยาวของแบบสอบร่วมเพื่อเปรียบเทียบขนาดของความคลาดเคลื่อนมาตรฐานของการเทียบมาตรา เมื่อใช้แบบสอบร่วมที่มีขนาด Vx Vy และ Vz เทียบกับ Vx ซึ่งเป็นวิธีที่ให้ค่า SEE น้อยที่สุดตามสมมติฐาน

$$RE_{Vz.\,Vx} = \frac{SEE}{SEE}\,\frac{X^*(Vx)}{X^*(Vz)} \text{ x } 100\%$$

และ

$$RE_{Vy.\,Vx} = \frac{SEE}{SEE}\,\frac{X^*(Vx)}{X^*(Vy)} \text{ x } 100\%$$

ค่าประสิทธิภาพสัมพัทธ์ เป็นค่าที่แสดงอัตราร้อยละของความแม่นยำของการเทียบมาตราที่ใช้แบบสอบร่วม Vy และ Vz เมื่อเทียบกับผลการเทียบของ Vx

4. การวัดความคลาดเคลื่อนของการปรับเทียบคะแนน (Equating error) ใช้สถิติ 2 ตัว คือ ค่าประมาณความลำเอียง (Estimate bias) และค่าประมาณที่ถ่วงน้ำหนักแล้วของรากที่สองของค่าเฉลี่ยยกกำลังสองของความคลาดเคลื่อนในการปรับเทียบคะแนน (Estimated weighted root mean squared error: RMSE) ซึ่งใช้เปรียบเทียบระหว่างวิธีที่แตกต่างกัน คำนวณจากสมการ

$$BIAS = \bar{x}' - \bar{x}$$

เมื่อ $\bar{x}$ คือ ค่าเฉลี่ยของคะแนนดิบของแบบสอบชุดหนึ่ง

$\bar{x}'$ คือ ค่าเฉลี่ยของคะแนนดิบที่เท่าเทียมกันโดยการเทียบแบบสอบชุดนั้นไปยังตัวเอง

และ $\text{RMSE} = \left[\frac{\sum N_i (x_i' - x_i)^2}{\sum N_i}\right]^{\frac{1}{2}}$

เมื่อ x_i คือ คะแนนดิบของคนที่ i

x_i' คือ คะแนนดิบที่เท่าเทียมกัน

N_i คือ จำนวนผู้สอบซึ่งได้คะแนน x_i

ในกรณีที่มีผู้สอบจำนวนมาก การคำนวณ RMSE จะง่ายขึ้นโดยใช้สมการ

$$\text{RMSE} = \left[(m-1)^2 \text{Var}(x) + (\text{BIAS})^2\right]^{\frac{1}{2}}$$

เมื่อ m คือ ความชันของฟังก์ชันการเทียบ $x' = mx + b$

Var(x) คือ ความแปรปรวนของ x

หรือหาได้จากสูตร RMSE (Root mean squared error)

$$\text{RMSE} = \sqrt{(Bias)^2 + SEE^2}$$

ซึ่ง RMSE เป็นดัชนีที่รวมทั้งความคลาดเคลื่อนเชิงสุ่มและความคลาดเคลื่อนเชิงระบบ (Random and systematic error) ไว้ด้วยกัน (Meng, 2012, p. 49)

5. ค่าดัชนีของแองกอฟ เป็นเกณฑ์ที่เสนอโดยแองกอฟ จำแนกเป็น 2 ประเภท คือ

5.1 ดัชนีที่ถ่วงน้ำหนักด้วยความถี่ ประกอบด้วย

5.1.1 ดัชนี RMS (Root mean square) จากสูตร

$$\text{RMS} = \sqrt{\frac{\sum f_i (A_i - B_i)^2}{\sum f_i}}$$

5.1.2 ดัชนี MAD (Mean absolute difference) จากสูตร

$$\text{MAD} = \frac{\sum f_i |A_i - B_i|}{\sum f_i}$$

5.1.3 ดัชนี MSD (Mean signed difference จากสูตร

$$\text{MSD} = \frac{\sum f_i (A_i - B_i)}{\sum f_i}$$

เมื่อ A_i แทน คะแนน i ที่ปรับจากแบบสอบใหม่ให้อยู่ในสเกลกับแบบสอบเดิม

B_i แทน คะแนนจริงหรือคะแนนที่ใช้เป็นเกณฑ์

f_i แทน ความถี่ของคะแนนดิบจากแบบสอบชุดใหม่

5.2 ดัชนีที่ไม่ได้ถ่วงน้ำหนักด้วยความถี่ ประกอบด้วย

5.2.1 ดัชนี RMS_u จากสูตร

$$RMS_u = \sqrt{\frac{\sum(A_i - B_i)^2}{k}}$$

5.2.2 ดัชนี MAD_u จากสูตร

$$MAD_u = \frac{\sum|A_i - B_i|}{k}$$

5.2.3 ดัชนี MSD_u จากสูตร

$$MSD_u = \frac{\sum(A_i - B_i)}{k}$$

(อดิศร ศรีบุญวงศ์, 2545, หน้า 41-42)

6. สัมประสิทธิ์การทำนาย (Coefficient of prediction) สัมประสิทธิ์การทำนายในการทำนายคะแนน y จากคะแนนเชื่อมโยง y* หาได้จากสูตร

$$R = \frac{n\sum(yy^*) - (\sum y)(y^*)}{\sqrt{[n\sum y^2 - (\sum y)^2][n\sum y^{*2} - (\sum y^*)^2]}}$$

ค่า R^2 มีค่าสูง แสดงว่าคะแนนปรับเทียบ y* สามารถทำนายคะแนนเกณฑ์ (y) ได้สูง (ศิริชัย กาญจนวาสี, 2551, หน้า 30)

7. ความเพียงพอของการปรับเทียบคะแนนความคลาดเคลื่อนของการปรับเทียบคะแนน

การปรับเทียบคะแนนแต่ละวิธีจะมีคุณภาพที่ดีที่สุดเมื่อคะแนนที่ได้จากแบบสอบเป็นไปตามเงื่อนไขต่าง ๆ ที่กำหนดไว้ในแต่ละรูปแบบของการปรับเทียบคะแนน แต่ในสถานการณ์จริงอาจมีข้อจำกัดทำให้ไม่สามารถเป็นไปตามเงื่อนไข จึงจำเป็นต้องตรวจสอบความเพียงพอ (Adequacy) ของการปรับเทียบคะแนนซึ่งเป็นการประเมินประสิทธิภาพของการปรับเทียบคะแนนซึ่งมีอยู่หลายวิธี เช่น วิธีดัชนีเปรียบเทียบเปอร์เซ็นไทล์ที่ใช้กลุ่มสอบทานในการหาคุณภาพการปรับเทียบคะแนน ดัชนีตรวจสอบความเพียงพอของเจเกอร์ ดัชนีความแตกต่าง (Discrepancy index) (อุทัยวรรณ พงศ์อร่าม, 2545, หน้า 82-85) รายละเอียด ดังนี้

7.1 วิธีดัชนีเปรียบเทียบเปอร์เซ็นไทล์ (The percentile comparison index)

เป็นการความไม่สอดคล้องระหว่างการแจกแจงคะแนนที่ได้จากกลุ่มตรวจสอบผลที่ทำแบบสอบทั้ง 2 ชุด ที่ทำการแปลงคะแนนให้อยู่ในสเกลเดียวกัน ดัชนีเปรียบเทียบเปอร์เซ็นไทล์หาได้จากค่าเฉลี่ยกำลังสองของความแตกต่าง ที่ได้จากการแจกแจงของคะแนนต่าง ๆ ของ X กับคะแนนแปลง X* ที่แปลงมาจาก Y ณ ตำแหน่งเปอร์เซ็นไทล์เดียวกัน ใช้ตรวจสอบความคลาด

เคลื่อนที่เกิดจากการปรับเทียบคะแนน เพราะเมื่อมีตารางปรับเทียบคะแนนจากแบบสอบ 2 ชุด แล้วนำคะแนนจากกลุ่มสอบทานผลไปปรับสเกลโดยใช้ตารางปรับเทียบเดิม ถ้าการปรับเทียบมีคุณภาพแล้ว ผลต่างระหว่างคะแนนที่ไม่ได้ปรับสเกลกับคะแนนที่ปรับสเกลจะมีค่าน้อย

ดัชนีเปรียบเทียบเปอร์เซ็นไทล์ คำนวณได้จาก

$$C = \sum (X_i - X_i^*)^2 / nk$$

เมื่อ n คือ จำนวนคะแนนดิบในกลุ่มตรวจสอบผล

k คือ จำนวนข้อสอบในแบบสอบร่วมที่ใช้

X คือ คะแนนจากแบบสอบชุด X ซึ่งเป็นคะแนนเกณฑ์

X* คือ คะแนนจากแบบสอบชุด Y ที่แปลงให้อยู่ในมาตราคะแนนของ X

7.2 ดัชนีตรวจสอบความเพียงพอของเจเกอร์ ใช้ตรวจสอบความเพียงพอของการใช้วิธีการปรับเทียบคะแนนเชิงเส้นตรง จากดัชนี 5 ตัว คือ 1) ดัชนีความคล้ายคลึงของการแจกแจงคะแนนสะสมของแบบสอบชุดเก่าและชุดใหม่ ด้วยการใช้การทดสอบ Kolmogorov-sminov two-sample test 2) รูปแบบของการแจกแจงคะแนนดิบกับคะแนนแปลงด้วยวิธีเชิงเส้นตรง 3) ความคงเส้นคงวาของผลลัพธ์ของการปรับเทียบคะแนนตามวิธีเชิงเส้นตรงกับวิธีเปอร์เซ็นไทล์ 4) ความคล้ายคลึงของการแจกแจงความยากของข้อสอบ โดยอาศัยคุณสมบัติความเป็นคู่ขนาน ถ้ามีการเบี่ยงเบนจากความเป็นคู่ขนานมากเท่าใด ต้องใช้วิธีการปรับเทียบคะแนนที่ซับซ้อนขึ้น 5) ความคล้ายคลึงของค่าอำนาจจำแนกของข้อสอบ

7.3 ดัชนีความแตกต่าง (Discrepancy index) ของปีเตอร์สัน มาร์โค และสตีเวอร์ท (Pertersen, Marco & Stewart, 1982, p. 91) เป็นค่าความคลาดเคลื่อนรวมในการปรับเทียบคะแนนโดยใช้คะแนนเกณฑ์ (t) กับคะแนนแปลง (t') จากสูตร

$$C = \frac{\sum f_j d_j^2}{nS_t^2} = \frac{\sum_{j=1}^{n} f_j (d_j - \bar{d})^2}{nS_t^2} + \frac{\bar{d}^2}{S_t^2}$$

เมื่อ $d_j = t_j - t'_j$

t'_j = คะแนนแปลง

t_j = คะแนนเกณฑ์

ตอนที่ 4 งานวิจัยที่เกี่ยวข้องกับการปรับเทียบคะแนน

เพเทียน (Patience, 1990) ได้ศึกษาเปรียบเทียบผลของการปรับเทียบคะแนนระหว่างระดับชั้นโดยใช้วิธีปรับเทียบคะแนน 5 วิธี คือวิธีการตามทฤษฎีการสอบแบบดั้งเดิม 2 วิธี คือ อิควิเปอร์เซ็นไทล์และวิธีของเธอร์สโตนและวิธีการตามทฤษฎีการตอบข้อสอบ 3 วิธี คือ วิธี 1, 2 และ 3 พารามิเตอร์ โดยศึกษาว่าวิธีใดให้ผลการปรับเทียบคะแนนที่เหมาะสมที่สุดแบบสอบที่ใช้แบ่งเป็น 3 ระดับ คือ ง่าย ปานกลาง และยาก ใช้กับนักเรียนระดับ 9, 10 และ 11 ตามลำดับ ผลการวิจัยพบว่า วิธีปรับเทียบคะแนนทั้ง 5 วิธี ให้ผลการเทียบคะแนนที่คล้ายคลึงกันเมื่อใช้กับแบบสอบฉบับที่ง่ายและฉบับปานกลาง ส่วนฉบับที่ยากผลการปรับเทียบคะแนนโดยใช้วิธีอิควิเปอร์เซ็นไทล์ วิธีเธอร์สโตนและวิธี 3 พารามิเตอร์ให้ผลที่คล้ายคลึงกับคะแนนที่ได้จากแบบสอบร่วมมากกว่าวิธี1 และ 2 พารามิเตอร์

กลอวากิ (Glowaki, 1991) ได้ตรวจสอบโมเดลของการปรับเทียบคะแนนที่มีความเหมาะสมกับการสอบของบัณฑิตวิทยาลัย ของมหาวิทยาลัยอลานามา โดยมีปัญหาการวิจัยว่าโมเดลของการปรับเทียบที่ตรวจสอบมีการแจกแจงของคะแนนดิบหรือคะแนนที่ผ่านจากการสอบแบบสอบการอ่านและคณิตศาสตร์แตกต่างกันหรือไม่โดยใช้โมเดลเชิงเส้นตรงอิควิเปอร์เซ็นไทส์และทฤษฎีการตอบข้อสอบชนิด 1 2 และ 3 พารามิเตอร์ ผลการวิจัยพบว่า วิธีการปรับเทียบคะแนนทั้ง 5 วิธีให้ผลคล้ายคลึงกัน แสดงว่าโมเดลทั้งหมดสามารถนำมาใช้กับการปรับเทียบคะแนนได้โดยไม่มีโมเดลที่ดีที่สุด

อเยอร์เว (Ayerve, 1992) ได้ศึกษาเปรียบเทียบประสิทธิภาพของการปรับเทียบคะแนนตามแนวตั้ง ด้วยวิธีอิควิเปอร์เซ็นไทล์ และวิธี IRT โมเดล 3 พารามิเตอร์ โดยใช้การจำลองข้อมูลภายใต้เงื่อนไขของขนาดกลุ่มตัวอย่างตามยาวของแบบสอบ และความยาวของแบบสอบร่วม โดยเปรียบเทียบ 3 พารามิเตอร์ โดยใช้การจำลองข้อมูลภายใต้เงื่อนไขของขนาดกลุ่มตัวอย่างความยาวของแบบสอบ และความยาวของแบบสอบร่วม โดยเปรียบเทียบ 3 กรณี คือ 1) เปรียบเทียบประสิทธิภาพของวิธีการปรับเทียบคะแนนทั้ง 2 ภายใต้เงื่อนไขทุกเงื่อนไข 2) เปรียบเทียบประสิทธิภาพของวิธีการปรับเทียบคะแนนแต่ละวิธีภายใต้เงื่อนไขของตัวแปรอิสระแต่ละตัว 3) ตรวจสอบผลของตัวแปรอิสระแต่ละตัวในการปรับเทียบคะแนนแต่ละวิธี ขนาดกลุ่มตัวอย่างที่ใช้มี 3 ขนาด คือ 200 500 และ 1,000 คน ความยาวของแบบสอบที่ใช้มี 2 ขนาด คือ 30 ข้อ และ 60 ข้อ ส่วนความยาวของแบบสอบร่วมที่ใช้ คือ 5 และ 10 ข้อ ผลการวิจัยพบว่า วิธีการปรับเทียบแบบอิควิเปอร์เซ็นไทล์และ IRT ไม่แตกต่างกันอย่างมีนัยสำคัญ และถ้าพิจารณาการปรับเทียบคะแนนตามวิธี IRT 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดใหญ่ คือ 500 และ 1,000 คน ให้ผลการปรับเทียบที่ถูกต้องกว่า

คาลด์เวลล์ (Caldwell, 1984) ได้ศึกษาการวัดประสิทธิผลสัมพัทธ์ในโมเดลการปรับเทียบความยากของแบบสอบ ด้วยวิธีการปรับเทียบเชิงเส้นตรงและรูปแบบของราสช์โดยใช้แบบสอบร่วม และคาดว่าการปรับเทียบเชิงเส้นตรงจะมีประสิทธิผลน้อยกว่าการปรับเทียบด้วยวิธีของราสช์และคาดว่าประสิทธิผลสัมพัทธ์ตามวิธีของราสช์จะเพิ่มขึ้นเมื่อความแตกต่างของค่าเฉลี่ยเพิ่มขึ้น การปรับเทียบคะแนนจะใช้แบบสอบร่วมสองชุด ชุดหนึ่งมีความยากในระดับปานกลางส่วนอีกชุดหนึ่งมีความยากสูงสุด กรณีการใช้แบบสอบร่วมที่มีความยากสูงสุดได้รับการพิจารณาแล้วว่าเป็นกรณีที่ให้ประสิทธิผลต่ำกว่าการใช้แบบสอบร่วมที่มีความยากปานกลาง ผลการวิจัย พบว่า คะแนนทั้งหมดจากรูปแบบการปรับเทียบของราสช์ดีกว่าการปรับเทียบเชิงเส้นตรง ทั้ง 2 กรณี ไม่ว่าจะใช้แบบสอบร่วมมีความยากในระดับปานกลางหรือมีความยากสูงสุด และเมื่อพิจารณาคะแนนจุดตัดที่แสดงถึงความสามารถต่ำสุดรูปแบบของราสช์ดีกว่าโดยมีความลำเอียงเพียงเล็กน้อย แต่รูปแบบการปรับเทียบเชิงเส้นตรงจะให้ค่าความคลาดเคลื่อนในระดับต่ำ ดัชนีความแตกต่างของการปรับเทียบของราสช์มีค่าเฉลี่ยเกือบเป็นศูนย์และมีการแปรแปลงในช่วงกว้าง ขณะที่การปรับเทียบเชิงเส้นตรงให้ค่าความลำเอียงด้านลบแต่มีการแปรเปลี่ยนที่น้อยกว่า

ฟิลลิปส์ (Phillips, 1986) ได้ศึกษาผลการปรับเทียบคะแนนในแนวตั้ง ที่ลดคนที่ไม่เหมาะสม (Missfitting) ออกโดยใช้รูปแบบการปรับเทียบของราสช์ เพื่อตรวจสอบว่าการตัดคนที่ไม่เหมาะสมออกไปก่อนที่จะนำข้อมูลไปวิเคราะห์ในการปรับเทียบคะแนน จะช่วยในการปรับปรุงข้อมูลให้มีความเหมาะสมกับรูปแบบหรือไม่ โดยการเปรียบเทียบจากค่าอำนาจจำแนกที่เหมาะสมกับรูปแบบ โดยใช้การวิเคราะห์ IRT 2 พารามิเตอร์ทั้งก่อนและหลังการตัดคนที่ไม่เหมาะสมออก หลังจากนั้นทำการตรวจสอบว่าก่อนและหลังการตัดคนที่ไม่เหมาะสมออกมีผลต่อคุณภาพของแบบสอบทั้ง 2 ฉบับหรือไม่ โดยพิจารณาเปรียบเทียบจากค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของค่าประมาณความยากของข้อสอบ และค่าอำนาจจำแนกของข้อสอบตลอดจนค่าความสอดคล้องภายในและจำนวนข้อสอบที่เหมาะสมกับโมเดลระหว่างก่อนและหลังการตัดคนที่ไม่เหมาะสมออก สุดท้ายตรวจสอบผลการปรับเทียบคะแนนโดยใช้วิธีของราสช์และวิธีอิควิเปอร์เซนไทล์ว่าให้ผลแตกต่างกันหรือไม่ โดยการเปรียบเทียบผลการปรับเทียบคะแนนก่อนและหลังการตัดคนที่ไม่เหมาะสมออก โดยใช้ข้อมูลของแบบสอบด้านคณิตศาสตร์และด้านการอ่าน จากนักเรียนเกรด 4 และเกรด 8 ระดับชั้นละ 300 และ 500 คนตามลำดับ ผลการวิจัยพบว่า การตัดคนที่ไม่เหมาะสมออกช่วยในการปรับปรุงข้อมูลให้เหมาะสมกับรูปแบบน้อย ส่วนการใช้วิธีปรับเทียบของราสช์ทั้งก่อนและหลังการตัดคนที่ไม่เหมาะสมออกให้คะแนนสมมูลคล้ายกันเมื่อร้อยละของการตัดคนที่ไม่เหมาะสมออกน้อย ค่าพารามิเตอร์ความยากของข้อสอบทั้งก่อนและหลังการตัดคนที่ไม่เหมาะสมออกมีแนวโน้มต่างกันน้อยกว่าค่าเบี่ยงเบนมาตรฐาน สำหรับผลการปรับเทียบคะแนนทั้ง 2 วิธีให้

ผลแตกต่างกัน การตัดคนที่มีความสามารถต่ำออกอาจลดการเดาแต่บางครั้งอาจทำให้โค้งลักษณะข้อสอบมีสัดส่วนของสเกลที่มีความถูกต้องน้อยลง

รีเบคกา และโวรัค (Rebecca & Dvorak, 2009) ได้เปรียบเทียบวิธีการปรับเทียบคะแนน 2 วิธีระหว่างวิธีเคอเนล และวิธี TCC: Test characteristic curve 2 พารามิเตอร์ โดยการจำลองข้อมูลตามเงื่อนไขของขนาดตัวอย่าง ความยาวของแบบสอบ เปอร์เซ็นต์ของข้อสอบร่วม และค่าเฉลี่ยของน้ำหนักองค์ประกอบ(Factor loading) ที่ต่างกันผลการวิจัยพบว่า ทั้ง2 วิธีให้ผลการปรับเทียบคะแนนดีพอ ๆ กัน และพบว่า เมื่อแบบสอบมีความยาว 75 ข้อ และค่าเฉลี่ยของน้ำหนักองค์ประกอบ = .62 การปรับเทียบด้วยวิธีเคอเนลจะให้ผลที่ถูกต้องมากกว่า วิธี TCC ไม่ว่าขนาดตัวอย่างและเปอร์เซ็นต์ของข้อสอบร่วมจะเป็นเท่าไรก็ตามแต่ถ้าค่าเฉลี่ยของน้ำหนักองค์ประกอบ = .5 เปอร์เซ็นต์ของข้อสอบร่วม 30% และกลุ่มตัวอย่างขนาดใหญ่การปรับเทียบด้วยวิธีTCC จะให้ผลที่ถูกต้องมากกว่า วิธีเคอเนลให้ผลการปรับเทียบที่ถูกต้องในช่วงคะแนนที่สูงกว่าค่าเฉลี่ยขณะที่วิธี TCC ให้ผลการปรับเทียบถูกต้องตลอดทุกช่วงคะแนน โดยภาพรวมของการวิจัยสรุปว่าวิธีเคอเนลให้ผลการปรับเทียบคะแนนดีกว่าวิธี TCC

เม็ง (Meng, 2012) ได้เปรียบเทียบวิธีการปรับเทียบคะแนน ระหว่างวิธีเคอเนล (Kernel post-stratification equating, Stocking-lord and mean/ sigma) และการปรับเทียบวิธี IRT-3 PL ด้วยการจำลองข้อมูล และการออกแบบเก็บรวบรวมข้อมูลโดยใช้ข้อสอบร่วมกับกลุ่มไม่เท่าเทียมกัน (Non-equivalent anchor test) กับกลุ่มตัวอย่างขนาด 200 500 และ 2,000 คน ตามลำดับ และใช้แบบสอบความยาว 40 ข้อ โดยมีข้อสอบร่วม 5 ข้อ และ 15 ข้อ ผลการวิจัยพบว่าความคลาดเคลื่อนของการปรับเทียบจะลดลงเมื่อใช้ข้อสอบร่วมและกลุ่มตัวอย่างที่เพิ่มขึ้นไม่ว่าจะปรับเทียบด้วยวิธีใด และกลุ่มตัวอย่างที่มีความสามารถที่ต่างกันจะทำให้ความความคลาดเคลื่อนของการปรับเทียบคะแนนสูงขึ้น โดยภาพรวมในทุกสถานการณ์ของการปรับเทียบคะแนนวิธีเคอเนลจะให้ความคงที่ในการปรับเทียบคะแนนมากกว่าแต่ก็ไม่ได้ถูกต้องมากไปกว่าวิธี IRT สำหรับการหาความคลาดเคลื่อนของการปรับเทียบคะแนน งานวิจัยส่วนใหญ่นิยมหาค่า RMSE (Root mean squared error)

ก็อดฟรีย์ และคิลลี (Godfrey & Kelly, 2007) ได้เปรียบเทียบวิธีการปรับเทียบคะแนนจริงระหว่างวิธีเคอเนล (Kernel) กับวิธี IRT 2PL โดยทำการเปรียบเทียบจากวิธีการปรับเทียบคะแนน 4 วิธี คือ Kernel, Traditional chained equipercentile, Concurrent calibration, and Stocking and lord transformation ด้วยการจำลองข้อมูล และใช้แบบสอบที่มีความยาวตั้งแต่ 20 ข้อ 60 ข้อ ถึง 100 ข้อ และใช้ข้อสอบร่วมขนาด 20% 35% และ 50% มีเฉพาะแบบสอบขนาด 20 ข้อ ที่ใช้ข้อสอบร่วมขนาด 20% โดยใช้แบบแผนกลุ่มตัวอย่างเดียว (Single group design) กับรูปแบบกลุ่มไม่เท่า

เทียมกันโดยใช้ข้อสอบร่วม (NEAT) ขนาด 1,000 10,000 100,000 คน ตามลำดับ ผลการวิจัยพบว่า เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น การปรับเทียบแต่ละวิธีจะให้ผลไม่ต่างกันจำนวนข้อสอบร่วมที่มากขึ้นจะทำให้การปรับเทียบมีความถูกต้อง ความยาวของข้อสอบร่วม และจำนวนข้อของแบบสอบที่น้อยกว่า 60 ข้อ ส่งผลต่อวิธีการปรับเทียบด้วยวิธี Stocking and lord โดยส่วนใหญ่แล้วการปรับเทียบด้วยวิธี Kernel จะให้ผลที่ถูกต้องดีกว่าวิธีอื่น ๆ สำหรับวิธี Chained equipercentile จะให้ผลการปรับเทียบที่ถูกต้องเมื่อกลุ่มตัวอย่างมีขนาดมากกว่า 1,000 คน

ฮอล์แลนด์ และซินฮาเรย์ (Holland & Sinharay, 2007) ได้พัฒนาการปรับเทียบคะแนนแบบโพสแตรทิฟิเคชั่นด้วยการใช้ข้อสอบร่วมที่มีความยากปานกลางผลการวิจัยพบว่า วิธีการนี้ช่วยลดความคลาดเคลื่อนในการปรับเทียบคะแนนได้ดี แต่มีข้อจำกัดว่าการสร้างข้อสอบร่วมที่มีความยากปานกลางทำได้ยาก เพราะข้อสอบที่มีความยากปานกลางอาจมีไม่เพียงพอ

ภัทราพร เกษสังข์ (2546) ได้ศึกษาพัฒนาการความสามารถทางคณิตศาตร์ของนักเรียน และศึกษาประสิทธิภาพของการเชื่อมโยงคะแนนตามแนวตั้ง ด้วยวิธีทฤษฎีตอบสนองข้อสอบแบบโลจิสติก 2 และ 3 พารามิเตอร์ที่ใช้แบบสอบร่วมภายในและแบบสอบร่วมภายนอกที่มีความยากต่างกัน คือ ระดับยากกับปานกลาง โดยใช้ข้อสอบร่วมขนาด 5 10 และ 15 ข้อ ของแบบสอบร่วม กับกลุ่มตัวอย่างชั้นมัธยมศึกษาปีที่ 1, 2 และ 3 ผลการวิจัยพบว่า ความคลาดเคลื่อนมาตรฐานของการเชื่อมโยงคะแนนของนักเรียนที่มีความสามารถแตกต่างกันทั้ง 3 ระดับ มีค่าระหว่าง 0.1011-0.1830 การเชื่อมโยงคะแนนตามแนวตั้งด้วยวิธีทฤษฎีตอบสนองข้อสอบแบบโลจิสติก 2 พารามิเตอร์ ที่มีแบบสอบร่วมภายใน 15 ข้อ ที่มีความยากเฉลี่ยระดับยากมาก ให้ค่าความคลาดเคลื่อนมาตรฐานของการเชื่อมโยงคะแนนต่ำสุด รวมทั้งมีความเพียงพอของการเชื่อมโยงคะแนนดีที่สุด

อดิศร ศรีบุญวงษ์ (2545) ได้พัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ มีทั้งหมด 7 เกณฑ์ คือ เกณฑ์รวมทุกเงื่อนไข เกณฑ์สำหรับโมเดล 1 พารามิเตอร์ เกณฑ์สำหรับโมเดล 3 พารามิเตอร์ เกณฑ์สำหรับการใช้กลุ่มสมมูล เกณฑ์สำหรับการใช้แบบสอบร่วม เกณฑ์สำหรับการปรับเทียบกลับสู่แบบสอบเดิม และเกณฑ์สำหรับการใช้กลุ่มสอบทานผล แต่ละเกณฑ์เป็นกลุ่มค่าดัชนี AMD, MAD และ RMS ที่บอกระดับคุณภาพการปรับเทียบคะแนน 4 ระดับ คือ การปรับเทียบน่าพอใจอย่างยิ่ง การปรับเทียบน่าพอใจ การปรับเทียบไม่น่าพอใจ และการปรับเทียบไม่น่าพอใจอย่างยิ่ง สำหรับผลการตัดสินคุณภาพการปรับเทียบคะแนนเมื่อใช้เกณฑ์ที่พัฒนาขึ้นเปรียบเทียบกับผลการใช้ตามเกณฑ์ของปิเตอร์เซน และคณะ พบว่าไม่สอดคล้องกัน แต่เมื่อเทียบกับเกณฑ์ความเสมอภาคของลอร์ด ปรากฏว่าผลการตัดสินคุณภาพการปรับเทียบมีความสอดคล้องกัน

อุทัยวรรณ พงศ์อร่าม (2545) ศึกษาขนาดกลุ่มตัวอย่างที่เหมาะสมสำหรับการปรับเทียบคะแนนด้วยวิธีอิควิเปอร์เซ็นไทล์ และวิธีเส้นตรงตามแบบจำลองคะแนนจริงสัมพันธ์ ที่มีแบบแผนการปรับเทียบและความยาวของแบบสอบแตกต่างกัน กลุ่มตัวอย่างเป็นนักเรียนชั้น ม.1 สังกัดกรมสามัญศึกษา จังหวัดชุมพร จำนวน 3,000 คน โดยใช้แบบสอบวัดผลสัมฤทธิ์วิชาคณิตศาสตร์ ค.101 ที่มีความยาวแบบสอบ 4 ขนาด คือ 30 40 50 และ 60 ข้อ โดยใช้ข้อสอบร่วมเรียงตามลำดับแบบสอบ คือ 8 10 13 และ 15 ข้อ กับกลุ่มตัวอย่างขนาด 7 ขนาด คือ 30 50 100 200 300 400 และ 500 คน ผลการวิจัยพบว่า ขนาดของกลุ่มตัวอย่างที่เหมาะสมในการปรับเทียบแต่ละวิธี ดังนี้

วิธี	ความยาวของแบบสอบ	ขนาดตัวอย่างที่เหมาะสม
เชิงเส้นตรงแบบแผนกลุ่มสมมูล	60, 50, 40	100
	30	200
อิควิเปอร์เซ็นไทล์แบบแผนกลุ่มสมมูล	60	200
	50, 40	300
	30	400
เชิงเส้นตรงแบบแผนข้อสอบร่วม	60, 50, 40	100
	30	300
อิควิเปอร์เซ็นไทล์แบบแผนข้อสอบร่วม	60, 50	200
	40, 30	500

จะเห็นว่าการปรับเทียบคะแนนด้วยวิธีเชิงเส้นตรงและอิควิเปอร์เซ็นไทล์ ทั้งแบบแผนกลุ่มสมมูลและแบบแผนข้อสอบร่วมในแต่ละความยาวของแบบสอบ วิธีเชิงเส้นตรงใช้กลุ่มตัวอย่างที่น้อยกว่า

พิชัย ละแมนชัย (2538) ได้ศึกษาเพื่อหาขนาดกลุ่มตัวอย่างขั้นต่ำที่ทำให้การปรับเทียบคะแนนในแนวระดับ ตามแนวทฤษฎีการตอบข้อสอบ โมเดลหนึ่ง และสามพารามิเตอร์ในแบบแผนกลุ่มสมมูลและแบบแผนข้อสอบร่วม ที่มีความยาวแบบสอบต่าง ๆ กัน คือ 30 50 70 90 110 130 และ 150 ข้อ เพื่อให้ผลการปรับเทียบคะแนนในแนวระดับเกิดประสิทธิภาพข้อมูลที่ใช้ในการวิจัยจำลองจากโปรแกรม IRTDATA เกณฑ์ที่ใช้ในการตัดสินคือการหาค่าดัชนีความแตกต่างที่อยู่ในระดับน่าพอใจอย่างมาก พบว่า โมเดลการตอบสนองข้อสอบรายข้อ 1 และ 3 พารามิเตอร์ทั้งแบบแผนกลุ่มสมมูลและแบบแผนข้อสอบร่วม ขนาดกลุ่มตัวอย่างขั้นต่ำในแต่ละความยาวแบบสอบต่างกัน โดยโมเดล 1 พารามิเตอร์ใช้ขนาดกลุ่มตัวอย่างขั้นต่ำน้อยกว่า และโมเดล

การตอบสนองรายข้อที่ใช้ 1 และ 3 พารามิเตอร์ทั้งแบบแผนกลุ่มสมมูลและแบบแผนข้อสอบร่วม ขนาดกลุ่มตัวอย่างขั้นต่ำในแต่ละความยาวแบบสอบแตกต่างกัน โดยแบบแผนข้อสอบร่วมใช้ขนาดกลุ่มตัวอย่างขั้นต่ำน้อยกว่า

จากการศึกษาจะเห็นได้ว่ายังไม่มีรูปแบบการปรับเทียบคะแนนรูปแบบใดที่ให้ผลในการประยุกต์ใช้ได้ดีที่สุดกับทุกสถานการณ์ที่กำหนด ผลการวิจัยให้ข้อค้นพบในเชิงประจักษ์เพียงเป็นแนวทางในการเลือกใช้ให้ใกล้เคียงเหมาะสมกับสภาพที่ต้องการ ดังนั้นเพื่อเป็นการพัฒนาความรู้เกี่ยวกับการปรับเทียบคะแนน ให้สารสนเทศแก่ผู้ตัดสินใจและเพื่อให้เกิดความยุติธรรมในระบบการทดสอบ ผู้วิจัยจึงเลือกใช้วิธีการปรับเทียบ 2 วิธี คือ วิธีเคอเนลที่ใช้ได้ดีกับกลุ่มตัวอย่างขนาดเล็กและวิธีตามทฤษฎีการตอบข้อสอบ 2 พารามิเตอร์ที่ใช้ได้ดีกับกลุ่มตัวอย่างขนาดใหญ่และให้ความถูกต้องของการปรับเทียบที่ดี ซึ่งสอดคล้องกับบริบทของสถานศึกษาที่ผู้วิจัยเลือกนำมาศึกษา เพราะวิชาที่เปิดสอน มีกลุ่มนักศึกษาเข้าสอบแตกต่างกันไป ตั้งแต่ขนาดเล็กจนถึงขนาดใหญ่มาก ๆ และเลือกใช้รูปแบบข้อสอบร่วมที่แตกต่างกันตามลักษณะการเป็นตัวแทนของเนื้อหา และความยากปานกลางของข้อสอบร่วมตามแนวคิดของ Kolen and Brennen และของ Holland and Sinharay

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการปรับเทียบคะแนน จากแบบสอบต่างฉบับที่วัดในเนื้อหาเดียวกัน มีโครงสร้างและรูปแบบแบบสอบอย่างเดียวกัน แต่สอบต่างเวลากันของชุดวิชาชุดหนึ่ง ประกอบด้วย 15 หน่วยการเรียนรู้ที่ผู้วิจัยเป็นผู้ดูแลอยู่ ด้วยการใช้วิธีการปรับเทียบ 2 วิธี ออกแบบแผนการเก็บรวบรวมข้อมูลรูปแบบกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบร่วมภายในจำนวน 15 ข้อ ที่มีรูปแบบแตกต่างกัน 2 รูปแบบ กับกลุ่มตัวอย่างที่มีขนาดแตกต่างกัน 3 ขนาด จากนั้นจึงปรับเทียบคะแนนด้วยการวิเคราะห์จากแบบสอบทั้งฉบับกับตัดข้อสอบที่ไม่มีคุณภาพทิ้ง วิธีการใดให้ผลการปรับเทียบที่มีคุณภาพ จะนำมาเปรียบเทียบผลการตัดเกรดระหว่างการตัดเกรดจากคะแนนก่อนการปรับเทียบคะแนนและการตัดเกรดจากคะแนนหลังการปรับเทียบคะแนน โดยใช้ข้อมูลจริงจากการสอบ มีการออกแบบและขั้นตอนในการศึกษา ประกอบด้วย วิธีดำเนินการ ประชากรและกลุ่มตัวอย่าง เครื่องมือที่ใช้ในการวิจัย การวิเคราะห์ข้อมูล รายละเอียด ดังนี้

วิธีดำเนินการ

ขั้นตอนการดำเนินงาน

การวิจัยครั้งนี้มีขั้นตอนการดำเนินการวิจัยตามลำดับ ดังนี้

1. ศึกษาค้นคว้าเกี่ยวกับทฤษฎีและหลักการของการปรับเทียบคะแนนระหว่างแบบสอบและงานวิจัยที่เกี่ยวข้องจากเอกสาร หนังสือ วารสาร และงานวิจัยทั้งภายในและต่างประเทศ

2. กำหนดกรอบการวิจัย ตั้งสมมติฐาน ออกแบบการเก็บรวบรวมข้อมูล ที่จะใช้ในการปรับเทียบคะแนน วางแผนจัดทำแบบสอบตามเงื่อนไขที่กำหนด

3. ดำเนินการสอบ/ ตรวจคะแนน บันทึกไฟล์ข้อมูลผลคำตอบของนักศึกษานำมาใช้ในการวิเคราะห์ข้อมูล

4. วิเคราะห์ข้อมูลการปรับเทียบคะแนน หาคุณภาพการปรับเทียบคะแนนวิธีต่าง ๆ

5. ตัดเกรด นำผลคะแนนที่ได้หลังการปรับเทียบคะแนนและคะแนนก่อนการปรับเทียบคะแนน มาตัดเกรดด้วยการใช้วิธีการวิเคราะห์จากแบบสอบทั้งฉบับกับตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

6. เปรียบเทียบผลการตัดเกรด พิจารณาความสอดคล้องของการตัดเกรดจากสัมประสิทธิ์แคปปา

ตัวแปรที่ใช้ในการวิจัย

การวิจัยครั้งนี้ มุ่งศึกษาใน 2ประเด็นหลัก คือ การปรับเทียบคะแนน กับการตัดเกรด มีรายละเอียด ดังนี้

1. การปรับเทียบคะแนน ประกอบด้วย ตัวแปร ดังนี้

1.1 ตัวแปรอิสระ ได้แก่

1.1.1 วิธีการปรับเทียบ 2 วิธี คือ วิธีตามทฤษฎีแนวใหม่ IRT 2 พารามิเตอร์ กับวิธีเคอเนล

1.1.2 รูปแบบของข้อสอบร่วม มี 2 รูปแบบ คือ สุ่มข้อสอบจากทุกหน่วยการเรียน จำนวน 15 ข้อโดยกำหนดค่าความยาก (ค่า *p*) ระหว่าง.4-.6 กับสุ่มอย่างง่ายโดยใช้ไม่กำหนดค่าความยาก

1.1.3 รูปแบบข้อมูลที่จะนำมาวิเคราะห์ มี 2 รูปแบบ คือ ใช้ข้อสอบทั้งฉบับ จำนวน 120 ข้อ มาวิเคราะห์ กับตัดข้อสอบบางข้อที่ไม่มีคุณภาพทิ้ง โดยพิจารณาจากผลการวิเคราะห์ข้อสอบรายข้อของภาคการศึกษาที่ 1/ 2556 ที่ใช้เป็นฐานในการปรับเทียบ เมื่อตัดข้อสอบข้อหนึ่งข้อใดที่ไม่ได้คุณภาพออก ก็จะไปตัดข้อที่ตรงกันของข้อสอบฉบับที่จะนำมาปรับให้เท่ากับฉบับของภาคการศึกษาที่ 1/ 2556

1.1.4 ขนาดกลุ่มตัวอย่างในการวิเคราะห์ มี 3 ขนาด คือ กลุ่มตัวอย่างขนาดเล็ก (100 คน) กลุ่มตัวอย่างขนาดกลาง (500 คน) และกลุ่มตัวอย่างขนาดใหญ่ (700 คน)

1.2 ตัวแปรตาม คุณภาพของการปรับเทียบคะแนน

2. การตัดเกรด ประกอบด้วย ตัวแปร ดังนี้

2.1 ตัวแปรอิสระ ได้แก่

2.1.1 วิธีการตัดเกรด 2 วิธี คือ แบ่งการตัดเกรดเป็น 3 เกรด กับแบ่งการตัดเกรดเป็น 8 เกรด ดังนี้

การตัดเกรด 3 เกรด กำหนดระดับ และคะแนน ดังนี้

เกรด	ระดับ	คะแนน
H	4.00	มากกว่า 75%
S	2.33	60-75%
U	0	ต่ำกว่า 60%

การตัดเกรด 8 เกรด กำหนดระดับ และคะแนน ดังนี้

เกรด	ระดับ	คะแนน
A	4.00	76-100%
B^{+}	3.50	70-75%

B	3.00	65-69%
C^+	2.50	60-64%
C	2.00	55-59%
D^+	1.50	50-54%
D	1.00	45-49%
F	0.00	ต่ำกว่า 45%

สำหรับการตัดเกรดโดยใช้คะแนนจากการปรับเทียบวิธีทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์ จะใช้การตัดเกรดจากคะแนนจริง (True score equating) จากสูตร (Crocker & Algina, 1986, p. 472)

$$T = \sum_{g=1}^{G} P_g(\theta)$$

เมื่อ $P_g(\theta) = \frac{e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}}$

การตัดเกรดจากการปรับเทียบวิธีทฤษฎีการตอบข้อสอบ จะตัดเกรดโดยอิงจากคะแนนจริง (True score equating) ที่เทียบเคียงมาจากคะแนนร้อยละ ในการตัดเกรด 3 เกรด และ 8 เกรด

2.1.2 คะแนนที่ใช้ในการตัดเกรด 2 แบบ คือ คะแนนก่อนการปรับเทียบกับคะแนนหลังจากการปรับเทียบแล้ว

2.2 ตัวแปรตาม ได้แก่ ความสอดคล้องของการตัดเกรด

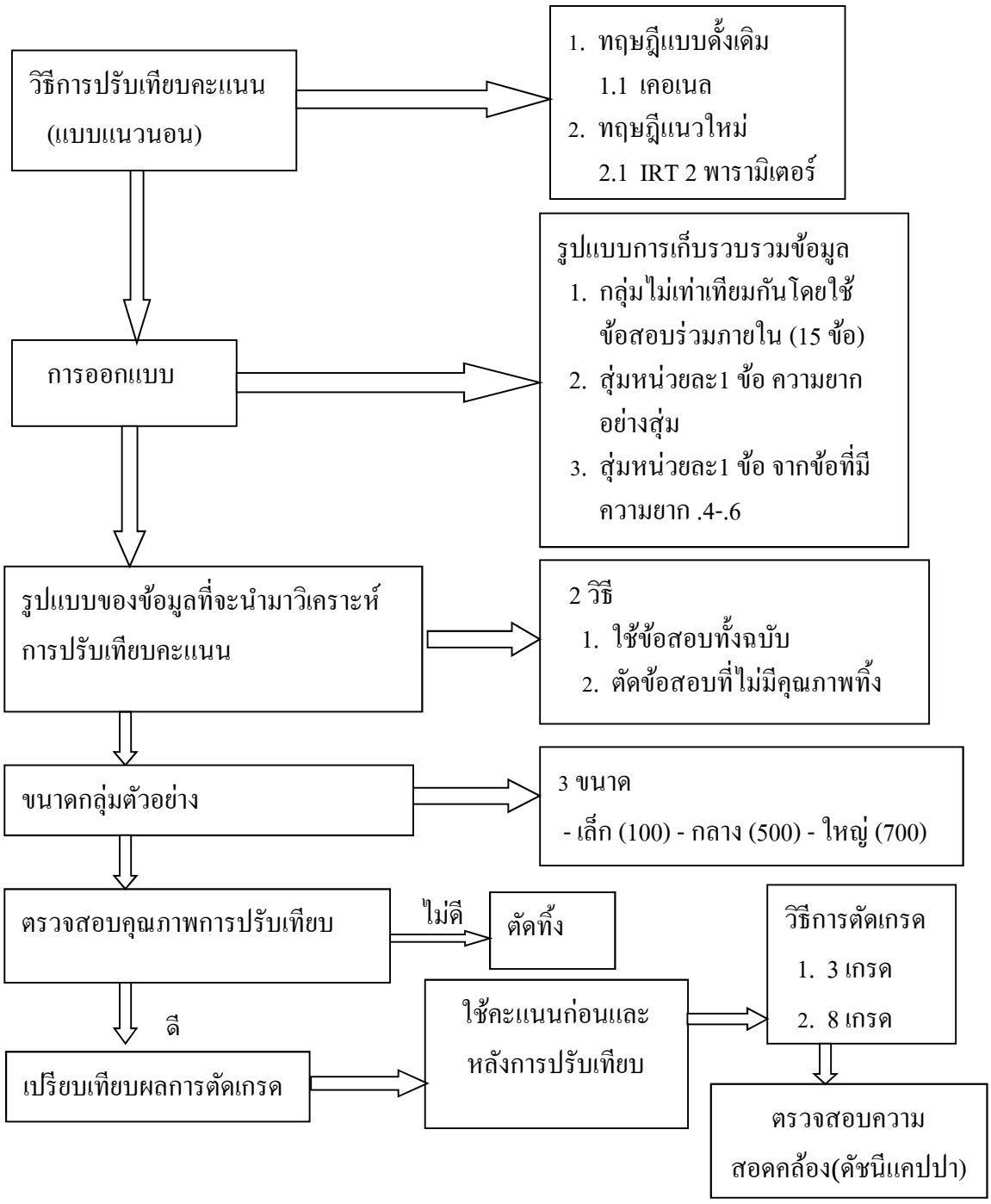
แบบแผนการเก็บรวบรวมข้อมูล

การปรับเทียบคะแนนครั้งนี้ได้ออกแบบแผนการเก็บรวบรวมข้อมูล โดยใช้แบบแผนกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบร่วม (Nonequivalent groups with anchor test: NEAT) ผู้วิจัยกำหนดกลุ่มผู้สอบเป็น 2 กลุ่ม รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้แบบสอบร่วมภายในดังตารางที่ 3-1

ตารางที่ 3-1 รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้แบบสอบร่วมภายใน

กลุ่มตัวอย่าง	**แบบสอบปรับเทียบคะแนน**	
	ฉบับที่ 1	ฉบับที่ 2
P1	XV	-
P2	-	VY

การดำเนินการปรับเทียบคะแนนภายใต้วิธีการปรับเทียบที่แตกต่างกัน จากเงื่อนไขต่าง ๆ ที่กำหนด จนกระทั่งถึงขั้นตอนของการตัดเกรดและตรวจสอบความสอดคล้องของนำผลคะแนนที่ได้หลังการปรับเทียบคะแนนและคะแนนก่อนการปรับเทียบคะแนน มาตัดเกรดตามเงื่อนไขที่แตกต่างกันตามระดับของการตัดเกรด แสดงได้ดังภาพ



ภาพที่ 3-1 ขั้นตอนการปรับเทียบคะแนนจนกระทั่งถึงการตรวจสอบความสอดคล้องของการตัดเกรด

การวิจัยครั้งนี้จะศึกษาจากสถานการณ์จริง จากคะแนนแบบสอบวัดผลสัมฤทธิ์ทางการเรียนตามหลักสูตรระดับปริญญาตรีของชุดวิชาหนึ่ง ศึกษาจากผลคะแนนนักศึกษาทั้งหมดที่ทำแบบสอบต่างฉบับของภาคการศึกษาที่ 1/ 2556 จำนวน 1,210 คน ภาคการศึกษาที่ 1/ 2557 จำนวน 927 คน และภาคการศึกษาที่1/ 2558 จำนวน 857 คน โดยใช้แบบสอบภาคการศึกษาที่ 1/ 2556 เป็นฐานในการปรับเทียบคะแนน โดยแบ่งนักศึกษาทั้งหมดออกเป็น 3 กลุ่ม อย่างสุ่ม เพื่อศึกษาการปรับเทียบและตัดเกรดตามเงื่อนไขต่าง ๆ ที่กำหนด ดังนี้

กลุ่มที่ 1 เป็นกลุ่มขนาดเล็ก จำนวน 100 คน

กลุ่มที่ 2 เป็นกลุ่มขนาดกลาง จำนวน 500 คน

กลุ่มที่ 3 เป็นกลุ่มขนาดใหญ่ จำนวน 700 คน

การสุ่มตัวอย่างขนาด 100 คน และ 500 คน อย่างสุ่มจากจำนวนผู้สอบในแต่ละภาคจะไม่มีการนับซ้ำ ยกเว้นตัวอย่างขนาด 700 คน เพราะจำนวนผู้สอบในแต่ละภาคการศึกษามีไม่เพียงพอ

เครื่องมือที่ใช้ในการวิจัย

แบบสอบวัดผลสัมฤทธิ์ทางการเรียนตามหลักสูตรระดับปริญญาตรีของชุดวิชาหนึ่ง เป็นแบบสอบปรนัย 5 ตัวเลือก จำนวน 120 ข้อ ที่เป็นตัวแทนจากเนื้อหาสาระในชุดวิชา จำนวน 15 หน่วยการเรียน ๆ ละ 8 ข้อ ที่สุ่มข้อสอบมาจากระบบคลังข้อสอบที่มีรูปแบบโครงสร้างในการสุ่มข้อสอบอย่างเดียวกัน ที่ใช้สอบในภาคการศึกษาที่ 1/ 2556 ภาคการศึกษาที่ 1/ 2557 และภาคการศึกษาที่ 1/ 2558 ใช้เวลาในการสอบ 3 ชั่วโมง แบบสอบของภาคการศึกษาที่ 1/ 2556 ใช้เป็นฐานสำหรับให้แบบสอบฉบับอื่น ๆ ปรับสู่ฉบับนี้ สำหรับข้อสอบร่วมได้มาจากการสุ่มข้อสอบภาคการศึกษาที่ 1/ 2556 ทำเป็นข้อสอบร่วมภายใน จัดเรียงลำดับข้อสอบร่วมในตำแหน่งที่ตรงกันทั้ง 2 ฉบับ มีขั้นตอนในการสร้าง ดังนี้

1. สุ่มข้อสอบจากระบบคลังข้อสอบ ที่ผ่านการตรวจสอบคุณภาพจากการวิเคราะห์ข้อสอบรายข้อด้วยระบบแบบดั้งเดิม (Classical test theory) และปรับปรุงคุณภาพจากคณะกรรมการที่เป็นผู้เชี่ยวชาญด้านเนื้อหาและตรวจสอบความถูกต้องตามหลักการสร้างข้อสอบ ความเป็นปรนัยของข้อคำถามโดยนักวัดผลการศึกษา

2. สุ่มข้อสอบตามโครงสร้างในการจัดฉบับ จำนวน 120 ข้อ จากทุกหน่วยการเรียนรู้ หน่วยละเท่า ๆ กัน ตามเกณฑ์การสุ่มของแต่หน่วย หน่วยละ 8 ข้อ 15 หน่วย รวม 120 ข้อ ดังตัวอย่าง

หน่วยที่ 1

ตอนที่	วัตถุประสงค์ข้อ	จำนวนข้อ
1	1	1
1	2, 3	1
2	1	2
2	2	1
2	3	1
3	1, 2	1
3	3	1

หน่วยอื่น ๆ ก็จะมีการกำหนดในลักษณะเดียวกัน ตามแผนผังการสร้างข้อสอบจนครบทุกหน่วย

1. จัดทำข้อสอบร่วมภายใน จำนวน 15 ข้อ จากข้อสอบของภาคการศึกษาที่ 1/ 2556 โดยมีรูปแบบที่แตกต่างกัน 2 รูปแบบ คือ

1.1 สุ่มข้อสอบร่วมอย่างง่ายจากทุกหน่วยในชุดวิชา ที่มีค่าความยากของข้อสอบร่วมอยู่ระหว่าง .4-.6 ซึ่งมีความยากปานกลาง หน่วยละ 1 ข้อ

1.2 สุ่มข้อสอบร่วมอย่างง่ายจากทุกหน่วยในชุดวิชา หน่วยละข้อ จำนวน 15 ข้อ โดยไม่กำหนดค่าความยากให้เป็นไปอย่างสุ่ม

จากนั้นนำข้อสอบร่วมเหล่านี้ไปแทนในแบบสอบของภาคการศึกษาที่ 1/ 2557 และภาคการศึกษาที่ 1/ 2558 ในตำแหน่งเลขข้อที่ตรงกันกับของภาคการศึกษาที่ 1/ 2556 ตามเงื่อนไขที่กำหนด สำหรับรายละเอียดของรูปแบบแบบสอบเป็น ดังนี้

ภาคการศึกษา	ข้อสอบร่วม (15 ข้อ)	ข้อสอบที่เหลือ	รวม
1/ 2556	V_1V_2	X	120 ข้อ
1/ 2557	V_1	Y	120 ข้อ
1/ 2558	V_2	Z	120 ข้อ

2. นำแบบสอบที่ได้ไปสอบตามสถานการณ์จริง ที่มีการบริหารการสอบ มีลำดับขั้นตอนการควบคุมการสอน ระยะเวลาการสอบเหมือนกันทุกครั้ง ต่างกันเพียงสถานที่สอบ

การเก็บรวบรวมข้อมูล

ข้อมูลในการวิจัยนี้เก็บรวบรวมข้อมูลจากสถานการณ์จริงของการสอบวัดผลสัมฤทธิ์ปลายภาคเรียนของนักศึกษาระดับปริญญาตรีของชุดวิชาชุดหนึ่งที่ผู้วิจัยนำมาศึกษา โดยเป็นผลการสอบภาคการศึกษาที่ 1/ 2556 ภาคการศึกษาที่ 1/ 2557 และภาคการศึกษา 1/ 2558 ของชุดวิชา

ชุดหนึ่ง ที่บริหารจัดการการทำแบบสอบจนกระทั่งถึงการบริหารการสอบที่เป็นไปในรูปแบบอย่างเดียวกัน ยกเว้นสถานที่สอบและวันที่สอบ

การวิเคราะห์ข้อมูล

1. วิเคราะห์คุณภาพข้อสอบรายข้อโดยใช้ทฤษฎีการทดสอบแบบดั้งเดิม เพื่อใช้ในการกำหนดข้อสอบร่วมที่มีค่าความยากอยู่ระหว่าง .4-.6 และใช้ในการตัดข้อที่ไม่มีคุณภาพทิ้งก่อนที่จะทำการปรับเทียบคะแนน ตามเงื่อนไขหนึ่งของการปรับเทียบคะแนนที่กำหนด

2. ตรวจสอบข้อตกลงเบื้องต้นของการปรับเทียบคะแนนตามแนวทฤษฎีการตอบข้อสอบ (Item response theory)

3. วิเคราะห์ข้อมูล โดยใช้โปรแกรมในการปรับเทียบคะแนน ตามวิธีเคอเนล (Kernel) และวิธีตามแนวทฤษฎีการตอบข้อสอบ (Item response theory) 2 พารามิเตอร์ โดยใช้วิธีการปรับค่าพารามีเตอร์พร้อมกัน ภายใต้เงื่อนไขที่กำหนดตามรูปแบบข้อสอบร่วม วิธีการวิเคราะห์ และขนาดของกลุ่มตัวอย่างจากนั้นหาคุณภาพการปรับเทียบ

4. หาคุณภาพของการปรับเทียบคะแนน ด้วยการหาความคลาดเคลื่อนของการปรับเทียบคะแนน

5. ตัดเกรดตามรูปแบบการตัดเกรดที่กำหนดเป็น 3 ระดับ และ 8 ระดับ จากคะแนนก่อนการปรับเทียบคะแนน และคะแนนหลังจากการปรับเทียบคะแนน จากวิธีการปรับเทียบที่มีคุณภาพตามเงื่อนไขการปรับเทียบคะแนนที่มีรูปแบบการสุ่มข้อสอบร่วมต่างกัน 2 รูปแบบ วิธีการวิเคราะห์ที่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง กับใช้ข้อสอบทั้งหมด และขนาดกลุ่มตัวอย่างที่แตกต่างกัน 3 ขนาด

6. ตรวจสอบความสอดคล้องของการตัดเกรด จากคะแนนก่อนการปรับเทียบคะแนน และหลังจากการปรับเทียบคะแนน ด้วยรูปแบบการตัดเกรดที่แตกต่างกัน 2 วิธี คือการตัดเกรด 3 ระดับ กับ 8 ระดับ หลังจากนั้นจึงตรวจสอบความสอดคล้องของการตัดเกรดจากค่าสัมประสิทธิ์แคปปา (Coefficient kappa) จากเกณฑ์ของแลนดิส (Landis & Koch, 1977, pp. 159-174) ดังนี้

0.0-0.20	สอดคล้องกันน้อยมาก
0.21-0.40	สอดคล้องกันพอสมควร
0.41-0.60	สอดคล้องกันปานกลาง
0.61-0.80	สอดคล้องกันมาก
0.81-1.00	สอดคล้องกันมากที่สุด

ซึ่งการตัดสินความสอดคล้องของการตัดเกรดจะใช้ ค่าแคปปา ตั้งแต่ .81ขึ้นไป จึงจะถือว่าวิธีการปรับเทียบตามเงื่อนไขนั้น ๆ คะแนนก่อนการปรับเทียบคะแนนกับคะแนนหลังการปรับเทียบ

คะแนน เมื่อนำมาใช้ในการตัดเกรดแล้วมีความสอดคล้องกันไม่จำเป็นจะต้องทำการปรับเทียบคะแนน แต่ถ้าค่าแคปปาต่ำกว่านี้จะต้องทำการปรับเทียบคะแนนก่อนที่จะตัดเกรด

7. หาความสัมพันธ์และระดับของความสัมพันธ์ของการตัดเกรดหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและ IRT 2 พารามิเตอร์ ด้วยการทดสอบไคสแควร์และหาค่าแคปปาตามลำดับ

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การศึกษาการปรับเทียบคะแนนในครั้งนี้ เป็นการศึกษาวิธีการปรับเทียบด้วยวิธีเคอเนล และวิธี IRT ภายใต้เงื่อนไขที่กำหนด กับการตัดเกรดว่าระหว่างการนำคะแนนที่ไม่มีการปรับเทียบคะแนนแล้วตัดเกรด กับการนำคะแนนที่ได้หลังจากการปรับเทียบคะแนนภายใต้เงื่อนไขต่าง ๆ แล้วมาตัดเกรด จะทำให้ผลการตัดเกรดสอดคล้องกันหรือไม่ ซึ่งการศึกษาดังกล่าวเป็นการศึกษาโดยใช้ข้อมูลจริงจากการสอบไม่ได้ใช้วิธีการจำลองข้อมูล ดังที่งานวิจัยส่วนใหญ่นิยมนำมาศึกษาในเรื่องของการปรับเทียบคะแนน เพื่อมุ่งประโยชน์ในการนำไปประยุกต์ใช้ในสถานการณ์จริง ผลการวิเคราะห์ข้อมูลแบ่งการนำเสนอออกเป็น 4 ตอน ดังนี้

ตอนที่ 1 ผลการวิเคราะห์ข้อมูลพื้นฐานของแบบสอบที่นำมาใช้ในการปรับเทียบคะแนน

ตอนที่ 2 ผลการปรับเทียบคะแนนด้วยวิธีเคอเนล

ตอนที่ 3 ผลการปรับเทียบคะแนนด้วยวิธี IRT

ตอนที่ 4 ผลการเปรียบเทียบคุณภาพการปรับเทียบคะแนนวิธีเคอเนลและวิธี IRT

ตอนที่ 5 เปรียบเทียบความสอดคล้องของการตัดเกรดระหว่างก่อนและหลังการปรับเทียบคะแนน

ตอนที่ 1 การวิเคราะห์ข้อมูลพื้นฐานของแบบสอบที่นำมาใช้ในการปรับเทียบคะแนน

แบบสอบที่นำมาใช้ในการปรับเทียบคะแนน ประกอบด้วย แบบสอบจำนวน 3 ฉบับ ที่มีความเป็นคู่ขนานด้านเนื้อหา โครงสร้าง รูปแบบของข้อสอบ และเวลาในการสอบ ที่เหมือนกันทั้ง 3 ฉบับ โดยเป็นแบบสอบชนิดเลือกตอบ 5 ตัวเลือก ที่วางโครงสร้างหลักเกณฑ์การสุ่มข้อสอบจากระบบคลังข้อสอบ จำนวน 120 ข้อต่อฉบับ มีข้อร่วมภายในจำนวน 15 ข้อ ใช้เวลาในการสอบ 3 ชั่วโมง แบบสอบที่ใช้ในการปรับเทียบคะแนนประกอบด้วยแบบสอบของภาคการศึกษา 1/ 2556 ภาคการศึกษา 1/ 2557 และภาคการศึกษา 1/ 2558 แบบสอบแต่ละฉบับจะนำไปปรับเทียบคะแนนให้อยู่บนสเกลเดียวกันกับแบบสอบของภาคการศึกษา 1/ 2556

การวิเคราะห์การปรับเทียบคะแนน ในเบื้องต้นผู้ศึกษาได้วิเคราะห์คุณภาพของข้อสอบตามทฤษฎีการตอบข้อสอบ (Item response theory) ด้วยโปรแกรม BILOG-MG v 3.0 การวิเคราะห์ข้อมูลถ้าข้อสอบข้อใดมีค่าอำนาจจำแนก (Biserial) ต่ำกว่า -0.15 ตามเงื่อนไขของโปรแกรมจะตัด

ข้อสอบข้อนั้นทิ้งไม่นำไปใช้ในการประมวลผล เช่นเดียวกันกับการวิเคราะห์การปรับเทียบคะแนนด้วย IRT ถ้าข้อสอบข้อใดมีค่า Initial slope ต่ำกว่า -0.15 โปรแกรมจะตัดข้อสอบข้อนั้นทิ้ง ไม่นำไปใช้ในประมวลค่าพารามิเตอร์ จึงทำให้การประมวลผลในครั้งนี้มีจำนวนข้อสอบที่ผ่านการวิเคราะห์ สามารถนำไปใช้ในการปรับเทียบคะแนน ดังตารางที่ 4-1

ตารางที่ 4-1 จำนวนข้อสอบต่อฉบับตามเงื่อนไขของการปรับเทียบคะแนน

เงื่อนไข/ ภาคการศึกษา	**จำนวนข้อสอบที่ผ่านการวิเคราะห์ต่อฉบับ จำแนกตามขนาดตัวอย่าง**		
	100 คน	**500 คน**	**700 คน**
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6			
1/ 2556	117	118	116
1/ 2557	114	119	116
2. ข้อสอบร่วมมีความยากอย่างสุ่ม			
1/ 2556	117	118	117
1/ 2558	114	119	119
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
1/ 2556	104	103	103
1/ 2557	102	105	105
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
1/ 2556	105	104	104
1/ 2558	103	108	108

สำหรับผลการวิเคราะห์คุณภาพรายข้อและรายฉบับ ของแบบสอบที่นำมาใช้ในการปรับเทียบคะแนน รายละเอียดดังนี้

1. ผลการวิเคราะห์ค่าสถิติพื้นฐานของแบบสอบที่นำมาใช้ในการปรับเทียบคะแนนทั้ง 3 ฉบับ ฉบับละ 120 ข้อ (คะแนนเต็ม 120 คะแนน) มีค่าใกล้เคียงกัน คือค่าเฉลี่ยอยู่ในช่วง 50-53 คะแนน ค่าส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11-12 คะแนน และค่าความเที่ยงประมาณ .80 ดังแสดงในตารางที่ 4-2

ตารางที่ 4-2 ค่าสถิติพื้นฐานของแบบสอบที่นำมาใช้ในการปรับเทียบคะแนน

ภาคการศึกษา	จำนวนผู้สอบ	*n* ในการวิเคราะห์	คะแนน		ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	ความเที่ยง
			สูงสุด	ต่ำสุด			
1/ 2556	1,209	796	91	18	50.60	11.71	0.83
1/ 2557	927	612	90	17	50.27	12.24	0.84
1/ 2558	857	564	89	26	52.88	11.33	0.81

หมายเหตุ ความเที่ยงใช้สูตร KR-20

2. ผลการวิเคราะห์คุณภาพของแบบสอบเป็นรายข้อด้วยทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory) และทฤษฎีการตอบข้อสอบ (Item response theory) ข้อสอบที่ถือว่ามีคุณภาพตามทฤษฎีการทดสอบแบบดั้งเดิม ข้อสอบควรจะต้องมีค่าความยาก (*p*) ระหว่าง .2-.8 และค่าอำนาจจำแนก (r) มากกว่า .2 ขณะที่ทฤษฎีการตอบข้อสอบ ค่าอำนาจจำแนก (*a*) และค่าความยาก (*b*) ที่เหมาะสม คือ a อยู่ระหว่าง 0.50-2.50 ค่า b อยู่ระหว่าง -2.50-2.50 (ศิริชัย กาญจนวาสี, 2555, หน้า 55) ผลการวิเคราะห์ข้อสอบ พบว่า การวิเคราะห์ข้อสอบด้วยทฤษฎีการทดสอบแบบดั้งเดิม จำนวนข้อสอบที่มีคุณภาพใช้ได้ทั้งค่าความยาก และอำนาจจำแนกของภาคการศึกษา 1/ 2557 มีจำนวนมากที่สุด (69) ข้อ) ส่วนข้อสอบที่ยากหรือง่ายเกินไปมีจำนวนใกล้เคียงกัน จำนวน 15 ข้อ ขณะที่ผลการวิเคราะห์ด้วยทฤษฎีการตอบข้อสอบ (IRT) ด้วยโมเดล 2 พารามิเตอร์ ผลการวิเคราะห์รายข้อพบว่า ข้อสอบข้อที่ 55 ภาคการศึกษา 1/ 2556 ข้อที่ 45 ภาคการศึกษา 1/ 2557 และข้อที่ 45 ภาคการศึกษา 1/ 2558 ได้ค่าอำนาจจำแนก (Biserial) ต่ำกว่า -0.15 ซึ่งจะส่งผลต่อการการปรับเทียบคะแนนและส่งผลต่อการวิเคราะห์ข้อสอบถ้าใช้โมเดล 3 พารามิเตอร์จะทำให้ไม่สามารถวิเคราะห์ข้อมูลได้ ในการปรับเทียบคะแนนครั้งนี้จึงตัดข้อสอบข้อดังกล่าวทิ้ง สำหรับผลการวิเคราะห์ด้วย IRT จำนวนข้อสอบที่มีคุณภาพใช้ได้ทั้งค่าความยาก (b) และอำนาจจำแนก (a) ของภาคการศึกษา 1/ 2556 มีจำนวนมากที่สุด (83 ข้อ) ส่วนจำนวนข้อสอบที่ทั้งค่าความยากและอำนาจจำแนกไม่เหมาะสมของภาคการศึกษา 1/ 2558 มีจำนวนมากที่สุด (37 ข้อ) ในการศึกษาครั้งนี้สำหรับเงื่อนไขการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จึงเลือกตัดทิ้งโดยใช้ผลการวิเคราะห์คุณภาพข้อสอบตามทฤษฎีดั้งเดิม เพื่อไม่ให้เหลือจำนวนข้อสอบที่แตกต่างจากแบบสอบต้นฉบับมากเกินไป ดังแสดงในตารางที่ 4-3

ตารางที่ 4-3 จำนวนข้อสอบจากการวิเคราะห์คุณภาพของแบบสอบด้วยทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบข้อสอบ

ภาคการศึกษา	จำนวนข้อสอบ/ฉบับ	ทฤษฎีการทดสอบดั้งเดิม			ทฤษฎีการตอบข้อสอบแนวใหม่		
		ใช้ได้	ไม่มีอำนาจจำแนก	ยากหรือง่ายเกินไป	a และ b ใช้ได้	a หรือ b ใช้ไม่ได้	ทั้ง a และ b ใช้ไม่ได้
1/ 2556	120	65	40	15	83	34	0
1/ 2557	120	69	37	14	64	31	25
1/ 2558	120	65	40	15	56	27	37

3. ผลการวิเคราะห์ข้อสอบที่ไม่มีคุณภาพทิ้ง ตามเงื่อนไขหนึ่งของการปรับเทียบคะแนนในครั้งนี้ พิจารณาจากค่าความยากของข้อสอบข้อที่ยากเกินไป ($p < 0.20$) และพิจารณาจากข้อสอบข้อที่ง่ายเกินไป ($p > 0.80$) ตามทฤษฎีการทดสอบแบบดั้งเดิม พบว่ามีข้อสอบที่จะต้องตัดทิ้งของภาคการศึกษา 1/ 2556 จำนวน 15 ข้อ ภาคการศึกษา 1/ 2557 จำนวน 14 ข้อ และภาคการศึกษา 1/ 2558 จำนวน 15 ข้อ แต่เนื่องจากข้อสอบที่จะต้องตัดทิ้งของภาคการศึกษา 1/ 2558 ซ้ำกับข้อสอบร่วมจำนวน 4 ข้อ จึงไม่สามารถตัดทิ้งได้เพราะจะทำให้เงื่อนไขของข้อสอบร่วมเปลี่ยนไป ดังนั้นจึงเหลือข้อสอบที่จะต้องตัดทิ้งของภาคการศึกษานี้เพียง 11 ข้อ ซึ่งข้อสอบทุกข้อที่ตัดทิ้งถ้าพิจารณาผลการวิเคราะห์ตามทฤษฎีการตอบข้อสอบ พบว่า ส่วนใหญ่สอดคล้องกัน คือมีค่า a หรือค่า b ไม่เหมาะสม ยกเว้นข้อ 13, 58 และ 94 ของภาคการศึกษา 1/ 2556 ข้อ 98 ของภาคการศึกษา 1/ 2551 และข้อ 30 ของภาคการศึกษา 1/ 2558 ดังแสดงในตารางที่ 4-4

ตารางที่ 4-4 ค่าพารามิเตอร์ของข้อสอบที่ตัดทิ้งตามเงื่อนไขการปรับเทียบคะแนน จำแนกตามภาคการศึกษา

ข้อ	ภาคการศึกษา 1/ 2556		ข้อ	ภาคการศึกษา 1/ 2557		ข้อ	ภาคการศึกษา 1/ 2558	
	ค่า *p*	ค่า a, b		ค่า *p*	ค่า a, b		ค่า *p*	ค่า a, b
13	0.82	1.42, -1.38	8	0.08	0.53, 5.02	1	0.12	0.19, 10.77
34	0.12	0.88, 4.30	27	0.10	0.34, 6.71	12	0.19	0.27, 5.71

ตารางที่ 4-4 (ต่อ)

ข้อ	ภาคการศึกษา 1/ 2556		ข้อ	ภาคการศึกษา 1/ 2557		ข้อ	ภาคการศึกษา 1/ 2558	
	ค่า *p*	ค่า a, b		ค่า *p*	ค่า a, b		ค่า *p*	ค่า a, b
36	0.19	1.29, 3.00	34	0.16	0.27, 6.24	26	0.17	0.30, 5.38
48	0.18	0.97, 3.90	46	0.17	0.56, 3.11	30	0.14	1.08, 2.27
58	0.81	0.67, -2.13	47	0.13	0.24, 8.68	32	0.17	0.22, 7.61
65	0.92	0.95, -2.90	59	0.14	0.38, 5.11	42	0.10	1.01, 2.71
72	0.06	0.00, 0.00	68	0.14	0.31, 6.44	49	0.14	0.43, 4.75
73	0.97	1.23, -3.44	83	0.10	0.20, 11.46	65	0.92	1.30, -2.60
79	0.09	0.00, 0.00	85	0.12	0.31, 6.33	72	0.05	0.65, 5.31
80	0.13	0.79, 4.10	90	0.19	0.33, 4.54	74	0.83	0.43, -3.76
85	0.13	1.18, 3.63	92	0.00	0.87, -3.18	78	0.08	0.25, 9.74
93	0.93	1.01, -2.98	98	0.88	1.18, -2.26	82	0.14	0.50, 4.06
94	0.18	1.67, 2.19	105	0.02	0.26, 12.98	101	0.19	0.37, 4.57
101	0.13	1.29, 4.02	111	0.15	0.40, 4.66	107	0.08	0.28, 5.65
120	0.14	1.64, 3.15				118	0.11	0.24, 8.83
รวม 15 ข้อ			รวม 14 ข้อ			รวม 15 ข้อ		

หมายเหตุ ข้อสอบของภาคการศึกษา 1/ 2558 ข้อ 26, 74, 101, 118 ซ้ำกับข้อสอบร่วมใน การปรับเทียบคะแนนจึงไม่ได้ตัดทิ้ง

ตอนที่ 2 ผลการปรับเทียบคะแนนด้วยวิธีเคอเนล

การปรับเทียบคะแนนด้วยวิธีเคอเนล โดยใช้โปรแกรม LOGLIN/ KE version 2.3 ภายใต้เงื่อนไขการใช้ข้อสอบร่วม จำนวน 15 ข้อ ที่ได้มาจากการสุ่มจากเนื้อหาวิชาที่ประกอบด้วย 15 หน่วยการเรียนรู้ หน่วยละ 1 ข้อ โดยวิธีที่หนึ่งใช้การสุ่มข้อสอบร่วมจากข้อสอบที่มีความยากอยู่ในช่วง .4-.6 ของทุกหน่วย กับวิธีที่สองใช้การสุ่มข้อสอบร่วมที่มีความยากอย่างสุ่มหน่วย ๆ ละ 1 ข้อ เช่นเดียวกัน จากนั้นนำมาปรับเทียบคะแนนโดยใช้วิธีเคอเนล กับข้อมูลที่ใช้จำนวนข้อสอบทั้งหมดกับตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ดำเนินการภายใต้กลุ่มตัวอย่างขนาดแตกต่างกัน 3 ขนาด คือ 100 คน 500 คนและ 700 คน ตามลำดับ ทั้งนี้จำนวนข้อสอบที่นำมาใช้ในการวิเคราะห์จะปรับ

ให้เท่ากับจำนวนข้อสอบที่ใช้ในการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ผลการวิเคราะห์ประกอบด้วย 4 เงื่อนไข คือ เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ข้อสอบร่วมมีความยากอย่างสุ่ม ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อที่ไม่มีคุณภาพทิ้ง และข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อที่ไม่มีคุณภาพทิ้ง รายละเอียด ดังนี้

1. การปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ใช้ข้อมูลผลการสอบของภาคการศึกษา 1/ 2556 และภาคการศึกษา 1/ 2557 โดยปรับเทียบคะแนนผลการสอบของภาคการศึกษา 1/ 2557 ให้อยู่บนสเกลเดียวกันกับคะแนนของภาคการศึกษา 1/ 2556 ผลการวิเคราะห์พบว่า คะแนนผลการสอบของภาคการศึกษา 1/ 2557 หลังจากการปรับเทียบคะแนน ส่วนใหญ่สูงกว่าคะแนนผลการสอบของภาคการศึกษา 1/ 2556 ไม่ว่าจะใช้ขนาดตัวอย่างจำนวน 100 คน 500 คน หรือ 700 คน ดังแสดงในตารางที่ 4-5

ตารางที่ 4-5 ผลการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

คะแนนภาคการศึกษา 1/ 2556	**คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดตัวอย่าง**		
	100 คน	**500 คน**	**700 คน**
0	5	0	-3
1	6	1	-2
2	7	3	-1
3	8	3	-1
4	8	5	0
5	9	6	1
6	9	7	2
7	10	9	3
8	11	9	3
9	12	11	5
10	13	12	6
11	14	13	8

ตารางที่ 4-5 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วม มีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดตัวอย่าง		
	100 คน	500 คน	700 คน
12	15	14	9
13	15	15	11
14	16	16	12
15	17	17	13
16	18	18	15
17	19	19	16
18	20	20	17
19	21	21	18
20	22	22	20
21	23	23	21
22	24	24	22
23	25	25	23
24	25	26	24
25	26	27	25
26	27	28	27
27	29	30	28
28	29	30	29
29	30	31	30
30	31	33	31
31	32	34	33
32	33	35	33
33	34	36	34
34	35	37	36
35	36	38	37
36	37	39	38

ตารางที่ 4-5 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดตัวอย่าง		
	100 คน	500 คน	700 คน
37	38	40	39
38	39	41	40
39	40	42	41
40	41	43	42
41	42	44	43
42	42	45	45
43	44	46	46
44	45	47	47
45	46	48	48
46	46	50	49
47	47	50	50
48	48	51	51
49	49	53	52
50	51	54	53
51	51	54	54
52	52	56	56
53	53	57	57
54	54	58	58
55	55	59	59
56	56	60	60
57	57	61	61
58	58	62	62
59	59	63	63
60	60	64	65
61	61	65	66

ตารางที่ 4-5 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วม มีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดตัวอย่าง		
	100 คน	500 คน	700 คน
62	62	66	67
63	63	67	68
64	63	68	69
65	64	69	71
66	66	70	72
67	67	71	73
68	68	72	74
69	68	73	76
70	69	74	77
71	70	75	78
72	71	76	79
73	72	78	81
74	73	78	82
75	74	79	84
76	75	80	86
77	76	82	87
78	76	82	88
79	77	84	90
80	78	85	91
81	79	86	92
82	79	87	95
83	80	88	97
84	80	88	98
85	81	91	99
86	82	92	101

ตารางที่ 4-5 (ต่อ)

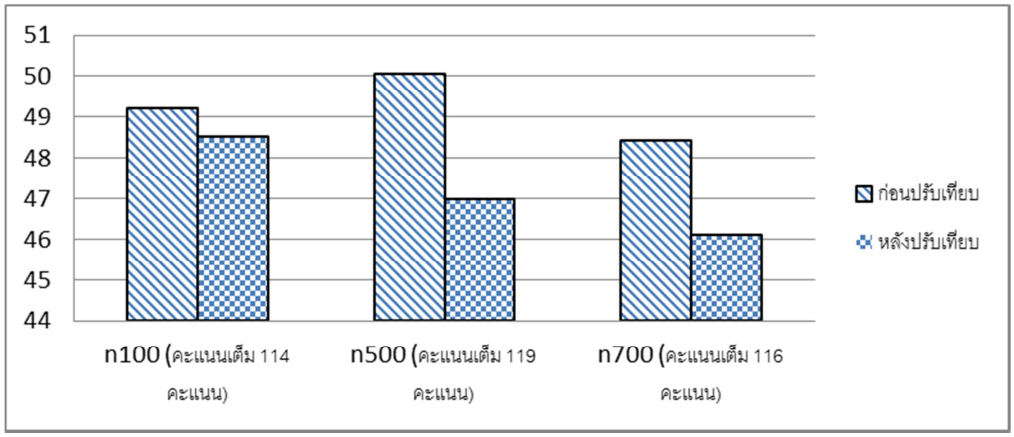
คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดตัวอย่าง		
	100 คน	500 คน	700 คน
87	82	93	101
88	83	94	102
89	83	95	103
90	84	96	103
91	84	96	104
92	85	99	104
93	85	99	104
94	86	99	104
95	86	246	104
96	87	246	104
97	87	246	104
98	88	246	104
99	88	246	104
100	88	246	104
101	89	246	104
102	91	246	104
103		246	104

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดตัวอย่าง หลังจากปรับเทียบคะแนนให้อยู่บนสเกลเดียวกันกับคะแนนของภาคการศึกษา 1/ 2556 พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลง 1-3 คะแนนในแต่ละเงื่อนไข โดยที่ก่อนการปรับเทียบคะแนนคะแนนเฉลี่ยอยู่ในช่วง 48.42-50.06 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 12.02-12.30 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนมีแนวโน้มลดลงอยู่ในช่วง 46.12-48.52 ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 10.91-12.77 แสดงให้เห็นว่าแบบสอบ

ที่นำมาปรับเทียบนี้เป็นแบบสอบที่ง่ายกว่าปีการศึกษา 1/ 2556 ดังแสดงในตารางที่ 4-6 และภาพที่ 4-1

ตารางที่ 4-6 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

เงื่อนไข/ กลุ่มตัวอย่าง	คะแนนเต็ม	คะแนน		$\bar{X}$	*SD*
		ต่ำสุด	สูงสุด		
แบบสอบเดิม ภาค 1/ 2557	120	17	90	50.27	12.24
1. ตัวอย่าง 100 คน	114				
ก่อนปรับเทียบ		16	89	49.21	12.25
หลังปรับเทียบ		13	90	48.52	12.77
2. ตัวอย่าง 500 คน	119				
ก่อนปรับเทียบ		16	90	50.06	12.30
หลังปรับเทียบ		14	85.5	46.99	11.83
3. ตัวอย่าง 700 คน	116				
ก่อนปรับเทียบ		15	87	48.42	12.02
หลังปรับเทียบ		16.5	80	46.12	10.91



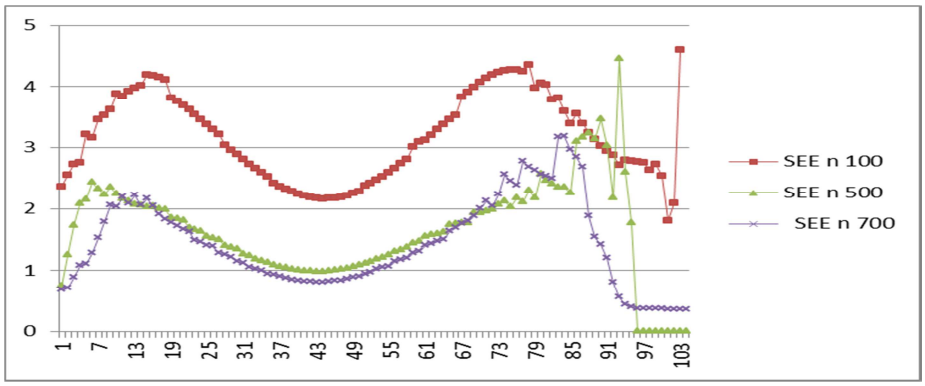
ภาพที่ 4-1 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

สำหรับค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดกลุ่มตัวอย่าง พบว่า เมื่อใช้ขนาดกลุ่มตัวอย่าง จำนวน 700 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด (1.42) รองลงมาคือ กลุ่มตัวอย่างขนาด 500 คน (1.63) และเมื่อขนาดกลุ่มตัวอย่างในการวิเคราะห์เพิ่มขึ้นค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนก็จะลดลง ดังแสดงในตารางที่ 4-7

ตารางที่ 4-7 ค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดตัวอย่าง

ขนาดกลุ่มตัวอย่าง	ความคลาดเคลื่อนมาตรฐาน				
	สูงสุด	ต่ำสุด	พิสัย	$\bar{X}$	SD
1. ขนาด 100 คน	4.60	1.81	2.80	3.17	0.70
2. ขนาด 500 คน	4.45	0.00	4.45	1.63	0.81
3. ขนาด 700 คน	3.17	0.36	2.81	1.42	0.72

เมื่อนำค่าความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดกลุ่มตัวอย่าง มาทำเป็นกราฟเส้น พบว่า กราฟมีลักษณะเป็นโค้งสองยอดในช่วงแรกและช่วงสุดท้าย ส่วนในช่วงกลาง ๆ ของคะแนนจะมีค่าคลาดเคลื่อนมาตรฐานในการปรับเทียบคะแนนต่ำ และเมื่อพิจารณาตามขนาดกลุ่มตัวอย่างการใช้กลุ่มตัวอย่างขนาด 700 คน ให้ค่า SEE ต่ำที่สุด ขณะที่ถ้าใช้กลุ่มตัวอย่างขนาด 100 คน จะให้ค่า SEE สูงสุด ดังแสดงในแผนภาพที่ 4-2



ภาพที่ 4-2 ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

2. การปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ใช้ข้อมูลผลการสอบของภาคการศึกษา 1/ 2556 และภาคการศึกษา 1/ 2558 โดยปรับคะแนนผลการสอบภาคของการศึกษา 1/ 2558 ให้อยู่บนสเกลเดียวกันกับคะแนนของภาคการศึกษา 1/ 2556 ผลการวิเคราะห์พบว่า คะแนนผลการสอบของภาคการศึกษา 1/ 2558 หลังจากการปรับเทียบคะแนน ส่วนใหญ่สูงกว่าคะแนนผลการสอบของภาคการศึกษา 1/ 2556 ไม่ว่าจะใช้ขนาดตัวอย่างจำนวน 100 คน 500 คน หรือ 700 คน ดังแสดงในตารางที่ 4-8

ตารางที่ 4-8 ผลการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

คะแนนภาคการศึกษา 1/ 2556	**คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม**		
	100 คน	**500 คน**	**700 คน**
0	8	2	0
1	11	4	2
2	13	6	3
3	15	7	5
4	16	8	7
5	17	9	9
6	18	10	10

ตารางที่ 4-8 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม		
	100 คน	500 คน	700 คน
7	19	11	11
8	20	13	13
9	20	13	14
10	21	15	15
11	22	16	16
12	23	17	17
13	23	17	18
14	24	18	19
15	25	19	20
16	26	20	22
17	27	21	23
18	27	22	24
19	28	23	25
20	29	24	26
21	30	25	27
22	31	26	28
23	31	27	29
24	32	28	29
25	33	29	30
26	34	30	31
27	35	31	32
28	35	32	33
29	36	33	34
30	37	34	35
31	38	35	36

ตารางที่ 4-8 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม		
	100 คน	500 คน	700 คน
32	39	36	37
33	39	37	38
34	40	38	39
35	41	39	40
36	42	40	41
37	43	42	42
38	43	43	43
39	44	44	44
40	45	45	45
41	46	46	46
42	47	47	47
43	47	48	48
44	48	49	49
45	49	50	50
46	50	51	51
47	51	52	52
48	51	53	53
49	52	54	54
50	53	55	55
51	54	56	56
52	54	57	57
53	55	57	58
54	56	58	59
55	57	59	60
56	58	60	61

ตารางที่ 4-8 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม		
	100 คน	500 คน	700 คน
57	58	61	62
58	59	62	63
59	60	63	64
60	61	64	65
61	61	65	66
62	62	67	67
63	63	68	68
64	64	69	69
65	64	70	70
66	65	71	71
67	66	72	72
68	67	73	73
69	68	74	74
70	68	75	75
71	69	76	76
72	70	77	77
73	71	78	78
74	71	79	79
75	72	80	80
76	73	81	81
77	74	82	82
78	75	83	83
79	75	84	84
80	76	85	85

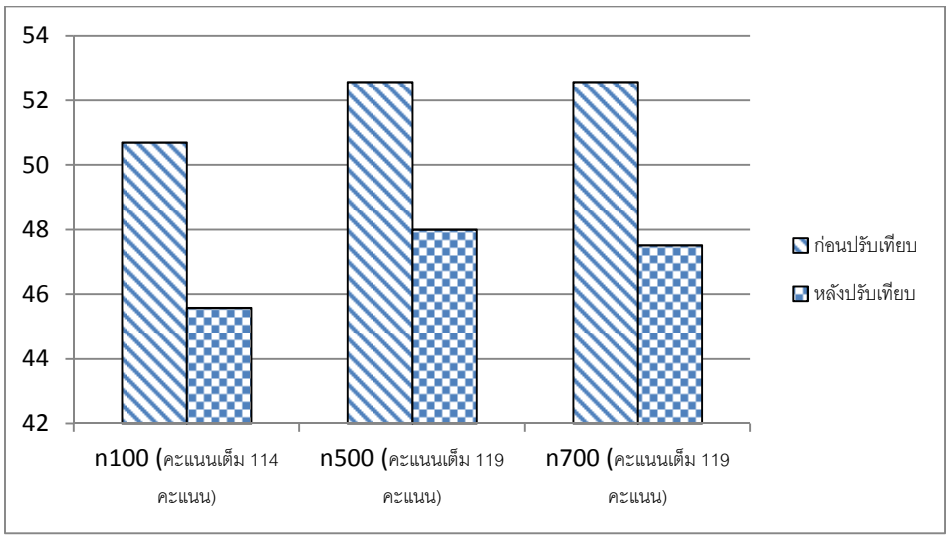
ตารางที่ 4-8 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	**คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม**		
	100 คน	**500 คน**	**700 คน**
81	77	85	86
82	78	86	87
83	78	87	88
84	79	89	89
85	80	89	90
86	81	90	92
87	82	91	93
88	83	92	94
89	83	93	96
90	84	94	96
91	85	95	97
92	86	96	98
93	86	96	98
94	87	98	98
95	87	246	98
96	88	246	98
97	88	246	100
98	89	246	100
99	89	246	100
100	89	246	100
101	89	246	100
102	89	246	100
		246	

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนนและหลังการปรับเทียบคะแนนให้อยู่บนสเกลเดียวกันกับ คะแนนของภาคการศึกษา 1/ 2556 ภายใต้เงื่อนไขข้อสอบร่วมอย่างสุ่มจำแนกตามขนาดตัวอย่าง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลง 4-5 คะแนน โดยที่ก่อนการปรับเทียบคะแนนคะแนนเฉลี่ยอยู่ในช่วง 50.70-52.56 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11.35-11.39 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนมีแนวโน้มลดลงอยู่ในช่วง 45.57-48.00 ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11.20-14.00 แสดงให้เห็นว่าแบบสอบที่นำมาปรับเทียบนี้เป็นแบบสอบที่ง่ายกว่าปีการศึกษา 1/ 2556 ดังแสดงในตารางที่ 4-9 และภาพที่ 4-3

ตารางที่ 4-9 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมที่มีความยากอย่างสุ่ม

เงื่อนไข/ กลุ่มตัวอย่าง	**คะแนนเต็ม**	**คะแนน**		$\bar{X}$	***SD***
		ต่ำสุด	**สูงสุด**		
แบบสอบเดิม ภาค 1/ 2558	120	26	89	52.88	11.33
1. ตัวอย่าง 100 คน	114				
ก่อนปรับเทียบ		24	87	50.70	11.35
หลังปรับเทียบ		13	91	45.57	14.00
2. ตัวอย่าง 500 คน	119				
ก่อนปรับเทียบ		26	89	52.56	11.39
หลังปรับเทียบ		22	84	48.00	11.20
3. ตัวอย่าง 700 คน	119				
ก่อนปรับเทียบ		26	89	52.56	11.39
หลังปรับเทียบ		20	84	47.52	11.46



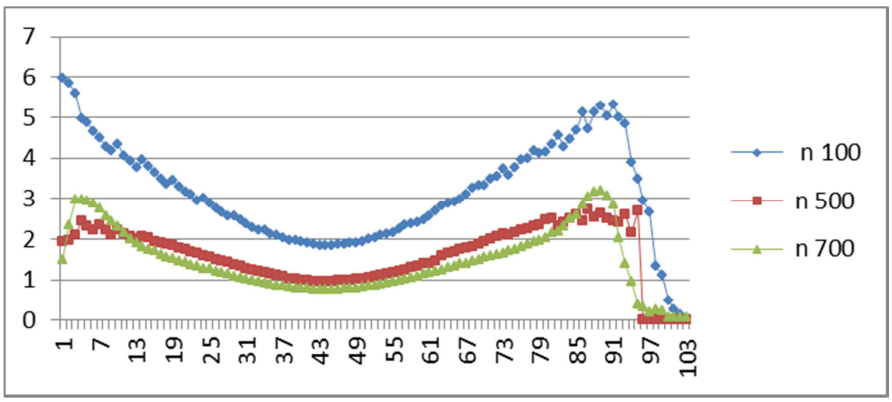
ภาพที่ 4-3 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

สำหรับค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมอย่างสุ่ม จำแนกตามขนาดกลุ่มตัวอย่าง พบว่าเมื่อใช้ขนาดกลุ่มตัวอย่างจำนวน 700 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด (1.45) ขณะที่กลุ่มตัวอย่างขนาด 100 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนสูงสุด (3.14) และเมื่อขนาดกลุ่มตัวอย่างในการวิเคราะห์เพิ่มขึ้นค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนก็จะลดลง ดังแสดงในตารางที่ 4-10

ตารางที่ 4-10 ค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มจำแนกตามขนาดตัวอย่าง

ขนาดกลุ่มตัวอย่าง	ความคลาดเคลื่อนมาตรฐาน				
	สูงสุด	ต่ำสุด	พิสัย	$\bar{X}$	*SD*
1. ขนาด 100 คน	5.97	0.05	5.92	3.14	1.26
2. ขนาด 500 คน	2.72	0.00	2.72	1.60	0.71
3. ขนาด 700 คน	3.19	0.08	3.11	1.45	0.78

และเมื่อนำค่าความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม จำแนกตามขนาดกลุ่มตัวอย่าง มาทำเป็นกราฟเส้น พบว่า ความคลาดเคลื่อนมาตรฐานในการปรับเทียบคะแนนเมื่อใช้กับกลุ่มตัวอย่างขนาด 500 คน และ 700 คน มีค่าต่ำสุดเกือบจะเท่า ๆ กัน ขณะที่เมื่อใช้กลุ่มตัวอย่างขนาด 100 คน ความคลาดเคลื่อนมาตรฐานในการปรับเทียบคะแนนสูงสุด เมื่อพิจารณาช่วงคะแนน พบว่า คะแนนช่วงแรกและช่วงสุดท้ายจะมีค่าความคลาดเคลื่อนมาตรฐานค่อนข้างสูง และมีค่าต่ำในช่วงกลาง ๆ ของคะแนน ดังแสดงในภาพที่ 4-4



ภาพที่ 4-4 ความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

3. การปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ในที่นี้ข้อสอบที่ไม่มีคุณภาพคือ ข้อที่มีค่าความยากตามทฤษฎีการทดสอบแบบดั้งเดิมที่ยากเกินไปหรือง่ายเกินไป ($p < .2$ หรือ $p > .8$) การปรับเทียบคะแนนใช้ข้อมูลผลการสอบของภาคการศึกษา 1/ 2556 และภาคการศึกษา 1/ 2557 โดยปรับคะแนนผลการสอบของภาคการศึกษา 1/ 2557 ให้อยู่บนสเกลเดียวกันกับคะแนนของภาคการศึกษา 1/ 2556 ผลการวิเคราะห์พบว่า หลังจากการปรับเทียบคะแนนของภาคการศึกษา 1/ 2557 เมื่อตัดข้อสอบที่ไม่มีคุณภาพทิ้ง 14 ข้อ คะแนนผลการสอบส่วนใหญ่สูงกว่าของภาคการศึกษา 1/ 2556 ไม่ว่าจะใช้ขนาดตัวอย่างจำนวน 100 คน 500 คน หรือ 700 คน ดังแสดงในตารางที่ 4-11

ตารางที่ 4-11 ผลการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

คะแนนภาคการศึกษา 1/ 2556	**คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง**		
	100 คน	**500 คน**	**700 คน**
0	3	1	1
1	5	3	2
2	6	4	3
3	7	6	4
4	8	7	5
5	8	8	6
6	9	9	7
7	11	10	8
8	12	11	9
9	13	12	10
10	13	13	11
11	14	14	12
12	15	15	14
13	16	17	15
14	17	17	16
15	18	18	17
16	19	20	18
17	20	21	20
18	21	22	20
19	22	23	22
20	23	24	23
21	24	25	24
22	25	26	25
23	26	27	26

ตารางที่ 4-11 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
	100 คน	500 คน	700 คน
24	27	28	27
25	28	29	29
26	29	30	29
27	30	31	31
28	31	32	32
29	32	33	33
30	33	35	34
31	34	35	35
32	35	37	36
33	36	38	38
34	37	39	39
35	38	40	40
36	39	41	41
37	40	42	42
38	41	43	43
39	41	44	45
40	43	45	46
41	44	46	47
42	44	47	48
43	45	49	49
44	47	49	50
45	48	51	51
46	48	51	53
47	49	53	53
48	51	54	55

ตารางที่ 4-11 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
	100 คน	500 คน	700 คน
49	51	55	56
50	52	56	57
51	53	57	58
52	54	58	59
53	55	59	60
54	56	60	62
55	57	61	63
56	58	62	64
57	59	63	65
58	60	64	66
59	61	65	67
60	62	66	68
61	63	67	70
62	64	68	71
63	65	69	71
64	66	70	72
65	67	71	73
66	68	72	74
67	69	73	76
68	69	73	77
69	70	75	77
70	71	75	78
71	72	76	79
72	73	77	80
73	73	78	81

ตารางที่ 4-11 (ต่อ)

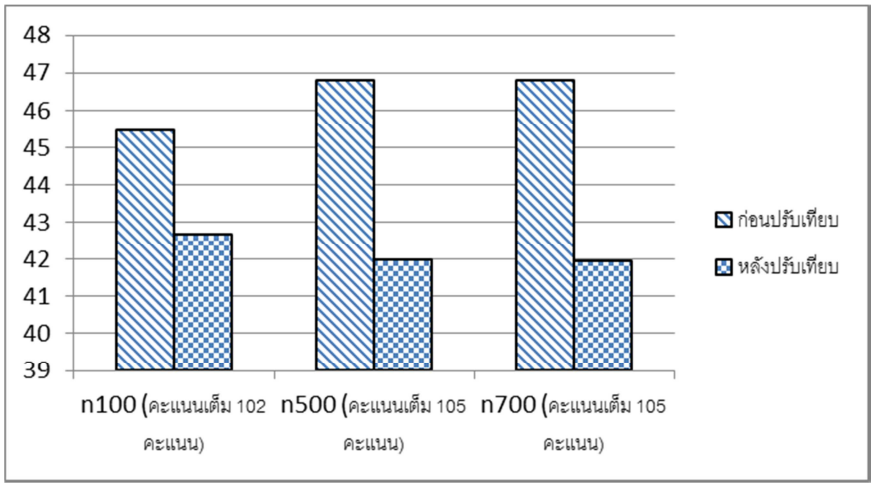
คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
	100 คน	500 คน	700 คน
74	74	79	82
75	75	80	82
76	76	80	83
77	76	81	84
78	77	82	84
79	78	83	85
80	78	84	86
81	79	84	87
82	79	85	87
83	80	86	87
84	80	87	88
85	81	88	88
86	81	88	88
87	82	89	88
88	82	89	88
89	83		

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดตัวอย่าง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนต่ำกว่าคะแนนเฉลี่ยก่อนการปรับเทียบคะแนน 2-4 คะแนน ไม่ว่าจะใช้กลุ่มตัวอย่างขนาดเท่าใดตามที่กำหนด โดยที่ก่อนการปรับเทียบคะแนนคะแนนเฉลี่ยอยู่ในช่วง 45.45-46.80 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11.80-11.84 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนอยู่ในช่วง 41.94-42.69 ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 10.58-12.10 แสดงให้เห็นว่า

แบบสอบที่นำมาปรับเทียบนี้เป็นแบบสอบที่ง่ายกว่าปีการศึกษา 1/ 2556 ดังแสดงในตารางที่ 4-12 และภาพที่ 4-5

ตารางที่ 4-12 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

เงื่อนไข/ กลุ่มตัวอย่าง	คะแนนเต็ม	คะแนน		$\bar{X}$	*SD*
		ต่ำสุด	สูงสุด		
แบบสอบเดิม ภาค 1/ 2557	120	17	90	50.27	12.24
1. ตัวอย่าง 100 คน	102				
ก่อนปรับเทียบ		16	84	45.45	11.80
หลังปรับเทียบ		13	83	42.69	12.10
2. ตัวอย่าง 500 คน	105				
ก่อนปรับเทียบ		16	85	46.80	11.84
หลังปรับเทียบ		13	79	41.98	11.22
3. ตัวอย่าง 700 คน	105				
ก่อนปรับเทียบ		16	85	46.80	11.84
หลังปรับเทียบ		14	77	41.94	10.58



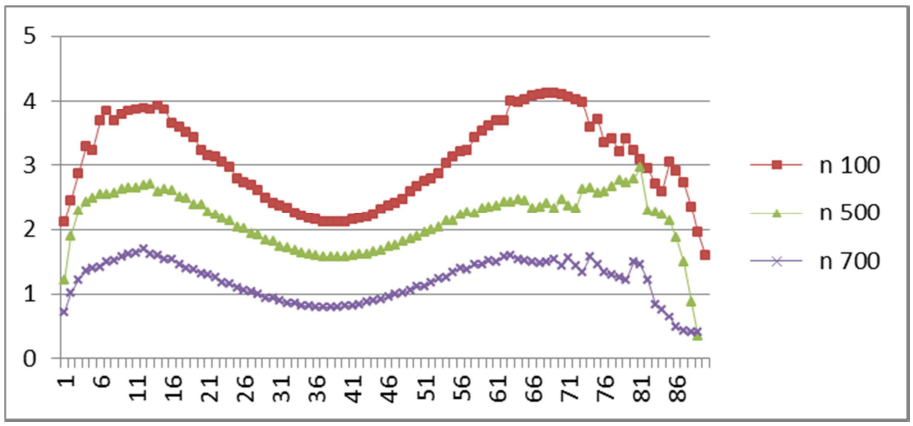
ภาพที่ 4-5 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

สำหรับค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดกลุ่มตัวอย่าง พบว่า เมื่อใช้ขนาดกลุ่มตัวอย่าง จำนวน 700 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด (1.19) ขณะที่กลุ่มตัวอย่างขนาด 100 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนสูงสุด (3.07) และเมื่อขนาดกลุ่มตัวอย่างในการวิเคราะห์เพิ่มขึ้นค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนก็จะลดลง ดังแสดงในตารางที่ 4-13

ตารางที่ 4-13 ค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ขนาดกลุ่มตัวอย่าง	ความคลาดเคลื่อนมาตรฐาน				
	สูงสุด	ต่ำสุด	พิสัย	$\bar{X}$	*SD*
1. ขนาด 100 คน	4.12	1.58	2.54	3.07	0.67
2. ขนาด 500 คน	2.97	0.35	2.62	2.14	0.46
3. ขนาด 700 คน	1.69	0.39	1.29	1.19	0.33

และเมื่อนำค่าความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดกลุ่มตัวอย่างมาทำเป็นกราฟเส้น พบว่า กราฟมีลักษณะเป็นโค้งสองยอดในช่วงแรกและช่วงสุดท้ายของคะแนน ส่วนในช่วงกลาง ๆ ของคะแนน จะมีค่าความคลาดเคลื่อนมาตรฐานค่อนข้างต่ำ และเมื่อพิจารณาตามขนาดตัวอย่างกลุ่มตัวอย่าง การใช้ขนาดตัวอย่าง 700 คน ในการวิเคราะห์จะให้ค่า SEE ต่ำสุด ดังแสดงในแผนภาพที่ 4-6



ภาพที่ 4-6 ความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง 4-6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

4. การปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ใช้ข้อมูลผลการสอบของภาคการศึกษา 1/ 2556 และภาคการศึกษา 1/ 2558 โดยปรับคะแนนผลการสอบของภาคการศึกษา 1/ 2558 ให้อยู่บนสเกลเดียวกันกับคะแนนของภาคการศึกษา 1/ 2556 ผลการวิเคราะห์พบว่า คะแนนผลการสอบของภาคการศึกษา 1/ 2558 หลังจากการปรับเทียบคะแนน ส่วนใหญ่สูงกว่าคะแนนผลการสอบของภาคการศึกษา 1/ 2556 ไม่ว่าจะใช้ขนาดตัวอย่างจำนวน 100 คน 500 คน หรือ 700 คน ดังแสดงในตารางที่ 4-14

ตารางที่ 4-14 ผลการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

คะแนนภาคการศึกษา 1/ 2556	**คะแนนหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง**		
	100 คน	**500 คน**	**700 คน**
0	12	4	3
1	13	7	5
2	15	8	7
3	16	10	9
4	17	11	11

ตารางที่ 4-14 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
	100 คน	500 คน	700 คน
5	18	12	12
6	19	13	14
7	20	15	15
8	21	15	16
9	22	16	17
10	23	18	18
11	24	19	19
12	25	20	20
13	25	21	21
14	26	22	22
15	27	23	23
16	28	24	24
17	29	24	26
18	29	25	27
19	30	27	28
20	31	28	28
21	32	29	29
22	33	30	30
23	34	31	31
24	34	32	32
25	35	33	34
26	36	34	35
27	37	35	36
28	37	36	36
29	38	37	37

ตารางที่ 4-14 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
	100 คน	500 คน	700 คน
30	39	38	38
31	40	39	39
32	41	40	41
33	42	41	42
34	42	42	43
35	43	43	44
36	44	44	44
37	45	45	45
38	46	46	46
39	46	47	47
40	47	48	49
41	48	49	50
42	49	50	51
43	49	51	52
44	50	52	52
45	51	53	53
46	52	54	54
47	53	55	56
48	54	56	57
49	54	57	58
50	55	58	59
51	56	59	60
52	57	60	60
53	57	61	61
54	58	62	62

ตารางที่ 4-14 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
	100 คน	500 คน	700 คน
55	59	64	64
56	60	65	65
57	60	66	66
58	61	66	67
59	62	67	67
60	63	68	68
61	64	70	69
62	65	71	70
63	66	72	72
64	66	73	73
65	67	74	74
66	68	74	75
67	69	75	76
68	69	77	76
69	70	78	77
70	71	79	78
71	72	80	80
72	72	81	81
73	73	82	82
74	74	83	83
75	75	84	84
76	76	84	85
77	77	85	86
78	77	86	87
79	78	87	89

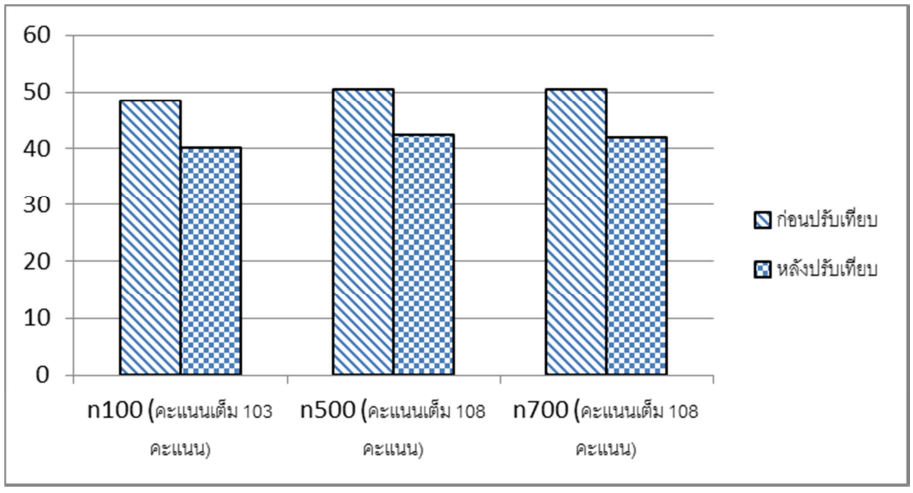
ตารางที่ 4-14 (ต่อ)

คะแนนภาคการศึกษา 1/ 2556	คะแนนหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
	100 คน	500 คน	700 คน
80	79	88	90
81	80	89	91
82	81	89	91
83	82	90	93
84	83	90	94
85	83	91	94
86	84	91	94
87	86	91	94
88	86	91	94
89	88	91	94
90	88		

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนน และหลังการปรับเทียบคะแนนภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดตัวอย่าง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนต่ำกว่าคะแนนเฉลี่ยก่อนการปรับเทียบคะแนนประมาณ 8 คะแนน ไม่ว่าจะใช้กลุ่มตัวอย่างขนาดเท่าใดตามที่กำหนด โดยคะแนนเฉลี่ยที่ได้ก่อนการปรับเทียบคะแนนอยู่ในช่วง 48.55-50.41 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 10.94-10.99 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนอยู่ในช่วง 40.18-42.40 ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 10.90-13.35 10 แสดงให้เห็นว่าแบบสอบที่นำมาปรับเทียบนี้เป็นแบบสอบที่ง่ายกว่าปีการศึกษา 1/ 2556 ดังแสดงในตารางที่ 4-15 และภาพที่ 4-7

ตารางที่ 4-15 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

เงื่อนไข/ กลุ่มตัวอย่าง	คะแนนเต็ม	คะแนน		$\bar{X}$	*SD*
		ต่ำสุด	สูงสุด		
แบบสอบเดิม ภาค 1/ 2558	120	26	89	52.88	11.33
1. ตัวอย่าง 100 คน	103				
ก่อนปรับเทียบ		23	83	48.55	10.94
หลังปรับเทียบ		10	82.5	40.18	13.35
2. ตัวอย่าง 500 คน	108				
ก่อนปรับเทียบ		25	85	50.41	10.99
หลังปรับเทียบ		17	76.5	42.40	10.90
3. ตัวอย่าง 700 คน	108				
ก่อนปรับเทียบ		25	85	50.41	10.99
หลังปรับเทียบ		17	76	41.90	10.94



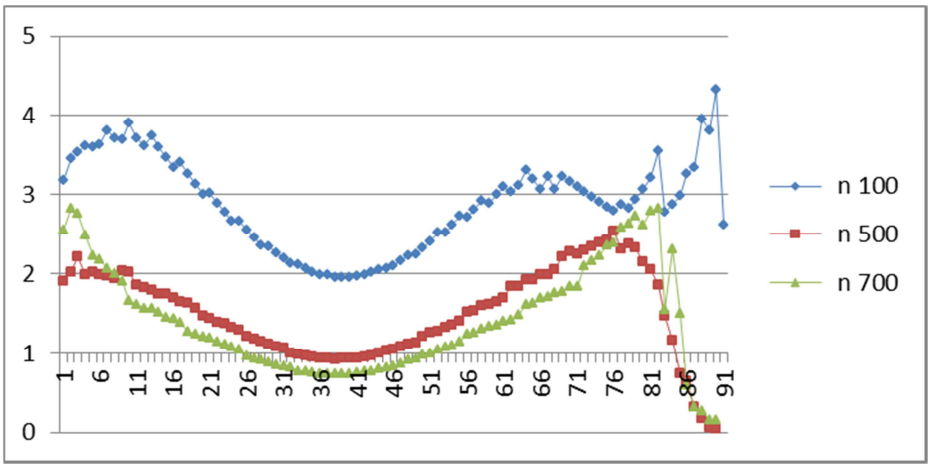
ภาพที่ 4-7 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

สำหรับค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดกลุ่มตัวอย่าง พบว่าเมื่อใช้ขนาดกลุ่มตัวอย่างจำนวน 700 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด ใกล้เคียงกับเมื่อใช้ขนาดกลุ่มตัวอย่างจำนวน 500 คน ขณะที่กลุ่มตัวอย่างขนาด 100 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนสูงสุด และเมื่อขนาดกลุ่มตัวอย่างในการวิเคราะห์เพิ่มขึ้นค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนก็จะลดลง ดังแสดงในตารางที่ 4-16

ตารางที่ 4-16 ค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดตัวอย่าง

ขนาดกลุ่มตัวอย่าง	ความคลาดเคลื่อนมาตรฐาน				
	สูงสุด	ต่ำสุด	พิสัย	$\bar{X}$	*SD*
1. ขนาด 100 คน	6.81	1.81	5.00	3.29	1.23
2. ขนาด 500 คน	2.53	0.05	2.49	1.51	0.56
3. ขนาด 700 คน	2.83	0.15	2.68	1.42	0.67

เมื่อนำค่าความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง มาทำเป็นกราฟเส้น พบว่า กราฟความคลาดเคลื่อนมาตรฐานในการปรับเทียบคะแนนเมื่อใช้กลุ่มตัวอย่างขนาด 500 คน และ 700 คน มีค่าต่ำใกล้เคียงกัน เมื่อพิจารณาช่วงคะแนน พบว่า คะแนนช่วงแรกและช่วงสุดท้ายจะมีค่าความคลาดเคลื่อนมาตรฐานค่อนข้างสูง และจะมีค่าต่ำในช่วงกลาง ๆ ของคะแนน ดังแสดงในภาพที่ 4-8



ภาพที่ 4-8 ความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

5. เปรียบเทียบผลการวิเคราะห์การปรับเทียบด้วยวิธีเคอเนล จำแนกตามขนาดตัวอย่าง

5.1 ขนาดกลุ่มตัวอย่าง 100 คน

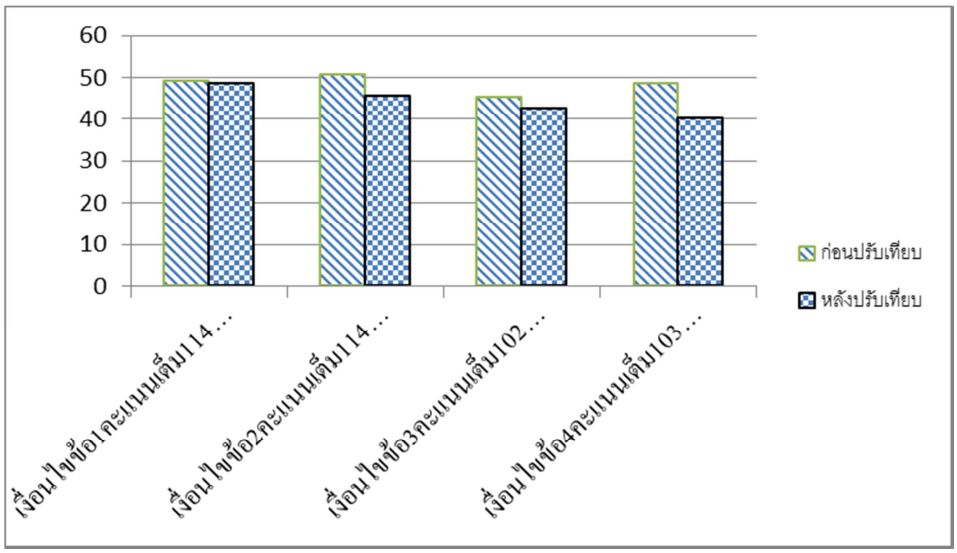
การปรับเทียบคะแนนด้วยวิธีเคอเนลเมื่อใช้กลุ่มตัวอย่างขนาด 100 คน เท่ากัน ด้วยเงื่อนไขข้อสอบร่วมมีความยาก .4-.6 ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ข้อสอบร่วมมีความยากอย่างสุ่ม และข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง พบว่าคะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลงจากก่อนปรับเทียบคะแนนทุกเงื่อนไข (1-8 คะแนน) แสดงให้เห็นว่าแบบสอบที่นำมาปรับเทียบของภาคการศึกษา 1/ 2557 และ 1/ 2558 เป็นแบบสอบที่ง่ายกว่าของ ภาค 1/ 2556 ดังแสดงในตารางที่ 4-17 และภาพที่ 4-9

ตารางที่ 4-17 ผลการเปรียบเทียบค่าสถิติพื้นฐานก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขที่ต่างกัน กับกลุ่มตัวอย่างขนาด 100 คน

เงื่อนไข	คะแนนเต็ม	Min	Max	$\bar{X}$	*SD*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	114				
1.1 ก่อนการปรับเทียบคะแนน		16	89	49.21	12.25
1.2 หลังการปรับเทียบคะแนน		13	90	48.52	12.77
ผลต่างของค่าสถิติ		-3	+1	-0.69	+0.52

ตารางที่ 4-17 (ต่อ)

เงื่อนไข	คะแนนเต็ม	Min	Max	$\bar{X}$	*SD*
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	114				
2.1 ก่อนการปรับเทียบคะแนน		24	87	50.70	11.35
2.2 หลังการปรับเทียบคะแนน		13	91	45.57	14.00
ผลต่างของค่าสถิติ		+11	+4	-5.13	2.65
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	102				
3.1 ก่อนการปรับเทียบคะแนน		16	84	45.45	11.80
3.2 หลังการปรับเทียบคะแนน		13	83	42.69	12.10
ผลต่างของค่าสถิติ		-3	-1	-2.76	+0.30
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	103				
4.1 ก่อนการปรับเทียบคะแนน		23	83	48.55	10.94
4.2 หลังการปรับเทียบคะแนน		10	82.5	40.18	13.35
ผลต่างของค่าสถิติ		-13	-0.5	-8.37	+2.41

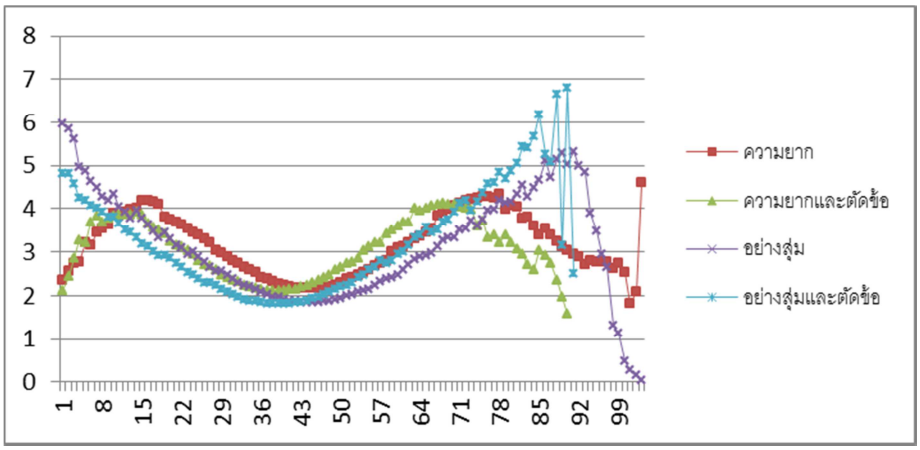


ภาพที่ 4-9 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้ตัวอย่างขนาด 100 คน

สำหรับค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขที่ต่างกันเมื่อใช้กลุ่มตัวอย่างขนาด 100 พบว่า ทุกเงื่อนไขค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานในการปรับเทียบคะแนนมีค่าใกล้เคียงกัน โดยเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้งมีค่าเฉลี่ยความคลาดเคลื่อนต่ำสุด (3.07) เมื่อนำค่าความคลาดเคลื่อนมาตรฐานในการปรับเทียบคะแนนมาทำกราฟเส้น พบว่ากราฟของความคลาดเคลื่อนมาตรฐานตามเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ที่ตัดและไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้งมีลักษณะเป็นโค้งสองยอด ขณะที่กราฟของเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ที่ตัดและไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้งมีลักษณะที่ช่วงแรกและช่วงสุดท้ายของคะแนนมีค่าความคลาดเคลื่อนมาตรฐานในการปรับเทียบคะแนนสูง ส่วนในช่วงกลาง ๆ ของคะแนนค่าความคลาดเคลื่อนมาตรฐานค่อนข้างต่ำ ดังแสดงในตารางที่ 4-18 และภาพที่ 4-10

ตารางที่ 4-18 ผลการเปรียบเทียบค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขที่ต่างกัน กับกลุ่มตัวอย่างขนาด 100 คน

เงื่อนไข	ความคลาดเคลื่อนมาตรฐาน				
	สูงสุด	ต่ำสุด	พิสัย	$\bar{X}$	*SD*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	4.60	1.81	2.80	3.17	0.70
2. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	4.12	1.58	2.54	3.07	0.67
3. ข้อสอบร่วมมีความยากอย่างสุ่ม	5.97	0.05	5.92	3.14	1.26
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	6.81	1.81	5.00	3.29	1.23



ภาพที่ 4-10 ความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขที่ต่างกัน กับกลุ่มตัวอย่างขนาด 100 คน

5.2 ขนาดกลุ่มตัวอย่าง 500 คน

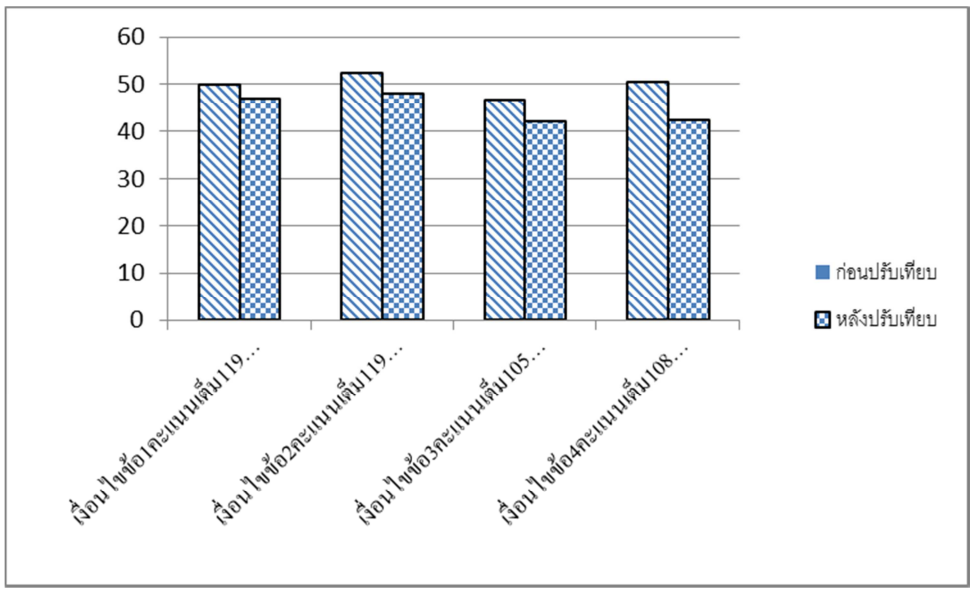
การปรับเทียบคะแนนเมื่อใช้กลุ่มตัวอย่างขนาด 500 คน เท่ากัน ด้วยเงื่อนไขข้อสอบร่วมมีความยาก .4-.6 ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ข้อสอบร่วมมีความยากอย่างสุ่ม และข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง พบว่าคะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลงจากก่อนปรับเทียบคะแนนทุกเงื่อนไข ในอัตราที่ใกล้เคียงกัน ยกเว้นเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้งที่ลดลงมากที่สุด (8 คะแนน) แสดงให้เห็นว่าแบบสอบที่นำมาปรับเทียบของภาคการศึกษา 1/ 2557 และ 1/ 2558 เป็นแบบสอบที่ง่ายกว่าของ ภาค 1/ 2556 ดังแสดงในตารางที่ 4-19 และภาพที่ 4-11

ตารางที่ 4-19 ผลการเปรียบเทียบค่าสถิติพื้นฐานก่อนและหลังการปรับเทียบคะแนนภายใต้เงื่อนไขที่ต่างกัน กับกลุ่มตัวอย่างขนาด 500 คน

เงื่อนไข	**คะแนนเต็ม**	**Min**	**Max**	$\bar{X}$	***SD***
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	119				
1.1 ก่อนการปรับเทียบคะแนน		16	90	50.06	12.30
1.2 หลังการปรับเทียบคะแนน		14	85.5	46.99	11.83
ผลต่างของค่าสถิติ		-2	-4.5	-3.07	-0.47

ตารางที่ 4-19 (ต่อ)

เงื่อนไข	คะแนนเต็ม	Min	Max	$\bar{X}$	*SD*
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	119				
2.1 ก่อนการปรับเทียบคะแนน		26	89	52.56	11.39
2.2 หลังการปรับเทียบคะแนน		22	84	48.00	11.20
ผลต่างของค่าสถิติ		-4	-5	-4.56	-0.19
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	105				
3.1 ก่อนการปรับเทียบคะแนน		16	85	46.80	11.84
3.2 หลังการปรับเทียบคะแนน		13	79	41.98	11.22
ผลต่างของค่าสถิติ		-3	-6	-4.82	-0.62
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	108				
4.1 ก่อนการปรับเทียบคะแนน		25	85	50.41	10.81
4.2 หลังการปรับเทียบคะแนน		17	76.5	42.40	12.21
ผลต่างของค่าสถิติ		-8	-8.5	-8.01	+1.4

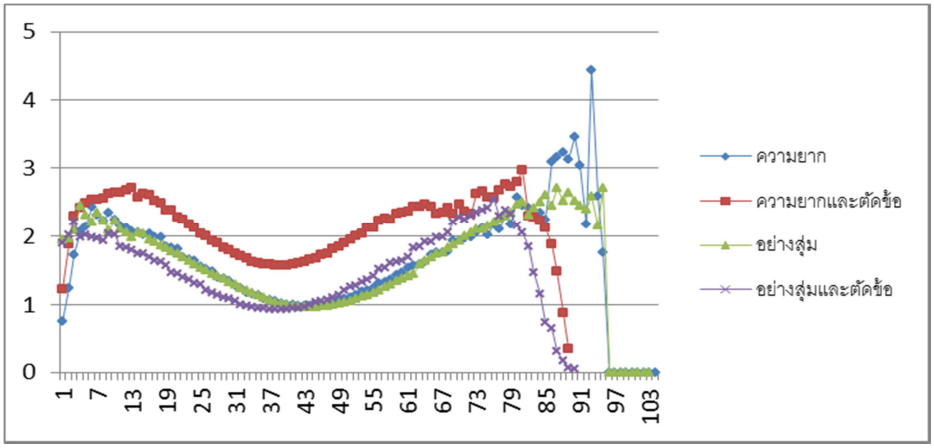


ภาพที่ 4-11 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้ตัวอย่างขนาด 500 คน

สำหรับค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนลกับกลุ่มตัวอย่างขนาด 500 คน ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และข้อสอบร่วมอย่างสุ่ม ระหว่างที่มีการตัดกับไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนที่จะนำไปปรับเทียบคะแนน พบว่าการปรับเทียบคะแนนตามเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง มีค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด (1.51) ซึ่งมีค่าใกล้เคียงกันกับการใช้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเมื่อพิจารณากราฟพบว่ามีลักษณะเป็นโค้งสองยอดในช่วงแรกและช่วงสุดท้ายของคะแนน ส่วนในช่วงกลาง ๆ ของคะแนนจะมีค่าความคลาดเคลื่อนมาตรฐานค่อนข้างต่ำ ดังแสดงในตารางที่ 4 -20 และภาพที่ 4-12

ตารางที่ 4-20 ค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนลกับกลุ่มตัวอย่างขนาด 500 คน

เงื่อนไข	ความคลาดเคลื่อนมาตรฐาน				
	สูงสุด	ต่ำสุด	พิสัย	$\bar{X}$	*SD*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	4.45	0	4.45	1.63	0.81
2. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	2.97	0.35	2.62	2.14	0.46
3. ข้อสอบร่วมมีความยากอย่างสุ่ม	2.72	0	2.72	1.60	0.71
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	2.53	0.05	2.49	1.51	0.56



ภาพที่ 4-12 ความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขที่ต่างกัน กับกลุ่มตัวอย่างขนาด 500 คน

5.3 ขนาดกลุ่มตัวอย่าง 700 คน

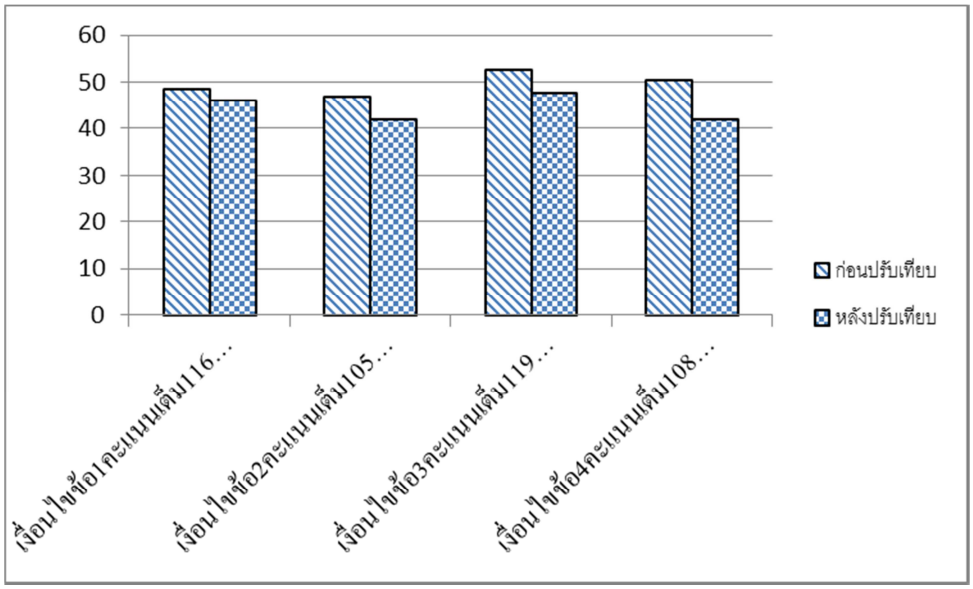
การปรับเทียบคะแนนเมื่อใช้กลุ่มตัวอย่างขนาด 700 คน เท่ากัน ด้วยเงื่อนไขข้อสอบร่วมมีความยาก .4-.6 ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ข้อสอบร่วมมีความยากอย่างสุ่ม และข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลงจากก่อนปรับเทียบคะแนนทุกเงื่อนไข (2-9 คะแนน) โดยการใช้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม คะแนนก่อนและหลังการปรับเทียบคะแนนลดลงในอัตราที่ใกล้เคียงกัน แสดงให้เห็นว่าแบบสอบที่นำมาปรับเทียบของภาคการศึกษา 1/ 2557 และ 1/ 2558 เป็นแบบสอบที่ง่ายกว่าของ ภาค 1/ 2556 ดังแสดงในตารางที่ 4-21 และภาพที่ 4-13

ตารางที่ 4-21 ผลการเปรียบเทียบค่าสถิติพื้นฐานก่อนและหลังการปรับเทียบคะแนนภายใต้เงื่อนไขที่ต่างกัน กับกลุ่มตัวอย่างขนาด 700 คน

เงื่อนไข	คะแนนเต็ม	Min	Max	$\bar{X}$	*SD*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	116				
1.1 ก่อนการปรับเทียบคะแนน		15	87	48.42	12.02
1.2 หลังการปรับเทียบคะแนน		16.5	80	46.12	10.91
ผลต่างของค่าสถิติ		1.5	-7	-2.3	-1.11
2. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	105				
2.1 ก่อนการปรับเทียบคะแนน		16	85	46.80	11.84
2.2 หลังการปรับเทียบคะแนน		14	77	41.94	10.58
ผลต่างของค่าสถิติ		-2	-8	-4.86	-1.26
3. ข้อสอบร่วมมีความยากอย่างสุ่ม	119				
3.1 ก่อนการปรับเทียบคะแนน		26	89	52.56	11.39
3.2 หลังการปรับเทียบคะแนน		20	84	47.52	11.46
ผลต่างของค่าสถิติ		-6	-5	-5.04	0.07

ตารางที่ 4-21 (ต่อ)

เงื่อนไข	คะแนนเต็ม	Min	Max	$\bar{X}$	*SD*
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	108				
4.1 ก่อนการปรับเทียบคะแนน		25	85	50.41	10.99
4.2 หลังการปรับเทียบคะแนน		17	76	41.90	10.94
ผลต่างของค่าสถิติ		-8	-9	-8.51	-.05



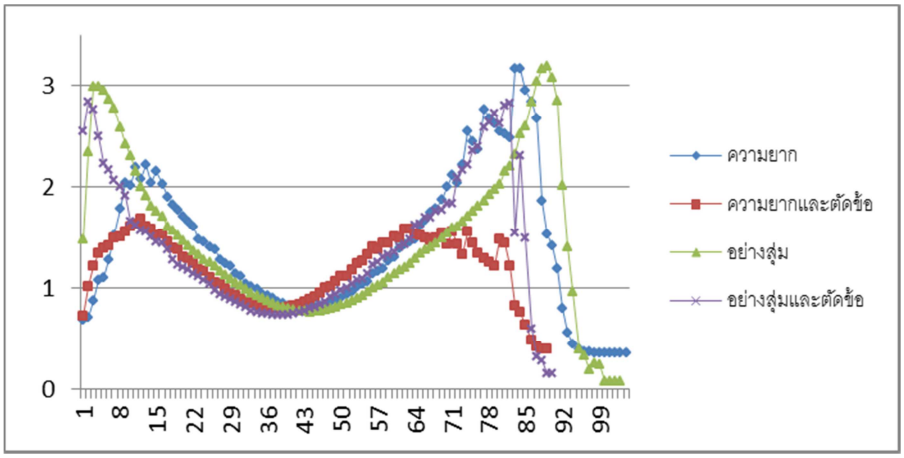
ภาพที่ 4-13 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้ตัวอย่างขนาด 700 คน

สำหรับค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนลกับกลุ่มตัวอย่างขนาด 700 คน ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และข้อสอบร่วมอย่างสุ่ม ระหว่างที่มีการตัดกับไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนที่จะนำไปปรับเทียบคะแนน พบว่าภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง มีค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด (1.19) ส่วนเงื่อนไขอื่น ๆ มีค่าใกล้เคียงกัน ประมาณ 1.4 และเมื่อพิจารณาจากกราฟจะเห็นได้ว่า ค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนของทุกเงื่อนไขมีค่าใกล้เคียงกัน

และเป็นไปในทิศทางเดียวกันมีลักษณะเป็นโค้งสองยอดในช่วงแรกและช่วงสุดท้ายของคะแนน ส่วนในช่วงกลางของคะแนนจะมีค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำ ดังแสดงในตารางที่ 4-22 และภาพที่ 4-14

ตารางที่ 4-22 ค่าสถิติพื้นฐานของความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนลกับกลุ่มตัวอย่างขนาด 700 คน

เงื่อนไข	ความคลาดเคลื่อนมาตรฐาน				
	สูงสุด	ต่ำสุด	พิสัย	$\bar{X}$	*SD*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	3.17	0.36	2.81	1.42	0.72
2. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	1.69	0.39	1.29	1.19	0.33
3. ข้อสอบร่วมมีความยากอย่างสุ่ม	3.19	0.08	3.11	1.45	0.78
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	2.83	0.15	2.68	1.42	0.67



ภาพที่ 4-14 ความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนด้วยวิธีเคอเนลภายใต้เงื่อนไขที่ต่างกัน กับกลุ่มตัวอย่างขนาด 700 คน

ตอนที่ 3 ผลการปรับเทียบคะแนนด้วยวิธี IRT

การปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ (IRT) โดยใช้โปรแกรม BILOG-Mg version 3.0 ภายใต้เงื่อนไขการใช้ข้อสอบร่วม จำนวน 15 ข้อ ที่ได้มาจากการสุ่มข้อสอบจากเนื้อหาวิชาที่ประกอบด้วย 15 หน่วยการเรียนรู้ หน่วยละ 1 ข้อ โดยวิธีที่หนึ่งใช้การสุ่มข้อสอบร่วมที่มีความยากอยู่ในช่วง .4-.6 จากทุกหน่วย กับวิธีที่สองใช้การสุ่มข้อสอบร่วมที่มีความยากอย่างสุ่ม หน่วยละ 1 ข้อ เช่นเดียวกัน จากนั้นนำมาปรับเทียบคะแนนโดยใช้วิธี IRT 2 พารามิเตอร์ โดยใช้ข้อมูลผลการสอบในสถานการณ์การสอบจริงของภาคการศึกษา 1/ 2556 และภาคการศึกษา 1/ 2558 ที่มีความเป็นคู่ขนานด้านเนื้อหา โครงสร้าง รูปแบบของข้อสอบ และเวลาในการสอบ ฉบับละ 120 ข้อ มีข้อสอบร่วมภายในฉบับละ 15 ข้อ แบบสอบแต่ละฉบับจะถูกปรับเทียบให้อยู่บนสเกลเดียวกันกับแบบสอบของภาคการศึกษา 1/ 2556 ในการวิเคราะห์ผลการปรับเทียบคะแนน จะใช้วิธีการวิเคราะห์โดยใช้ผลการตอบข้อสอบทั้งหมดกับตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดแตกต่างกัน 3 ขนาด คือ 100 คน 500 คน และ 700 คน เช่นเดียวกันกับวิธีเคอเนล

การปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ (IRT) ด้วยโมเดล 2 พารามิเตอร์ ดำเนินการใน 3 ขั้นตอน คือ 1) ตรวจสอบข้อมูลเบื้องต้นตามเงื่อนไขของการปรับเทียบคะแนนด้วย IRT 2) วิเคราะห์ปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ และ 3) ประมาณคะแนนดิบจากการปรับเทียบคะแนนด้วยโปรแกรม IRT-CLASS นำมาใช้ในการสร้างสมการพยากรณ์คะแนนดิบ รายละเอียด ดังนี้

1. ตรวจสอบข้อมูลเบื้องต้นตามเงื่อนไขของการปรับเทียบคะแนนด้วย IRT ซึ่งจะต้องมีคุณสมบัติสำคัญ คือ มีลักษณะเด่นลักษณะเดียว (Unidimensionality) มีความเป็นคู่ขนานของโครงสร้างของแบบสอบ และมีความเที่ยงทัดเทียมกัน (ศิริชัย กาญจนวาสี, 2555, หน้า 159)

ผลการตรวจสอบการมีลักษณะเด่นลักษณะเดียว (Unidimensionality) ของแบบสอบ 3 ฉบับ ที่นำมาใช้ในการปรับเทียบคะแนน ใช้วิธีการวิเคราะห์องค์ประกอบ (Factor analysis) ด้วยการวิเคราะห์ตัวประกอบสำคัญ (Principal component analysis) พบว่า ค่าไอเกนที่เกิน 1 ของแบบสอบทั้ง 3 ฉบับ กระจายเป็นจำนวนมาก แต่มีตัวประกอบหลักตัวแรกที่มีค่าสูง เด่นกว่าตัวประกอบอื่น ๆ รวมทั้งมีเพียงตัวประกอบหลักตัวแรก ตัวประกอบเดียวที่มีตัวแปร 3 ตัวขึ้นไป และแต่ละตัวแปรมีค่าน้ำหนักปัจจัยเกิน 0.40 ที่สามารถจัดเป็นองค์ประกอบได้ (บุญใจ ศรีสถิตย์นรากูล, 2547, หน้า 498) ส่วนตัวประกอบอื่นๆ ค่าน้ำหนักปัจจัยมีค่าน้อยมากไม่เกิน 0.40 หรือถ้าเกินก็ไม่ถึง 3 ตัวแปร ไม่เพียงพอที่จะเป็นตัวประกอบอีกตัวประกอบได้ จึงพอสรุปได้ว่าแบบสอบทั้งหมดที่ใช้มีความเป็นมิติเดียว ดังแสดงในตารางที่ 4-23 และตารางที่ 4-24

ตารางที่ 4-23 ค่าไอเกนและร้อยละของความแปรปรวนของตัวประกอบของแบบสอบ

แบบสอบภาคการศึกษา 1/ 2556			แบบสอบภาคการศึกษา 1/ 2557			แบบสอบภาคการศึกษา 1/ 2558		
ตัวประกอบ	ค่าไอเกน	%ความแปรปรวน	ตัวประกอบ	ค่าไอเกน	%ความแปรปรวน	ตัวประกอบ	ค่าไอเกน	%ความแปรปรวน
1	7.84	6.59	1	7.42	6.24	1	7.47	6.22
2	2.40	2.02	2	2.56	2.15	2	2.65	2.21
3	1.87	1.57	3	2.22	1.87	3	2.24	1.87
4	1.81	1.52	4	2.03	1.70	4	2.03	1.69
5	1.66	1.40	5	1.76	1.48	5	1.77	1.47
6	1.58	1.33	6	1.73	1.46	6	1.74	1.45
7	1.55	1.30	7	1.69	1.42	7	1.69	1.40
8	1.55	1.30	8	1.62	1.36	8	1.62	1.35
9	1.51	1.27	9	1.60	1.35	9	1.61	1.34
10	1.49	1.25	10	1.56	1.31	10	1.57	1.31
⋮			⋮			⋮		
46	1.01	.85	47	1.00	.84	49	1.00	.83

ตารางที่ 4-24 จำนวนตัวแปรในแต่ละตัวประกอบจำแนกตามแบบสอบ

แบบสอบภาค 1/ 2556			แบบสอบภาค1/ 2557		แบบสอบภาค 1/ 2558	
ตัวประกอบ	จำนวนตัวแปร > .4	.35 < จำนวนตัวแปร < .4	จำนวนตัวแปร >.4	.35 < จำนวนตัวแปร < .4	จำนวนตัวแปร > .4	.35 < จำนวนตัวแปร < .4
1	15	9	14	5	13	7
2		3	2	1	2	2
3				2	2	
4				1		
5			1	1		

ตารางที่ 4-24 (ต่อ)

แบบสอบภาค 1/ 2556			แบบสอบภาค1/ 2557		แบบสอบภาค 1/ 2558	
ตัวประกอบ	จำนวนตัวแปร > .4	.35 < จำนวนตัวแปร < .4	จำนวนตัวแปร >.4	.35 < จำนวนตัวแปร < .4	จำนวนตัวแปร > .4	.35 < จำนวนตัวแปร < .4
6				1		1
7	1			1		1
8						
9						
10						
⋮						
46						

สำหรับความเป็นคู่ขนานของโครงสร้างของแบบสอบ แบบสอบทั้ง 3 ฉบับ มีการจัดโครงสร้างและเกณฑ์ในการสุ่มข้อสอบที่เหมือนกัน สำหรับใช้ในการสุ่มข้อสอบจากระบบคลังข้อสอบ ที่แบ่งโครงสร้างการสุ่มข้อสอบจากเนื้อหาชุดวิชาเดียวกัน จำแนกเป็น 15 หน่วยการเรียน ๆ ละ 8 ข้อ และในแต่ละหน่วยจะแบ่งเนื้อหาออกเป็นตอน แต่ละตอนแบ่งตามวัตถุประสงค์ กำหนดจำนวนข้อตามความสำคัญของแต่ละวัตถุประสงค์ ตามแผนผังการสร้างข้อสอบ แบบสอบทุกฉบับถูกสุ่มภายใต้โครงสร้างเดียวกัน นำมาจัดฉบับเป็นแบบสอบชนิดเลือกตอบ 5 ตัวเลือก ฉบับละ 120 ข้อ ใช้เวลาในการสอบ 3 ชั่วโมง ดำเนินการสอบตามคู่มือการปฐมนิเทศ เพื่อให้ทุกสนามสอบดำเนินการในแนวทางเดียวกัน จึงถือได้ว่าแบบสอบมีความเป็นคู่ขนานด้านโครงสร้างของแบบสอบ ประกอบกับเมื่อตรวจสอบค่าความเที่ยง (Reliability) ของแบบสอบ ซึ่งเป็นเงื่อนไขหนึ่งของวิธีการปรับเทียบตามทฤษฎี IRT ที่กำหนดว่าแบบสอบควรมีความเที่ยงทัดเทียมกัน ผลจากการวิเคราะห์ข้อสอบของแบบสอบทั้ง 3 ฉบับ พบว่า แบบสอบของภาคการศึกษา 1/ 2556 ภาคการศึกษา 1/ 2557 และภาคการศึกษา 1/ 2558 มีค่าความเที่ยง 0.83 0.84 และ 0.81 ตามลำดับ ถือได้ว่าแบบสอบที่นำมาใช้ในการปรับเทียบคะแนนมีความเที่ยงทัดเทียมกันเป็นไปตามเงื่อนไขของการปรับเทียบคะแนนด้วย IRT (ศิริชัย กาญจนวาสี, 2555, หน้า 159)

2. วิเคราะห์ปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ โดยใช้โปรแกรม BILOG MG version 3.0 ภายใต้เงื่อนไขการใช้ข้อสอบร่วมภายใน จำนวน 15 ข้อ ที่ได้มาจากการสุ่มข้อสอบที่มีความยากอยู่ในช่วง .4-.6 กับการสุ่มข้อสอบร่วมที่มีความยากอย่างสุ่มจากทุกหน่วย ๆ ละ 1 ข้อ จากนั้นนำมาปรับเทียบคะแนนโดยใช้วิธี IRT 2 พารามิเตอร์ ด้วยวิธีการประมาณค่าพารามิเตอร์พร้อมกัน (Concurrent calibration) วิเคราะห์ผลโดยใช้ข้อมูลผลการสอบทั้งหมดกับตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ภายใต้กลุ่มตัวอย่างขนาดแตกต่างกัน 3 ขนาด คือ 100 คน 500 คน และ 700 คน ตามลำดับ หลังจากประมาณค่าพารามิเตอร์ของข้อสอบรายข้อจากการวิเคราะห์ดังกล่าวแล้ว นำค่าพารามิเตอร์ของข้อสอบ ค่าความสามารถและค่าถ่วงน้ำหนัก (Theta points and weights) ไปใช้ในการประมาณคะแนนดิบ (Raw scores) ของผู้สอบ โดยใช้โปรแกรม IRT-CLASS ที่พัฒนาโดย Won-Chan Lee and Michael J. Kolen ในปี ค.ศ. 2008 จากภาษาฟอร์แทรน รายละเอียดผลการวิเคราะห์ ดังนี้

2.1 ผลจากการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ของแบบสอบทุกเงื่อนไข ข้อสอบข้อใดมีค่า initial slope ต่ำกว่า -0.15 โปรแกรมจะตัดข้อสอบข้อนั้นทิ้งไม่นำไปใช้ในการคำนวณค่าพารามิเตอร์ ปรากฏว่าเหลือข้อสอบในแต่ละเงื่อนไขของการปรับเทียบคะแนนทั้งที่เป็นข้อสอบร่วม ข้อสอบปกติที่ไม่ใช่ข้อสอบร่วม ดังแสดงในตารางที่ 4-25

ตารางที่ 4-25 จำนวนข้อสอบที่นำไปใช้ในการปรับเทียบคะแนนของแต่ละเงื่อนไข

เงื่อนไข/ ขนาดตัวอย่าง	ข้อสอบร่วม	ข้อสอบปกติ	รวมทั้งฉบับ
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6			
1.1 100 คน	15	99	114
1.2 500 คน	15	104	119
1.3 700 คน	13	103	116
2. ข้อสอบร่วมมีความยากอย่างสุ่ม			
2.1 100 คน	15	99	114
2.2 500 คน	15	104	119
2.3 700 คน	15	104	119
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
3.1 100 คน	15	87	102

ตารางที่ 4-25 (ต่อ)

เงื่อนไข/ ขนาดตัวอย่าง	ข้อสอบร่วม	ข้อสอบปกติ	รวมทั้งฉบับ
3.2 500 คน	15	90	105
3.3 700 คน	15	90	105
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
4.1 100 คน	15	88	103
4.2 500 คน	15	93	108
4.3 700 คน	15	93	108

2.2 คุณภาพของแบบสอบหลังจากปรับเทียบ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ระหว่างที่มีการตัดกับไม่มีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อใช้กลุ่มตัวอย่างที่แตกต่างกัน 3 ขนาด คือ 100 คน 500 คน และ 700 คน ผลการวิเคราะห์พบว่า ค่าเฉลี่ยของค่าอำนาจจำแนก (a) ของทุกเงื่อนไขมีค่าใกล้เคียงกันอยู่ในช่วง 0.51-0.59 ถือได้ว่าค่าอำนาจจำแนกอยู่ในเกณฑ์ของการคัดเลือกข้อสอบที่ดี (0.5-2.5) ที่นิยมใช้กันทั่วไป ส่วนค่าเฉลี่ยของความยากอยู่ในช่วง 0.24-1.16 ถือได้ว่าความยากอยู่ในเกณฑ์ของการคัดเลือกข้อสอบที่ดี (-2.5-2.5) ที่นิยมใช้เช่นเดียวกัน (ศิริชัย กาญจนวาสี, 2555, หน้า 55) ดังแสดงในตารางที่ 4-26

ตารางที่ 4-26 ค่าอำนาจจำแนกและค่าความยากของแบบสอบจากการปรับเทียบคะแนนด้วยวิธี IRT จำแนกตามเงื่อนไขการปรับเทียบคะแนน

เงื่อนไข/ ขนาดตัวอย่าง	ค่าอำนาจจำแนก (a)					ค่าความยาก (b)				
	Max	**Min**	**Range**	$\bar{X}$	***SD***	**Max**	**Min**	**Range**	$\bar{X}$	***SD***
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6										
1.1 100 คน	1.54	0.31	1.23	0.58	0.19	4.78	-4.04	8.82	0.49	1.80
1.2 500 คน	1.17	0.18	0.99	0.53	0.21	8.12	-5.16	13.28	0.92	2.41
1.3 700 คน	1.17	0.13	1.04	0.52	0.24	9.33	-6.58	15.91	1.07	2.66

ตารางที่ 4-26 (ต่อ)

เงื่อนไข/ ขนาดตัวอย่าง	ค่าอำนาจจำแนก (a)					ค่าความยาก (b)				
	Max	**Min**	**Range**	$\bar{X}$	***SD***	**Max**	**Min**	**Range**	$\bar{X}$	***SD***
2. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง										
2.1 100 คน	1.54	0.28	1.26	0.59	0.20	3.74	-2.56	6.30	0.48	1.51
2.2 500 คน	1.05	0.18	0.87	0.53	0.22	6.33	-5.19	11.52	0.57	1.95
2.3 700 คน	1.28	0.14	1.15	0.53	0.26	7.32	-6.52	13.84	0.91	2.22
3. ข้อสอบร่วมมีความยากอย่างสุ่ม										
3.1 100 คน	1.17	0.31	0.86	0.56	0.17	4.70	-3.20	7.90	0.53	1.99
3.2 500 คน	1.37	0.17	1.20	0.51	0.25	7.6	-5.10	12.70	0.81	2.59
3.3 700 คน	1.44	0.14	1.31	0.52	0.27	10.77	-5.12	15.89	1.16	2.72
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง										
4.1 100 คน	1.08	0.30	0.78	0.57	0.17	4.67	-2.94	7.61	0.24	1.71
4.2 500 คน	1.35	0.17	1.19	0.51	0.25	6.15	-5.08	11.23	0.48	2.30
4.3 700 คน	1.39	0.14	1.25	0.52	0.26	6.79	-5.04	11.83	0.87	2.32

3. การประมาณค่าคะแนนดิบจากการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ นำผลการวิเคราะห์การปรับเทียบคะแนน ค่าอำนาจจำแนก (a) และค่าความยาก (b) ค่าระดับความสามารถและค่าถ่วงน้ำหนัก (Theta points and weights) ของแบบสอบทั้ง 2 ฉบับ ที่มีข้อสอบร่วมภายในมาวิเคราะห์ด้วย IRT-CLASS เพื่อสร้างสมการพยากรณ์คะแนนดิบ จากค่าระดับความสามารถ (Theta) ค่าคาดหวังของคะแนนดิบ (Expected raw scores) ตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน โดยกำหนดสัญลักษณ์ในสมการพยากรณ์ ดังนี้

$\hat{Y}$ แทน คะแนนดิบที่ได้จากการพยากรณ์

X แทน ระดับความสามารถ (Theta)

3.1 สมการพยากรณ์คะแนนดิบหลังการปรับเทียบคะแนน ตามเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 จำแนกตามขนาดกลุ่มตัวอย่างที่ใช้ในการปรับเทียบคะแนน ได้ผลดังนี้ (Lord, 1982, cited in Davier & Holland and Thayer, 2003, p. 68)

$\hat{Y} = 94.39 + 25.94X$ ใช้กับกลุ่มตัวอย่างขนาด 100 คน

$\hat{Y} = 90.61 + 23.52X$ ใช้กับกลุ่มตัวอย่างขนาด 500 คน

$\hat{Y} = 89.78 + 23.70X$ ใช้กับกลุ่มตัวอย่างขนาด 700 คน

3.2 สมการพยากรณ์คะแนนดิบหลังการปรับเทียบคะแนน ตามเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม จำแนกตามขนาดกลุ่มตัวอย่างที่ใช้ในการปรับเทียบคะแนน ได้ผล ดังนี้

$\hat{Y} = 95.26 + 25.21X$ ใช้กับกลุ่มตัวอย่างขนาด 100 คน

$\hat{Y} = 93.33 + 22.99X$ ใช้กับกลุ่มตัวอย่างขนาด 500 คน

$\hat{Y} = 92.61 + 22.87X$ ใช้กับกลุ่มตัวอย่างขนาด 700 คน

3.3 สมการพยากรณ์คะแนนดิบหลังการปรับเทียบคะแนน ตามเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดกลุ่มตัวอย่างที่ใช้ในการปรับเทียบคะแนน ได้ผล ดังนี้

$\hat{Y} = 85.10 + 23.84X$ ใช้กับกลุ่มตัวอย่างขนาด 100 คน

$\hat{Y} = 82.88 + 22.14X$ ใช้กับกลุ่มตัวอย่างขนาด 500 คน

$\hat{Y} = 82.05 + 21.33X$ ใช้กับกลุ่มตัวอย่างขนาด 700 คน

3.4 สมการพยากรณ์คะแนนดิบหลังการปรับเทียบคะแนน ตามเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง จำแนกตามขนาดกลุ่มตัวอย่างที่ใช้ในการปรับเทียบคะแนน ได้ผล ดังนี้

$\hat{Y} = 87.81 + 23.97X$ ใช้กับกลุ่มตัวอย่างขนาด 100 คน

$\hat{Y} = 86.41 + 21.80X$ ใช้กับกลุ่มตัวอย่างขนาด 500 คน

$\hat{Y} = 86.01 + 21.44X$ ใช้กับกลุ่มตัวอย่างขนาด 700 คน

สำหรับความคลาดเคลื่อนมาตรฐานของการประมาณค่าคะแนนดิบ (Raw scores) ด้วย IRT-CLASS เมื่อวิเคราะห์ภายใต้เงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน พิจารณาจากความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบ (Overall error variance for raw

scores) และค่าความเที่ยง (Reliability) พบว่า ความเที่ยงมีค่าสูงมากเกือบเท่ากันทุกเงื่อนไข มีค่าอยู่ในช่วง 0.96-0.97 เช่นเดียวกันกับค่าความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบที่มีค่าใกล้เคียงกันระหว่าง 29.80-32.88 และเมื่อพิจารณาค่าความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบของปรับเทียบคะแนนกรณีวิเคราะห์โดยตัดข้อสอบที่ไม่มีคุณภาพทิ้ง พบว่ามีคุณภาพในการปรับเทียบคะแนนสูงกว่าการไม่ตัดข้อสอบทิ้งไม่ว่าจะวิเคราะห์กับกลุ่มตัวอย่างตามที่กำหนดขนาดใดก็ตาม โดยภาพรวมการประมาณค่าคะแนนดิบตามเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้งที่ศึกษากับกลุ่มตัวอย่างทั้ง 3 ขนาด ให้ค่าความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบต่ำสุด รองลงมาคือ การประมาณคะแนนดิบเมื่อใช้ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้งกับกลุ่มตัวอย่างทั้ง 3 ขนาด ดังแสดงในตารางที่ 4-27

ตารางที่ 4-27 ค่าความเที่ยงและความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบจำแนกตามเงื่อนไขของการปรับเทียบคะแนน

เงื่อนไข/ กลุ่มตัวอย่าง	ความเที่ยง	ความคลาดเคลื่อนของความแปรปรวนโดยรวม
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6		
1.1 ตัวอย่าง 100 คน	0.97	31.06
1.2 ตัวอย่าง 500 คน	0.96	30.90
1.3 ตัวอย่าง 700 คน	0.96	32.88
2. ข้อสอบร่วมมีความยากอย่างสุ่ม		
2.1 ตัวอย่าง 100 คน	0.97	31.17
2.2 ตัวอย่าง 500 คน	0.96	31.53
2.3 ตัวอย่าง 700 คน	0.96	31.58
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
3.1 ตัวอย่าง 100 คน	0.96	30.66
3.2 ตัวอย่าง 500 คน	0.96	29.80
3.3 ตัวอย่าง 700 คน	0.96	29.84

ตารางที่ 4-27 (ต่อ)

เงื่อนไข/ กลุ่มตัวอย่าง	ความเที่ยง	ความคลาดเคลื่อนของความแปรปรวนโดยรวม
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง		
4.1 ตัวอย่าง 100 คน	0.97	30.11
4.2 ตัวอย่าง 500 คน	0.96	30.54
4.3 ตัวอย่าง 700 คน	0.96	30.64

หลังจากได้สมการพยากรณ์คะแนนดิบ ขั้นตอนต่อไปจะเป็นการพยากรณ์หาคะแนนดิบหลังการปรับเทียบคะแนนของแต่ละเงื่อนไข นำไปใช้ในการตัดเกรดในตอนที่ 5 ต่อไป

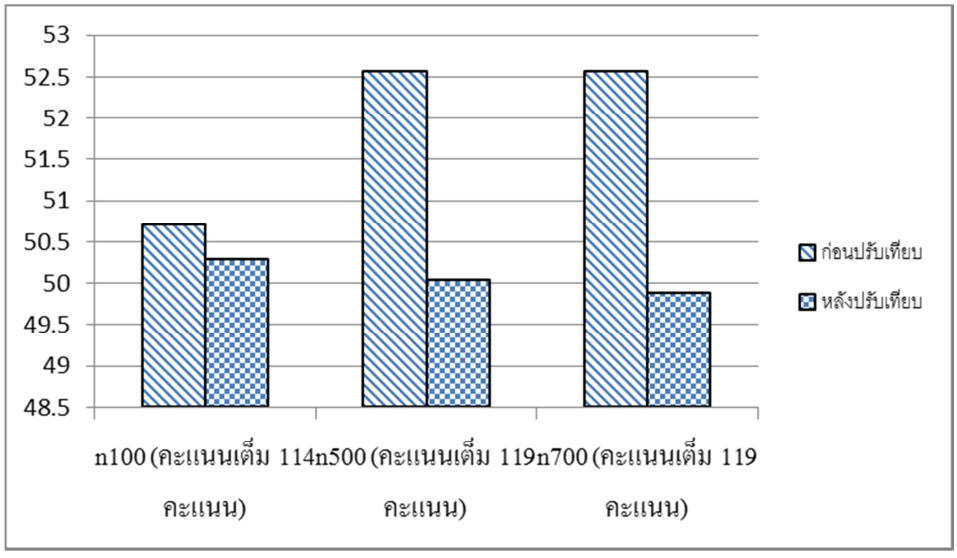
4. ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ภายใต้เงื่อนไขต่าง ๆ จากสมการพยากรณ์คะแนนดิบ ภายใต้เงื่อนไขต่าง ๆ ได้ผลดังนี้

4.1 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนนและหลังการปรับเทียบคะแนนจำแนกตามขนาดตัวอย่าง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลงเล็กน้อย ยกเว้นเมื่อวิเคราะห์ด้วยกลุ่มตัวอย่างขนาด 100 คนที่เพิ่มขึ้นเล็กน้อยโดยที่ก่อนการปรับเทียบคะแนนคะแนนเฉลี่ยอยู่ในช่วง 48.42-50.06 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 12.02-12.30 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนมีแนวโน้มลดลงอยู่ในช่วง 46.56-49.82 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11.61-12.66 คะแนนโดยรวมแสดงให้เห็นว่าหลังจากปรับเทียบคะแนนของแบบสอบภาคการศึกษา 1/ 2557 ให้อยู่บนสเกลเดียวกันกับแบบสอบภาคการศึกษา 1/ 2556 คะแนนลดลง นั่นคือ แบบสอบแบบสอบภาคการศึกษา 1/ 2557 เป็นแบบสอบที่ง่ายกว่าแบบสอบของภาคการศึกษา 1/ 2556 ดังแสดงในตารางที่ 4-28 และภาพที่ 4-15

ตารางที่ 4-28 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

เงื่อนไข/ กลุ่มตัวอย่าง	คะแนนเต็ม	คะแนน		$\bar{X}$	*SD*
		ต่ำสุด	สูงสุด		
แบบสอบเดิม ภาค 1/ 2557	120	17	90	50.27	12.24
1. ตัวอย่าง 100 คน	114				
ก่อนปรับเทียบ		16	89	49.21	12.25
หลังปรับเทียบ		17	92	49.82	12.66
2. ตัวอย่าง 500 คน	119				
ก่อนปรับเทียบ		16	90	50.06	12.30
หลังปรับเทียบ		18	88	48.57	11.66
3. ตัวอย่าง 700 คน	116				
ก่อนปรับเทียบ		15	87	48.42	12.02
หลังปรับเทียบ		17	87	46.56	11.61



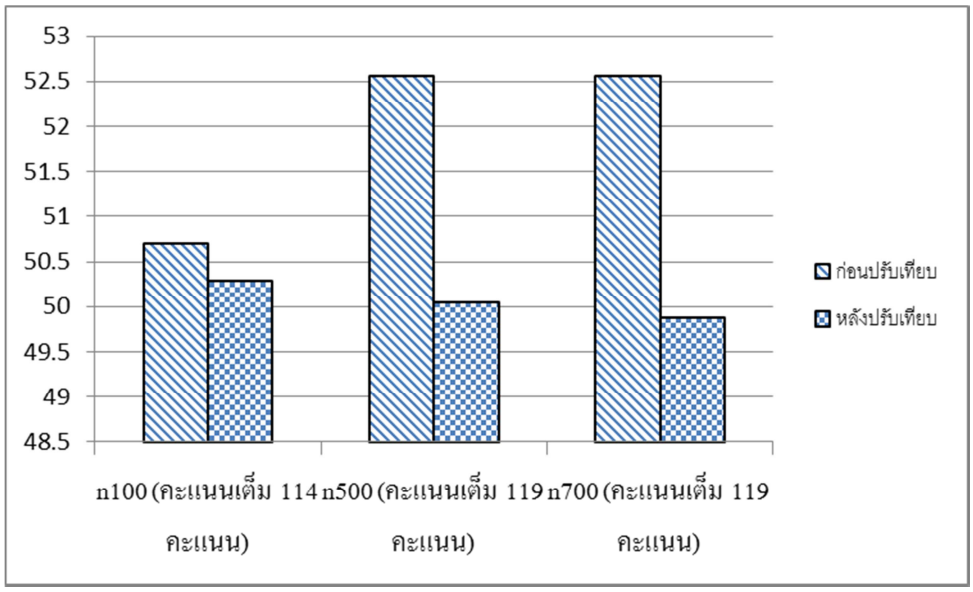
ภาพที่ 4-15 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

4.2 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนนและหลังการปรับเทียบคะแนนจำแนกตามขนาดตัวอย่าง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลง โดยที่ก่อนการปรับเทียบคะแนนคะแนนเฉลี่ยอยู่ในช่วง 50.70-52.56 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11.35-11.39 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนมีแนวโน้มลดลงอยู่ในช่วง 49.88-50.29 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11.16-12.16 คะแนน แสดงให้เห็นว่าหลังจากปรับเทียบคะแนนของแบบสอบภาคการศึกษา 1/ 2558 ให้อยู่บนสเกลเดียวกันกับแบบสอบภาคการศึกษา 1/ 2556 คะแนนลดลง นั่นคือ แบบสอบแบบสอบภาคการศึกษา 1/ 2558 เป็นแบบสอบที่ง่ายกว่าแบบสอบของภาคการศึกษา 1/ 2556 ดังแสดงในตารางที่ 4-29 และภาพที่ 4-16

ตารางที่ 4-29 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

เงื่อนไข/ กลุ่มตัวอย่าง	**คะแนนเต็ม**	**คะแนน**		$\bar{X}$	***SD***
		ต่ำสุด	**สูงสุด**		
แบบสอบเดิม ภาค 1/ 2558	120	26	89	52.88	11.33
1. ตัวอย่าง 100 คน	114				
ก่อนปรับเทียบ		24	87	50.70	11.35
หลังปรับเทียบ		21	90	50.29	12.16
2. ตัวอย่าง 500 คน	119				
ก่อนปรับเทียบ		26	89	52.56	11.39
หลังปรับเทียบ		22	8	50.04	11.26
3. ตัวอย่าง 700 คน	119				
ก่อนปรับเทียบ		26	89	52.56	11.39
หลังปรับเทียบ		23	86	49.88	11.16



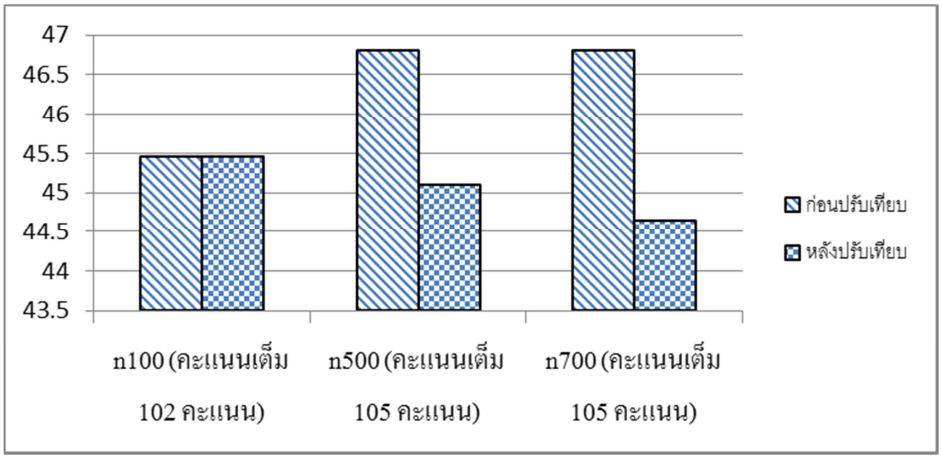
ภาพที่ 4-16 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

4.3 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง.4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนนและหลังการปรับเทียบคะแนนจำแนกตามขนาดตัวอย่าง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลงเล็กน้อยมาก โดยที่ก่อนการปรับเทียบคะแนนคะแนนเฉลี่ยอยู่ในช่วง 45.45-46.80 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 11.80-11.84 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนมีแนวโน้มลดลงอยู่ในช่วง 44.64-45.45 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 10.73-11.77 คะแนน แสดงให้เห็นว่าหลังจากปรับเทียบคะแนนของแบบสอบภาคการศึกษา 1/ 2557 ให้อยู่บนสเกลเดียวกันกับแบบสอบภาคการศึกษา 1/ 2556 คะแนนลดลงเล็กน้อย นั่นคือแบบสอบแบบสอบภาคการศึกษา 1/ 2557 เป็นแบบสอบที่ง่ายกว่าแบบสอบของภาคการศึกษา 1/ 2556 เล็กน้อย ดังแสดงในตารางที่ 4-30 และภาพที่ 4-17

ตารางที่ 4-30 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

เงื่อนไข/ กลุ่มตัวอย่าง	คะแนนเต็ม	คะแนน		$\bar{X}$	*SD*
		ต่ำสุด	สูงสุด		
แบบสอบเดิม ภาค 1/ 2557	120	17	90	50.27	12.24
1. ตัวอย่าง 100 คน	102				
ก่อนปรับเทียบ		16	84	45.45	11.80
หลังปรับเทียบ		15	85	45.45	11.77
2. ตัวอย่าง 500 คน	105				
ก่อนปรับเทียบ		16	85	46.80	11.84
หลังปรับเทียบ		16	82	45.09	11.14
3. ตัวอย่าง 700 คน	105				
ก่อนปรับเทียบ		16	85	46.80	11.84
หลังปรับเทียบ		17	81	44.64	10.73



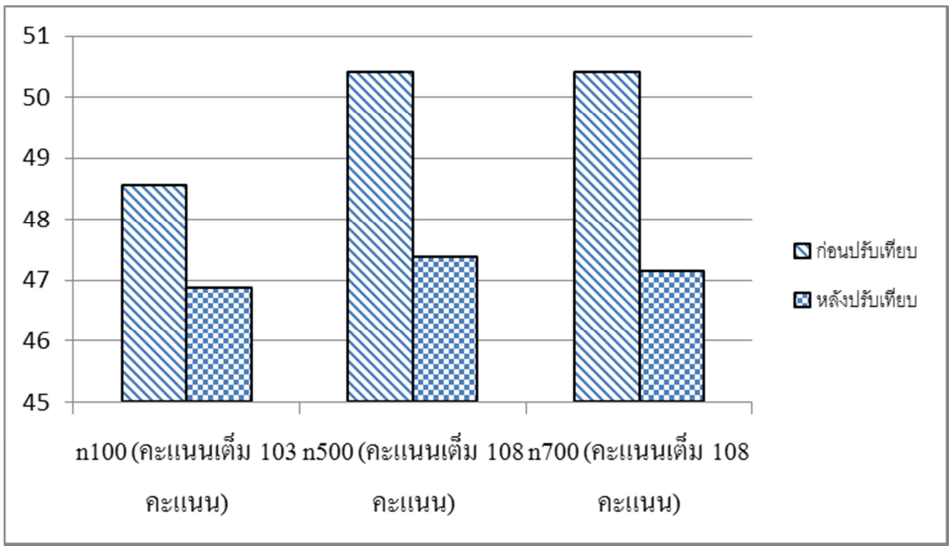
ภาพที่ 4-17 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

4.4 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

เมื่อพิจารณาค่าสถิติพื้นฐานของแบบสอบก่อนการปรับเทียบคะแนน และหลังการปรับเทียบคะแนนจำแนกตามขนาดตัวอย่าง พบว่า คะแนนเฉลี่ยหลังการปรับเทียบคะแนนลดลง โดยที่ก่อนการปรับเทียบคะแนนคะแนนเฉลี่ยอยู่ในช่วง 48.55-50.41 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 10.94-10.99 คะแนน ขณะที่คะแนนเฉลี่ยหลังการปรับเทียบคะแนนมีแนวโน้มลดลงอยู่ในช่วง 46.87-47.38 คะแนน ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 10.74-11.69 คะแนน แสดงให้เห็นว่าหลังจากปรับเทียบคะแนนของแบบสอบภาคการศึกษา 1/ 2558 ให้อยู่บนสเกลเดียวกันกับแบบสอบภาคการศึกษา 1/ 2556 คะแนนลดลง นั่นคือแบบสอบแบบสอบภาคการศึกษา 1/ 2558 เป็นแบบสอบที่ง่ายกว่าแบบสอบของภาคการศึกษา 1/ 2556 ดังแสดงในตารางที่ 4-31 และภาพที่ 4-18

ตารางที่ 4-31 ค่าสถิติพื้นฐานของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

เงื่อนไข/ กลุ่มตัวอย่าง	**คะแนนเต็ม**	**คะแนน**		$\bar{X}$	***SD***
		ต่ำสุด	**สูงสุด**		
แบบสอบเดิม ภาค 1/ 2558	120	26	89	52.88	11.33
1. ตัวอย่าง 100 คน	103				
ก่อนปรับเทียบ		23	83	48.55	10.94
หลังปรับเทียบ		19	85	46.87	11.69
2. ตัวอย่าง 500 คน	108				
ก่อนปรับเทียบ		25	85	50.41	10.99
หลังปรับเทียบ		21	83	47.38	10.92
3. ตัวอย่าง 700 คน	108				
ก่อนปรับเทียบ		25	85	50.41	10.99
หลังปรับเทียบ		21	82	47.16	10.74



ภาพที่ 4-18 ค่าเฉลี่ยของแบบสอบก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ตอนที่ 4 ผลการเปรียบเทียบคุณภาพการปรับเทียบคะแนนวิธีเคอเนลและวิธี IRT

การตรวจสอบคุณภาพการปรับเทียบคะแนน มีด้วยกันหลายวิธีแต่ด้วยข้อจำกัดของการศึกษาครั้งนี้ที่ศึกษาในสถานการณ์การสอบจริง จึงไม่สามารถใช้วิธีการสอบทานผลด้วยการให้ผู้สอบส่วนหนึ่งทำแบบสอบ 2 ฉบับ แล้วนำมาหาคุณภาพของการปรับเทียบคะแนนได้ จึงใช้วิธีการหาคุณภาพของการปรับเทียบคะแนน จากค่าสถิติที่ได้จากโปรแกรมการวิเคราะห์ โดยจะแบ่งการนำเสนอออกเป็น 2 ประเด็น ประเด็นแรกจะเป็นการเปรียบเทียบคุณภาพของการปรับเทียบคะแนนจำแนกตามวิธีเคอเนลและวิธี IRT 2 กับประเด็นที่สองเป็นการเปรียบเทียบคุณภาพการปรับเทียบคะแนนด้วยบูทแทรบ รายละเอียด ดังนี้

1. การเปรียบเทียบคุณภาพการปรับเทียบคะแนนจำแนกตามวิธีการปรับเทียบวิธีเคอเนล

การตรวจสอบคุณภาพโดยรวมของการปรับเทียบคะแนนภายใต้เงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน วิธีเคอเนลจะพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน (SEE) ส่วนวิธี IRT 2 พารามิเตอร์จะพิจารณาจากความเที่ยงและความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบ

คุณภาพโดยรวมของการปรับเทียบคะแนนด้วยวิธีเคอเนล พิจารณาจากค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนน พบว่าค่าเฉลี่ยของ SEE ของทุกเงื่อนไขอยู่ในช่วง 1.19-3.29 ส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 0.33-1.26 โดยการปรับเทียบคะแนนตามเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบ

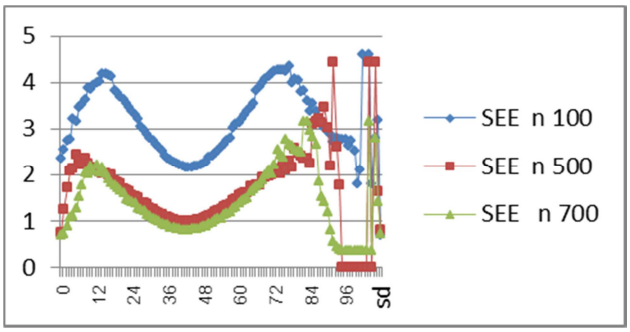
ที่ไม่มีคุณภาพทิ้ง เมื่อวิเคราะห์โดยใช้กลุ่มตัวอย่างขนาด 700 คน ให้ค่าเฉลี่ยของ SEE ต่ำสุด (1.19) ขณะที่การปรับเทียบคะแนนตามเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อวิเคราะห์โดยใช้กลุ่มตัวอย่างขนาด 100 คน ให้ค่าเฉลี่ยของ SEE สูงสุด (3.29)

เมื่อพิจารณาการปรับเทียบคะแนนวิธีเคอเนลจำแนกตามขนาดตัวอย่างพบว่า เมื่อใช้กลุ่มตัวอย่างขนาด 100 คน การปรับเทียบคะแนนด้วยเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง มีความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด ส่วนเงื่อนไขอื่น ๆ มีความคลาดเคลื่อนแตกต่างกันเล็กน้อย ส่วนการวิเคราะห์กับกลุ่มตัวอย่างขนาด 500 คน ทุกเงื่อนไขมีความคลาดเคลื่อนมาตรฐานพอ ๆ กัน ยกเว้นเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ที่มีความคลาดเคลื่อนมาตรฐานค่อนข้างสูง สำหรับการวิเคราะห์กับกลุ่มตัวอย่างขนาด 700 คน การปรับเทียบคะแนนด้วยเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง มีความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนต่ำสุด ส่วนเงื่อนไขอื่น ๆ มีความคลาดเคลื่อนมาตรฐานต่ำพอ ๆ กัน

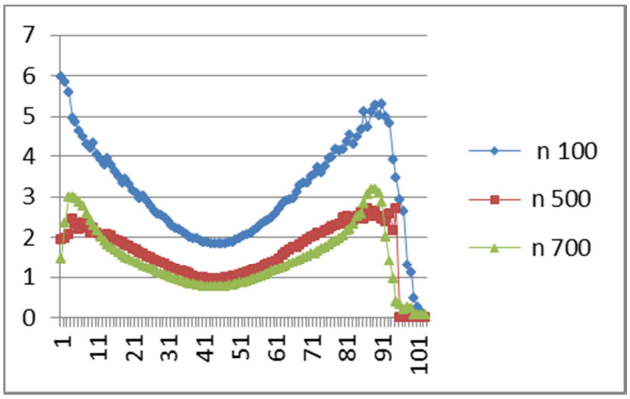
นอกจากนี้พบว่าเมื่อตัดข้อสอบที่ไม่มีคุณภาพทิ้งของทั้งเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนจะลดต่ำลงไม่ว่าจะใช้กลุ่มตัวอย่างขนาดใดก็ตาม ยกเว้นการตัดข้อสอบที่ไม่มีคุณภาพทิ้งของเงื่อนไขการตัดข้อสอบที่ไม่มีคุณภาพทิ้งของเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อวิเคราะห์กับกลุ่มตัวอย่างขนาด 500 คน กับเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม เมื่อวิเคราะห์กับกลุ่มตัวอย่างขนาด 100 คน ที่เพิ่มขึ้นเล็กน้อย นอกจากนี้พบว่าในแต่ละเงื่อนไขของการปรับเทียบคะแนน เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้นความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนจะมีแนวโน้มลดลง ส่วนกรณีการปรับเทียบคะแนนโดยใช้ข้อสอบร่วมที่มีความยากปานกลาง (.4-.6) กับความยากอย่างสุ่ม ให้ค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนใกล้เคียงกัน ดังแสดงในตารางที่ 4-32 และภาพที่ 4-19

ตารางที่ 4-32 ภาพรวมความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนวิธีเคอเนล

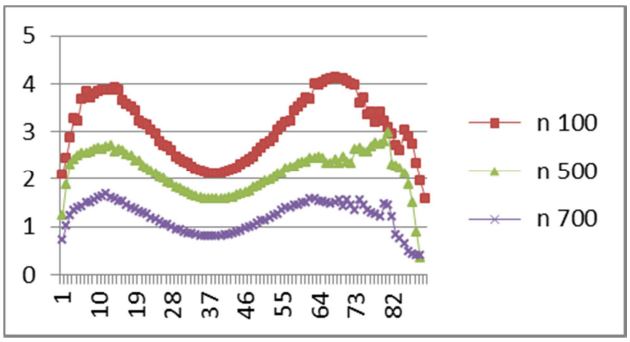
เงื่อนไข	ความคลาดเคลื่อนมาตรฐาน					
	100 คน		500 คน		700 คน	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	3.17	0.70	1.63	0.81	1.42	0.72
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	3.14	1.26	1.60	0.71	1.45	0.78
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	3.07	0.67	2.14	0.46	1.19	0.33
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	3.29	1.23	1.51	0.56	1.42	0.67



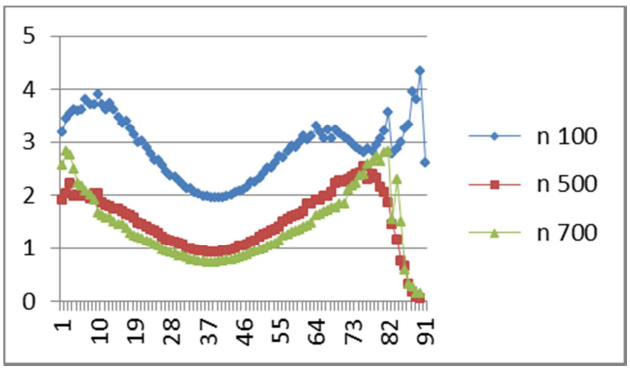
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6



2. ข้อสอบร่วมมีความยากอย่างสุ่ม



3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง



4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ภาพที่ 4-19 ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนวิธีเคอเนลจำแนกตามเงื่อนไข

2. การเปรียบเทียบคุณภาพของการแปลงเป็นคะแนนดิบ (Raw scores) ด้วยโปรแกรม IRT CLASS จากการนำผลการวิเคราะห์การปรับเทียบคะแนนวิธี IRT 2 พารามิเตอร์โดยการสร้างสมการพยากรณ์ สามารถเปรียบเทียบคุณภาพของการแปลงเป็นคะแนนดิบ จากความเที่ยงและความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบ ผลการวิเคราะห์พบว่า ไม่ว่าใช้เงื่อนไขใดกับทุกขนาดกลุ่มตัวอย่างได้ค่าความเที่ยงสูงมากใกล้เคียงกันตั้งแต่ .9 ขึ้นไป ส่วนความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบก็มีค่าใกล้เคียงกันอยู่ในช่วง 29.80-32.88 โดยการแปลงเป็นคะแนนดิบตามเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อวิเคราะห์กับกลุ่มตัวอย่างขนาด 500 คน มีความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบต่ำสุด ขณะที่การแปลงเป็นคะแนนดิบด้วยเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 เมื่อวิเคราะห์กับกลุ่มตัวอย่างขนาด 700 คน มีความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบสูงสุด สำหรับการปรับเทียบคะแนนกรณีใช้ข้อสอบร่วมที่มีความยากปานกลาง (.4-.6) กับความยากอย่างสุ่ม ค่าความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบแตกต่างกันเล็กน้อยมาก

นอกจากนี้พบว่าเมื่อตัดข้อสอบที่ไม่มีคุณภาพทิ้งของทั้งเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ค่าความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบจะลดต่ำลงไม่ว่าจะใช้กลุ่มตัวอย่างขนาดใดก็ตาม ดังแสดงในตารางที่ 4-33

ตารางที่ 4-33 ค่าความเที่ยงและความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบจำแนกตามเงื่อนไขของการปรับเทียบคะแนน

เงื่อนไข	100 คน		500 คน		700 คน	
	reli	var	reli	var	reli	var
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	0.97	31.06	0.96	30.90	0.96	32.88
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	0.97	31.17	0.96	31.53	0.96	31.58
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	0.97	30.66	0.96	29.80	0.96	29.84
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	0.97	30.11	0.96	30.54	0.96	30.64

หมายเหตุ reli: reliability for raw score, var: overall error variance for raw scores

3. การเปรียบเทียบคุณภาพการปรับเทียบคะแนนของวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ด้วยการหาความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนจากคะแนนรวมเดิมกับคะแนนหลังการปรับเทียบจากสูตร

$$SEE = \sqrt{RMSE^2 - BIAS^2}$$

โดยภาพรวมพบว่า การปรับเทียบคะแนนวิธีเคอเนลมีคุณภาพสูงกว่าวิธี IRT 2 พารามิเตอร์ และเมื่อพิจารณาตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน พบว่าการใช้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้งมีคุณภาพในการปรับเทียบคะแนนสูงกว่าเงื่อนไขอื่น ๆ ทั้ง 3 ขนาดตัวอย่าง ซึ่งให้ผลสอดคล้องกันไม่ว่าจะวิเคราะห์ด้วยวิธีเคอเนลหรือวิธี IRT 2 พารามิเตอร์ โดยที่การวิเคราะห์ด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มกับตัวอย่างขนาด 700 คน และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้งกับตัวอย่างขนาด 500 คน มีคุณภาพในการปรับเทียบคะแนนสูงที่สุด (SEE = 0.18) ดังแสดงในตารางที่ 4-34

ตารางที่ 4-34 เปรียบเทียบความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนระหว่างวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์

เงื่อนไข	เคอเนล			IRT 2 พารามิเตอร์		
	n100	n500	n700	n100	n500	n700
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	2.21	3.58	3.61	14.37	13.83	13.63
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	2.68	0.48	0.18	2.67	2.59	2.58
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	3.24	5.33	5.49	13.54	13.22	13.00
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	2.44	0.18	0.43	2.75	2.69	2.67

ตอนที่ 5 เปรียบเทียบความสอดคล้องของการตัดเกรดระหว่างก่อนและหลังการปรับเทียบคะแนน

การเปรียบเทียบความสอดคล้องของผลการตัดเกรด จากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังการปรับเทียบคะแนน เพื่อตรวจสอบดูว่าระหว่างที่ทำการปรับเทียบคะแนนกับไม่มีการปรับเทียบคะแนน จะส่งผลต่อการตัดเกรดหรือไม่ ทำให้ผู้เรียนที่ทำแบบสอบต่างฉบับที่มีโครงสร้างของแบบสอบ และบริหารการสอบแบบเดียวกัน เกิดการได้เปรียบหรือเสียเปรียบกันหรือไม่ โดยใช้วิธีการปรับเทียบคะแนน 2 วิธี คือ เคอเนล และวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ข้อสอบร่วมมีความยากอย่างสุ่ม และวิเคราะห์ข้อมูลโดยการตัดและไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้งกับกลุ่มตัวอย่างที่มีขนาดแตกต่างกัน 3 ขนาด คือ 100 คน 500 คน และ 700 คน

การศึกษาครั้งนี้ใช้วิธีการตัดเกรด 2 วิธี คือ ตัดเกรด 3 ระดับ กับ 8 ระดับ เปรียบเทียบความสอดคล้องของผลการตัดเกรดตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน จากดัชนีของแคปปา สำหรับการนำเสนอผลการวิเคราะห์การตัดเกรดแบ่งออกเป็น 3 ตอนย่อย คือตอนที่ 5.1 การเปรียบเทียบผลการตัดเกรด ก่อนและหลังการปรับเทียบคะแนนด้วยวิธี เคอเนล ตอนที่ 5.2 การเปรียบเทียบผลการตัดเกรด ก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ และตอนที่ 5.3 การเปรียบเทียบความสัมพันธ์ของการตัดเกรดระหว่างวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ รายละเอียด ดังนี้

ตอนที่ 1 การเปรียบเทียบผลการตัดเกรด ก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล

การนำเสนอผลการตัดเกรดก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ในที่นี้จะแยกการนำเสนอออกเป็น 2 ข้อย่อย ตามระดับของการตัดเกรด คือ 3 ระดับ และ 8 ระดับ ดังนี้

1. การเปรียบเทียบผลการตัดเกรด 3 ระดับ ระหว่างการใช้คะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล

การตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ข้อสอบร่วมมีความยากอย่างสุ่ม ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และข้อสอบร่วมมีความยากอย่างสุ่มระหว่างการตัดและไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง กับกลุ่มตัวอย่างขนาด 100 คน 500 คน และ 700 คน โดยกำหนดเกณฑ์คะแนนจุดตัด ในการตัดเกรด ดังนี้

1.1 เกรด H ระดับ 4.00 เมื่อได้คะแนนมากกว่า 75%

1.2 เกรด S ระดับ 2.33 เมื่อได้คะแนน 60-75%

1.3 เกรด U ระดับ 0 เมื่อได้คะแนนต่ำกว่า 60%

จากข้อกำหนดดังกล่าว นำมาสร้างเป็นเกณฑ์การตัดเกรดของแต่ละเงื่อนไขของการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-35

ตารางที่ 4-35 คะแนนจุดตัดในการตัดเกรด 3 ระดับ จำแนกตามเงื่อนไขการปรับเทียบคะแนนด้วยวิธีเคอเนล

เงื่อนไข/ ขนาดตัวอย่าง	**เกณฑ์ที่ใช้ในการตัดเกรด**		
	< ร้อยละ60	**ร้อยละ 60-75**	**> ร้อยละ 75**
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6			
1.1 100 คน	< 68	68-86	> 86
1.2 500 คน	< 71	71-89	> 89
1.3 700 คน	< 70	70-87	> 87
2. ข้อสอบร่วมมีความยากอย่างสุ่ม			
2.1 100 คน	< 68	68-86	> 86
2.2 500 คน	< 71	71-89	> 89
2.3 700 คน	< 71	71-89	> 89
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
3.1 100 คน	< 61	61-76	> 77
3.2 500 คน	< 63	63-79	> 79
3.3 700 คน	< 63	63-79	> 79
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
4.1 100 คน	< 62	62-77	> 77
4.2 500 คน	< 65	65-81	> 81
4.3 700 คน	< 65	65-81	> 81

สำหรับเกณฑ์การพิจารณาความสอดคล้องตามดัชนีแคปปา ประยุกต์มาจากเกณฑ์ของ Landis and Koch ที่กำหนดค่าดัชนีแคปปาตั้งแต่ .81-1.00 ว่ามีความสอดคล้องในระดับมาก (Landis & Koch, 1977 อ้างถึงใน ประสพชัย พสุนนท์, หน้า 8) ซึ่งในการนำมาใช้ในเรื่องของการตัดเกรดหลังจากการปรับเทียบคะแนน ถ้าหากการตัดเกรดก่อนที่จะปรับเทียบคะแนนกับการตัดเกรดหลังจากปรับเทียบคะแนนค่าดัชนีแคปปาตั้งแต่ .81 ขึ้นไป แสดงว่าจะทำการปรับเทียบคะแนนหรือไม่ก็ตามไม่มีผลทำให้เกรดที่ได้เปลี่ยนแปลง ถือว่าแบบสอบต่างฉบับที่นำมาใช้มีความเท่าเทียมกันไม่จำเป็นจะต้องทำการปรับเทียบคะแนน เพราะแบบสอบต่างฉบับกันนี้ไม่ได้ทำให้ผู้สอบเกิดความเสียเปรียบหรือเสียเปรียบสามารถยอมรับได้ในทางสถิติ การศึกษาครั้งนี้จึงใช้ดัชนีแคปปา ตั้งแต่ .81 ขึ้นไปเป็นเกณฑ์ในการตัดสินผลการตัดเกรดก่อนและหลังการปรับเทียบคะแนนว่ามีความสอดคล้องกัน ไม่จำเป็นที่จะต้องทำการปรับเทียบคะแนน แต่ถ้าดัชนีแคปปาต่ำกว่า .81 แสดงว่าผู้สอบเกิดการได้เปรียบเสียเปรียบกันในการทำข้อสอบต่างฉบับ จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนที่จะตัดเกรด การวิเคราะห์ความสอดคล้องของการตัดเกรด 3 ระดับจากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล จำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ได้ผลดังนี้

1. ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 เมื่อใช้กลุ่มตัวอย่างในการปรับเทียบคะแนนขนาด 100 คน 500 คน และ 700 คน ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่าง 500 คน และ 700 คน ในการปรับเทียบคะแนนทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด จำเป็นจะต้องทำการปรับเทียบคะแนน (ค่าแคปปาต่ำกว่า .81) แต่ถ้าใช้กลุ่มตัวอย่างในการปรับเทียบ 100 คน ไม่จำเป็นจะต้องทำการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-36

ตารางที่ 4-36 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อนปรับเทียบ	U	853	0	0	853
		S	0	69	1	70
		H	0	0	4	4
		รวม	853	69	5	927
		แคปปา = 0.99				
n = 500	ก่อนปรับเทียบ	U	869	0	0	869
		S	24	33	0	57
		H	0	1	0	1
		รวม	893	34	0	927
		แคปปา = .72				
n = 700	ก่อนปรับเทียบ	U	874	0	0	874
		S	28	25	0	53
		H	0	0	0	0
		รวม	902	25	0	927
		แคปปา = .63				

2. ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ผลการวิเคราะห์ พบว่า การใช้กลุ่มตัวอย่างขนาด 500 คน และ 700 คน ในการปรับเทียบคะแนนทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด จำเป็นจะต้องทำการปรับเทียบคะแนน (ค่าแคปปาต่ำกว่า .81) แต่ถ้าใช้กลุ่มตัวอย่างในการปรับเทียบ 100 คน ไม่จำเป็นจะต้องทำการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-37

ตารางที่ 4-37 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อนปรับเทียบ	U	784	0		784
		S	8	64		72
		H	0	0	1	1
		รวม	792	64	1	857
		แคปปา = .94				
n = 500	ก่อนปรับเทียบ	U	797	0		797
		S	25	35		60
		H				
		รวม	822	35		857
		แคปปา = .72				
n = 700	ก่อนปรับเทียบ	U	797	0	0	797
		S	27	33	0	60
		H	0	0	0	0
		รวม	824	33	1	857
		แคปปา = .69				

3. ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างขนาด 500 คน และ 700 คน ในการปรับเทียบคะแนนทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด จำเป็นจะต้องทำการปรับเทียบคะแนน (ค่าแคปปาต่ำกว่า .81) แต่ถ้าใช้กลุ่มตัวอย่างในการปรับเทียบ 100 คน ไม่จำเป็นจะต้องทำการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-38

ตารางที่ 4-38 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อนปรับเทียบ	U	829	0	0	829
		S	15	77	0	92
		H	0	1	5	6
		รวม	844	78	5	927
		แคปปา = .90				
n = 500	ก่อนปรับเทียบ	U	832	0	0	832
		S	48	42	0	90
		H	0	5	0	5
		รวม	880	47	0	927
		แคปปา = .60				
n = 700	ก่อนปรับเทียบ	U	832	0	0	832
		S	54	36	0	90
		H	0	5	0	5
		รวม	886	41	7	927
		แคปปา = .54				

4. ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างทั้ง 3 ขนาด ในการปรับเทียบคะแนนทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-39

ตารางที่ 4-39 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อนปรับเทียบ	U	744	0	0	744
		S	52	60	0	112
		H	0	0	1	1
		รวม	796	60	1	857
		แคปปา = .67				
n = 500	ก่อนปรับเทียบ	U	759	0	0	759
		S	71	26	0	97
		H	0	1	0	1
		รวม	830	27	0	857
		แคปปา = .39				
n = 700	ก่อนปรับเทียบ	U	759	0	0	759
		S	71	26	0	97
		H	0	1	0	1
		รวม	830	27	0	857
		แคปปา = .39				

สรุปความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนลจำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ส่วนใหญ่เกือบทุกเงื่อนไขของการปรับเทียบคะแนน เมื่อนำผลคะแนนก่อนและหลังการปรับเทียบคะแนนมาวิเคราะห์ความสอดคล้องของการตัดเกรด ผลการวิเคราะห์ พบว่า โดยภาพรวมการตัดเกรดไม่มีความสอดคล้องกัน โดยที่การวิเคราะห์กับกลุ่มตัวอย่างขนาด 500 คน และ 700 คน ในทุกเงื่อนไขจะทำให้ผู้สอบเกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด จึงจำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ยกเว้นกรณีตัวอย่างขนาด 100 คน หรือถ้ามีผู้สอบประมาณ 100 คน

ที่ไม่จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด มีเพียงเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ที่ถึงแม้จะมีผู้สอบประมาณ 100 คน ก็จะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด และเมื่อพิจารณาแต่ละเงื่อนไขของการปรับเทียบคะแนนขนาดของกลุ่มตัวอย่างยิ่งมากขึ้น จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากยิ่งขึ้น

นอกจากนี้พบว่า เมื่อตัดข้อสอบที่ไม่มีคุณภาพทิ้งของทั้งเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ค่าดัชนีแคปปาลดลง ไม่ว่าจะวิเคราะห์โดยใช้กลุ่มตัวอย่างตามที่กำหนดขนาดใดก็ตาม การตัดข้อสอบที่ไม่มีคุณภาพทิ้ง แสดงให้เห็นความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากยิ่งขึ้น ส่วนความไม่สอดคล้องของการตัดเกรดเมื่อใช้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และข้อสอบร่วมมีความยากอย่างสุ่มกับกลุ่มตัวอย่างตามที่กำหนด ไม่สามารถระบุได้ว่าวิธีใดให้ผลชัดเจนกว่ากัน ดังแสดงในตารางที่ 4-40

ตารางที่ 4-40 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขของการปรับเทียบคะแนน

เงื่อนไข	ค่าแคปปาตามเงื่อนไขการปรับเทียบ		
	100 คน	500 คน	700 คน
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	.99	.72	.63
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	.94	.72	.69
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.90	.60	.54
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.67	.39	.39

2. การเปรียบเทียบผลการตัดเกรด 8 ระดับ ระหว่างการใช้คะแนนก่อน และหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล

การตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ข้อสอบร่วมมีความยากอย่างสุ่ม ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และข้อสอบร่วม

มีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง กับกลุ่มตัวอย่างขนาด 100 คน 500 คน และ 700 คน ในการตัดเกรดจะกำหนดเกณฑ์คะแนนจุดตัด ดังนี้

2.1 เกรด A ระดับ 4.00 เมื่อได้คะแนน 76-100%

2.2 เกรด B^+ ระดับ 3.50 เมื่อได้คะแนน 70-75%

2.3 เกรด B ระดับ 3.00 เมื่อได้คะแนน 65-69%

2.4 เกรด C^+ ระดับ 2.50 เมื่อได้คะแนน 60-64%

2.5 เกรด C ระดับ 2.00 เมื่อได้คะแนน 55-59%

2.6 เกรด D^+ ระดับ 1.50 เมื่อได้คะแนน 50-54%

2.7 เกรด D ระดับ 1.00 เมื่อได้คะแนน 45-49%

2.8 เกรด F ระดับ 0.00 เมื่อได้คะแนนต่ำกว่า 45%

จากข้อกำหนดดังกล่าว นำมาสร้างเป็นเกณฑ์ของแต่ละเงื่อนไขของการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-41

ตารางที่ 4-41 คะแนนจุดตัดในการตัดเกรด 8 ระดับ จำแนกตามเงื่อนไขการปรับเทียบคะแนนด้วยวิธีเคอเนล

เงื่อนไข/	**เกณฑ์ที่ใช้ในการตัดเกรด (คะแนน)**							
ขนาดตัวอย่าง	**F**	**D**	**D+**	**C**	**C+**	**B**	**B+**	**A**
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6								
1.1 100 คน	< 51	51-56	57-62	63-67	68-73	74-79	80-86	> 86
1.2 500 คน	< 54	54-58	59-64	65-70	71-76	77-82	83-89	> 89
1.3 700 คน	< 52	52-57	58-63	64-69	70-74	75-80	81-87	> 87
2. ข้อสอบร่วมมีความยากอย่างสุ่ม								
2.1 100 คน	< 51	51-56	57-62	63-67	68-73	74-79	80-86	> 86
2.2 500 คน	< 54	54-58	59-64	65-70	71-76	77-82	83-89	> 89
2.3 700 คน	< 54	54-58	59-64	65-70	71-76	77-82	83-89	> 89
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง								
3.1 100 คน	< 46	46-50	51-55	56-60	61-65	66-70	71-77	> 77
3.2 500 คน	< 47	47-51	52-57	58-62	63-67	68-72	73-79	> 79
3.3 700 คน	< 47	47-51	52-57	58-62	63-67	68-72	73-79	> 79

ตารางที่ 4-41 (ต่อ)

เงื่อนไข/ ขนาดตัวอย่าง	เกณฑ์ที่ใช้ในการตัดเกรด (คะแนน)							
	F	**D**	**D+**	**C**	**C+**	**B**	**B+**	**A**
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง								
4.1 100 คน	< 46	46-50	51-56	57-61	62-66	67-71	72-77	> 77
4.2 500 คน	< 49	49-53	54-58	59-64	65-69	70-75	76-81	> 81
4.3 700 คน	< 49	49-53	54-58	59-64	65-69	70-75	76-81	> 81

การวิเคราะห์ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล จำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ได้ผล ดังนี้

1. ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ผลการวิเคราะห์พบว่า เมื่อใช้กลุ่มตัวอย่างในการปรับเทียบคะแนนขนาด 100 คน การใช้แบบสอบต่างฉบับไม่ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด ไม่จำเป็นจะต้องทำการปรับเทียบคะแนน (ค่าแคปตั้งแต่ .81 ขึ้นไป) แต่ถ้าใช้กลุ่มตัวอย่างในการปรับเทียบ 500 คน และ 700 คน จำเป็นจะต้องทำการปรับเทียบคะแนน (ค่าแคปปาต่ำกว่า .81) ดังแสดงในตารางที่ 4-42

ตารางที่ 4-42 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	F	532	0	0	0	0	0	0		532
	ปรับเทียบ	D	16	138	0	0	0	0	0		154
	คะแนน	D^+	0	0	111	0	0	0	0		111
		C	0	0	0	56	0	0	0		56
		C^+	0	0	0	0	36	3	0		39
		B	0	0	0	0	0	22	1		23
		B^+	0	0	0	0	0	0	8	1	9
		A	0	0	0	0	0	0	0	3	3
แคปปา = .96		รวม	548	138	111	56	36	25	9	4	927
n = 500	ก่อนการ	F	589	0	0	0	0	0	0	0	589
	ปรับเทียบ	D	78	47	0	0	0	0	0	0	125
	คะแนน	D^+	0	62	39	0	0	0	0	0	101
		C	0	0	37	17	0	0	0	0	54
		C^+	0	0	0	24	9	0	0	0	33
		B	0	0	0	0	11	4	0	0	15
		B^+	0	0	0	0	0	6	3	0	9
		A	0	0	0	0	0	0	1	0	1
แคปปา = .54		รวม	667	109	76	41	20	10	4	0	927
n = 700	ก่อนการ	F	582	0	0	0	0	*0*	0	0	582
	ปรับเทียบ	D	79	74	0	0	0	0	0	0	153
	คะแนน	D^+	0	53	37	0	0	0	0	0	90
		C	0	0	37	12	0	0	0	0	49
		C^+	0	0	0	28	0	0	0	0	28
		B	0	0	0	0	14	4	0	0	18
		B^+	0	0	0	0	0	7	0	0	7
		A									
แคปปา = .55		รวม	661	127	74	40	14	11	4	0	927

2. ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มในการปรับเทียบคะแนน ผลการวิเคราะห์พบว่าการใช้การใช้กลุ่มตัวอย่างทั้ง 3 ขนาด ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-43

ตารางที่ 4-43 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

ขนาดตัวอย่าง	**เงื่อนไข**	**หลังการปรับเทียบคะแนน**								**รวม**
		เกรด	**F**	**D**	**D$^+$**	**C**	**C$^+$**	**B**	**B$^+$**	**A**	
n = 100	ก่อนการ	F	442	0	0	0	0	0	0	0	442
	ปรับเทียบ	D	119	45	0	0	0	0	0	0	164
	คะแนน	D$^+$	0	62	55	0	0	0	0	0	117
		C	0	0	23	38	0	0	0	0	61
		C$^+$	0	0	0	8	31	1	0	0	40
		B	0	0	0	0	0	21	6	0	27
		B$^+$	0	0	0	0	0	0	5	0	5
		A	0	0	0	0	0	0	0	1	1
แคปปา = .59		รวม	561	107	78	46	31	22	11	1	857
n = 500	ก่อนการ	F	478	0	0	0	0	0	0		478
	ปรับเทียบ	D	133	0	0	0	0	0	0		133
	คะแนน	D$^+$	0	96	16	0	0	0	0		112
		C	0	0	50	24	9	0	0		74
		C$^+$	0	0	0	25	8	0	0		33
		B	0	0	0	0	20	1	1		21
		B$^+$	0	0	0	0	0	5	1		6
		A									
แคปปา = .33		รวม	511	96	66	49	28	6	6		857

ตารางที่ 4-43 (ต่อ)

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน								รวม
		เกรด	**F**	**D**	**D$^+$**	**C**	**C$^+$**	**B**	**B$^+$**	**A**
n = 700	ก่อนการ	F	478	0	0	0	0	0	0	478
	ปรับเทียบ	D	133	0	0	0	0	0	0	133
	คะแนน	D$^+$	0	97	15	0	0	0	0	112
		C	0	0	65	9	0	0	0	74
		C$^+$	0	0	0	27	6	0	0	33
		B	0	0	0	0	20	1	0	21
		B$^+$	0	0	0	0	0	5	1	6
		A								
แคปปา = .29		รวม	611	97	80	36	26	6	1	857

3. ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้งในการปรับเทียบคะแนน ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างทั้ง 3 ขนาด ทำให้ผู้สอบเกิดการได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-44

ตารางที่ 4-44 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน								รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	F	496	0	0	0	0	0	0	0	496
	ปรับเทียบ	D	73	66	0	0	0	0	0	0	139
	คะแนน	D^+	0	63	48	0	0	0	0	0	111
		C	0	0	43	40	0	0	0	0	83
		C^+	0	0	0	15	28	0	0	0	43
		B	0	0	0	0	10	13	0	0	23
		B^+	0	0	0	0	0	12	15	0	27
		A	0	0	0	0	0	0	2	3	5
แคปปา = .63		รวม	569	129	91	55	38	25	17	3	927
n = 500	ก่อนการ	F	487	0	0	0	0	0	0	0	487
	ปรับเทียบ	D	141	0	0	0	0	0	0	0	141
	คะแนน	D^+	0	114	19	0	0	0	0	0	133
		C	0	0	71	0	8	0	0	0	71
		C^+	0	0	6	38	0	5	0	0	44
		B	0	0	0	4	22	0	0	0	26
		B^+	0	0	0	0	2	17	1	0	20
		A	0	0	0	0	0	0	5	0	5
แคปปา = .25		รวม	628	114	96	42	24	17	6	0	927
n = 700	ก่อนการ	F	487	0	0	0	0	0	0	0	487
	ปรับเทียบ	D	141	0	0	0	0	0	0	0	141
		D^+	7	126	0	0	0	0	0	0	133
		C	0	0	71	0	0	0	0	0	71
		C^+	0	0	12	32	0	0	0	0	44
		B	0	0	0	10	16	0	0	0	26

ตารางที่ 4-44 (ต่อ)

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 700	ก่อนการ	B^+	0	0	0	0	11	9	0	0	20
	ปรับเทียบ	A	0	0	0	0	0	1	4	0	5
แคปปา = .21		รวม	635	126	83	42	27	10	4	0	927

4. ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างทั้ง 3 ขนาด ทำให้ผู้สอบเกิดการได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-45

ตารางที่ 4-45 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	F	359	0	0	0	0	0	0	0	359
	ปรับเทียบ	D	149	0	0	0	0	0	0	0	149
	คะแนน	D^+	58	88	0	0	0	0	0	0	146
		C	0	8	82	0	0	0	0	0	90
		C^+	0	0	5	47	8	0	0	0	60
		B	0	0	0	0	20	11	0	0	31
		B^+	0	0	0	0	0	11	10	0	21
		A	0	0	0	0	0	0	0	1	1
แคปปา = .19		รวม	566	96	87	47	28	22	10	1	857

ตารางที่ 4-45 (ต่อ)

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 500	ก่อนการ	F	393	0	0	0	0	0	0	0	393
	ปรับเทียบ	D	146	0	0	0	0	0	0	0	146
	คะแนน	D^+	79	38	0	0	0	0	0	0	117
		C	0	60	43	0	0	0	0	0	103
		C^+	0	0	24	29	0	0	0	0	53
		B	0	0	0	18	13	0	10	0	31
		B^+	0	0	0	0	9	4	0	0	13
		A	0	0	0	0	0	0	1	0	1
แคปปา = .14		รวม	618	98	67	47	22	4	1	0	857
n = 700	ก่อนการ	F	393	0	0	0	0	0	0	0	3693
	ปรับเทียบ	D	146	0	0	0	0	0	0	0	146
	คะแนน	D^+	90	27	0	0	0	0	0	0	117
		C	0	61	42	0	0	0	0	0	103
		C^+	0	0	32	21	0	0	0	0	53
		B	0	0	0	18	13	0	0	0	31
		B^+	0	0	0	0	9	4	0	0	13
		A	0	0	0	0	0	0	1	0	1
แคปปา = .14		รวม	629	88	74	39	22	4	1	0	857

สรุปความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล จำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน โดยภาพรวม พบว่าไม่มีความสอดคล้องกัน การใช้แบบสอบต่างฉบับทำให้ผู้สอบเกิดการได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ยกเว้นการใช้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 กับตัวอย่างขนาด 100 คน หรือผู้สอบประมาณ 100 คน เพียงเงื่อนไขเดียวที่ไม่จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ส่วนการใช้เงื่อนไขอื่น ๆ จะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด และเมื่อพิจารณา

แต่ละเงื่อนไขของการปรับเทียบคะแนนเมื่อขนาดกลุ่มตัวอย่างในการวิเคราะห์เพิ่มขึ้น จะเห็นความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากยิ่งขึ้น

นอกจากนี้ พบว่า การวิเคราะห์ความสอดคล้องของการตัดเกรดภายใต้เงื่อนไขการตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนที่จะนำไปปรับเทียบคะแนนของทั้งเงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม พบว่า เงื่อนไขข้อสอบร่วมที่มีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ค่าแคปปาจะต่ำกว่าการวิเคราะห์ด้วยเงื่อนไขที่ไม่มีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง แสดงให้เห็นถึงความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากกว่า และเมื่อพิจารณาความสอดคล้องของการตัดเกรดของเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม พบว่า การใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม ค่าแคปปาต่ำกว่าการใช้ข้อสอบร่วมที่มีความยากอยู่ในช่วง .4-.6 ไม่ว่าจะใช้กลุ่มตัวอย่างตามที่กำหนดขนาดเท่าใดก็ตาม แสดงให้เห็นถึงความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากกว่าเช่นเดียวกัน ดังแสดงในตารางที่ 4-46

ตารางที่ 4-46 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขของการปรับเทียบคะแนน

เงื่อนไข	ค่าแคปปาตามเงื่อนไขการปรับเทียบ		
	100 คน	500 คน	700 คน
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	.96	.54	.55
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	.59	.33	.29
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.63	.25	.21
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.19	.14	.14

ตอนที่ 5.2 การเปรียบเทียบผลการตัดเกรด ก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์

การนำเสนอผลการตัดเกรดก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ในที่นี้จะแยกการนำเสนอออกเป็น 2 ข้อย่อย ตามระดับของการตัดเกรด คือ 3 ระดับ และ 8 ระดับ ดังนี้

1. การเปรียบเทียบผลการตัดเกรด 3 ระดับ ระหว่างการใช้คะแนนก่อน และหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์

การตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT ภายใต้เงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ข้อสอบร่วมมีความยากอย่างสุ่ม ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อใช้กลุ่มตัวอย่างขนาด 100 คน 500 คน และ 700 คน การตัดเกรดจะกำหนดเกณฑ์คะแนนจุดตัด ดังนี้

1.1 เกรด H ระดับ 4.00 เมื่อได้คะแนนมากกว่า 75%

1.2 เกรด S ระดับ 2.33 เมื่อได้คะแนน 60-75%

1.3 เกรด U ระดับ 0 เมื่อได้คะแนนต่ำกว่า 60%

จากข้อกำหนดดังกล่าว นำมาสร้างเป็นเกณฑ์ของแต่ละเงื่อนไขของการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-47

ตารางที่ 4-47 คะแนนจุดตัดในการตัดเกรด 3 ระดับ จำแนกตามเงื่อนไขการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์

เงื่อนไข/ ขนาดตัวอย่าง	**เกณฑ์ที่ใช้ในการตัดเกรด**		
	< ร้อยละ 60	**ร้อยละ 60-75**	**> ร้อยละ 75**
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6			
1.1 100 คน	< 68	68-86	> 86
1.2 500 คน	< 71	71-89	> 89
1.3 700 คน	< 70	70-87	> 87
2. ข้อสอบร่วมมีความยากอย่างสุ่ม			
2.1 100 คน	< 68	68-86	> 86
2.2 500 คน	< 71	71-89	> 89
2.3 700 คน	< 71	71-89	> 89
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
3.1 100 คน	< 61	61-76	> 77
3.2 500 คน	< 63	63-79	> 79
3.3 700 คน	< 63	63-79	> 79

ตารางที่ 4-47 (ต่อ)

เงื่อนไข/ ขนาดตัวอย่าง	เกณฑ์ที่ใช้ในการตัดเกรด		
	< ร้อยละ 60	**ร้อยละ 60-75**	**> ร้อยละ 75**
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง			
4.1 100 คน	< 62	62-77	> 77
4.2 500 คน	< 65	65-81	> 81
4.3 700 คน	< 65	65-81	> 81

สำหรับเกณฑ์การพิจารณาความสอดคล้องของการตัดเกรดใช้ดัชนีแคปปา เป็นเกณฑ์เดียวกันกับที่ใช้ตามวิธีเคอเนลคือแคปปาตั้งแต่ .81-1.00 ถือว่ามีความสอดคล้องในระดับมาก การใช้แบบสอบต่างฉบับไม่ได้ทำให้เกิดความเสียเปรียบกันในการตัดเกรด เป็นค่าที่ยอมรับได้ในทางสถิติ แต่ถ้าค่าดัชนีแคปปาต่ำกว่า .81 แสดงว่าผู้สอบเกิดการได้เปรียบเสียเปรียบกัน ในการทำแบบสอบต่างฉบับจำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด การวิเคราะห์ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ จำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ได้ผล ดังนี้

1.1 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ผลการวิเคราะห์ พบว่า การใช้กลุ่มตัวอย่างทั้ง 3 ขนาด ทำให้เกิดความได้เปรียบเสียเปรียบกัน ในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-48

ตารางที่ 4-48 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อนปรับเทียบ	U	799	49	5	853
		S	37	33	0	70
		H	3	0	1	4
		รวม	839	82	6	927
		แคปปา = .37				
n = 500	ก่อนปรับเทียบ	U	848	21	0	869
		S	41	16	0	57
		H	1	0	0	1
		รวม	890	37	0	927
		แคปปา = 0.30				
n = 700	ก่อนปรับเทียบ	U	853	21	0	874
		S	37	16	0	53
		H	0	0	0	0
		รวม	890	37	0	927
		แคปปา = .32				

1.2 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มในการปรับเทียบคะแนน ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างขนาด 100 คน ไม่ได้ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาตั้งแต่ .81 ขึ้นไป) จึงไม่จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด แต่ถ้าใช้กลุ่มตัวอย่างขนาด 500 คน และ 700 คน ในการวิเคราะห์จะทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-49

ตารางที่ 4-49 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

ก่อนการปรับเทียบคะแนน/ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อน	U	763	21	0	784
	ปรับเทียบ	S	4	68	0	72
		H	0	0	1	1
		รวม	767	89	1	857
		แคปปา = .83				
n = 500	ก่อน	U	796	1		797
	ปรับเทียบ	S	21	39		60
		H				
		รวม	817	40		857
		แคปปา = .77				
n = 700	ก่อน	U	787	1		788
	ปรับเทียบ	S	33	35		68
		H	0	1		1
		รวม	820	37		857
		แคปปา = .65				

1.3 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ผลการวิเคราะห์พบว่า การใช้กลุ่มตัวอย่างในการปรับเทียบคะแนนทั้ง 3 ขนาด ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-50

ตารางที่ 4-50 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ก่อนการปรับเทียบคะแนน/ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อน	U	764	59	6	829
	ปรับเทียบ	S	49	43	0	92
		H	4	1	1	6
		รวม	817	103	7	927
			แคปปา = .36			
n = 500	ก่อน	U	804	25	3	832
	ปรับเทียบ	S	85	25	0	90
		H	4	0	1	5
		รวม	873	50	4	927
			แคปปา = .30			
n = 700	ก่อน	U	806	25	1	832
	ปรับเทียบ	S	68	22	0	90
		H	4	1	0	5
		รวม	878	48	1	927
			แคปปา = .26			

1.4 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้งในการปรับเทียบคะแนน ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างขนาด 100 คน ไม่ได้ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาตั้งแต่ .81 ขึ้นไป) จึงไม่จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด แต่ถ้าใช้กลุ่มตัวอย่างขนาด 500 คน และ 700 คน ในการวิเคราะห์จะทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-51

ตารางที่ 4-51 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ก่อนการปรับเทียบคะแนน/ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน				รวม
		เกรด	U	S	H	
n = 100	ก่อน	U	730	14		744
	ปรับเทียบ	S	14	95	3	112
		H			1	1
		รวม	744	109	4	857
			แคปปา = .84			
n = 500	ก่อน	U	759			759
	ปรับเทียบ	S	42	55		97
		H			1	1
		รวม	801	55	1	857
			แคปปา = .70			
n = 700	ก่อน	U	759			759
	ปรับเทียบ	S	45	52		97
		H			1	1
		รวม	804	52	1	857
			แคปปา = .68			

สรุปความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ จำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ส่วนใหญ่เกือบทุกเงื่อนไขของการปรับเทียบคะแนน เมื่อนำผลคะแนนก่อนและหลังการปรับเทียบคะแนนมาวิเคราะห์ความสอดคล้องของการตัดเกรด ผลการวิเคราะห์พบว่า โดยภาพรวมการตัดเกรดไม่มีความสอดคล้องกัน โดยที่เมื่อใช้กลุ่มตัวอย่างขนาด 500 คน และ 700 คน ในการวิเคราะห์ทุกเงื่อนไขจะทำให้ผู้สอบเกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ขณะที่เมื่อวิเคราะห์กับกลุ่มตัวอย่างขนาด 100 คน หรือ

ถ้าผู้เข้าสอบประมาณ 100 คน ไม่จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ยกเว้นถ้าวิเคราะห์ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มทั้งที่มีการตัดและไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ที่จะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด และเมื่อพิจารณาแต่ละเงื่อนไขขนาดของกลุ่มตัวอย่างยิ่งมากขึ้น จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากยิ่งขึ้น

นอกจากนี้พบว่า การวิเคราะห์ความสอดคล้องภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ทั้งที่มีการตัดและไม่มีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ค่าแคปปาจะต่ำกว่าการวิเคราะห์ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ทั้งที่มีการตัดและไม่มีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง การตัดข้อสอบที่ไม่มีคุณภาพทิ้งจะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากกว่าการไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และการวิเคราะห์ความสอดคล้องภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ไม่ว่าจะใช้กลุ่มตัวอย่างขนาดเท่าใดก็ตามจะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนกว่าการวิเคราะห์ด้วยเงื่อนไข ข้อสอบร่วมมีความยากอย่างสุ่ม ดังแสดงในตารางที่ 4-52

ตารางที่ 4-52 ความสอดคล้องของการตัดเกรด 3 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขของการปรับเทียบคะแนน

เงื่อนไข	ค่าแคปปาตามเงื่อนไขการปรับเทียบ		
	100 คน	500 คน	700 คน
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	.37	.30	.32
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	.83	.77	.65
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.36	.30	.26
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.84	.70	.68

2. การเปรียบเทียบผลการตัดเกรด 8 ระดับ ระหว่างการใช้คะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์

การตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ข้อสอบร่วมมีความยากอย่างสุ่ม ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง และข้อสอบร่วม

มีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อใช้กลุ่มตัวอย่างขนาด 100 คน 500 คน และ 700 คน ในการตัดเกรดจะกำหนดเกณฑ์คะแนนจุดตัด เช่นเดียวกันกับวิธีเคอเนล โดยมีเกณฑ์การตัดเกรดของแต่ละเงื่อนไขของการปรับเทียบคะแนน ดังแสดงในตารางที่ 4-53

ตารางที่ 4-53 คะแนนจุดตัดในการตัดเกรด 8 ระดับ จำแนกตามเงื่อนไขการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์

เงื่อนไข/ขนาดตัวอย่าง	เกณฑ์ที่ใช้ในการตัดเกรด (คะแนน)							
	F	**D**	**D+**	**C**	**C+**	**B**	**B+**	**A**
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6								
1.1 100 คน	< 51	51-56	57-62	63-67	68-73	74-79	80-86	> 86
1.2 500 คน	< 54	54-58	59-64	65-70	71-76	77-82	83-89	> 89
1.3 700 คน	< 52	52-57	58-63	64-69	70-74	75-80	81-87	> 87
2. ข้อสอบร่วมมีความยากอย่างสุ่ม								
2.1 100 คน	< 51	51-56	57-62	63-67	68-73	74-79	80-86	> 86
2.2 500 คน	< 54	54-58	59-64	65-70	71-76	77-82	83-89	> 89
2.3 700 คน	< 54	54-58	59-64	65-70	71-76	77-82	83-89	> 89
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง								
3.1 100 คน	< 46	46-50	51-55	56-60	61-65	66-70	71-77	> 77
3.2 500 คน	< 47	47-51	52-57	58-62	63-67	68-72	73-79	> 79
3.3 700 คน	< 47	47-51	52-57	58-62	63-67	68-72	73-79	> 79
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง								
4.1 100 คน	< 46	46-50	51-56	57-61	62-66	67-71	72-77	> 77
4.2 500 คน	< 49	49-53	54-58	59-64	65-69	70-75	76-81	> 81
4.3 700 คน	< 49	49-53	54-58	59-64	65-69	70-75	76-81	> 81

การวิเคราะห์ความสอดคล้องของการตัดเกรด 8 ระดับจากคะแนนก่อน และหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ จำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนนได้ผล ดังนี้

2.1 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ผลการวิเคราะห์ พบว่า การใช้กลุ่มตัวอย่างในการปรับเทียบคะแนนทั้ง 3 ขนาด ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ดังแสดงในตารางที่ 4-54

ตารางที่ 4-54 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	F	377	79	25	21	16	5	6	3	532
	ปรับเทียบ	D	60	54	22	11	5	1	1	0	154
	คะแนน	D^+	32	18	41	12	4	1	1	2	111
		C	15	13	8	11	8	0	1		56
		C^+	13	4	1	0	17	3	1		39
		B	7	4	0	3	0	6	3		23
		B^+	5		1		1	1	1		9
		A	1		1					1	3
แคปปา = .29		รวม	510	172	99	58	51	17	14	6	927
n = 500	ก่อนการ	F	458	68	30	18	7	5	3		589
	ปรับเทียบ	D	59	49	6	8	1		2		125
	คะแนน	D^+	41	36	19	3	2				101
		C	21	8	7	17	1				54
		C^+	14	2	0	10	7				33
		B	5	1	2	0	4	3			15
		B^+	5	1	0	1	1		1		9
		A			1						1
แคปปา = .27		รวม	603	165	65	57	23	8	6		927

ตารางที่ 4-54 (ต่อ)

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 700	ก่อนการ	F	436	86	29	17	3	8	3		582
	ปรับเทียบ	D	60	72	8	9	1	1	2		153
	คะแนน	D^+	35	29	20	4	1	1			90
		C	15	11	5	17	1	1			49
		C^+	12	2	0	7	6	1			28
		B	5	1	2	2	3	5			18
		B^+	4	1	1				1		7
		A									
แคปปา = .30		รวม	567	202	65	56	14	17	6		927

2.2 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างในการปรับเทียบคะแนนทั้ง 3 ขนาด ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จึงจำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด เช่นเดียวกันกับการตัดเกรดโดยใช้เงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ดังแสดงในตารางที่ 4-55

ตารางที่ 4-55 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	F	419	23							442
	ปรับเทียบ	D	25	122	17						164
	คะแนน	D^+		28	75	13	1				117
		C			12	29	20				61
		C^+				4	32	4			40
		B					3	15	9		27
		B^+						1	4		5
		A								1	1
แคปปา = .72		รวม	444	173	104	46	56	20	13	1	857
n = 500	ก่อนการ	F	461	17							478
	ปรับเทียบ	D	47	86							133
	คะแนน	D^+	4	74	28	6					112
		C		7	23	43	1				74
		C^+				20	12	1			33
		B				1	8	12			21
		B^+						5	1		6
		A									
แคปปา = .60		รวม	512	184	51	70	21	18	1		857
n = 700	ก่อนการ	F	463	15							478
	ปรับเทียบ	D	50	83							133
	คะแนน	D^+	4	74	30						112
		C		7	24	42	1				74
		C^+				23	9	1			33
		B				1	10	10			21

ตารางที่ 4-55 (ต่อ)

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
		B^+						5	1		6
		A									
แคปปา = .58		รวม	517	179	54	70	20	16	1		857

2.3 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อไม่มีคุณภาพทิ้ง ผลการวิเคราะห์ พบว่า การใช้กลุ่มตัวอย่างในการปรับเทียบคะแนนทั้ง 3 ขนาด ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จึงจำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด เช่นเดียวกันกับการตัดเกรดโดยใช้เงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม ดังแสดงในตารางที่ 4-56

ตารางที่ 4-56 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	F	350	48	49	13	21	3	9	3	496
	ปรับเทียบ	D	59	33	28	10	5	1	3	0	139
	คะแนน	D^+	38	10	46	8	5	2	0	2	111
		C	24	9	26	13	9	0	2		83
		C^+	8	4	7	3	18	2	1		43
		B	9	3	1	0	2	6	2		23
		B^+	8	2	3	1	2	1	10		27
		A	3			1	0			1	5
แคปปา = .27		รวม	499	109	160	49	62	15	27	6	927

ตารางที่ 4-56 (ต่อ)

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 500	ก่อนการ	F	357	64	22	26	6	6	5	1	487
	ปรับเทียบ	D	89	47	12	10	0	2	1	0	141
	คะแนน	D^+	45	49	24	9	3	1		2	133
		C	23	14	18	15	0	1			71
		C^+	11	8	3	17	4	1			44
		B	11	2	0	1	5	7			26
		B^+	6	3	1	2	1	5	2	1	20
		A	3		1					4	5
แคปปา = .22		รวม	525	187	81	80	19	23	8	6	927
n = 700	ก่อนการ	F	359	68	20	23	5	8	4		487
	ปรับเทียบ	D	72	46	10	10	0	3	0		141
	คะแนน	D^+	45	55	20	8	2	1	1	1	133
		C	24	14	19	13	0	1			71
		C^+	12	7	3	19	2	1			44
		B	11	2	0	2	6	5			26
		B^+	6	3	1	2	1	7			20
		A	3		1				1		5
แคปปา = .20		รวม	532	195	74	77	16	26	6	1	927

2.4 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้งในการปรับเทียบคะแนน ผลการวิเคราะห์พบว่าการใช้กลุ่มตัวอย่างในการปรับเทียบคะแนนทั้ง 3 ขนาด ทำให้เกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จึงจำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด และเมื่อใช้ขนาดตัวอย่างเพิ่มขึ้นจะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากยิ่งขึ้น ดังแสดงในตารางที่ 4-57

ตารางที่ 4-57 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	F	354	5							359
	ปรับเทียบ	D	65	56	28						149
	คะแนน	D^+	7	25	109	5					146
		C		1	51	24	14				90
		C^+			3	11	39	7			60
		B					15	10	6		31
		B^+						1	17	3	21
		A								1	1
แคปปา = .60		รวม	426	87	191	40	68	18	23	4	857
n = 500	ก่อนการ	F	386	7							393
	ปรับเทียบ	D	66	78	2						146
	คะแนน	D^+	12	80	23	2					117
		C		16	52	35					103
		C^+			1	39	13				53
		B				2	13	16			31
		B^+						10	3		13
		A								1	1
แคปปา = .49		รวม	464	181	78	78	26	26	3	1	857
n = 700	ก่อนการ	F	386	7							393
	ปรับเทียบ	D	67	78	1						146
	คะแนน	D^+	12	83	20	1					116
		C		18	52	33					103
		C^+			1	41	11				53

ตารางที่ 4-57 (ต่อ)

ขนาดตัวอย่าง	เงื่อนไข	หลังการปรับเทียบคะแนน									รวม
		เกรด	F	D	D^+	C	C^+	B	B^+	A	
n = 100	ก่อนการ	B				3	14	14			31
	ปรับเทียบ	B^+						10	3		13
	คะแนน	A								1	1
แคปปา = .47		รวม	465	186	74	78	25	24	3	1	857

สรุปภาพรวมของความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ จำแนกตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน พบว่า ทุกเงื่อนไขของการปรับเทียบคะแนนทำให้ผู้สอบเกิดความได้เปรียบเสียเปรียบกันในการตัดเกรด (ค่าแคปปาต่ำกว่า .81) จำเป็นจะต้องทำการปรับเทียบคะแนนเพื่อให้ความยุติธรรมกับผู้สอบทุกคน และเมื่อขนาดกลุ่มตัวอย่างในการวิเคราะห์เพิ่มขึ้น จะเห็นความไม่สอดคล้องของการตัดเกรดชัดเจนมากยิ่งขึ้นทุกเงื่อนไข ยกเว้นเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 เมื่อวิเคราะห์กับตัวอย่างขนาด 700 คน นอกจากนี้พบว่า การใช้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ทั้งที่มีการตัดและไม่มีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ค่าแคปปาจะต่ำกว่าการวิเคราะห์ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มทั้งที่มีและไม่มีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง แสดงให้เห็นถึงความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากกว่า และเมื่อพิจารณาความสอดคล้องของการตัดเกรดของเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม พบว่าการใช้ข้อสอบร่วมที่มีความยากอยู่ในช่วง .4-.6 ค่าแคปปาต่ำกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม ไม่ว่าจะใช้กลุ่มตัวอย่างตามที่กำหนดขนาดเท่าใดก็ตาม แสดงให้เห็นถึงความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากกว่า ดังแสดงในตารางที่ 4-58

ตารางที่ 4-58 ความสอดคล้องของการตัดเกรด 8 ระดับ จากคะแนนก่อนและหลังการปรับเทียบคะแนนด้วยวิธี IRT 2 พารามิเตอร์ ภายใต้เงื่อนไขของการปรับเทียบคะแนน

เงื่อนไข	ค่าแคปปาตามเงื่อนไขการปรับเทียบ		
	100 คน	500 คน	700 คน
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	.29	.27	.30
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	.72	.60	.58
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.27	.22	.20
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.60	.49	.47

ตอนที่ 5.3 การเปรียบเทียบความสัมพันธ์ของการตัดเกรดหลังการปรับเทียบคะแนนระหว่างวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์

การเปรียบเทียบความสัมพันธ์ของการตัดเกรดระหว่างการใช้คะแนนหลังการปรับเทียบคะแนนวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ โดยใช้การทดสอบความเป็นอิสระด้วยไคสแควร์ เพื่อตรวจสอบว่าการตัดเกรดจากคะแนนหลังการปรับเทียบระหว่างการใช้วิธีเคอเนลกับวิธี IRT 2 พารามิเตอร์มีความสัมพันธ์กันหรือไม่ ผลการทดสอบ พบว่า การตัดเกรดจากทุกเงื่อนไขของการปรับเทียบคะแนนจากทั้งสองวิธีมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ .01 ไม่ว่าจะตัดเกรด 3 ระดับ หรือ 8 ระดับ ดังแสดงในตารางที่ 4-59 และตารางที่ 4-60

ตารางที่ 4-59 การทดสอบความเป็นอิสระของการตัดเกรด 3 ระดับหลังการปรับเทียบคะแนนระหว่างวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ด้วยไคสแควร์ ภายใต้เงื่อนไขต่าง ๆ

เงื่อนไข	n100			n500			n700		
	χ^2	*df*	*Sig*	χ^2	*df*	Sig	χ^2	*df*	*Sig*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	169.9	4	.00	74.1	4	.00	68.7	1	.00
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	1393.0	4	.00	617.3	1	.00	667.3	1	.00

ตารางที่ 4-59 (ต่อ)

เงื่อนไข	n100			n500			n700		
	χ^2	*df*	*Sig*	χ^2	*df*	*Sig*	χ^2	*df*	*Sig*
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	164.1	4	.00	123.0	2	.00	115.1	2	.00
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	642.9	4	.00	407.7	2	.00	430.9	2	.00

ตารางที่ 4-60 การทดสอบความเป็นอิสระของการตัดเกรด 8 ระดับหลังการปรับเทียบคะแนนระหว่างวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ด้วยไคสแควร์ ภายใต้เงื่อนไขต่าง ๆ

เงื่อนไข	n100			n500			n700		
	χ^2	*df*	*Sig*	χ^2	*df*	*Sig*	χ^2	*df*	*Sig*
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	522.5	49	.00	378.6	36	.00	353.1	30	.00
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	3074.9	49	.00	2676.9	36	.00	2724.8	36	.00
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	422.3	49	.00	416.1	42	.00	408.0	42	.00
4. ข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	2044.7	49	.00	2870.7	42	.00	2896.6	42	.00

สำหรับระดับของความสัมพันธ์ของการตัดเกรด 3 ระดับ หลังการปรับเทียบคะแนนด้วยวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ ที่ได้จากดัชนีแคปปาพบว่าหลังการปรับเทียบคะแนนด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม กับทุกขนาดตัวอย่างทั้งสองวิธีมีความสัมพันธ์กันในระดับดีมาก (. 75 ขึ้นไป ตามเกณฑ์ของ Fleiss Levin & Paik, 2003 อ้างถึงใน ประสพชัย พสุนันท์, หน้า 8) ส่วนหลังการปรับเทียบคะแนนด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ความสัมพันธ์ของการตัดเกรดทั้งสองวิธี อยู่ในระดับดี (.40-.74) ที่เหลือสัมพันธ์กันในระดับต่ำ

ส่วนความสัมพันธ์ของการตัดเกรด 8 ระดับ หลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ พบว่า หลังการปรับเทียบคะแนนมีเพียงเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มเพียงเงื่อนไขเดียว ที่การตัดเกรดหลังการปรับเทียบคะแนนของทั้งสองวิธีมีความสัมพันธ์กันอยู่ในระดับดี ส่วน เงื่อนไขที่เหลือสัมพันธ์กันในระดับต่ำ ดังแสดงในตารางที่ 4-61

ตารางที่ 4-61 ระดับความสัมพันธ์ของการตัดเกรดหลังการปรับเทียบคะแนน ด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์จากค่าดัชนีแคปปา

เงื่อนไข	ตัดเกรด 3 ระดับ			ตัดเกรด 8 ระดับ		
	n100	n500	n700	n100	n500	n700
1. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6	.37	.28	.27	.29	.28	.26
2. ข้อสอบร่วมมีความยากอย่างสุ่ม	.78	.85	.88	.55	.58	.58
3. ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.36	.35	.35	.24	.24	.24
4. ข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง	.65	.62	.65	.39	.35	.33

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาคุณภาพของวิธีการปรับเทียบคะแนนภายใต้เงื่อนไขรูปแบบข้อสอบร่วม ขนาดกลุ่มตัวอย่างและรูปแบบของข้อมูลที่จะนำมาวิเคราะห์ที่แตกต่างกัน และ 2) เพื่อเปรียบเทียบความสอดคล้องของการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนการปรับเทียบคะแนนตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ภายใต้การตัดเกรด 3 ระดับ และ 8 ระดับ ตามลำดับ

ข้อมูลที่ใช้ในการวิจัย การวิจัยครั้งนี้เป็นการศึกษาวิธีการปรับเทียบคะแนนโดยใช้ข้อมูลจากสถานการณ์การสอบจริง จากแบบสอบต่างฉบับที่ผู้วิจัยดำเนินการวางแผนจัดการแบบสอบตามเงื่อนไขของการปรับเทียบคะแนน ของชุดวิชาชุดหนึ่งที่มีนักศึกษาสอบเป็นจำนวนมากเพื่อให้มีจำนวนนักศึกษาเพียงพอในการสุ่มตัวอย่างตามเงื่อนไขของการปรับเทียบคะแนน จำนวน 3 ภาคการศึกษา โดยแบบสอบทั้ง 3 ฉบับ เป็นแบบสอบแบบชนิดเลือกตอบ 5 ตัวเลือก ที่มีเนื้อหาและโครงสร้างของแบบสอบแบบเดียวกัน และใช้วิธีการสุ่มมาจากระบบคลังข้อสอบ ที่มีเกณฑ์การสุ่มข้อสอบแบบเดียวกัน และมีการบริหารจัดการการสอบที่เป็นไปในทิศทางเดียวกัน แบบสอบแต่ละฉบับมีจำนวน 120 ข้อ มีข้อสอบร่วมภายใน จำนวน 15 ข้อ ใช้เวลาในการสอบ 3 ชั่วโมง ข้อมูลของแบบสอบมีลักษณะ ดังนี้

1. แบบสอบของภาคการศึกษาที่ 1/ 2556 ใช้เป็นแบบสอบฐานให้แบบสอบฉบับอื่น ๆ ปรับเทียบคะแนนเข้าสู่แบบสอบฉบับนี้ เพื่อปรับคะแนนให้อยู่บนสเกลเดียวกัน
2. แบบสอบของภาคการศึกษาที่ 1/ 2557 เป็นแบบสอบที่มีข้อสอบร่วมภายในที่มีความยากอยู่ในช่วง 0.4-0.6 จำนวน 15 ข้อ ที่เหมือนกันกับข้อสอบของภาคการศึกษาที่ 1/ 2556 กระจายตามเนื้อหาชุดวิชาหน่วยละ 1 ข้อ
3. แบบสอบของภาคการศึกษาที่ 1/ 2558 เป็นแบบสอบที่มีข้อสอบร่วมภายในที่ได้มาอย่างสุ่ม จำนวน 15 ข้อ ที่เหมือนกันกับข้อสอบของภาคการศึกษาที่ 1/ 2556 กระจายตามเนื้อหาหน่วยละ 1 ข้อ เช่นเดียวกัน

ประชากรและตัวอย่าง เป็นนักศึกษาทั้งหมดที่เข้าสอบในชุดวิชาที่ผู้วิจัยทำการศึกษา ประกอบด้วย นักศึกษาของภาคการศึกษาที่ 1/ 2556 จำนวน 1,210 คน ภาคการศึกษาที่ 1/ 2557 จำนวน 927 คน และภาคการศึกษาที่ 1/ 2558 จำนวน 857 คน สำหรับกลุ่มตัวอย่างที่นำมาใช้เป็นเงื่อนไขหนึ่งของการปรับเทียบคะแนน แบ่งออกเป็น 3 ขนาด คือ 100 คน 500 คน และ 700 คน ที่ได้มาด้วยวิธีการสุ่มอย่างง่าย

วิธีดำเนินการ แบ่งเป็น 2 ขั้นตอนใหญ่ คือ ขั้นตอนของปรับเทียบคะแนน กับขั้นตอนการเปรียบเทียบผลการตัดเกรดระหว่างการใช้คะแนนก่อนทำการปรับเทียบคะแนนกับคะแนนหลังการปรับเทียบคะแนน

ขั้นตอนแรก เป็นการศึกษาวิธีการปรับเทียบคะแนน 2 วิธี คือ วิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ โดยวิธีเคอเนลวิเคราะห์จากโปรแกรม LOGLIN/ KE version 2.0 ก่อนทำการวิเคราะห์ปรับเทียบคะแนนจะต้องเตรียมข้อมูลการแจกแจงของคะแนนรวมหลังจากตัดข้อสอบร่วมออกกับคะแนนของข้อสอบร่วมด้วยโปรแกรม SAS จากนั้นปรับเทียบคะแนนภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และข้อสอบร่วมอย่างสุ่ม ที่วิเคราะห์การปรับเทียบคะแนนด้วยการตัดและไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง กับกลุ่มตัวอย่าง ขนาด 100 คน 500 คน และ 700 คน ตามลำดับ หลังจากนั้นตรวจสอบคุณภาพของการปรับเทียบคะแนนวิธีเคอเนล โดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน (SEE) ส่วนวิธี IRT 2 พารามิเตอร์ วิเคราะห์ภายใต้เงื่อนไขและขนาดตัวอย่าง เช่นเดียวกันกับวิธีเคอเนลด้วยโปรแกรม BILOG MG version 3.0 และตรวจสอบคุณภาพการประมาณคะแนนดิบจากความเที่ยงและจากค่าความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบที่ได้จากการปรับเทียบคะแนนของแต่ละเงื่อนไข ด้วยโปรแกรม IRT-CLASS สุดท้ายทำการตรวจสอบความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน (SEE) จากคะรวมเดิมกับคะแนนรวมหลังการปรับเทียบคะแนนของทั้งวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์

ขั้นตอนที่สอง เป็นการศึกษาเปรียบเทียบผลการตัดเกรดระหว่างการใช้คะแนนก่อนทำการปรับเทียบคะแนนกับคะแนนหลังการปรับเทียบคะแนน จากการปรับเทียบคะแนนทั้ง 2 วิธี ภายใต้เงื่อนไขต่าง ๆ ของการปรับเทียบ โดยแบ่งการตัดเกรด เป็น 3 ระดับ (H S และ U) และ 8 ระดับ (A B^+ B C^+ C D^+ D และ F) ด้วยการพิจารณาความสอดคล้องของการตัดเกรดจากค่าดัชนีแคปปา โดยกำหนดดัชนีแคปปาตั้งแต่ .81 ขึ้นไปในทางสถิติสามารถยอมรับได้ว่าแบบสอบต่างฉบับที่มีเนื้อหาสาระและมีโครงสร้างของแบบสอบแบบเดียวกัน มีความเท่าเทียมกัน ไม่ทำให้ผู้สอบเกิดความได้เปรียบหรือเสียเปรียบกัน จึงไม่จำเป็นจะต้องทำการปรับเทียบคะแนน แต่ถ้าค่าแคปปามีค่าต่ำกว่า .81 ในทางสถิติถือว่าการตัดเกรดโดยใช้คะแนนก่อนทำการปรับเทียบคะแนนกับคะแนนหลังการปรับเทียบคะแนนไม่สอดคล้องกัน การใช้แบบสอบต่างฉบับกันก่อให้เกิดความได้เปรียบหรือเสียเปรียบกัน จำเป็นจะต้องทำการปรับเทียบคะแนนให้แบบสอบต่างฉบับกันนั้น มีความเทียบเท่ากันก่อนที่จะนำไปตัดเกรด

การวิเคราะห์ข้อมูล ประกอบด้วย

1. การวิเคราะห์ค่าสถิติพื้นฐานของแบบสอบเพื่ออธิบายลักษณะของแบบสอบ และวิเคราะห์เพื่อดูคุณภาพของแบบสอบตามทฤษฎีการตอบข้อสอบแบบดั้งเดิม และทฤษฎีการตอบ

ข้อสอบแนวใหม่

2. การวิเคราะห์การปรับเทียบคะแนนวิธีเคอเนล มีการเตรียมข้อมูลการแจกแจงของคะแนนรวมกับคะแนนของข้อสอบร่วมด้วยโปรแกรม SAS ก่อนที่จะวิคราะห์การปรับเทียบคะแนนวิธีเคอเนลด้วยโปรแกรม LOGLIN/ KE

3. การวิเคราะห์การปรับเทียบคะแนนวิธี IRT 2 พารามิเตอร์ ด้วยโปรแกรม BILOG-MG version 3.0 หลังจากนั้นทำการประมาณคะแนนดิบหลังการปรับเทียบคะแนนด้วยโปรแกรม IRT-CLASS นำมาสร้างสมการพยากรณ์คะแนนดิบของการปรับเทียบคะแนนวิธี IRT 2 พารามิเตอร์ด้วยการวิเคราะห์การถดถอย

4. วิเคราะห์เปรียบเทียบคุณภาพของการปรับเทียบคะแนนจำแนกตามวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ตามเงื่อนไขต่าง ๆ ที่กำหนดจากค่าความคลาดเคลื่อนมาตรฐาน (SEE)

5. วิเคราะห์เปรียบเทียบความสอดคล้องของการตัดเกรดก่อนการปรับเทียบคะแนนและหลังการปรับเทียบคะแนน ของวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ตามเงื่อนไขต่าง ๆ ที่กำหนดด้วยการวิเคราะห์ดัชนีแคปปา และเปรียบเทียบความสัมพันธ์ของการตัดเกรดระหว่างการใช้คะแนนหลังการปรับเทียบคะแนนวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ โดยใช้การทดสอบไคสแควร์ ตลอดจนวิเคราะห์ระดับความสัมพันธ์ของการตัดเกรดระหว่างการใช้คะแนนหลังการปรับเทียบคะแนนวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ของเงื่อนไขต่าง ๆ ที่กำหนด

สรุปผลการวิจัย

1. การปรับเทียบคะแนนวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วม ลักษณะของข้อมูลที่นำมาวิคราะห์และกลุ่มตัวอย่างที่มีขนาดแตกต่างกัน ด้วยโปรแกรม KERNEL สรุปได้ว่า

1.1 การปรับเทียบคะแนนด้วยวิธีเคอเนล เมื่อใช้กลุ่มตัวอย่างขนาด 500 คน และ 700 คน ไม่ว่าจะปรับเทียบคะแนนภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ข้อสอบร่วมมีความยากอย่างสุ่ม ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง หรือข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบต่ำใกล้เคียงกันแสดงถึงคุณภาพในการปรับเทียบคะแนนที่พอ ๆ กัน โดยที่การปรับเทียบคะแนนโดยใช้ข้อสอบร่วมที่มีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง เมื่อวิเคราะห์กับขนาดตัวอย่าง 700 คน มีคุณภาพมากที่สุด ขณะการที่ปรับเทียบคะแนนเมื่อใช้ตัวอย่างขนาด 100 คน ให้ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการปรับเทียบค่อนข้างสูง และเมื่อพิจารณาการปรับเทียบคะแนนกรณีวิเคราะห์โดยตัดข้อสอบที่ไม่มีคุณภาพทิ้ง พบว่า มีคุณภาพในการปรับเทียบคะแนนสูงกว่าการไม่ตัดข้อสอบทิ้ง

1.2 การปรับเทียบคะแนนด้วยวิธีเคอเนล ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และข้อสอบร่วมมีความยากอย่างสุ่มทั้งที่ตัด และไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง โดยภาพรวม พบว่า ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐาน (SEE) ของการปรับเทียบคะแนนมีแนวโน้มลดลงเมื่อใช้ขนาดกลุ่มตัวอย่างในการปรับเทียบคะแนนเพิ่มขึ้น แสดงให้เห็นว่ากลุ่มตัวอย่างยิ่งมากคุณภาพในการปรับเทียบคะแนนก็จะสูงตาม

2. การประมาณค่าคะแนนดิบหลังจากการปรับเทียบคะแนน หลังจากการวิเคราะห์ด้วย IRT-CLASS นำมาสร้างสมการถดถอยในการพยากรณ์คะแนนดิบหลังจากการปรับเทียบคะแนนของแต่เงื่อนไขของการปรับเทียบคะแนน ความคลาดเคลื่อนมาตรฐานของการประมาณค่าคะแนนดิบ เมื่อพิจารณาจากค่าความเที่ยงและความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบ สรุปได้ว่า ค่าความเที่ยงในการประมาณค่าคะแนนดิบของทุกเงื่อนไขมีค่าสูงใกล้เคียงกันตั้งแต่ .9 ขึ้นไป สำหรับค่าความคลาดเคลื่อนส่วนใหญ่แล้วความแปรปรวนของคะแนนดิบของทุกเงื่อนไขก็มีค่าใกล้เคียงกัน เมื่อพิจารณาค่าความคลาดเคลื่อนโดยรวมของความแปรปรวนของคะแนนดิบของปรับเทียบคะแนนกรณีวิเคราะห์โดยตัดข้อสอบที่ไม่มีคุณภาพทิ้ง พบว่ามีคุณภาพในการปรับเทียบคะแนนสูงกว่าการไม่ตัดข้อสอบทิ้งไม่ว่าจะวิเคราะห์กับกลุ่มตัวอย่างตามที่กำหนดขนาดใดก็ตาม

โดยภาพรวมการประมาณค่าคะแนนดิบด้วยเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้งให้ค่าความแปรปรวนของคะแนนดิบต่ำสุดไม่ว่าจะวิเคราะห์กับตัวอย่างตามที่กำหนดขนาดใดก็ตาม ยกเว้นวิเคราะห์กับตัวอย่างขนาด 100 คน ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้งที่สูงกว่าเล็กน้อย

3. การเปรียบเทียบคุณภาพของวิธีการปรับเทียบคะแนนระหว่างวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ วิธีเคอเนลมีคุณภาพในการปรับเทียบคะแนนสูงกว่าวิธี IRT 2 พารามิเตอร์ ทั้งสองวิธีนี้ให้ผลสอดคล้องกันว่าการปรับเทียบคะแนนภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้งเมื่อวิเคราะห์กับกลุ่มตัวอย่างทั้ง 3 ขนาดมีคุณภาพในการปรับเทียบคะแนนสูงกว่าเงื่อนไขอื่น ๆ โดยที่การวิเคราะห์ด้วยวิธีเคอเนลภายใต้เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มกับตัวอย่างขนาด 700 คน และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้งกับตัวอย่างขนาด 500 คน มีคุณภาพในการปรับเทียบคะแนนสูงที่สุด

4. การเปรียบเทียบความสอดคล้องของการตัดเกรด จากการใช้คะแนนก่อนการปรับเทียบคะแนน และคะแนนหลังจากการปรับเทียบคะแนนมาใช้ในการตัดเกรด 3 ระดับ และ 8 ระดับ ตามวิธีและเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน สรุปได้ว่า

โดยภาพรวมการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนหลังจากการปรับเทียบคะแนนด้วยวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ ตามเงื่อนไขต่าง ๆ ภายใต้การตัดเกรด 3 ระดับ และ 8 ระดับ ส่วนใหญ่ให้ผลตรงกันว่าการตัดเกรดไม่มีความสอดคล้องกันจำเป็นจะต้องทำการปรับเทียบคะแนนก่อนที่จะตัดเกรด และการใช้ตัวอย่างขนาด 500 คน และ 700 คน ให้ผลสอดคล้องกันทุกเงื่อนไขว่าจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด ขณะที่การใช้ตัวอย่างขนาด 100 คน โดยเงื่อนไขส่วนใหญ่ (5 เงื่อนไข จาก 8 เงื่อนไข) ถ้ามีผู้เข้าสอบจำนวน 100 คน แล้วในการตัดเกรดไม่จำเป็นจะต้องทำการปรับเทียบคะแนน ผลการวิเคราะห์ความสอดคล้องของการตัดเกรดจำแนกตามระดับของการตัดเกรดสรุปได้ดังนี้

4.1 ความสอดคล้องของการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนหลังจากการปรับเทียบคะแนนด้วยวิธีเคอเนล ตามเงื่อนไขต่าง ๆ ภายใต้การตัดเกรด 3 ระดับ โดยภาพรวมแบบสอบที่นำมาปรับเทียบคะแนนกันนั้นไม่มีความเท่าเทียมกัน จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนที่จะตัดเกรด และเมื่อตัวอย่างในการวิเคราะห์ขนาดใหญ่ขึ้น ความไม่สอดคล้องของการตัดเกรดก็จะชัดเจนมากขึ้น ไม่ว่าจะปรับเทียบคะแนนโดยใช้เงื่อนไขใดก็ตาม เช่นเดียวกันกับการวิเคราะห์ด้วยวิธี IRT 2 พารามิเตอร์ และเมื่อพิจารณาความสอดคล้องของการตัดเกรดกรณีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง การปรับเทียบคะแนนวิธีเคอเนล จะเห็นความชัดเจนของความไม่สอดคล้องของการตัดเกรดมากกว่าการไม่ตัดข้อสอบ ขณะที่การวิเคราะห์ด้วยวิธี IRT 2 พารามิเตอร์การตัดหรือไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้งไม่สามารถระบุได้อย่างชัดเจน

ส่วนความสอดคล้องของการตัดเกรดหลังการปรับเทียบคะแนนการใช้วิธีเคอเนลและ IRT 2 พารามิเตอร์ ด้วยการใช้ข้อสอบร่วมที่มีความยากปานกลาง (.4-.6) สามารถระบุความไม่สอดคล้องของการตัดเกรดได้ชัดเจนกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม ยกเว้นเมื่อวิเคราะห์กับตัวอย่างขนาด 100 คน ด้วยวิธีเคอเนลที่ไม่สามารถระบุได้

4.2 การเปรียบเทียบความสอดคล้องของการตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ภายใต้การตัดเกรด 8 ระดับ ด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ โดยภาพรวมแบบสอบที่นำมาปรับเทียบคะแนนไม่มีความเท่าเทียมกัน จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนที่จะตัดเกรด การใช้กลุ่มตัวอย่างในการวิเคราะห์ขนาดใหญ่ขึ้น จะยิ่งเห็นถึงความไม่สอดคล้องของการตัดเกรดชัดเจนมากขึ้นไม่ว่าจะปรับเทียบคะแนนโดยใช้เงื่อนไขใดก็ตาม สำหรับการใช้ตัวอย่างขนาด 500 คน และ 700 คน ของแต่ละเงื่อนไขในการวิเคราะห์ไม่ว่าจะใช้วิธีเคอเนลหรือวิธี IRT 2 พารามิเตอร์ให้ผลใกล้เคียงกัน ส่วนการใช้ข้อสอบร่วมที่มีความยากปานกลาง

(.4-.6) และการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม การปรับเทียบคะแนน ด้วยวิธี IRT 2 พารามิเตอร์ สามารถระบุได้ว่าการใช้ข้อสอบร่วมที่มีความยากปานกลางบอกระดับความไม่สอดคล้องของการตัดเกรดได้ชัดเจนกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม ส่วนวิธีเคอเนลให้ผลตรงกันข้ามกัน สำหรับความสอดคล้องของการตัดเกรด กรณีการตัดข้อสอบที่ไม่มีคุณภาพทิ้ง การปรับเทียบคะแนนทั้งวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ จะเห็นความชัดเจนของความไม่สอดคล้องของการตัดเกรดมากกว่าการไม่ตัดข้อสอบ

4.3 ความสอดคล้องของการตัดเกรดทั้ง 3 ระดับ และ 8 ระดับ เมื่อวิเคราะห์ด้วยตัวอย่างขนาด 500 คน และ 700 คน ด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ ให้ผลสอดคล้องกันทุกเงื่อนไข คือ จำเป็นจะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด แต่ถ้าใช้ตัวอย่างขนาด 100 คน โดยเงื่อนไขส่วนใหญ่แล้ว (5 เงื่อนไข จาก 8 เงื่อนไข) ไม่จำเป็นจะต้องปรับเทียบคะแนนก่อนการตัดเกรด

4.4 ความสอดคล้องของการตัดเกรดทั้ง 3 ระดับ และ 8 ระดับ เมื่อวิเคราะห์ด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ เมื่อนำมาหาความสัมพันธ์ของการตัดเกรดหลังการปรับเทียบคะแนน ทุกเงื่อนไขของการปรับเทียบทั้งสองวิธีมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ โดยที่การตัดเกรด 3 ระดับ หลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ด้วยเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม กับทุกขนาดตัวอย่างมีความสัมพันธ์กันในระดับดีมาก ส่วนเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบที่ไม่มีคุณภาพทิ้ง มีความสัมพันธ์อยู่ในระดับดีที่เหลือสัมพันธ์กันในระดับต่ำ ส่วนการตัดเกรด 8 ระดับ หลังการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ มีเพียงเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มเพียงเงื่อนไขเดียว ที่การตัดเกรดหลังการปรับเทียบคะแนนของทั้งสองวิธีมีความสัมพันธ์กันอยู่ในระดับดี ส่วนเงื่อนไขที่เหลือสัมพันธ์กันในระดับ

อภิปรายผลการวิจัย

ผลการวิจัยดังกล่าวข้างต้น สามารถนำมาอภิปรายผลได้ ในประเด็นดังต่อไปนี้

1. วิธีการปรับเทียบคะแนน ด้วยวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ เมื่อวิเคราะห์กับตัวอย่างขนาดเล็ก 100 คน วิธีที่มีคุณภาพมากที่สุดคือวิธีเคอเนลโดยใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 ซึ่งสอดคล้องกับตำราของโคเลน และแบรนแนน และรีเบคกา (Kolen & Brennan, 2004, pp. 293-294; Rebecca & Dvorak, 2009) และสมมติฐานข้อที่ 1 แต่เมื่อตัวอย่างมีขนาดใหญ่ในที่นี้ใช้ 700 คน การวิเคราะห์ด้วยวิธี IRT 2 พารามิเตอร์ไม่ได้มีคุณภาพสูงกว่าการวิเคราะห์ด้วยวิธีเคอเนล ซึ่งไม่ได้เป็นไปตามสมมติฐานข้อที่ 2 ทั้งนี้อาจจะมีสาเหตุจากการใช้ตัวอย่างขนาด 700 คน ที่ไม่ใหญ่เพียงพอ ซึ่งแฮร์ริสกล่าวว่า ความคลาดเคลื่อนของการปรับเทียบคะแนนที่

แสดงถึงคุณภาพของการปรับเทียบคะแนน ขึ้นอยู่กับขนาดของตัวอย่างที่ใช้ ถ้าวิเคราะห์ด้วย 1 พารามิเตอร์สามารถใช้ตัวอย่างขนาด 400 คน แต่ถ้าเป็น 3 พารามิเตอร์จะต้องใช้ถึง 1500 คน จึงจะเพียงพอ (Harris, 1993 cited in Kolen & Brennan, 2004, p. 288)

2. วิธีการปรับเทียบคะแนน ด้วยวิธีเคอเนลส่วนใหญ่พบว่าคะแนนในช่วงต้นและช่วงปลายค่อนข้างแกว่งไม่ว่าจะปรับเทียบโดยใช้เงื่อนไขใดก็ตาม ซึ่งทำให้ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนมีแนวโน้มสูงกว่าช่วงกลาง ๆ ของข้อมูล ซึ่งเป็นไปตามคุณลักษณะของการวิเคราะห์ของวิธีนี้ ที่ทำการวิเคราะห์ด้วยวิธี Gaussian สเกลของคะแนนจะเริ่มตั้งแต่-∞ ถึง + ∞ ทำให้คะแนนในช่วงต้นและปลายมีความคลาดเคลื่อนสูง (Mao et al., 2006 cited in Meng, 2012, p. 74) ส่วนความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน (SEE) ของวิธีเคอเนลจากข้อค้นพบจะลดลงทุกเงื่อนไขของการปรับเทียบคะแนน เมื่อกลุ่มตัวอย่างที่นำมาใช้ในการวิเคราะห์มีขนาดใหญ่สอดถล้องกับข้อค้นพบของยู เม็ง ที่ว่าตัวอย่างในการวิเคราะห์มีขนาดใหญ่ขึ้นย่อมเป็นตัวแทนประชากรที่ดีกว่า และทำให้การประมาณค่าได้ดีกว่า (Meng, 2012, p. 71) ซึ่งสอดคล้องกับที่แฮนสัน และบีแกน และลีได้กล่าวไว้ (Hanson & Beguin, 2002; Lee, 2007 cited in Meng, 2012, p. 27)

3. การตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนที่จะนำไปปรับเทียบคะแนนของทั้งเงื่อนไขการใช้ข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มของทั้งวิธีเคอเนลและวิธี IRT 2 พารามิเตอร์ ส่วนใหญ่ไม่ได้ทำให้คุณภาพของการปรับเทียบคะแนนดีขึ้น ทั้งนี้อาจจะเป็นผลมาจากการตัดข้อสอบทิ้งทำให้ความยาวของแบบสอบสั้นลง ซึ่งความยาวของแบบสอบมีผลต่อการประมาณค่าพารามิเตอร์ และมีผลต่อคุณภาพของการปรับเทียบคะแนน ซึ่งความยาวของแบบสอบที่เพิ่มขึ้น สามารถลดขนาดตัวอย่างในการวิเคราะห์ให้น้อยลงได้ โดยยังคงไว้ซึ่งคุณภาพที่ไม่แปรเปลี่ยน (พิชัย ละแมนชัย, 2538, หน้า 135)

4. การใช้ข้อสอบร่วมที่มีความยากปานกลาง (.4-.6) ในการปรับเทียบคะแนนทั้งวิธีเคอเนล และ IRT 2 พารามิเตอร์ ไม่ได้มีคุณภาพในการปรับเทียบคะแนนสูงกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม ทั้งนี้อาจจะเนื่องมาจากการใช้ข้อสอบร่วมที่มียากปานกลาง (.4-.6) มีช่วงที่แคบกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่มที่มีความเป็นตัวแทนของแบบสอบทั้งฉบับได้ดีกว่า จึงทำให้ความคลาดเคลื่อนต่ำกว่า

5. การตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ภายใต้การตัดเกรด 3 ระดับ ด้วยวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ ส่วนใหญ่ให้ผลตรงกันว่าไม่มีความสอดคล้องกัน แสดงว่าแบบสอบทั้ง 2 ฉบับที่นำมาปรับเทียบคะแนนไม่มีความเท่าเทียมกันทำให้ผู้สอบเกิดการได้เปรียบ

เสียเปรียบกัน เช่นเดียวกันกับการตัดเกรด 8 ระดับ ซึ่งการตัดเกรด 8 ระดับ เป็นการยืนยันและชี้ให้เห็นถึงความไม่สอดคล้องของการตัดเกรดได้ชัดเจนมากยิ่งขึ้น ว่าทำให้ผู้สอบที่ทำแบบสอบต่างฉบับเกิดการได้เปรียบเสียเปรียบกัน จำเป็นอย่างยิ่งที่จะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด จึงจะทำให้ผู้สอบที่ทำแบบสอบต่างฉบับภายใต้เนื้อหาสาระโครงสร้างแบบสอบและการบริหารการสอบแบบเดียวกันได้รับความยุติธรรม

นอกจากนี้การตัดเกรดจากการใช้คะแนนก่อนการปรับเทียบคะแนนกับคะแนนที่ได้หลังจากการปรับเทียบคะแนนตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน ภายใต้การตัดเกรด 3 ระดับ และ 8 ระดับ ด้วยวิธีเคอเนล และวิธี IRT 2 พารามิเตอร์ทุกขนาดตัวอย่าง ชี้ให้เห็นถึงความไม่สอดคล้องของการตัดเกรดว่า การปรับเทียบคะแนนโดยการตัดข้อสอบที่ไม่มีคุณภาพออกไปก่อนให้ผลชัดเจนกว่าไม่ตัด สำหรับการใช้ข้อสอบร่วมที่มีความยากปานกลาง (.4-.6) ในการปรับเทียบคะแนนด้วย IRT 2 พารามิเตอร์ เมื่อพิจารณาความไม่สอดคล้องของการตัดเกรดสามารถบอกได้อย่างชัดเจนว่าดีกว่าการใช้ข้อสอบร่วมที่มีความยากอย่างสุ่ม ขณะที่การปรับเทียบคะแนนด้วยวิธีเคอเนลไม่สามารถระบุได้ อาจจะมาจากสาเหตุความสามารถของตัวอย่างแต่ละกลุ่มที่นำมาวิเคราะห์มีความแตกต่างกัน ซึ่งวิธีการวิเคราะห์ด้วยเคอเนลจะมีความคลาดเคลื่อนสูงเมื่อตัวอย่างในการวิเคราะห์มีความสามารถแตกต่างกัน (Meng, 2012, p. 74) ความสามารถเฉลี่ยของกลุ่มที่นำมาวิเคราะห์อยู่ในช่วง -.07 ถึง .04

ข้อเสนอแนะ

ข้อเสนอแนะในการนำไปใช้

1. ควรนำวิธีการปรับเทียบคะแนนที่ศึกษาในครั้งนี้ ไปใช้กับแบบสอบที่มีนักศึกษาลงทะเบียนเรียนเป็นจำนวนมาก ซึ่งนักศึกษาจำนวน 500 คนขึ้นไป ตามผลการวิจัยนี้จะเป็นจำนวนที่ให้ผลการปรับเทียบคะแนนมีความคงที่ ไม่ว่าจะใช้วิธีเคอเนลหรือวิธี IRT 2 พารามิเตอร์ภายใต้เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และเงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มทั้งที่มีการตัดหรือไม่ตัดข้อสอบที่ไม่มีคุณภาพทิ้ง ซึ่งจะให้ผลสอดคล้องกัน โดยเงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง มีคุณภาพในการปรับเทียบคะแนนมากที่สุด ขณะที่แบบสอบใดมีนักศึกษาลงทะเบียนเรียนราว 100 คน ไม่จำเป็นจะต้องทำการปรับเทียบคะแนน

2. ควรนำวิธีการปรับเทียบคะแนนวิธีเคอเนลไปใช้ มากกว่าจะใช้วิธี IRT 2 พารามิเตอร์ เพราะว่ามีคุณภาพในการปรับเทียบคะแนนสูงกว่า รวมทั้งมีกระบวนการขั้นตอนการวิเคราะห์ที่ไม่ซับซ้อน สะดวกและง่ายต่อการใช้งานมากกว่า

3. ควรทำการตรวจสอบชุดวิชาอื่น ๆ ว่าแบบสอบต่างฉบับมีความเท่าเทียมกันหรือไม่ เพื่อการยืนยันหากแบบสอบต่างฉบับก่อให้เกิดการได้เปรียบหรือเสียเปรียบกัน ก็ควรทำการปรับเทียบคะแนนก่อนการตัดเกรดเพื่อสร้างความเป็นมาตรฐานเดียวกันของแบบสอบและการตัดเกรด เป็นการให้ความยุติธรรมกับผู้สอบทุกคน

4. ผลจากการวิจัย การตัดเกรด 3 ระดับ ถึงแม้จะมีช่วงคะแนนในการตัดเกรดแต่ละระดับกว้างยังก่อให้เกิดความได้เปรียบกันในการตัดเกรด การขยายระดับของการตัดเกรดถี่ขึ้นเป็น 8 ระดับ จะยิ่งเพิ่มความได้เปรียบเสียเปรียบของการตัดเกรดชัดเจนมากขึ้น ดังนั้นถ้าจะขยายการตัดเกรดเป็น 8 ระดับ จำเป็นอย่างยิ่งที่จะต้องทำการปรับเทียบคะแนนก่อนการตัดเกรด

ข้อเสนอแนะสำหรับการวิจัยครั้งต่อไป

1. ควรทำการศึกษาการปรับเทียบคะแนนด้วยวิธีเคอเนลและวิธี IRT ภายใต้เงื่อนไขที่ต่างกัน ด้วยการจำลองข้อมูลเพื่อลดข้อจำกัดด้านการใช้สถานการณ์จริง ที่ไม่สามารถนำแบบสอบไปทดลองใช้ เพื่อวิเคราะห์คุณภาพข้อสอบตัดข้อสอบที่ไม่มีคุณภาพทิ้งก่อนนำไปใช้จริงได้ จึงทำให้การใช้สถานการณ์จริงเมื่อปรับเทียบคะแนนด้วย IRT ของแต่ละเงื่อนไขทำให้ข้อสอบบางข้อถูกตัดทิ้งไปตามเงื่อนไขของโปรแกรม เป็นเหตุให้จำนวนข้อสอบในแบบสอบของแต่ละเงื่อนไขไม่เท่ากัน ซึ่งความยาวของแบบสอบมีผลต่อคุณภาพของการปรับเทียบคะแนน

2. ควรขยายผลการศึกษาครั้งต่อไปด้วยการเปรียบเทียบคุณภาพการปรับเทียบคะแนนวิธีเคอเนล โดยใช้วิธีการวิเคราะห์ 2 แบบ คือ CE (Chain equating) และ PSE (Post-stratification equating) ส่วน IRT ใช้วิธีการวิเคราะห์ 1 2 และ 3 พารามิเตอร์ ภายใต้เงื่อนไขขนาดตัวอย่างที่ต่างกัน และควบคุมความยาวของแบบสอบให้เท่ากัน เพื่อศึกษาว่าวิธีการใดจะดีกว่ากัน

3. งานวิจัยครั้งนี้ยังไม่ได้ข้อค้นพบว่าวิธีการปรับเทียบใดมีคุณภาพและให้ผลที่สอดคล้องกับการตัดเกรดกับข้อมูลที่มีขนาดเล็กที่ชัดเจนเพียงพอ จึงควรทำการศึกษาซ้ำโดยเน้นไปที่การหาวิธีการปรับเทียบคะแนนที่มีคุณภาพเหมาะสมกับกลุ่มขนาดเล็ก

บรรณานุกรม

ชยุตม์ ภิรมย์สมบัติ. (2556). *คู่มือปฏิบัติการครูในการประเมินสู่การเรียนการสอน "การปฏิรูปการศึกษาโดยใช้การประเมินเป็นฐาน" สำนักงานกองทุนสนับสนุนการวิจัย (สกว.)* . กรุงเทพฯ: พริกหวานกราฟฟิค.

ประสพชัย พสุนนท์. (2558). การประเมินความเชื่อมั่นระหว่างผู้ประเมินโดยใช้สถิติแคปปา. *วารสารวิชาการศิลปศาสตร์ประยุกต์*, *8*(1), 18.

พจน์ สะเพียรชัย. (2549). *รวมบทความทางการประเมินโครงการ เรื่องการวิเคราะห์ระบบการประเมินผล*. กรุงเทพฯ: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

พระราชบัญญัติข้อมูลข่าวสารทางราชการ พ.ศ. 2540. (2540). *ความหมายของข้อมูลข่าวสารราชการ*. เข้าถึงได้จาก www.rd.go.th

พัชรี จันทร์เพ็ง. (2550). *การเปรียบเทียบคุณภาพของวิธีการเชื่อมโยงคะแนนตามทฤษฎีการตอบข้อสอบแบบพหุมิติภายใต้การหมุนแกนโครงสร้างเชิงมิติและระดับความสัมพันธ์ที่แตกต่างกัน*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

พิชัย ละแมนชัย. (2538). *ขนาดกลุ่มตัวอย่างขั้นต่ำสำหรับการปรับเทียบคะแนนในแนวระดับ*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

ภาวิณี ศรีสุขวัฒนานันท์. (2528). *การเปรียบเทียบผลจากการใช้รูปแบบการเทียบมาตราที่ต่างกันเมื่อแบบสอบร่วมมีความยาวต่างกัน*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

ภัทราพร เกษสังข์. (2546). *การศึกษาพัฒนาการความสามารถทางคณิตศาสตร์ของนักเรียนชั้นมัธยมศึกษาตอนต้นโดยการปรับเทียบคะแนนในแนวตั้งที่ใช้วิธีการที่เหมาะสม*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

ศจี จิระโรและศศิธร บัวทอง. (2557). *แนวทางสำหรับระบบการตัดเกรดของนักศึกษาระดับปริญญาตรีมหาวิทยาลัยสุโขทัยธรรมาธิราช*. นนทบุรี: ศูนย์วิชาการประเมินผล สำนักทะเบียนและวัดผล มหาวิทยาลัยสุโขทัยธรรมาธิราช.

ศิริชัย กาญจนวาสี. (2551). *การพัฒนาวิธีการปรับเทียบผลการเรียนเฉลี่ยสะสมตามกลุ่มสาระการเรียนรู้ โดยใช้คะแนน O-NET ของนักเรียนมัธยมศึกษาตอนปลาย*. กรุงเทพฯ: สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ.

ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่ (ฉบับปรับปรุง)*. กรุงเทพฯ: ศูนย์ตำราเอกสารทางวิชาการ คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

สำนักงานคณะกรรมการข้อมูลข่าวสารของทางราชการ. (2558). *พระราชบัญญัติข้อมูลข่าวสารทางราชการ พ.ศ. 2540*. เข้าถึงได้จาก http:// www.oic.go.th / web2014/ ACTOfficial_ Information.ht

สำนักทะเบียนและวัดผล. (2550). *สารสนเทศงานทะเบียนและวัดผล ประจำปี 2549*. นนทบุรี: มหาวิทยาลัยสุโขทัยธรรมาธิราช.

สุนิสา จุ้ยม่วงศรี. (2537). *ผลของความยาวของแบบสอบร่วมที่มีต่อคุณภาพของวิธีการเทียบมาตราเชิงเส้นตรง*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาวิจัยการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

สุรชัย มีชาญ. (2547). สัมประสิทธิ์แคปป้า (Coefficient kappa): แนวคิดและการประยุกต์ใช้. *วารสารวัดผลการศึกษา*, *26*(78), 55-65.

สุวิมล ว่องวาณิช. (2550). *รวมบทความการประเมินผลการเรียนรู้แนวใหม่*. กรุงเทพฯ: ศูนย์ตำราและเอกสารทางวิชาการ คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

อดิศร ศรีบุญวงศ์. (2545). *การพัฒนาเกณฑ์ตัดสินคุณภาพการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

อุทัยวรรณ พงศ์อร่าม. (2545). *การศึกษาขนาดกลุ่มตัวอย่างที่เหมาะสมสำหรับการปรับเทียบคะแนนด้วยวิธีอิควิเปร์เซ็นไทล์ และวิธีเชิงเส้นตรงตามแบบจำลองคะแนนจริง สัมพันธ์ที่มีแบบแผนการปรับเทียบและความยาวของแบบสอบแตกต่างกัน*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาการทดสอบและการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

Alina, A., Davier, V., Holland, P. W., & Thayer, D. T. (2003). *The Kernel method of test equating*. New York Berlin Heidelberg.

Angoff, W. H. (1971). Scales, norms and equivalent score. *Educational Measurement*, *8*(2), 151-170.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Ayerve, R. I. (1992). *The effectiveness of the equipercentile method and IRT three parameter model on vertical equating under verying condition of sample size, test length and anchor test length: A simulation study*. Doctoral Dissertation. Columbia University.

Baker, F. B., & Al-karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 157-172.

Bastari, B. (2000). *Linking multiple-choice and constructed-response items toba common proficiency scale*. Dissertation, Ed. D. UEA: Graduate School, Massachusetts Amherst University. Available: UMI/ Dissertation, 2001, pp. 1-78.

Budescu, D. V. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement*, *22*, 183-196.

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.

Cook, L. 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.

Cizek, G. J. (1994). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Erlbaum.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, Florida: Holt, Rinehart and Winston.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT Verbal and Mathematical sections*. Princeton, NJ: Educational Testing Service.

Glowaki, M. L. (1991). *The analysis of test equating models for the alabama high school graduation examination.* Doctoral Dissertation, University of Alabama, *Dissertation Abstracts International*, *52*, 1722.

Godfrey, K. E., & Kelly, K. (2007). *A comparison of kernel equating and IRT true score equating methods*. Greensboro: University of North Carolina.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item respose theory item parameters using scparate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New Jersey: Education Testing Service, Princeeton, Academic Press.

Holland, P. W., & 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.

Huang, J. (2009). RAD18 transmits DNA damage signalling to elicit homologous recombination repair. *Nat Cell Biol*, *11*(5), 592-603.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131-143.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating Test. *Journal of Educational Measurement*, *18*(1), 1-11.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, *27*, 27-39.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*,

Lee, M. S. (2007). *A study on relationship among leadership, organizational culture, the operation of learning organization and employees' job satisfaction*. Retriered from http://www.emeraldinsight.com/doi/pdfplus/10.1108/09696470710727014

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates.

Lord, F. M., & Winggerskey, D. (1984). Comparison of IRT true-score and equipercentile observed-score “eqatings.” *Applied psychological measurement*, *8*, 452-461.

Meng, X. (2012). Labor Market Outcomes and Reforms in China. *Journal of Economic Perspectives*, *26*(4), 75-102.

Patience, W. (1990). *A comparison of latent trait and equipercentile method of vertical equting test, paper presented at annual meeting of the national council on measurement in* education, Los Angeles. Los Angeles: Muthén & Muthén.

Paul, W. M., Donald, B. R. (1982). *Test equating*. New York: Academic Press.

Petersen., N. S., Kolen, M. J., & Hoover, H. D. (1989). *Scaling, norming, and equating* (3rd ed.). New York: Macmillan.

Pertersen, N. S., Marco, G. L., & Stewart, E. E. (1982). *A test of the adequacy of linear score equating models*. In P.W. Holland and D.B. Rubin (Eds.) *Test Equating.* New York: Academic Press, pp. 71-135.

Phillips, S. E. (1986). Comparison of equipercentile and item response theory equating when the scaling test method is applied to a multilevel achievement battery. *Applied Psychological Measurement*, *7*(3), 267-281.

Rebecca, L., & Dvorak, N. (2009). *A comparison of kernel equating to the test characteristic curve method*. Lincoln, Nebraska: University of Nebraska.

Ricker, K. L., & Von Davier, A. A. (2007). *The impact of anchor test length on equating results in a nonequivalent groups design*. Retrieved from http://www.ets.org/research/policy_ research_reports/publications/report/2007/hsof

Sinharay, S., & Holland, P. W. (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*, 249-275.

Slinda, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the rasch model for the problem of vertical equating. *Journal of Educational Measurement*. *15*(1), 23-35.

Stocking, M. L., & 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*. Austin: The University of Texas.

Wang, X., Haiming, X., Donghong, J., & Youming, X. (2008). Disruption of Rpn4-induced proteasome expression in saccharomyces cerevisiae reduces cell viability under stressed conditions. *Genetics*, *180*(4), 1945-53.

Wendy, L. (2009). *Linking current and future score scales for the AICPA uniform CPA exam*. Amherst: University of Massachusetts.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicaco: Mesa Press.

Yu Meng. (2012). *Comparison of kernel equating and item response theory equating methods*. Massachusetts: United States, University of Massachusetts.

ภาคผนวก

ภาคผนวก ก

ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนด้วยวิธีเคอเนล
ภายใต้เงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน

1. เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

x	n 100	n 500	n 700
0	2.362991	0.748958	0.683312
1	2.548275	1.248084	0.709413
2	2.732805	1.717945	0.879495
3	2.762187	2.073558	1.076605
4	3.21038	2.140066	1.105671
5	3.165313	2.43328	1.282176
6	3.45726	2.318693	1.525607
7	3.529534	2.223663	1.784713
8	3.627059	2.352722	2.043199
9	3.886885	2.236589	2.020489
10	3.847559	2.160573	2.19183
11	3.923498	2.12301	2.079007
12	3.98242	2.06926	2.214935
13	4.019479	2.052654	2.044851
14	4.195192	2.02398	2.15431
15	4.187329	2.032832	2.032547
16	4.159854	1.990247	1.89843
17	4.116387	1.984516	1.824712
18	3.81315	1.848899	1.77107
19	3.755615	1.830752	1.707051
20	3.69007	1.809417	1.661653
21	3.618478	1.687289	1.613369
22	3.542794	1.662388	1.480123
23	3.464244	1.633845	1.451796
24	3.383243	1.546673	1.40119
25	3.300098	1.517833	1.380496

x	n 100	n 500	n 700
26	3.215638	1.487353	1.28091
27	3.040316	1.392718	1.248695
28	2.962297	1.376497	1.213698
29	2.884926	1.344799	1.137225
30	2.808419	1.267313	1.112765
31	2.733553	1.238388	1.043883
32	2.661313	1.180652	1.020729
33	2.592292	1.154748	0.995372
34	2.526644	1.1303	0.939289
35	2.421194	1.086953	0.920208
36	2.368395	1.065916	0.901378
37	2.320857	1.046749	0.864588
38	2.279029	1.01887	0.846922
39	2.243	1.007175	0.82669
40	2.212981	0.997265	0.818384
41	2.189546	0.985087	0.811638
42	2.173339	0.983166	0.805198
43	2.162342	0.984582	0.806317
44	2.165396	0.991452	0.816698
45	2.175358	0.99935	0.826417
46	2.192387	1.018531	0.835184
47	2.216572	1.034842	0.86165
48	2.247536	1.053121	0.878798
49	2.284516	1.087979	0.896282
50	2.377444	1.110954	0.936416
51	2.412066	1.135985	0.962605
52	2.469583	1.186317	1.018055
53	2.531487	1.214581	1.042206

x	n 100	n 500	n 700
54	2.596831	1.24258	1.063514
55	2.664975	1.303302	1.137603
56	2.735776	1.33475	1.161643
57	2.809131	1.365116	1.19746
58	3.008358	1.434083	1.274412
59	3.094209	1.463309	1.307358
60	3.126529	1.541182	1.395053
61	3.209129	1.569711	1.431382
62	3.292047	1.594553	1.463082
63	3.374511	1.615979	1.488663
64	3.455434	1.733064	1.627967
65	3.534085	1.751529	1.678304
66	3.831295	1.763946	1.75347
67	3.915847	1.76901	1.788422
68	3.995835	1.934158	1.878471
69	4.069325	1.92653	2.000195
70	4.134716	1.961343	2.11344
71	4.190655	1.985753	2.033763
72	4.235176	2.056777	2.220022
73	4.265296	2.11936	2.546701
74	4.277814	2.017074	2.440906
75	4.271204	2.174784	2.372
76	4.247037	2.104271	2.766296
77	4.353865	2.294354	2.669907
78	3.978761	2.170698	2.620765
79	4.063797	2.568115	2.547984
80	4.035183	2.452856	2.526146
81	3.789202	2.408787	2.483367

x	n 100	n 500	n 700
82	3.814304	2.345409	3.167947
83	3.590926	2.352846	3.171655
84	3.391058	2.250805	2.954885
85	3.552724	3.091587	2.835516
86	3.389534	3.166517	2.66844
87	3.248837	3.227721	1.868081
88	3.128973	3.132843	1.531623
89	3.028608	3.462735	1.417087
90	2.94598	3.032506	1.193963
91	2.881414	2.175944	0.796596
92	2.713717	4.447058	0.554021
93	2.798242	2.594958	0.440385
94	2.776144	1.764875	0.390797
95	2.766855	0	0.371638
96	2.760038	0	0.36443
97	2.631113	0	0.361753
98	2.730429	0	0.360781
99	2.534538	0	0.360401
100	1.805949	0	0.360242
101	2.09656	0	0.36017
102	4.60196	0	0.360137
max	4.60196	4.447058	3.171655
min	1.805949	0	0.360122
range	2.796011	4.447058	2.811533
mean	3.173733	1.632232	1.422358
SD	0.70095	0.806936	0.715662

2. เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

x	n 100	n 500	n 700
0	5.967073	1.943247	1.47782
1	5.854012	1.964343	2.349428
2	5.605328	2.080326	2.983907
3	4.959766	2.445067	2.985786
4	4.872545	2.312837	2.948596
5	4.637377	2.206512	2.868032
6	4.485598	2.355989	2.768044
7	4.280741	2.233011	2.582399
8	4.189469	2.093831	2.425764
9	4.328206	2.221351	2.305122
10	4.054563	2.109813	2.160196
11	3.913279	2.056017	2.007827
12	3.771793	1.993039	1.909796
13	3.954033	2.057597	1.811204
14	3.791296	2.031978	1.759675
15	3.635236	1.946722	1.706855
16	3.486301	1.911237	1.616082
17	3.344686	1.871416	1.566173
18	3.45115	1.82824	1.51677
19	3.302258	1.782627	1.468129
20	3.161959	1.735437	1.420449
21	3.089101	1.687364	1.373772
22	2.959167	1.638867	1.328039
23	3.013411	1.590262	1.283247
24	2.882636	1.541859	1.259092
25	2.760721	1.494005	1.214828

x	n 100	n 500	n 700
26	2.647117	1.447019	1.172024
27	2.574839	1.421531	1.130672
28	2.560525	1.374774	1.090778
29	2.485725	1.329458	1.052462
30	2.387478	1.285806	1.026283
31	2.297628	1.244083	0.990431
32	2.233499	1.204503	0.956704
33	2.213764	1.167216	0.925164
34	2.138473	1.140908	0.895957
35	2.082482	1.107372	0.874164
36	2.022294	1.076874	0.849427
37	1.971139	1.030003	0.827739
38	1.962271	1.009684	0.80921
39	1.921228	0.992913	0.793974
40	1.889746	0.980944	0.782988
41	1.866284	0.970956	0.774214
42	1.852961	0.965163	0.769271
43	1.848562	0.963621	0.768157
44	1.84772	0.966318	0.770832
45	1.857109	0.973235	0.774471
46	1.864794	0.98436	0.783917
47	1.892487	0.993752	0.797032
48	1.894734	1.011902	0.813633
49	1.934063	1.033764	0.833489
50	1.98301	1.059105	0.856408
51	2.041545	1.087744	0.874246
52	2.083039	1.119529	0.901846
53	2.111454	1.140663	0.932028

x	n 100	n 500	n 700
54	2.156964	1.176474	0.964553
55	2.239147	1.214628	0.999226
56	2.330194	1.254952	1.021842
57	2.387301	1.297346	1.059105
58	2.416716	1.34171	1.098237
59	2.473073	1.387868	1.139076
60	2.582227	1.408859	1.181463
61	2.700375	1.455683	1.203123
62	2.82794	1.580458	1.246593
63	2.894713	1.635743	1.291541
64	2.913974	1.692386	1.337873
65	2.975046	1.750208	1.385447
66	3.117204	1.763959	1.40175
67	3.269898	1.820584	1.449316
68	3.334679	1.878094	1.49814
69	3.33479	1.936302	1.54818
70	3.496621	1.994915	1.599284
71	3.553995	2.053643	1.606613
72	3.730252	2.112242	1.656768
73	3.58861	2.101572	1.707989
74	3.758897	2.154815	1.76029
75	3.940846	2.206272	1.813507
76	3.98411	2.25506	1.867417
77	4.180665	2.300308	1.92183
78	4.127496	2.341123	1.976363
79	4.153882	2.469161	2.029838
80	4.350761	2.501246	2.155814
81	4.561418	2.317295	2.201562

x	n 100	n 500	n 700
82	4.284543	2.416725	2.318331
83	4.474377	2.51144	2.515533
84	4.673649	2.599051	2.591001
85	5.112532	2.446052	2.840388
86	4.726463	2.71964	3.045437
87	5.129087	2.52598	3.171468
88	5.277477	2.633706	3.189919
89	5.024152	2.513366	3.08145
90	5.317217	2.445114	2.855238
91	4.984659	2.394625	2.017418
92	4.841202	2.586791	1.410428
93	3.901522	2.161397	0.960423
94	3.474278	2.70292	0.389755
95	2.937383	0	0.332329
96	2.658997	0	0.195127
97	1.320907	0	0.25075
98	1.128675	0	0.242148
99	0.497416	0	0.082058
100	0.275836	0	0.080814
101	0.147968	0	0.080457
102	0.049838	0	0.080376
max	5.967073	2.71964	3.189919
min	0.049838	0	0.080376
range	5.917235	2.71964	3.109543
mean	3.144055	1.598807	1.454157
SD	1.263672	0.714914	0.780461

3. เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อที่ไม่มีคุณภาพทิ้ง

x	n 100	n 500	n 700
0	2.101874	1.223392	0.718608
1	2.440545	1.885438	1.013645
2	2.872161	2.296921	1.220454
3	3.27683	2.418596	1.347031
4	3.218384	2.484246	1.392061
5	3.686963	2.540721	1.423767
6	3.841949	2.537272	1.495789
7	3.688799	2.56127	1.512322
8	3.780008	2.617617	1.565318
9	3.839231	2.641871	1.604832
10	3.868561	2.64107	1.629785
11	3.871659	2.684659	1.685753
12	3.85331	2.710038	1.604029
13	3.916065	2.578563	1.586913
14	3.862758	2.629026	1.526938
15	3.648456	2.61372	1.5302
16	3.579382	2.512319	1.460138
17	3.504852	2.484541	1.391308
18	3.426646	2.380755	1.379055
19	3.230238	2.382355	1.310294
20	3.152371	2.277868	1.293461
21	3.125077	2.237504	1.246878
22	3.041265	2.167347	1.182748
23	2.95817	2.123749	1.160868
24	2.795602	2.031935	1.102259
25	2.720271	2.010486	1.048337

x	n 100	n 500	n 700
26	2.677512	1.926979	1.036732
27	2.603287	1.901997	0.987371
28	2.478204	1.827686	0.943602
29	2.416584	1.802491	0.929986
30	2.358809	1.740353	0.892258
31	2.320531	1.710091	0.860816
32	2.239427	1.666707	0.849441
33	2.200868	1.62775	0.826286
34	2.168418	1.612169	0.809683
35	2.142503	1.587892	0.803836
36	2.115219	1.579918	0.797446
37	2.108966	1.569235	0.797797
38	2.108624	1.571601	0.801693
39	2.11557	1.573892	0.813399
40	2.142377	1.585289	0.827003
41	2.15976	1.61044	0.84005
42	2.191692	1.62087	0.867245
43	2.230052	1.66114	0.892567
44	2.308505	1.67508	0.912323
45	2.363547	1.727863	0.952691
46	2.400913	1.744803	0.998618
47	2.463255	1.809875	1.00897
48	2.587157	1.850904	1.060071
49	2.661958	1.899352	1.115832
50	2.739986	1.943608	1.120585
51	2.779513	1.995607	1.17965
52	2.858654	2.039469	1.243217
53	3.029192	2.128458	1.263469

x	n 100	n 500	n 700
54	3.116215	2.130925	1.330156
55	3.204203	2.223759	1.401262
56	3.227624	2.26223	1.381432
57	3.431566	2.245281	1.447705
58	3.519042	2.330547	1.450145
59	3.604693	2.353487	1.509495
60	3.688271	2.368357	1.495799
61	3.67504	2.435387	1.581577
62	4.007887	2.42713	1.589479
63	3.969353	2.475882	1.537465
64	4.027094	2.442888	1.522626
65	4.073939	2.333984	1.498905
66	4.106726	2.340699	1.470657
67	4.122015	2.410877	1.497516
68	4.117561	2.334175	1.541876
69	4.093801	2.466435	1.427141
70	4.054733	2.361107	1.55382
71	4.00895	2.327383	1.436745
72	3.971735	2.621754	1.33041
73	3.58737	2.652799	1.564391
74	3.708581	2.574463	1.448954
75	3.346366	2.579592	1.342112
76	3.402212	2.673976	1.297716
77	3.20247	2.761779	1.255258
78	3.396746	2.728156	1.213634
79	3.227854	2.792308	1.484809
80	3.078509	2.966105	1.446429
81	2.945514	2.297069	1.212097

x	**n 100**	**n 500**	**n 700**
82	2.706456	2.275441	0.827391
83	2.595349	2.223513	0.759341
84	3.044003	2.127674	0.640065
85	2.908908	1.879154	0.47951
86	2.724946	1.489324	0.416892
87	2.344996	0.871741	0.397961
88	1.953573	0.3489	0.393881
89	1.583877		
max	4.122015	2.966105	1.685753
min	1.583877	0.3489	0.393881
range	2.538138	2.617204	1.291872
mean	3.06723	2.14381	1.191574
SD	0.674733	0.456054	0.325533

4. เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่มและตัดข้อสอบข้อที่ไม่มีคุณภาพทิ้ง

x	n 100	n 500	n 700
0	4.806682	1.903392	2.547044
1	4.803917	2.024143	2.834678
2	4.575512	2.216798	2.760908
3	4.259289	1.985381	2.492288
4	4.203362	2.013015	2.22666
5	4.06697	1.979795	2.171918
6	4.013842	1.973807	2.06844
7	3.92338	1.93904	1.998708
8	3.811136	2.031112	1.910012
9	3.79378	2.019381	1.664491
10	3.655656	1.854345	1.619717
11	3.516359	1.823501	1.572343
12	3.4643	1.786326	1.561215
13	3.323043	1.744669	1.510003
14	3.18706	1.741353	1.459234
15	3.122971	1.691759	1.439451
16	2.992662	1.6413	1.388187
17	2.924517	1.624253	1.27909
18	2.931063	1.571099	1.234313
19	2.854042	1.465837	1.21097
20	2.731088	1.442467	1.185664
21	2.655719	1.393391	1.141519
22	2.544396	1.36661	1.114581
23	2.472221	1.318798	1.072313
24	2.373283	1.289784	1.044388
25	2.306056	1.210645	0.973401

x	n 100	n 500	n 700
26	2.28878	1.170439	0.938903
27	2.221541	1.143635	0.914546
28	2.141062	1.107091	0.890345
29	2.082138	1.081488	0.860755
30	2.016799	1.056412	0.838848
31	1.968006	1.004525	0.814225
32	1.918616	0.98614	0.7804
33	1.881307	0.969418	0.76668
34	1.871232	0.952872	0.753844
35	1.841972	0.941373	0.745151
36	1.821698	0.932594	0.738995
37	1.806976	0.926321	0.736479
38	1.805634	0.926676	0.73626
39	1.805675	0.933248	0.741056
40	1.823435	0.939848	0.755283
41	1.838272	0.954533	0.766014
42	1.855634	0.966909	0.778935
43	1.879353	1.001274	0.799741
44	1.928087	1.029953	0.816905
45	1.964049	1.05137	0.84334
46	2.029912	1.073707	0.863777
47	2.076756	1.110389	0.917164
48	2.159178	1.134723	0.941744
49	2.215199	1.203455	0.979704
50	2.247016	1.250345	1.005228
51	2.304379	1.277947	1.046624
52	2.408774	1.328258	1.07229
53	2.472062	1.35476	1.096882

x	n 100	n 500	n 700
54	2.59104	1.407609	1.141156
55	2.658996	1.510629	1.23417
56	2.792468	1.536424	1.259657
57	2.747541	1.597211	1.311727
58	2.810312	1.619379	1.334858
59	2.949665	1.638312	1.355652
60	3.014193	1.699086	1.408598
61	3.167534	1.833458	1.426386
62	3.332525	1.847005	1.480741
63	3.40099	1.91376	1.617761
64	3.582248	1.920341	1.632303
65	3.454778	1.985775	1.69392
66	3.509487	1.984188	1.704013
67	3.68923	2.04674	1.766529
68	3.741281	2.211409	1.771587
69	3.935458	2.275374	1.834147
70	4.143836	2.2486	1.833092
71	4.19436	2.30119	2.09466
72	3.965741	2.348055	2.162451
73	4.162162	2.386882	2.223477
74	4.370948	2.415097	2.362079
75	4.592714	2.538895	2.388649
76	4.610731	2.301642	2.584486
77	4.83799	2.379353	2.631981
78	4.689013	2.329678	2.727224
79	4.877244	2.152659	2.619291
80	5.044151	2.046386	2.795802
81	5.44223	1.851603	2.821245

x	**n 100**	**n 500**	**n 700**
82	5.405152	1.460255	1.549922
83	5.676639	1.156127	2.3045
84	6.16647	0.742473	1.496287
85	5.259206	0.650938	0.595108
86	5.079285	0.313768	0.322005
87	6.655946	0.173568	0.276654
88	3.164348	0.069115	0.159143
89	6.809434	0.046947	0.154918
90	2.501389		
max	6.809434	2.538895	2.834678
min	1.805634	0.046947	0.154918
range	5.0038	2.491948	2.679759
mean	3.285853	1.508907	1.416643
SD	1.225713	0.560484	0.67138

ภาคผนวก ข

ค่า Theta และ Expected raw scores จากโปรแกรม IRT-CLASS
ตามเงื่อนไขต่าง ๆ ของการปรับเทียบคะแนน

1. เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4-.6

Theta	Exp_Raw		
	n 100	**n 500**	**n 700**
-4	9.01157	12.30947	178.518
-3.43	13.41191	16.65519	17.20989
-2.86	19.59205	22.67769	22.65799
-2.29	28.05223	30.80905	30.09149
-1.71	39.31145	41.2961	40.39492
-1.14	53.03536	53.88991	121.4063
-0.571	69.1855	68.64341	69.2318
0	87.35539	85.07471	86.40728
0.571	106.8834	102.2998	104.1897
1.14	126.6002	119.4381	121.4063
1.71	145.3004	135.7811	137.0204
2.29	162.035	150.766	150.6143
2.86	175.5247	163.3491	161.7118
3.43	186.1508	173.7766	170.8837
4	194.3372	182.3638	178.518

2. เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม

Theta	Exp_Raw		
	n 100	**n 500**	**n 700**
-4	10.53784	14.43958	14.57915
-3.43	15.50213	19.55461	19.58425
-2.86	22.32882	26.40005	26.17627
-2.29	31.28969	35.15195	34.69792
-1.71	42.7945	46.04225	45.54097
-1.14	56.47326	58.84508	58.47096
-0.571	72.17896	73.59471	73.46682
0	89.48948	89.74272	89.82287
0.571	107.8247	106.26556	106.3938
1.14	126.3429	122.27643	122.31204
1.71	144.0728	137.37892	137.10265
2.29	160.1566	151.37776	150.42296
2.86	173.3939	163.40173	161.51299
3.43	184.0512	173.54329	170.71816
4	192.4172	181.9407	178.37217

3. เงื่อนไขข้อสอบร่วมมีความยากอยู่ในช่วง .4 -.6 และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

Theta	Exp_Raw		
	n 100	**n 500**	**n 700**
-4	7.45256	10.63291	11.73824
-3.43	11.00585	14.15124	15.34044
-2.86	16.12714	18.99942	20.21144
-2.29	23.29752	25.69652	26.82632
-1.71	33.25105	34.98706	35.94286
-1.14	46.09778	46.86426	47.63036
-0.571	61.70381	61.21287	61.73702
0	79.2068	77.33227	77.47336
0.571	97.60176	94.19405	93.81024
1.14	115.7656	110.8529	109.8027
1.71	132.7645	126.5489	124.623
2.29	147.8398	140.6257	137.6652
2.86	159.8183	152.0299	148.1552
3.43	168.9435	161.0272	156.5529
4	175.611	168.0023	163.2755

4. เงื่อนไขข้อสอบร่วมมีความยากอย่างสุ่ม และตัดข้อสอบที่ไม่มีคุณภาพทิ้ง

Theta	Exp_Raw		
	n 100	**n 500**	**n 700**
-4	8.17089	12.72662	13.20686
-3.43	12.1666	16.98825	17.39743
-2.86	17.92597	22.73398	23.03355
-2.29	25.93029	30.34167	30.50158
-1.71	36.7295	40.35342	40.37523
-1.14	49.9851	52.63222	52.62305
-0.571	65.47552	67.09832	67.22017
0	82.6484	83.04477	83.3039
0.571	100.7298	99.31729	99.58899
1.14	118.8134	114.929	115.1028
1.71	135.8726	129.3918	129.2598
2.29	150.8884	142.4326	141.6295
2.86	162.6134	153.2317	151.5382
3.43	171.41	161.9766	159.4659
4	177.796	168.9259	165.8411

ประวัติย่อของผู้วิจัย

ชื่อ สกุล	นางสาวศศิธร ชุตินันทกุล
วัน เดือน ปี เกิด	2 เมษายน พ.ศ. 2505
สถานที่เกิด	อำเภอปากพนัง จังหวัดนครศรีธรรมราช
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 100/ 1436 ถนนติวานนท์
	อำเภอปากเกร็ด จังหวัดนนทบุรี 11120
ตำแหน่งและประวัติการทำงาน	
พ.ศ. 2530-2537	เจ้าหน้าที่วิเคราะห์นโยบายและแผน สำนักงาน
	คณะกรรมการการศึกษาเอกชน กระทรวงศึกษาธิการ
พ.ศ. 2537-2544	เจ้าหน้าที่วิเคราะห์นโยบายและแผน
	มหาวิทยาลัยสุโขทัยธรรมาธิราช
พ.ศ. 2544-ปัจจุบัน	อาจารย์ประจำสำนักทะเบียนและวัดผล
	มหาวิทยาลัยสุโขทัยธรรมาธิราช
	- รักษาการในตำแหน่งหัวหน้าศูนย์วิจัยและพัฒนา
	แบบทดสอบ
	- ผู้ทรงคุณวุฒิในการจัดทำข้อสอบกลาง
	มหาวิทยาลัยมหาจุฬาลงกรณราชวิทยาลัย
ประวัติการศึกษา	
พ.ศ. 2528	วิทยาศาสตรบัณฑิต (คณิตศาสตร์)
	มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี
พ.ศ. 2536	การศึกษามหาบัณฑิต (การวัดผลการศึกษา)
	มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร
พ.ศ. 2560	ปรัชญาดุษฎีบัณฑิต (วิจัย วัดผลและสถิติการศึกษา)
	มหาวิทยาลัยบูรพา