

การเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับขั้นที่มีตัวแปรและลำดับในการพิจารณาที่แตกต่างกัน

พัชรญา ตันติพงศ์อนันต์^{1*} วุฒิพงษ์ ชินศรี²

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อการเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับขั้นที่มีตัวแปรและลำดับในการพิจารณาที่แตกต่างกัน โดยนำเสนอวิธีการทดสอบแบบปรับเหมาะตามตัวแปรต่าง ๆ 4 รูปแบบ ดังนี้ รูปแบบที่ 1 พิจารณาค่าความยากง่าย (P) เพียงค่าเดียว รูปแบบที่ 2 พิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว รูปแบบที่ 3 พิจารณาค่าตัวแปรที่มีการถ่วงน้ำหนัก (T) ซึ่งสามารถกระจายออกเป็น 3 รูปแบบย่อย ได้แก่ ถ่วงน้ำหนักความยากง่ายสำคัญสุด (Tp) ถ่วงน้ำหนักอำนาจจำแนกสำคัญสุด (Tr) ถ่วงน้ำหนักความยากง่ายและอำนาจจำแนกสำคัญเสมอกัน (Tpr) รูปแบบที่ 4 พิจารณาค่าถ่วงน้ำหนักรวม (S) สามารถวิเคราะห์ได้โดยการเพิ่มตัวชี้วัดเข้ามาช่วยถ่วงน้ำหนักให้กับตัวแปรในแต่ละตัว เรียงลำดับความสำคัญ ได้แก่ ค่าอำนาจจำแนก (R) ค่าความยากง่าย (P) ค่าโอกาสตอบถูกในกลุ่มนักเรียนที่ได้คะแนนต่ำ (K_L) นำทั้ง 4 รูปแบบมาเปรียบเทียบเพื่อหาว่าวิธีการในรูปแบบใดที่มีประสิทธิภาพมากที่สุด โดยอาศัยทฤษฎีการทดสอบแบบมาตรฐานเดิม ในการวิเคราะห์หาค่าตัวแปรต่าง ๆ เพื่อนำมาใช้งานวิจัยฉบับนี้

ผลการวิจัยพบว่า 1) รูปแบบที่ 4 การพิจารณาค่าถ่วงน้ำหนักรวม (S) สามารถประเมินผลโดยการเทียบเกรดกับระดับขั้นได้ตรงกันได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 70.59 2) รูปแบบที่ 1 การพิจารณาค่าความยากง่าย (P) เพียงค่าเดียว สามารถประเมินผลผ่านหรือไม่ผ่านโดยเทียบผลการประเมินเกรด B ขึ้นไปกับการประเมินระดับ 4 ขึ้นไปได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 94.12 3) รูปแบบที่ 3 โดยใช้รูปแบบย่อยพิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเสมอกัน (Tpr) สามารถประเมินผลผ่านหรือไม่ผ่านเทียบกับร้อยละที่ทำข้อสอบถูกได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 88.82 และสามารถประเมินผลผ่านหรือไม่ผ่านทั้ง 2 แนวทาง ได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 89.51 4) รูปแบบที่พิจารณาค่าตัวแปรที่มีการถ่วงน้ำหนัก 2 ตัวแปรขึ้นไปผู้สอบจะต้องทำข้อสอบจำนวนมากกว่ารูปแบบที่พิจารณาค่าตัวแปรเพียงค่าเดียว โดยรูปแบบที่ 2 การพิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว ผู้สอบจะทำข้อสอบโดยเฉลี่ยน้อยที่สุด

คำสำคัญ : การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ การวิเคราะห์ข้อสอบ รูปแบบปรับระดับขั้น

¹ หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการจัดการเทคโนโลยีสารสนเทศ วิทยาลัยเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยรังสิต จังหวัดปทุมธานี e-mail: appletanti@gmail.com

² หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการจัดการเทคโนโลยีสารสนเทศ วิทยาลัยเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยรังสิต จังหวัดปทุมธานี e-mail: wutthipong.c@rsu.ac.th

* ผู้นิพนธ์หลัก e-mail: appletanti@gmail.com

A COMPARISON OF COMPUTERIZED ADAPTIVE TESTING BY USING MULTISTAGE STRATEGIES UNDER DIFFERENT PRIORITIZATION OF CONDITIONS AND VARIABLES

Pacharaya Tantipongnant^{1*} Wutthipong Chinnasri²

Abstract

This research is aimed to compare the computerized adaptive testing by using multistage strategies with different variables and prioritization. Four different types of the adaptive testing were studied as follows: Type 1: the determination of only one difficulty index (P); Type 2: the determination of only one discriminant index (R); Type 3: the determination of weighted index (T), which could be disseminated into three sub-categories including most-important weighted difficulty index (Tp), most-important weighted discriminant index (Tr), and equally-important weighted difficulty and discriminant index (Tpr); and Type 4: the determination of good quality test index (S). The fourth index could be analyzed by the addition of indicators to help weighting each variable. The prioritized indices were discriminant index (R), difficulty index (P), and chance of correct answer index in low-scored student (KL), respectively. The four types of adaptive testing were compared to identify the most efficient type of method for testing. The Classical Test Theory was employed to analyze various variables used in this research.

The results showed that 1) the assessment by Type 4: the determination of good quality test index (S) could be able to assess by comparing the grading with the study level at mostly agreed, in which the average percentage of 70.59 was obtained; 2) the assessment by Type 1: the determination of only one difficulty index (P) could be able to assess the results from passing or failing the test by comparing the assessment of higher than grade B with the assessment at level 4 at mostly agreed, in which the average percentage of 94.12 was obtained; 3) the assessment by Type 3 by using the sub-category equally-important weighted difficulty and discriminant index (Tpr) could be able to assess the results from passing or failing the test relative to the percentage of marking correct answers at mostly agreed, in which the average percentage of 88.82 was obtained. In addition, it could be able to assess the results from passing or failing the test from both approaches at mostly agreed, in which the average percentage of 89.51 was obtained; and 4) in the type in which at least 2 weighted difficulty indices were determined, the testers must complete more number of tests than the type in which only one index was determined. It was found that in Type 2: the determination of only one discriminant index (R), the testers would complete the least mean number of tests.

Keywords : Computerized Adaptive Testing, Item Analysis, Stradaptive Model

¹ Master of Science in Information Technology Management, College of Information and Communication Technology, Rangsit University, e-mail: appletanti@gmail.com

² Master of Science in Information Technology Management, College of Information and Communication Technology, Rangsit University, e-mail: wutthipong.c@rsu.ac.th

* Corresponding author, email: appletanti@gmail.com

บทนำ

การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) เริ่มเป็นที่รู้จักแพร่หลายและได้รับการยอมรับกันมากขึ้น ดังจะเห็นได้ว่าตามสนามสอบสำคัญระดับประเทศหรือระดับโลก มีการนำระบบทดสอบนี้มาใช้ในการทดสอบเพื่อวัดความรู้ความสามารถกันมากขึ้น เช่น การสอบวัดผลทางด้านวิชาการสำหรับการศึกษาในระดับบัณฑิตศึกษาศาखाวิชาธุรกิจและการบริหารที่มีมาตรฐานและใช้ได้ทั่วโลก เรียกว่า สอบ GMAT หรือ Graduate Management Admission Test การทดสอบเพื่อประเมินความสามารถทางด้านภาษาอังกฤษของผู้ที่ไม่ได้ใช้ภาษาอังกฤษเป็นภาษาประจำชาติ เรียกว่า สอบ TOEFL หรือ The Test of English as a Foreign Language เป็นต้น (เบญจมาภรณ์ เสนารัตน์, สมประสงค์ เสนารัตน์, และสมบัติ ท้ายเรือคำ, 2557) สำหรับประเทศไทยเริ่มมีการนำคอมพิวเตอร์มาช่วยในการทดสอบวัดผลมากขึ้น เนื่องจากมีความสะดวก รวดเร็ว แม่นยำ และตรวจสอบได้ง่าย เหมาะสำหรับจะนำมาใช้ทดแทนการสอบโดยใช้กระดาษ (Paper and Pencil Test) โดยเฉพาะการสอบเพื่อวัดผลระหว่างเรียนและการสอบซ่อมเสริมรายวิชา (พิมพ์สิริ เจริญรเศรษฐ์, 2549)

การทดสอบแบบปรับเหมาะตามความสามารถของผู้เรียนนั้น ข้อสอบแต่ละข้อจะมีการปรับเปลี่ยนไปตามการทำข้อสอบของผู้เรียน การตอบถูกหรือตอบผิดจะส่งผลต่อการเริ่มทำแบบทดสอบข้อถัดไป โดยหลักการที่นิยมของการทดสอบแบบปรับเหมาะคือ ถ้าผู้สอบตอบข้อคำถามในแบบทดสอบข้อที่ผ่านมาถูกต้องข้อคำถามข้อถัดไปจะยากขึ้น ถ้าผู้สอบตอบข้อคำถามในแบบทดสอบข้อที่ผ่านมาผิดข้อคำถามข้อถัดไปจะง่ายขึ้น และจะปฏิบัติเช่นนี้ไปเรื่อย ๆ จนกระทั่งสิ้นสุดการสอบตามเกณฑ์ที่กำหนด ส่งผลให้การทำแบบทดสอบของผู้สอบแต่ละคนเป็นอิสระจากกัน ผู้สอบที่มีความสามารถสูงจึงไม่จำเป็นต้องเสียเวลาทำข้อสอบที่ง่ายเกินไปและสำหรับผู้สอบที่มีความสามารถต่ำก็ไม่จำเป็นต้องทำข้อสอบที่ยากเกินความสามารถของตนเอง การทดสอบแบบปรับเหมาะตามความสามารถจึงช่วยลดจำนวนข้อในการทำข้อสอบและลดระยะเวลาในการทำข้อสอบได้ (อุทัยรัตน์ รื่นจิตร, 2558) เหมือนกับการทดสอบตามแบบประเพณีนิยม (Conventional Testing) การทดสอบที่ดีและมีประสิทธิภาพจึงไม่ควรนำข้อสอบที่ง่ายหรือยากเกินไปสำหรับผู้สอบมาให้ทดสอบ (รังสรรค์ มณีเล็ก, 2540) หรือควรมีจำนวนข้อสอบดังกล่าวในปริมาณที่เหมาะสม ไม่มากจนเกินไป

กลยุทธ์ในการทดสอบแบบปรับเหมาะตามความสามารถ (Adaptive Testing) ออกเป็นสองกลยุทธ์หลัก (Weiss, 1974; Hambleton and Swaminathan, 1985) ดังนี้ แบบที่หนึ่ง กลยุทธ์สองขั้นตอน (Two-Stage Strategies) ทำแบบทดสอบแบบแยกทิศทาง (Routing Test) ให้ครบทุกข้อ แล้วนำคะแนนสอบมาประมาณค่าความสามารถเพื่อเลือกข้อสอบเริ่มต้นจากแบบทดสอบเพื่อประเมินผล (Measurement Test) ผู้ที่ทำแบบทดสอบในขั้นแรกแล้วได้คะแนนต่ำจะได้รับแบบทดสอบชุดที่ง่าย ส่วนผู้ที่ทำคะแนนได้สูงจะได้รับแบบทดสอบชุดที่ยาก (ญานิศรา มุนินทร์สาคร, 2558) สำหรับแบบที่สอง คือแบบกลยุทธ์หลายขั้นตอน (Multi-Stage Strategies) รูปแบบแยกทางคงที่ (Fixed Branching Mode) และรูปแบบแยกทางแปรผัน (Variable Branching Model) แต่ละขั้นตอนอาจจะมีข้อเดียวหรือหลายข้อก็ได้ โดยจะมีรูปแบบหรือเส้นทางในการทดสอบที่ชัดเจนคงที่แน่นอน

กลยุทธ์หลายขั้นตอนรูปแบบแยกทางคงที่แบ่งย่อยเป็น 3 รูปแบบ ดังนี้ 1) รูปแบบพีระมิด (Pyramid Model) จะจัดการทดสอบในรูปแบบของพีระมิดมีเงื่อนไขในการกำหนดจำนวนข้อของแต่ละลำดับขั้นที่ตายตัวและมีการกำหนดช่วงระดับความยากของข้อสอบที่ติดกันในแต่ละขั้นให้มีค่าเท่ากันหรือมีค่าไม่แน่นอน จะขึ้นอยู่กับรูปแบบของพีระมิดที่นำมาใช้ทดสอบ 2) รูปแบบยืดหยุ่น (Flexi-level Model) เป็นการทดสอบโดยให้ข้อสอบแต่ละข้อมีช่วงห่างของค่าความยากของข้อสอบที่เท่ากัน เรียงลำดับจากง่ายไปหายาก กำหนดให้ผู้สอบ

เริ่มต้นข้อแรกที่ข้อสอบมีความยากในระดับปานกลางหรือตำแหน่งมัธยฐาน ซึ่งการให้คะแนนสอบของการทดสอบรูปแบบนี้ Lord (1971) เสนอว่า หากผู้สอบตอบคำถามข้อสุดท้ายถูก คะแนนจะเท่ากับจำนวนข้อสอบที่ตอบถูกทั้งหมด หากตอบข้อคำถามข้อสุดท้ายผิด คะแนนจะเท่ากับจำนวนข้อสอบที่ทำทั้งหมดบวกอีกครั้ง คะแนน 3) รูปแบบปรับระดับขั้น (Stradaptive Model) เป็นการทดสอบที่พัฒนารูปแบบมาจากการทดสอบแบบยืดหยุ่น (Flexilevel Model) ซึ่งการทดสอบรูปแบบนี้จะมีข้อสอบในแต่ละระดับความยาก มากกว่า 1 ข้อ แบ่งเป็นระดับขั้น แต่ละขั้นจะบรรจุกลุ่มข้อสอบที่มีความยากง่ายแตกต่างกันไปในแต่ละขั้น (ญาณิศรา มุนินทร์สาร, 2558)

กลยุทธ์หลายขั้นตอนรูปแบบแยกทางแปรผัน (Variable Branching Model) เป็นรูปแบบการทดสอบที่ไม่ได้กำหนดเงื่อนไขการทดสอบที่แน่นอน แต่จะให้ความสำคัญกับการประมาณค่าความสามารถของผู้สอบก่อน แล้วเลือกข้อสอบข้อถัดไปที่เหมาะสมกับระดับความสามารถของผู้เข้าสอบ ส่วนมากจะใช้คอมพิวเตอร์เป็นผู้เลือกข้อสอบจากกลุ่มผู้สอบ (Item Pool) โดยให้มีค่าความยากของข้อสอบใกล้เคียงกับค่าความสามารถของผู้สอบหลังจากตอบข้อสอบไปแล้วในแต่ละข้อ ซึ่งการประมาณค่าความสามารถของผู้เข้าสอบที่ได้รับความนิยมจะมี 2 วิธี วิธีที่หนึ่ง การประมาณค่าแบบเบส์ (Bayesian Estimation) ได้มาจากทฤษฎีของเบส์ วิธีที่สอง การประมาณค่าความความเป็นไปได้สูงสุด (Maximum Likelihood Estimation) (นุภาพรณ ปลื้มใจ, 2558)

การทดสอบแบบปรับเหมาะนั้น จำเป็นต้องสร้างคลังข้อสอบขึ้นมาก่อน ซึ่งคลังข้อสอบนั้นจะมีข้อสอบที่ผ่านการวิเคราะห์ข้อสอบมาแล้ว โดยทฤษฎีที่นิยมนำมาใช้เพื่อวิเคราะห์คุณภาพข้อสอบมี 2 ทฤษฎีที่ได้รับความนิยมอย่างแพร่หลาย (David J. Weiss, 2004) ได้แก่ ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) อธิบายความสัมพันธ์ในรูปของฟังก์ชันทางคณิตศาสตร์ที่เรียกว่า ฟังก์ชันการตอบข้อสอบ (Item Response Function: IRF) เป็นโมเดลที่แสดงความสัมพันธ์ระหว่างระดับความสามารถ คุณลักษณะของข้อสอบ และสัมประสิทธิ์การเดาหรือโอกาสในการตอบข้อสอบถูก รวมเรียกว่า พารามิเตอร์ของข้อสอบ (Item Parameter) แล้วแสดงเป็นกราฟ เรียกว่า โค้งลักษณะข้อสอบ (Item Characteristic Curve: ICC) กราฟจะมี 2 ลักษณะ ดังนี้ 1) ฟังก์ชันปกติสะสม (Normal Ogive Function) ซึ่งมีความซับซ้อนมาก 2) ฟังก์ชันโลจิสติก (Logistic Function) ได้รับความนิยมมากเพราะคำนวณได้ง่ายกว่า ซึ่งทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) จะมุ่งเน้นความเป็นเอกมิติหรือมิติเดียวของความรู้วัดคุณลักษณะแฝงหรือองค์ประกอบหลักเพียงลักษณะเดียว วิเคราะห์โดยคำนวณค่าไอเกน (Eigen Value) การวิเคราะห์ข้อสอบทั้งฉบับจะวิเคราะห์หาค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function) และความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถ หากมีค่าต่ำแสดงว่ามีความแม่นยำสูง (นันทพร หาญวิทย์สกุล, 2552) ส่วนอีกทฤษฎีที่ได้รับความนิยมมานานและยังคงใช้อยู่อย่างต่อเนื่องตามสถานศึกษาโดยทั่วไป เรียกว่า ทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical Test Theory: CTT) เป็นการวิเคราะห์ข้อสอบเป็นรายข้อจะวิเคราะห์หาค่าความยาก (Item Difficulty) ค่าอำนาจจำแนก (Discrimination Power) และประสิทธิภาพของตัวลวง (Effective of Distracters) ส่วนการวิเคราะห์คุณภาพข้อสอบทั้งฉบับจะวิเคราะห์ค่าความเที่ยงตรง (Validity) ค่าความเที่ยงหรือค่าความเชื่อมั่น (Reliability) ของแบบทดสอบ

ในงานวิจัยนี้ผู้วิจัยทำการพัฒนาระบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับขั้น โดยแบ่งขั้นทั้งหมด 5 ขั้น โดยใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิมในการวิเคราะห์ข้อสอบเพื่อสร้างคลังข้อสอบ และกำหนดตัวแปรและลำดับในการพิจารณาที่แตกต่างกัน 4 รูปแบบ โดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในรูปแบบต่าง ๆ

ซึ่งจะช่วยให้ทราบว่ารูปแบบใดเหมาะสมที่จะนำมาใช้ในการประเมินผล ในกรณีที่จะนำมาใช้แทนการทดสอบตามแบบประเพณีนิยม

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับขั้น ที่มีตัวแปรและลำดับในการพิจารณาที่แตกต่างกัน

วิธีดำเนินการวิจัย

1. ขอบเขตของการวิจัย

1.1 การวิจัยนี้ศึกษาโดยใช้วิธีการจำลองสถานการณ์จากข้อมูลจริง (Post-Hoc Simulations) อาศัยแนวทางการทดสอบตามหลักการทฤษฎีการทดสอบแบบมาตรฐานเดิม ผู้วิจัยใช้รูปแบบการทดสอบแบบปรับเหมาะตามตัวแปรและลำดับในการพิจารณาที่แตกต่างกัน 4 รูปแบบ ดังนี้ รูปแบบที่ 1 พิจารณาค่าความยากง่าย (P) รูปแบบที่ 2 พิจารณาค่าอำนาจจำแนก (R) รูปแบบที่ 3 พิจารณาค่าตัวแปรที่มีการถ่วงน้ำหนัก 2 ตัวแปร (T) ซึ่งกระจายออกเป็นอีก 3 รูปแบบย่อย ได้แก่ พิจารณาค่าความยากง่ายสำคัญสุด (Tp) พิจารณาค่าอำนาจจำแนกสำคัญสุด (Tr) พิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเสมอกัน (Tpr) รูปแบบที่ 4 พิจารณาค่าถ่วงน้ำหนักกรรม (S) สามารถวิเคราะห์ได้โดยการเพิ่มตัวชี้วัดเข้ามาช่วยถ่วงน้ำหนักให้กับตัวแปรสำคัญในแต่ละตัวเรียงตามลำดับความสำคัญ ใช้ 3 ตัวแปรในการพิสูจน์ ได้แก่ ค่าอำนาจจำแนก (R) ค่าความยากง่าย (P) ค่าโอกาสตอบถูกในกลุ่มนักเรียนที่ได้คะแนนต่ำ (K_L)

1.2 คลังข้อสอบ ผู้วิจัยใช้ชุดแนวข้อสอบวิชาคณิตศาสตร์ (PAT 1) ระดับชั้นมัธยมศึกษาตอนปลาย ลักษณะแบบทดสอบเป็นข้อสอบแบบเลือกตอบ (ปรนัย) 4 ตัวเลือก (R-TOP TEST) จำนวน 198 ข้อ โดยมีกลุ่มตัวอย่างที่ใช้ในการศึกษาครั้งนี้ คือ นักเรียนระดับชั้นมัธยมศึกษาตอนปลาย จำนวน 51 คน ที่กำลังศึกษาวิชาคณิตศาสตร์ (PAT 1) แคมป์วิชาการเดือนตุลาคม ปีการศึกษา 2559 ที่สถาบันกวดวิชาอาร์ทีโอคาเดเทททดลองทำข้อสอบด้วยการทดสอบตามแบบประเพณีนิยม เมื่อวันที่ 8 ตุลาคม 2559 แล้วนำผลการทดสอบมาทำการวิเคราะห์คุณภาพและคัดเลือกข้อสอบที่มีคุณภาพตามเกณฑ์ของทฤษฎีการทดสอบแบบมาตรฐานเดิม โดยกำหนดค่าอำนาจจำแนกอยู่ระหว่าง 0.20-1.00 ค่าความยากง่ายอยู่ระหว่าง 0.00-1.00 มาจัดเก็บในคลังข้อสอบ (R-TOP BANK) และนำมาแบ่งออกเป็น 2 คลังย่อย ดังนี้

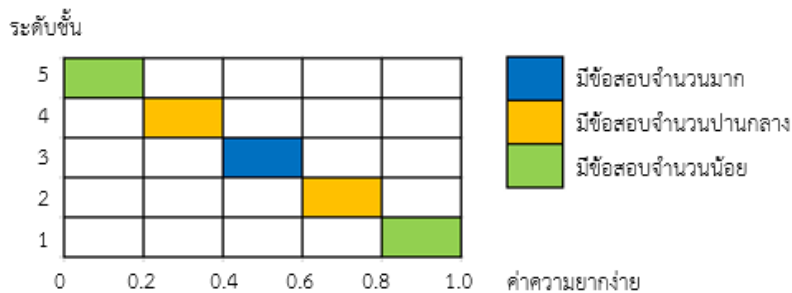
1.2.1. คลังข้อสอบ A จะมีข้อสอบทั้งหมด 93 ข้อ จัดแบ่งตามระดับ 5 ชั้น (Level) ชั้นที่ 1 ข้อสอบจำนวน 13 ข้อ ชั้นที่ 2 ข้อสอบจำนวน 15 ข้อ ชั้นที่ 3 ข้อสอบจำนวน 31 ข้อ ชั้นที่ 4 ข้อสอบจำนวน 23 ข้อ และชั้นที่ 5 ข้อสอบจำนวน 11 ข้อ

1.2.2. คลังข้อสอบ B จะมีข้อสอบทั้งหมด 107 ข้อ จัดแบ่งตามระดับ 5 ชั้น (Level) ชั้นที่ 1 ข้อสอบจำนวน 11 ข้อ ชั้นที่ 2 ข้อสอบจำนวน 15 ข้อ ชั้นที่ 3 ข้อสอบจำนวน 39 ข้อ ชั้นที่ 4 ข้อสอบจำนวน 29 ข้อ และชั้นที่ 5 ข้อสอบจำนวน 13 ข้อ

2. ขั้นตอนการวิจัย

ผู้วิจัยได้ทำการพัฒนาระบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับขั้น โดยแบ่งขั้นทั้งหมด 5 ขั้น ข้อสอบข้อแรกจะขึ้นอยู่กับการสุ่มแบบสุ่มและตัวแปรที่นำมาพิจารณา ซึ่งจะเริ่มในระดับขั้นที่มีค่าความยากปานกลาง (Level 3) เมื่อผู้สอบตอบคำถามถูกหรือผิดจะส่งผลต่อการพิจารณาเลื่อนข้อสอบข้อถัดไป โดยจะทำแบบนี้ไปเรื่อย ๆ จนทำข้อสอบถูกหรือผิดครบจำนวนข้อที่

กำหนดไว้ในแต่ละชั้น ซึ่งคำนวณได้จากสูตร $(n-1)/2$ โดย n หมายถึงจำนวนข้อทั้งหมดในชั้นนั้นและค่า n ต้องเป็นเลขคี่เท่านั้น ถ้าทำถูกครบจำนวนจะเลื่อนระดับขึ้น 1 ระดับ ถ้าทำผิดครบจำนวนจะลดระดับลง 1 ระดับ อีกกรณีหนึ่งหากทำข้อสอบครบทุกข้อของระดับชั้นนั้นแล้วให้พิจารณาว่าผู้สอบตอบถูกหรือผิดมากกว่ากัน โดยหากตอบถูกมากกว่าก็จะเลื่อนระดับขึ้น 1 ระดับ และหากตอบผิดมากกว่าก็จะลดระดับลง 1 ระดับ



ภาพที่ 1 การจัดเก็บข้อสอบในชั้นต่าง ๆ ตามกลยุทธ์หลายขั้นตอนแบบปรับระดับชั้น 5 ชั้น

จากภาพที่ 1 ผู้วิจัยได้แบ่งข้อสอบในคลังข้อสอบออกเป็น 5 ชั้น ตามความยากง่ายของข้อสอบนั้น ชั้นแรก (Level 1) บรรจุข้อสอบระดับง่ายมาก มีค่าความยากง่ายตั้งแต่ 0.81-1.00 ชั้นที่สอง (Level 2) บรรจุข้อสอบระดับค่อนข้างง่าย มีค่าความยากง่ายตั้งแต่ 0.61-0.80 ชั้นที่สาม (Level 3) บรรจุข้อสอบระดับปานกลาง (ยากง่ายพอเหมาะ) มีค่าความยากง่ายตั้งแต่ 0.41-0.60 ชั้นที่สี่ (Level 4) บรรจุข้อสอบระดับค่อนข้างยาก มีค่าความยากง่ายตั้งแต่ 0.20-0.40 ชั้นที่ห้า (Level 5) บรรจุข้อสอบระดับยากมาก มีค่าความยากง่ายตั้งแต่ 0.00-0.19 และแต่ละระดับชั้นจะมีค่าอำนาจจำแนกของข้อสอบคละกันไปตั้งแต่ 0.20-1.00

การพิจารณาเลือกข้อสอบข้อแรกและข้อถัดไปในแต่ละรูปแบบมีดังนี้

1) รูปแบบที่ 1 พิจารณาค่าความยากง่าย (P) เพียงค่าเดียว จะกำหนดให้ผู้สอบเริ่มต้นข้อแรกที่ข้อสอบที่มีความยากง่ายในตำแหน่งมัธยฐานของชั้นที่ 3 ถ้าผู้สอบตอบถูก ข้อถัดไปจะมีความยากมากขึ้น ถ้าผู้สอบตอบผิด ข้อถัดไปจะมีความยากน้อยลง ทำเช่นนี้ไปจนกระทั่งจำนวนข้อถูกหรือผิดครบตามจำนวนข้อที่กำหนด จึงทำการปรับเลื่อนชั้นขึ้นหรือลงตามเงื่อนไขการปรับเหมาะที่กำหนด โดยจุดเริ่มต้นข้อแรกของชั้นใหม่จะเริ่มที่ข้อสอบที่มีความยากง่ายในตำแหน่งมัธยฐานของชั้นนั้นเสมอ

2) รูปแบบที่ 2 พิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว จะกำหนดให้ผู้สอบเริ่มต้นข้อแรกโดยเลือกข้อสอบที่มีค่าอำนาจจำแนกสูงสุดของชั้นที่ 3 ซึ่งไม่ว่าผู้สอบจะตอบถูกหรือผิดข้อถัดไปจะต้องเริ่มต้นที่ข้อสอบที่มีค่าอำนาจจำแนกสูงสุดของข้อสอบที่เหลืออยู่ในชั้นนั้นเสมอ ทำเช่นนี้ไปจนกระทั่งจำนวนข้อถูกหรือผิดครบตามจำนวนข้อที่กำหนด จึงทำการปรับเลื่อนชั้นขึ้นหรือลงตามเงื่อนไขการปรับเหมาะที่กำหนด โดยจุดเริ่มต้นของข้อสอบแรกของชั้นใหม่จะเริ่มที่ข้อสอบที่มีค่าอำนาจจำแนกสูงสุดในชั้นนั้นเสมอ

3) รูปแบบที่ 3 พิจารณาค่าตัวแปรที่มีการถ่วงน้ำหนัก 2 ตัวแปร (T) ซึ่งกระจายออกเป็นอีก 3 รูปแบบย่อย ได้แก่ พิจารณาค่าความยากง่ายสำคัญสุด (Tp) พิจารณาค่าอำนาจจำแนกสำคัญสุด (Tr) พิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเสมอ (Tpr) การพิจารณาค่าตัวแปรที่มีการถ่วงน้ำหนักสองตัวแปรนั้นได้มาจากการเพิ่มตัวคูณเข้ามาช่วยถ่วงน้ำหนักให้กับตัวแปรที่ต้องการให้ความสำคัญ ตามสูตรด้านล่างนี้

$$Tp = \frac{(Rx1.5)+(Px2)}{2+1.5} \text{ หรือ } Tr = \frac{(Rx2)+(Px1.5)}{2+1.5} \text{ หรือ } Tpr = \frac{(Rx2)+(Px2)}{2+2}$$

โดยในทุกรูปแบบย่อย เมื่อคำนวณแล้วจะได้ค่า ๆ หนึ่งคือ ค่า T ซึ่งจะนำมาใช้ในการพิจารณาในการเลือกข้อถัดไป โดยในรูปแบบนี้จะกำหนดให้ผู้สอบเริ่มต้นข้อแรกที่ข้อสอบมีค่า T ในตำแหน่งมัธยฐานของชั้นที่ 3 ถ้าผู้สอบตอบถูก ข้อถัดไปจะมีค่า T มากขึ้น ถ้าผู้สอบตอบผิด ข้อถัดไปจะมีค่า T น้อยลง ทำเช่นนี้ไปจนกระทั่งจำนวนข้อถูกหรือผิดครบตามจำนวนข้อที่กำหนด จึงทำการปรับเลื่อนชั้นขึ้นหรือลงตามเงื่อนไขการปรับเหมาะที่กำหนด โดยจุดเริ่มต้นข้อแรกของชั้นใหม่จะเริ่มที่ข้อสอบที่มีค่า T ในตำแหน่งมัธยฐานของชั้นนั้นเสมอ

4) รูปแบบที่ 4 พิจารณาค่าถ่วงน้ำหนักรวม (S) สามารถวิเคราะห์ได้โดยใช้ตัวแปรทั้งหมด 3 ตัว ได้แก่ ค่าอำนาจจำแนก (R) ค่าความยากง่าย (P) ค่าโอกาสตอบถูกในกลุ่มนักเรียนที่ได้คะแนนต่ำ (K_L) มาใช้ในการคำนวณหาถ่วงน้ำหนักรวม (S) โดย ค่า K_L ได้มาจากการวิเคราะห์ผู้สอบกลุ่มคะแนนรวมต่ำ (กลุ่มอ่อน) หากกลุ่มนี้ตอบข้อสอบข้อนั้นถูกมากกว่าผิดค่าก็จะสูง หากตอบผิดมากกว่าถูกค่าจะต่ำ ซึ่งข้อสอบที่ดีควรจะมีความ K_L น้อยกว่า 0.5 โดยค่า K_L คำนวณได้ตามสูตรด้านล่างนี้

$$K_L = \frac{L}{H+L}$$

โดยค่า K_L คือ ค่าของโอกาสในการตอบถูกของผู้สอบกลุ่มคะแนนรวมต่ำ (กลุ่มอ่อน) ค่า L คือ จำนวนผู้สอบในกลุ่มที่ได้คะแนนรวมต่ำที่ตอบตัวเลือกนั้น และค่า H คือ จำนวนผู้สอบในกลุ่มที่ได้คะแนนรวมสูงที่ตอบตัวเลือกนั้น โดยค่า S นั้นจะคำนวณได้จากสูตรด้านล่างนี้

$$S = \frac{(Rx2) + (Px2) + (K_Lx1)}{2 + 2 + 1}$$

ค่า S จะนำมาใช้ในการพิจารณาในการเลือกข้อถัดไป โดยในรูปแบบนี้จะกำหนดให้ผู้สอบเริ่มต้นข้อแรกที่ข้อสอบมีค่า S ในตำแหน่งมัธยฐานของชั้นที่ 3 ถ้าผู้สอบตอบถูก ข้อถัดไปจะมีค่า S มากขึ้น ถ้าผู้สอบตอบผิด ข้อถัดไปจะมีค่า S น้อยลง ทำเช่นนี้ไปจนกระทั่งจำนวนข้อถูกหรือผิดครบตามจำนวนข้อที่กำหนด จึงทำการปรับเลื่อนชั้นขึ้นหรือลงตามเงื่อนไขการปรับเหมาะที่กำหนด โดยจุดเริ่มต้นข้อแรกของชั้นใหม่จะเริ่มที่ข้อสอบที่มีค่า S ในตำแหน่งมัธยฐานของชั้นนั้นเสมอ

สำหรับเงื่อนไขการหยุดทำข้อสอบและการประเมินผลจะพิจารณานั้น ระบบจะจบการทดสอบในกรณีต่าง ๆ ดังนี้

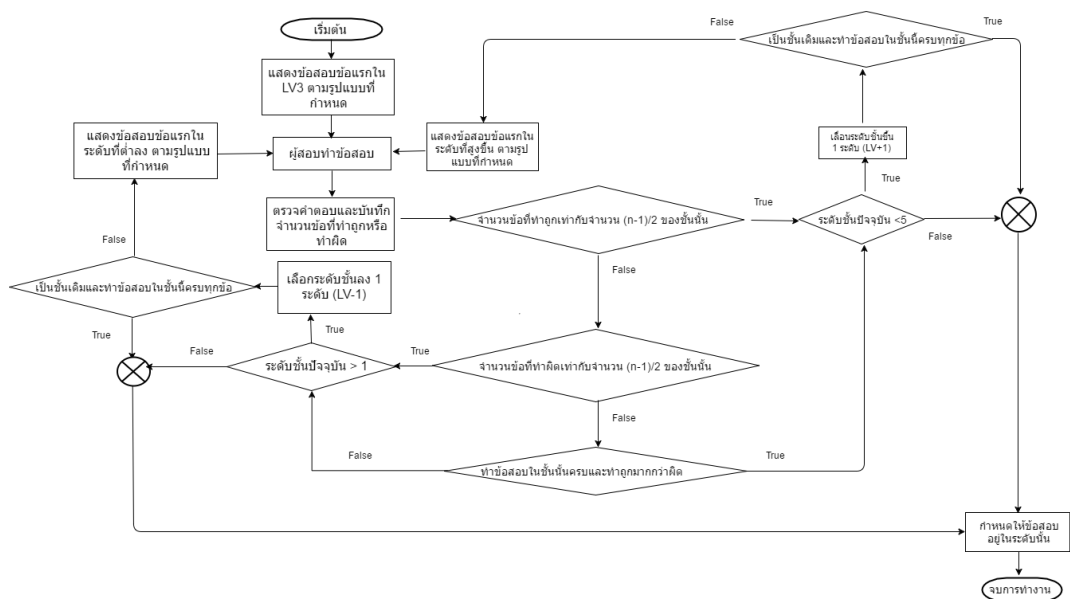
1) ผู้สอบสามารถทำข้อสอบในระดับชั้นที่ 5 ได้ถูกต้องครบตามจำนวนที่คำนวณไว้ในชั้นนั้นการทดสอบจะจบลงและประเมินผู้สอบอยู่ในระดับ 5

2) ผู้สอบทำข้อสอบในระดับชั้นที่ 1 ผิดครบตามจำนวนที่คำนวณไว้ในชั้นนั้น การทดสอบจะจบลงและประเมินผู้สอบอยู่ในระดับ 1

3) ในกรณีถ้าผู้สอบถูกปรับเลื่อนระดับขึ้นหรือลงมาซ้ำในระดับชั้นเดิม และได้ทำข้อสอบในชั้นนั้นครบทุกข้อแล้วให้สรุปผลว่าอยู่ที่ระดับชั้นนั้น ตัวอย่างเช่น

3.1) เริ่มทำข้อสอบในระดับชั้นที่ 3 โดยผู้สอบตอบถูกครบตามจำนวนที่กำหนดไว้ จะได้รับการปรับระดับขึ้นไปชั้นที่ 4 สมมติว่าชั้นที่ 4 ทำข้อสอบจนครบแล้วตอบผิดมากกว่าตอบถูกจึงปรับระดับกลับมาระดับชั้นที่ 3 กรณีนี้ยังทำข้อสอบไม่ครบในชั้นที่ 3 ให้ทำต่อให้ครบ แล้วตรวจสอบว่าทำผิดมากกว่าถูกหรือถูกมากกว่าผิด กรณีที่ผิดมากกว่าถูกจะเลื่อนระดับลงไประดับที่ 2 และทำข้อสอบต่อไป แต่หากทำถูกมากกว่าผิดจะเลื่อนขึ้นไประดับที่ 4 ซึ่งทำข้อสอบครบแล้ว จะจบการทดสอบและประเมินว่าผู้สอบอยู่ในระดับที่ 4

3.2) เริ่มทำข้อสอบในระดับชั้นที่ 3 โดยผู้สอบทำข้อสอบครบทุกข้อในชั้นที่ 3 แล้วตรวจสอบว่าทำผิดมากกว่าถูกหรือถูกมากกว่าผิด กรณีที่ผิดมากกว่าถูกจะเลื่อนระดับลงไประดับที่ 2 และทำข้อสอบต่อไป แต่หากทำถูกมากกว่าผิดจะเลื่อนขึ้นไประดับที่ 4 และทำข้อสอบต่อไป สมมติว่าชั้นที่ 4 ทำข้อสอบจนครบแล้วตอบผิดมากกว่าตอบถูกจึงปรับระดับกลับมาระดับชั้นที่ 3 ซึ่งทำข้อสอบครบทุกข้อแล้ว จึงจบการทดสอบและประเมินว่าผู้สอบอยู่ในระดับที่ 3 โดยกระบวนการทำงานของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับชั้น จำนวนทั้งหมด 5 ขั้นตอน แสดงในภาพที่ 2



ภาพที่ 2 กระบวนการทำงานของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับชั้น จำนวนทั้งหมด 5 ขั้นตอน

ผู้วิจัยนำคลังข้อสอบ A และคลังข้อสอบ B มาใช้ในการทดลอง โดยวิธีในการทดลอง ผู้วิจัยได้กำหนดตัวแปรและลำดับในการพิจารณาที่ละรูปแบบ แล้วนำผลการทำข้อสอบของนักเรียนทั้ง 51 คน มาใช้เป็นข้อมูลในการจำลองการทำข้อสอบในรูปแบบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ จากนั้นทำการเก็บรวบรวมข้อมูลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่พัฒนาขึ้น เพื่อนำมาใช้ในการเปรียบเทียบ เช่น จำนวนข้อสอบที่นักเรียนแต่ละคนทำ และผลการประเมินระดับ เพื่อนำมาเปรียบเทียบกับผลการประเมินที่ได้จากการทดสอบตามแบบประเพณีนิยม

ผลการวิจัย

ผู้วิจัยได้กำหนดเกณฑ์เพื่อเทียบเคียงการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยจะแบ่งคะแนนของการสอบตามแบบประเพณีนิยมเป็น 5 ช่วง แบ่งเป็นเกรด A ถึง F ตามช่วงคะแนนสูงไปต่ำตามลำดับ สำหรับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์จะประเมินผลออกมาเป็นระดับชั้น ตั้งแต่ 1 ถึง 5 โดยชั้นที่ 5 หมายถึงผู้สอบอยู่ในระดับเก่งมาก ชั้นที่ 1 หมายถึงผู้สอบอยู่ในระดับอ่อนมาก ดังตารางที่ 1

ตารางที่ 1 ตารางเทียบเคียงเกรดจากการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับระดับชั้นจากการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ของคลังข้อสอบ A และคลังข้อสอบ B

แบบประเพณีนิยม			แบบปรับเหมาะ ด้วยคอมพิวเตอร์	การแปลความหมาย	
ช่วงคะแนน คลังข้อสอบ A	ช่วงคะแนน คลังข้อสอบ B	เกรด	ระดับชั้น (Level)	ความยาก ของข้อสอบ	ผู้เรียนอยู่ใน ระดับ
74.4 - 93.0	85.6 - 107.0	A	5	ยากมาก	เก่งมาก
55.8 - 74.3	64.2 - 85.5	B	4	ค่อนข้างยาก	ค่อนข้างเก่ง
37.2 - 55.7	42.8 - 64.1	C	3	ปานกลาง	ปานกลาง
18.6 - 37.1	21.4 - 42.7	D	2	ค่อนข้างง่าย	ค่อนข้างอ่อน
0.0 - 18.5	0.0 - 21.3	F	1	ง่ายมาก	อ่อนมาก

ผลการวิจัยพบว่า เมื่อพิจารณาเปรียบเทียบระหว่างการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ เช่น ผลการประเมินจากทดสอบตามแบบประเพณีนิยมประเมินว่าผู้สอบได้เกรด B ผลการประเมินจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ประเมินว่าผู้สอบอยู่ในระดับชั้นที่ 4 ตัวอย่างดังกล่าว หมายถึงสามารถประเมินผลได้ตรงกัน โดยผลการเปรียบเทียบการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยแสดงเป็นร้อยละของการประเมินผลโดยการเทียบเกรดกับระดับชั้นได้ตรงกันในแต่ละรูปแบบแสดงดังตารางที่ 2

ตารางที่ 2 ตารางแสดงร้อยละของการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยการเทียบเกรดกับระดับชั้นตรงกัน

คลังข้อสอบ	ร้อยละของการประเมินผลโดยการเทียบเกรดกับระดับชั้นได้ตรงกันในแต่ละรูปแบบ					
	(P)	(R)	(Tp)	(Tr)	(Tpr)	(S)
คลังข้อสอบ A	72.55	35.29	52.94	47.06	47.06	64.71
คลังข้อสอบ B	64.71	31.37	47.06	49.02	50.98	76.47
เฉลี่ย	68.63	33.33	50.00	48.04	49.02	70.59

จากตารางที่ 2 เมื่อพิจารณาจากค่าเฉลี่ย พบว่า รูปแบบที่ 4 การพิจารณาค่าถ่วงน้ำหนักรวม (S) สามารถประเมินผลได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 70.59 รองลงมาคือ รูปแบบที่ 1 การพิจารณาค่าความยากง่าย (P) เพียงค่าเดียว โดยมีค่าเฉลี่ยร้อยละ 68.63 สำหรับรูปแบบที่ประเมินผลได้ตรงกันน้อยที่สุด คือ รูปแบบที่ 2 การพิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว โดยมีค่าเฉลี่ยร้อยละ 33.33

เมื่อพิจารณาเปรียบเทียบระหว่างการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยพิจารณาการประเมินผ่านหรือไม่ผ่าน เช่น ผลการประเมินจากทดสอบตามแบบประเพณีนิยมประเมินว่าผู้สอบได้ตั้งแต่เกรด B ขึ้นไป ผลการประเมินจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ประเมินว่าผู้สอบอยู่ตั้งแต่ระดับชั้นที่ 4 ขึ้นไป ตัวอย่างดังกล่าว หมายถึงสามารถประเมินผลได้ตรงกัน ซึ่งผู้วิจัยกำหนดให้เกณฑ์การตัดสินผ่านไม่ผ่าน พิจารณาว่าต้องประเมินผลได้ตั้งแต่เกรด B ขึ้นไปหรือระดับชั้นที่ 4 ขึ้นไป จึงแปลความหมายว่าผ่าน (เกณฑ์นี้สามารถปรับเปลี่ยนได้ตามความเหมาะสม) โดยผลการเปรียบเทียบการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยแสดงเป็นร้อยละของการประเมินผล โดยพิจารณาการประเมินผ่านหรือไม่ผ่านตรงกัน แสดงดังตารางที่ 3

ตารางที่ 3 ตารางแสดงร้อยละของการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยพิจารณาการประเมินผ่านหรือไม่ผ่านตรงกัน

คลังข้อสอบ	ร้อยละของการประเมินผล โดยการเทียบผ่านหรือไม่ผ่านได้ตรงกันในแต่ละรูปแบบ					
	(P)	(R)	(Tp)	(Tr)	(Tpr)	(S)
คลังข้อสอบ A	98.04	84.31	90.20	86.27	88.24	88.24
คลังข้อสอบ B	90.20	84.31	90.20	90.20	92.16	90.20
เฉลี่ย	94.12	84.31	90.20	88.24	90.20	89.22

จากตารางที่ 3 พบว่า รูปแบบที่ 1 การพิจารณาค่าความยากง่าย (P) เพียงค่าเดียว สามารถประเมินผลผ่านหรือไม่ผ่านได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 94.12 รองลงมาคือ รูปแบบที่ 3 โดยใช้รูปแบบย่อยพิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเสมอ (Tpr) และใช้รูปแบบย่อยพิจารณาค่าความยากง่ายสำคัญสุด (Tp) โดยมีค่าเฉลี่ยร้อยละ 90.20 ส่วนรูปแบบที่ประเมินผลผ่านหรือไม่ผ่านได้ตรงกันน้อยที่สุด คือ รูปแบบที่ 2 การพิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว โดยมีค่าเฉลี่ยร้อยละ 84.31

เมื่อพิจารณาเปรียบเทียบระหว่างการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยพิจารณาการประเมินผ่านหรือไม่ผ่าน โดยไม่ได้อ้างอิงจากตารางที่ 1 แต่ใช้วิธีเทียบกับร้อยละที่ทำข้อสอบถูก เช่น ผลการประเมินจากทดสอบตามแบบประเพณีนิยมประเมินว่าผู้สอบได้ตั้งแต่ตั้งแต่ร้อยละ 70 ขึ้นไป ถือว่าผ่าน ผลการประเมินจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ประเมินว่าผู้สอบอยู่ตั้งแต่ระดับชั้นที่ 4 ขึ้นไป ตัวอย่างดังกล่าว หมายถึงสามารถประเมินผลได้ตรงกัน โดยผลการเปรียบเทียบการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยแสดงเป็นร้อยละของการประเมินผล โดยพิจารณาการประเมินผ่านหรือไม่ผ่านจากร้อยละที่ทำข้อสอบถูกตรงกัน แสดงดังตารางที่ 4

ตารางที่ 4 ตารางแสดงร้อยละของการประเมินผลจากทดสอบตามแบบประเพณีนิยมกับการประเมินผลจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยพิจารณาการประเมินผ่านหรือไม่ผ่านจากร้อยละที่ทำข้อสอบถูกต้องตรงกัน

คลังข้อสอบ	ร้อยละของการประเมินผล ผ่านหรือไม่ผ่าน เทียบกับ ร้อยละที่ทำข้อสอบถูกต้องตรงกันในแต่ละรูปแบบ					
	(P)	(R)	(Tp)	(Tr)	(Tpr)	(S)
คลังข้อสอบ A (ร้อยละ 80)	56.86	62.75	56.86	70.59	84.31	76.47
คลังข้อสอบ A (ร้อยละ 75)	56.86	62.75	56.86	70.59	84.31	76.47
คลังข้อสอบ A (ร้อยละ 70)	58.82	64.71	58.82	72.55	86.27	78.43
คลังข้อสอบ A (ร้อยละ 65)	58.82	64.71	58.82	74.51	88.24	80.39
คลังข้อสอบ A (ร้อยละ 60)	60.78	66.67	60.78	78.43	92.16	84.31
คลังข้อสอบ B (ร้อยละ 80)	90.20	72.55	98.04	78.43	92.16	82.35
คลังข้อสอบ B (ร้อยละ 75)	90.20	76.47	94.12	82.35	96.08	86.27
คลังข้อสอบ B (ร้อยละ 70)	92.16	82.35	92.16	88.24	94.12	88.24
คลังข้อสอบ B (ร้อยละ 65)	88.24	86.27	84.31	88.24	90.20	92.16
คลังข้อสอบ B (ร้อยละ 60)	78.43	96.08	74.51	90.20	80.39	86.27
เฉลี่ย	73.14	73.53	73.53	79.41	88.82	83.14

จากตารางที่ 4 พบว่า รูปแบบที่ 3 โดยใช้รูปแบบย่อยพิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเหมือนกัน (Tpr) สามารถประเมินผลผ่านหรือไม่ผ่านเทียบกับร้อยละที่ทำข้อสอบถูกต้องตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 88.82 รองลงมาคือ รูปแบบที่ 4 การพิจารณาค่าถ่วงน้ำหนักรวม (S) โดยมีค่าเฉลี่ยร้อยละ 83.14 ส่วนรูปแบบที่ประเมินผลผ่านหรือไม่ผ่านเทียบกับร้อยละที่ทำข้อสอบถูกต้องตรงกันน้อยที่สุด คือ รูปแบบที่ 1 พิจารณาค่าความยากง่าย (P) เพียงค่าเดียว โดยมีค่าเฉลี่ยร้อยละ 73.14

เมื่อทำการเฉลี่ยค่าร้อยละของการประเมินผลโดยพิจารณาผ่านหรือไม่ผ่านทั้ง 2 แนวทาง จากตารางที่ 3 และ ตารางที่ 4 พบว่า รูปแบบที่ 3 โดยใช้รูปแบบย่อยพิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเหมือนกัน (Tpr) สามารถประเมินผลผ่านหรือไม่ผ่านทั้ง 2 แนวทาง ได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 89.51 รองลงมาคือ รูปแบบที่ 4 การพิจารณาค่าถ่วงน้ำหนักรวม (S) โดยมีค่าเฉลี่ยร้อยละ 86.18 ส่วนรูปแบบที่ประเมินผลผ่านหรือไม่ผ่านทั้ง 2 แนวทาง ได้ตรงกันน้อยที่สุด คือ รูปแบบที่ 2 การพิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว โดยมีค่าเฉลี่ยร้อยละ 78.92 ซึ่งแสดงในตารางที่ 5

ตารางที่ 5 ตารางค่าเฉลี่ยค่าร้อยละของการประเมินผลโดยพิจารณาผ่านหรือไม่ผ่าน จากตารางที่ 3 และ ตารางที่ 4

ข้อมูลร้อยละใน ตาราง	ค่าเฉลี่ย ค่าร้อยละของการประเมินผลโดยพิจารณาผ่านหรือไม่ผ่าน ตรงกันในแต่ละรูปแบบ					
	(P)	(R)	(Tp)	(Tr)	(Tpr)	(S)
ตารางที่ 3	94.12	84.31	90.2	88.24	90.2	89.22
ตารางที่ 4	73.14	73.53	73.53	79.41	88.82	83.14
เฉลี่ย	83.63	78.92	81.86	83.83	89.51	86.18

สำหรับค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของจำนวนข้อสอบที่ผู้สอบทำข้อสอบด้วยการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แสดงดังตารางที่ 6

ตารางที่ 6 ตารางค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของจำนวนข้อสอบที่ผู้สอบทำข้อสอบด้วยการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

คลังข้อสอบ	ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของจำนวนข้อสอบที่ผู้สอบทำข้อสอบ					
	(P)	(R)	(Tp)	(Tr)	(Tpr)	(S)
คลังข้อสอบ A	47.04	36.78	54.73	64.76	65.65	64.08
	(8.50)	(28.82)	(17.14)	(22.85)	(21.96)	(27.13)
คลังข้อสอบ B	57.18	41.22	61.80	70.53	73.67	61.80
	(12.25)	(34.36)	(34.40)	(28.78)	(32.13)	(34.40)

จากตารางที่ 6 พบว่า รูปแบบที่ 3 โดยใช้รูปแบบย่อยพิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเสมอ (Tpr) ผู้สอบต้องทำข้อสอบโดยเฉลี่ยจำนวนมากที่สุด รองลงมาคือ รูปแบบที่ 3 โดยใช้รูปแบบย่อยพิจารณาค่าอำนาจจำแนกสำคัญสุด (Tr) ส่วนรูปแบบที่ผู้สอบต้องทำข้อสอบโดยเฉลี่ยน้อยที่สุด คือ รูปแบบที่ 2 การพิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว

สรุปผลและอภิปรายผล

จากผลการเปรียบเทียบประสิทธิภาพการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับขั้นที่มีตัวแปรและลำดับในการพิจารณาที่แตกต่างกัน สามารถสรุปผลตามวัตถุประสงค์ของการวิจัยได้ ดังนี้ 1) รูปแบบที่ 4 การพิจารณาค่าถ่วงน้ำหนักรวม (S) สามารถประเมินผลโดยการเทียบเกรดกับระดับขั้นได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 70.59 2) รูปแบบที่ 1 การพิจารณาค่าความยากง่าย (P) เพียงค่าเดียว สามารถประเมินผลผ่านหรือไม่ผ่านโดยเทียบผลการประเมินเกรด B ขึ้นไป กับผลการประเมินระดับ 4 ขึ้นไป ได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 94.12 3) รูปแบบที่ 3 โดยใช้รูปแบบย่อยพิจารณาค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเสมอ (Tpr) สามารถประเมินผลผ่านหรือไม่ผ่านเทียบกับร้อยละที่ทำข้อสอบถูกได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 88.82 และสามารถประเมินผลผ่านหรือไม่ผ่านทั้ง 2 แนวทาง ได้ตรงกันมากที่สุด โดยมีค่าเฉลี่ยร้อยละ 89.51 4) รูปแบบที่พิจารณาค่าตัวแปรที่มีการถ่วงน้ำหนัก 2 ตัวแปรขึ้นไปผู้สอบจะต้องทำข้อสอบจำนวนมากกว่ารูปแบบที่พิจารณาค่าตัวแปรเพียงค่าเดียว โดยรูปแบบที่ 2 การพิจารณาค่าอำนาจจำแนก (R) เพียงค่าเดียว ผู้สอบจะทำข้อสอบโดยเฉลี่ยน้อยที่สุด

ในงานวิจัยนี้ใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิม มาวิเคราะห์ค่าตัวแปรทั้ง 4 รูปแบบ ซึ่งค่าความยากง่าย (P) และค่าอำนาจจำแนก (R) เป็นค่าตัวแปรที่สำคัญและมีการนำมาใช้อย่างแพร่หลาย ส่วนค่าตัวแปรผู้วิจัยได้คิดค้นขึ้นมาเพื่อพิจารณาว่ามีผลต่อการปรับระดับเลื่อนขั้นของผู้สอบมากน้อยเพียงใด และดูความแม่นยำว่าตรงกับคะแนนสอบจริงหรือไม่นั้นพบว่า ค่าถ่วงน้ำหนักรวม (S) ซึ่งเกิดจากการวิเคราะห์ค่าตัวแปรปกติคือ ค่า P และ ค่า R และเพิ่มตัวแปรพิเศษอีก 1 ตัว คือค่า K_L สามารถประเมินผลได้ตรงกับผลการประเมินในการทดสอบตามแบบประเพณีนิยมมากที่สุด ผู้วิจัยเห็นว่า ตัวแปรนี้มีความเหมาะสมในการนำมาใช้กับการทดสอบแบบปรับเหมาะในรูปแบบที่นำเสนอ เพราะค่า K_L ได้มาจากการวิเคราะห์ผู้สอบกลุ่มคะแนนรวมต่ำ (กลุ่มอ่อน) หากกลุ่มนี้ตอบข้อสอบข้ออื่นถูกมากกว่าผิดค่า K_L ก็สูง ซึ่งหมายถึงข้อสอบนั้นง่ายต่อการเดา ขณะที่รูปแบบ

การพิจารณาค่าตัวแปรแบบถ่วงน้ำหนักค่าความยากง่ายและค่าอำนาจจำแนกสำคัญเหมือนกัน (Tpr) สามารถประเมินผลผ่านหรือไม่ผ่านเทียบกับร้อยละที่ทำข้อสอบถูกต้องตรงกันมากที่สุด ซึ่งสูงกว่าการพิจารณาค่า (P) หรือค่า (R) เพียงค่าเดียวเช่นกัน ผู้วิจัยเห็นว่า การจะนำตัวแปรมาใช้เปรียบเทียบความแม่นยำนั้นจำเป็นต้องมีค่าตัวแปรหลักอย่างน้อย 2 ตัวขึ้นไป จึงจะช่วยให้การวิเคราะห์สามารถทำได้แม่นยำมากยิ่งขึ้น สำหรับเรื่องของจำนวนข้อสอบที่ผู้สอบต้องทำนั้น หากพิจารณาเฉพาะค่า P จะช่วยให้ผู้สอบทำข้อสอบได้น้อยลง ดังนั้น ค่า P จึงน่าจะเป็นตัวแปรที่สำคัญหากต้องการลดจำนวนข้อสอบในการทดสอบแบบปรับเหมาะ เพราะสามารถวัดระดับของผู้สอบได้โดยตรง โดยเฉพาะสามารถวัดผู้สอบที่เก่งมากหรืออ่อนมากได้ดี แต่จะไม่สามารถวัดผู้สอบที่อยู่ในระดับกลางๆ ได้ชัดเจน

ข้อเสนอแนะ

1. ระดับชั้นในงานวิจัยนี้แบ่งเป็น 5 ระดับชั้นตามค่าความยากง่ายที่งานวิจัยส่วนใหญ่อ้างอิง แต่ในบางบทความมีการแบ่งเป็น 7 ระดับชั้น (มลิวัลย์ ผิวคราม, ม.ป.ป.) ซึ่งหากปรับเป็น 7 ระดับชั้นน่าจะช่วยให้การประเมินผลมีความละเอียดขึ้น

2. ตัวคุณในการคำนวณค่าถ่วงน้ำหนักในรูปแบบต่าง ๆ ที่ใช้ตัวแปร 2 ตัวแปรขึ้นไปนั้น หากพิจารณาค่าตัวคุณให้มีความเหมาะสมกับตัวแปร เช่น ค่า P ปานกลาง ตัวคุณจะมีค่าสูง เพราะมีค่าความยากง่ายพอเหมาะ ค่า R สูง ตัวคุณจะมีค่าสูง เพราะมีอำนาจจำแนกดี ส่วน ค่า K_d ถ้ามีค่าปานกลางไปทางน้อย ตัวคุณจะมีค่าสูง เพราะข้อสอบจะเดายาก ดังนั้น หากปรับค่าตัวคุณให้เหมาะสม น่าจะได้ค่าถ่วงน้ำหนักรวม ที่สามารถบอกถึงข้อสอบที่มีคุณภาพดีได้

3. การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับชั้น หากต้องการให้ผู้สอบทำข้อสอบในจำนวนไม่มากจนเกินไปและสามารถประเมินผลได้รวดเร็ว ผู้วิจัยคิดว่า ควรใช้การถ่วงน้ำหนักระหว่างค่า P รวมกับค่าของข้อสอบคุณภาพดี ดังที่ได้นำเสนอไว้ข้างต้น น่าจะช่วยให้การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยใช้กลยุทธ์หลายขั้นตอนแบบปรับระดับชั้นมีประสิทธิภาพมากยิ่งขึ้น

เอกสารอ้างอิง

- ญาณิศรา มุนินทร์สาคร. (2558). การพัฒนาโปรแกรมทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 6. วิทยานิพนธ์หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวัดและเทคโนโลยีทางวิทยาการปัญญา มหาวิทยาลัยบูรพา.
- นันทรพร หาญวิทย์สกุล (2552). รายงานผลการวิเคราะห์ข้อสอบ วิชาหลักการประชาสัมพันธ์ รหัส 2201-270. แผนกวิชาการตลาด วิทยาลัยบริหารธุรกิจและการท่องเที่ยว. กรุงเทพฯ: กระทรวงศึกษาธิการ.
- นุภาพรณ ปลื้มใจ. (2558). การพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-Net ระดับชั้นมัธยมศึกษาปีที่ 6. วิทยานิพนธ์ สาขาวิชาการวัดและเทคโนโลยีทางวิทยาการปัญญา มหาวิทยาลัยบูรพา.
- เบญจมาภรณ์ เสนารัตน์, สมประสงค์ เสนารัตน์, และสมบัติ ท้ายเรือคำ. (2557) การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ. วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม. 20(2): 22-34.
- พิมพ์สิริ เจริญนเรศวร. (2549). การเปรียบเทียบค่าความสามารถ ลักษณะของการเปลี่ยนคำตอบ และเวลาที่ใช้ในการทำข้อสอบของผู้สอบ ในการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบด้วยคอมพิวเตอร์เมื่อเงื่อนไขการทดสอบและระดับความสามารถของผู้สอบแตกต่างกัน. วิทยานิพนธ์ ครุศาสตรมหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย.

มลิวัลย์ ผิวคราม. (ม.ป.ป.). การวิเคราะห์ข้อสอบแบบอิงกลุ่ม. ค้นเมื่อ 10 มกราคม 2560, จากเว็บไซต์:

<http://www.ipecp.ac.th/ipecp/cgi-binn/webpili/unit9/level9-2.html>

รังสรรค์ มณีเล็ก. (2540). ผลของตัวแปรบางตัวต่อความเที่ยงตรงเชิงสภาพและจำนวนข้อสอบที่ใช้ในการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบด้วยคอมพิวเตอร์. ปรินญาณพนธ์ หลักสูตร ปรินญาการศึกษาคุุณบัณฑิต สาขาการทดสอบและวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ. อุทัยรัตน์ รื่นจิตร. (2558). การศึกษาการประเมินเชิงวินิจัยโดยการประยุกต์ใช้โมเดลลำดับขั้นของ คุณลักษณะและการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์. วารสารวิชาการบัณฑิตวิทยาลัย มหาวิทยาลัยราชภัฏชัยภูมิ. 7(2): 23-30.

David J. Weiss. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. Measurement and Evaluation in Counseling and Development. 37(2): 70-84.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer Nijhoff Publishing.

Lord, F. M. (1971). A theoretical study of two-stage testing. Psychometrika. 36. 227-242.

Weiss, D. J. (1974). Strategies of Adaptive Ability Measurement. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.