Test Construction System with Optimizing Number of Overlapping Items for Standard Evaluation

ระบบสร้างชุดข้อสอบที่มีจำนวนคำถามร่วมที่เหมาะสมสำหรับการประเมินมาตรฐาน

Received 23 Apr 20
Reviewed 7 May 20
Revised 4 Jun 20
Accepted 23 Jun 20

Sarunya Deachnatee and Pokpong Songmuang
ศรัญญา เดชนที และ ปกป้อง ส่องเมือง
Department of Computer Science, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand
สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมสาสตร์ จังหวัดปทุมธานี

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 080-4524676 อีเมล: sarunya.deachnatee@gmail.com
*Corresponding Author, Tel. +66-80-4524676, E-mail: sarunya.deachnatee@gmail.com

บทคัดย่อ

ในปัจจุบันการสร้างข้อสอบหลายชุดให้อยู่บนมาตรฐานเดียวกันสามารถทำได้โดยใช้ทฤษฎีการตอบสนองข้อสอบ(Item Response Theory, IRT) แต่การสร้างชุดข้อสอบที่มีขนาดเล็กหลายชุด เพื่อลดความผิดพลาดจากการอ่อนล้าของผู้สอบ โดยชุดข้อสอบต้องมีจำนวนคำถามร่วมที่เหมาะสม เพื่อให้ชุดข้อสอบนั้นยังคงสามารถประเมินได้อย่างมีมาตรฐาน การสร้างนี้มีความซับซ้อน และมีค่าใช้จ่ายสูง ทางผู้วิจัยจึงได้ประยุกต์ใช้การจำลองการสร้างชุดข้อสอบและการสอบ เพื่อหาจำนวนคำถามร่วมที่เหมาะสม และนำข้อมูลไปพัฒนาระบบสร้างชุดข้อสอบ โดยระบบจะรับเงื่อนไขจากผู้ใช้งาน เช่น จำนวนชุดข้อสอบ จำนวนคำถามในข้อสอบ และสร้างชุดข้อสอบที่มีจำนวนคำถามร่วมที่เหมาะสมสำหรับการประเมินมาตรฐาน

คำสำคัญ: ทฤษฎีการตอบสนองข้อสอบ, การเทียบมาตราเชิงเส้นตรง, ข้อสอบ

Abstract

Recently, a multiple test construction with the same standard can be done using the Item Response Theory (IRT). We construct multiple small tests for decreasing the examinee's error from exhausting. The proper number of overlapping items between the tests is required in order to effectively estimate the IRT parameters. However, this is complicated and high costs. Therefore, we apply simulation technique to construct tests and optimize the number of overlapping items for standard evaluation. We also develop the test construction system based on the simulation results. The system constructs the test based on the user's constrains such as the number of tests, the number of items and sends the constructed multiple test with the same standard to the user.

Keywords: Item Response Theory, Linear Equating, Test

1. บทนำ

ปัจจุบันในสถาบันการศึกษา สถาบันจัดสอบ หรือสถานที่ทำงานได้ทำการคัดเลือกบุคคลากรเพื่อเข้ามาศึกษาทำงานโดยการจัดการสอบขึ้นสำหรับวัดระดับความสามารถ เช่น การสอบเข้ามหาวิทยาลัย การสอบเข้าทำงาน ซึ่งในแต่ละปีจะมีการจัดสอบอยู่บ่อยครั้งและมีผู้เข้าร่วมการสอบเป็นจำนวนมาก ดังนั้นมาตรฐานการวัดผลสอบจึงเป็นสิ่งที่สำคัญ โดยชุดข้อสอบที่ใช้จัดสอบนั้นจะต้องสามารถวัดระดับของผู้เข้าสอบได้อย่างมีมาตรฐานโดยไม่มีความลำเอียงหรือเกิดข้อวิพากษ์วิจารณ์จากผู้สอบว่าการจัดสอบแต่ละครั้งนั้นใช้ชุดข้อสอบที่มีระดับความยากไม่เท่าเทียมกัน ดั้งนั้นการจัดการคุณภาพของชุดข้อสอบที่ใช้ในแต่ละครั้งของการจัดสอบจึงจำเป็นต้องใช้ข้อสอบที่มีคุณภาพที่ใกล้เคียงกัน ซึ่งในการวิจัยนี้คุณภาพของชุดข้อสอบนั้นสามารถใช้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) [1][2] โดยข้อสอบแต่ละข้อจะมีความยากเป็นตัวแปรสำคัญของระบบ ซึ่งได้มาจากการเก็บข้อมูลจากการทำการทดสอบ หรือเป็นข้อสอบที่ผ่านการทดสอบมาแล้ว

ในปัจจุบันการออกข้อสอบโดยทั่วไปจะใช้ประสบการณ์และสัญชาตญาณจากผู้เชี่ยวชาญในการออกข้อสอบเท่านั้น ทำให้ข้อสอบที่ออกมาแต่ละชุดนั้นไม่ได้มาตรฐาน ทำให้มาตรฐานในการสอบแต่ละครั้งนั้นมีมาตรฐานไม่เท่าเทียมกัน [10]

นอกจากนั้นยังมีคำถามจำนวนมากที่ต้องการประเมินค่าคุณภาพของคำถาม เพื่อให้การประเมินค่าคุณภาพของคำถามนั้นอยู่บนมาตรฐานเดียวกัน โดยทั่วไปจะใช้กลุ่มประชากรกลุ่มเดียวในการประเมินค่าคุณภาพของคำถาม แต่คำถามที่ต้องการประเมินค่าคุณภาพนั้นมีจำนวนมาก ปัญหาที่ตามมาคือ ต้องเสียค่าใช้จ่ายราคาสูง และได้การปะเมินค่าคุณภาพของคำถามที่ไม่มีประสิทธิภาพ เนื่องจากในการประเมินคุณภาพนั้น หากคำถามมีจำนวนมากจำนวนผู้เข้าสอบจำเป็นที่ต้องมีจำนวนมากตามเพื่อให้ได้ปริมาณข้อมูลที่เพียงพอต่อการประเมินค่าคุณภาพของคำถามที่มีประสิทธิภาพและต้องแบ่งการตอบคำถามออกเป็นหลาย ๆ ครั้ง ทำให้ต้องเสียค่าใช้จ่ายราคาสูงในการจ้างผู้เข้าสอบ และใช้ระยะเวลาที่นานในการตอบคำถามจนครบ และในการประเมินค่าคุณภาพของคำถามนั้นหากให้ผู้เข้าสอบนั้นทำการตอบคำถามที่มีจำนวนมากจนเกินไปจะทำให้คุณภาพของคำถามที่ได้นั้นไม่มีประสิทธิภาพ เช่น หากให้ผู้เข้าสอบที่คำถามจำนวนมากในครั้งเดียว ผู้เข้าสอบนั้นอาจเกิดอาการล้าในการทำข้อสอบได้ ซึ่งทำให้ข้อมูลที่จะนำมาใช้เพื่อประเมินค่าคุณภาพคำถามที่ได้มานั้นไม่มีประสิทธิภาพ

เพื่อแก้ไขปัญหาที่กล่าวมาด้านต้น สามารถทำได้โดยการแบ่งคำถามจำนวนมากเป็นชุดคำถามหลายๆชุด และแจกจ่ายให้กับกลุ่มผู้เข้าสอบหลายๆกลุ่มทำการตอบคำถามเพื่อประเมินค่าคุณภาพของข้อสอบ แต่การใช้กลุ่มผู้เข้าสอบหลายกลุ่มนั้นจะทำให้คุณภาพของคำถามนั้นไม่อยู่บนมาตรฐานเดียวกัน มีความกระจายของความสามารถไม่เท่ากัน จึงแก้ไขให้ชุดคำถามแต่ละชุดนั้นแบ่งออกเป็น 2 ส่วน คือคำถามที่ต้องการประเมินค่าคุณภาพของคำถามและคำถามจำนวนหนึ่งที่คาบเกี่ยวกันเป็นคำถามหลักในข้อสอบทุกๆชุด เพื่อนำไปใช้ในการปรับคุณภาพของชุดคำถามในแต่ละชุดให้สมดุลและอยู่บนมาตรฐานเดียวกัน (Equation) [4][7][14]

วิธีการปรับให้ชุดคำถามทุกชุดให้อยู่บนมาตรฐานเดียวกันนั้นมีความสำคัญเป็นอย่างมาก เนื่องจากหากทำการปรับค่าคุณภาพของข้อสอบผิดพลาดหรือคลาดเคลื่อนในหนึ่งชุดจะทำให้การปรับค่าคุณภาพในชุดต่อๆไปมีความผิดพลาดที่สูงขึ้นเรื่อยๆ และทำให้ค่าคุณภาพของชุดคำถามและค่าความสามารถของผู้สอบที่ได้มาหลังปรับชุดคำถามเข้าหากันไม่มีประสิทธิภาพ

ผู้พัฒนาจึงจัดทำโมเดลเพื่อหาวิธีการปรับชุดคำถามเข้าหากันที่จะทำให้มีค่าความคลาดเคลื่อนของคุณภาพชุดคำถามที่น้อยที่สุด โดยเปรียบเทียบจากค่าคุณภาพของคำถาม ก่อน และหลัง การปรับชุดคำถาม

2. วัสดุ อุปกรณ์ และวิธีการวิจัย

2.1. ทฤษฎีการตอบสนองของข้อสอบ (Item Response Theory) [1][2]

ทฤษฎีการตอบสนองของข้อสอบ (Item Response Theory) หรือ ในที่นี้จะเรียกย่อๆว่า IRT เป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในตัวบุคคล กับพฤติกรรมการตอบสนองของข้อสอบของบุคคลนั้นว่ามีโอกาสตอบข้อสอบ ได้ถูกต้องมากน้อยเพียงใด โดยมีแนวคิดว่า พฤติกรรมการตอบสนองต่อข้อสอบของผู้สอบจะถูก กำหนดโดยความสามารถที่มีอยู่ภายในตัวบุคคลหรือความสามารถแฝง ที่ไม่สามารถสังเกตได้โดยตรง โดย IRT จะสร้างตัวแบบทางคณิตศาสตร์ที่แสดงความสัมพันธ์ที่กำหนดโอกาสที่ผู้สอบคนใดคนหนึ่งตอบสนองต่อข้อสอบข้อใดข้อหนึ่งที่เป็นฟังก์ชันของระดับความสามารถของผู้สอบ กับค่าพารามิเตอร์ของข้อสอบ โดยเรียกฟังก์ชันนี้ว่า ฟังก์ชันการตอบสนองข้อสอบ ซึ่งมีลักษณะฟังกชันเป็นแบบฟังก์ชันโลจิสติก (Logistic Function) หรือฟังก์ชันใกล้เคียงกับฟังก์ชันปกติสะสม (Normal Ogive Function) ตัวแบบของ IRT มีหลายรูปแบบขึ้นอยู่กับประเภทของผลตอบสนอง และจำนวนพารามิเตอร์ของข้อสอบที่อยู่ในตัวแบบ

ถ้าแยกตามประเภทของผลตอบสนองสามารถแยกตัวแบบ IRT ได้เป็น 2 ประเภทใหญ่ๆ คือ ตัวแบบ IRT ที่ใช้กับผลตอบสนองที่มีค่าคำตอบที่เป็นไปได้แค่ 2 ค้า หรือที่เรียกว่า ตัวแบบ IRT ทวิภาค (Dichotomous IRT Models) กับตัวแบบ IRT ที่ใช้กับผลตอบสนองที่มีค่าคำตอบที่เป็นไปได้มากกว่า 2 ค่า หรือ ที่เรียกตัวแบบนี้ว่า ตัวแบบ IRT พหุวิภาค (Polytomous IRT Models) นอกจากนี้ตัวแบบ IRT อาจจะแยกตามจำนวนพารามิเตอร์ในตัวแบบ เช่น ในกรณีตัวแบบ IRT ทวิภาคที่ในตัวแบบมีพารามิเตอร์เพียงตัวเดียว คือ ความยากของข้อสอบ เรียกตัวแบบนี้ว่าตัวแบบหนึ่งพารามเตอร์ซึ่งรู้จักกันทั่วไปในชื่อตัวแบบราสช์ ถ้าในตัวแบบมีพารามิเตอร์เกี่ยวกับความชัน (Slope Parameter) ที่ระบุอำนาจในการจำแนกของข้อสอบเพิ่มขึ้น เรียกตัวแบบนี้ว่า ตัวแบบสองพารามิเตอร์ และถ้าในตัวแบบมีพารามิเตอร์เกี่ยวกับการคาดเดา (Guessing Parameter) เพิ่มขึ้น จะเรียกตัวแบบนี้ว่าเป็นตัวแบบสามพารามิเตอร์ โดยที่ใน IRT ทุกแบบนั้นจะมีค่าความสามารถของผู้สอบ และค่าพารามิเตอร์ของข้อสอบมีความสัมพันธ์กัน นั่นคือการประมาณค่าของพารามิเตอร์ต่างๆของข้อสอบ ซึ่งจะต้องพิจารณาร่วมกับความสามารถของผู้สอบ ดังนั้นถ้ากลุ่มตัวอย่างของผู้สอบเป็นตัวแทนที่ดีของประชากรแล้ว จะทำให้ค่าประมาณที่ประเมิณมาได้นั้นมีความน่าเชื่อถือที่สามารถสรุปผลหรืออนุมานไปถึงกลุ่มประชากร และค่าประมาณที่ได้จากตัวแบบ IRT มีค่าไม่เปลี่ยนแปลง (Invariance) นั่น คือค่าประมาณพารามิเตอร์ของข้อสอบไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ และค่าประมาณพารามิเตอร์ของผู้สอบไม่แปรเปลี่ยนไปตามชุดของข้อสอบ ซึ่งจะต้องพิจารณาร่วมกับความสามารถของผู้สอบ ดังนั้นถ้ากลุ่มตัวอย่างของผู้สอบเป็นตัวแทนที่ดีของประชากรแล้ว จะทำให้ค่าประมาณที่ประเมิณมาได้นั้นมีความน่าเชื่อถือที่สามารถสรุปผลหรืออนุมานไปถึงกลุ่มประชากร และค่าประมาณที่ได้จากตัวแบบ IRT มีค่าไม่เปลี่ยนแปลงนั่น คือค่าประมาณพารามิเตอร์ของข้อสอบไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ และค่าประมาณพารามิเตอร์ของผู้สอบไม่แปรเปลี่ยนไปตามชุดของข้อสอบ

2.1.1. ข้อตกลงเบื้องต้นของ IRT

IRT มีข้อตกลงเบื้องต้นที่สำคัญ 2 ประการคือ

1 .ความเปนเอกมิติของแบบทดสอบ (Unidimension) หมายความว่าขอสอบ แตละขอในแบบทดสอบจะต้องมุงวัดคุณภาพหรือความสามารถเพียงอยางเดียวเท่านั้น

2. ความเป็นอิสระเฉพาะที่ (Local Independence) โดยจะแยกออกเปนความเป็นอิสระเฉพาะที่ของข้อสอบ (Local Item Independence) และความเปนอิสระเฉพาะที่ของผูสอบ (Local Person Independence)

2.1.2. ตัวแบบ IRT ทวิภาค (Dichotomous IRT Models) ตัวแบบ IRT ทวิภาคเป็นตัวแบบ IRT ที่ใช้ในกรณีผลตอบสนองของข้อสอบมีค่าคำตอบที่เป็นไปได้เพียง 2 ค้า เช่น 0 หรือ 1, ถูกหรือผิด, ใช่หรือไม่ใช่ เป็นต้น ตัวแบบ IRT มีชื่อเรียกตามจำนวนพารามิเตอร์ที่มีอยู่ในตัวแบบ ตามรูปแบบฟังก์ชัน และตัวแบบที่นำมาใช้ในการวิจัยครั้งนี้คือ ตัวแบบโลจิสติกแบบสองพารามิเตอร์(Two-Parameter Logistic Model : 2PL)โดยที่ตัวแบบสามารถเขียนเป็นสมการได้ดังต่อไปนี้

$$P_j(\theta_i) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp\,[a_j(\theta_i - b_j]} \qquad (1)$$

เมื่อ $P_j(\theta_i)$ คือความนาจะเปนที่ผูสอบที่มีความสามารถ θ จะสามารถตอบข้อสอบข้อที่ i ได ถูกตองโดยจะมีค่าระหว่าง -3 ถึง 3

b_j คือ คาความยาก (Difficulty Parameter) ของข้อสอบข้อที่ j ที่เป็นคาของระดับความสามารถที่ผูสอบมีโอกาส 50% ที่จะตอบข้อสอบข้อนั้นถูกตอง หรือเรียกว่าจุดเปลี่ยนโค้งของเสนโค้งลักษณะของข้อสอบ (Item Characteristic Curve: ICC) โดยทั่วไปคาความยากจะถูกทำให้เปนมาตรฐานที่มีคาอยูระหว่าง -3 ถึง 3 และคาที่ใกล -3 แสดงวาเปนข้อสอบที่ง่าย แตถามี คาใกล 3 แสดงวาเปนขอสอบที่ยาก

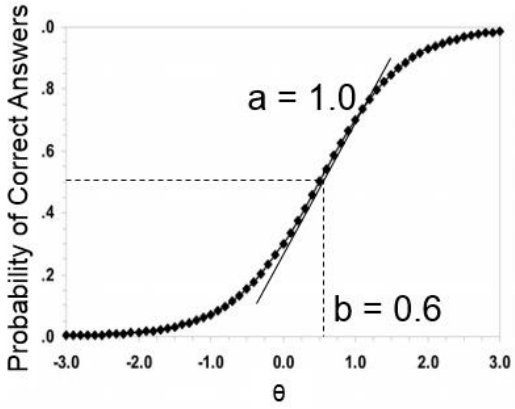
a_j คือ ค่าอำนาจการจำแนกของข้อสอบ (Discrimination Parameter) ของข้อสอบข้อที่ j เป็นค่าระบุว่า ข้อสอบข้อที่ j มีความสามารถมากน้อยแค่ไหนในการจำแนก ผู้สอบที่มีความสามารถต่ำและสูงออกจากกัน โดยค่านี้จะมีผลต่อความชันของเสนโค้งลักษณะ ของข้อสอบที่ตำแหน่งของ b_j ถ้าค่า a_j มีค่ายิ่งมากจะแสดงถึงอำนาจการจำแนกสูง โดยทั่วไปค่านี้ จะมีค่าอยู่ระหว่าง 0 ถึง 1

จากสมการที่(1) สามารถนำมาใช้ศึกษาความสัมพันธ์ระหว่างความน่จะเป็นในความน่าจะเป็นที่จะตอบข้อสอบได้ถูกต้อง กับระดับความสามารถของผู้สอบที่วัดได้ เมื่อนำมาเขียนกราฟ ICC โดยที่ให้ค่า a เป็น1 และค่า b เป็น 0.6 จะได้ดังรูปที่1 โดยกำหนดให้

แกนนอน คือ ระดับความสามารถของผู้สอบ (θ)

แกนตั้ง คือ ความน่าจะเป็นที่ผู้สอบจะทำข้อสอบได้ถูก

จะเห็นว่าค่า a เป็นความชันของกราฟ ถ้า a มีค่ามากขึ้น กราฟจะมีความชันที่มากขึ้นเช่นเดียวกัน และค่า b เป็นจุดตัดของแกนนอนที่ความน่าจะเป็นมีค่าเท่ากับ 0.5 ถ้า b มีค่ามากขึ้น กราฟจะเลื่อนไปทางขวา



รูปที่ 1 กราฟ โค้งลักษณะข้อสอบ (Item Characteristic Curve) ของคำถาม

2.2. การเทียบมาตราเชิงเส้นตรง(LE)

แองกอฟฟ์(Angoff 1971 : 569-586) ได้ทำการรวบรวมและนำเสนอ การเทียบมาตราเชิงเส้นตรงที่มี 6 รูปแบบ ซึ่งแต่ละรูปแบบจะมีวิธีการประมาณค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของแบบสอบแตกต่างกันออกไปตามเงื่อนไขของรูปแบบการรวบรวมข้อมูล

$$\bar{x} = \frac{\sum x}{n} \qquad (2)$$

เมื่อ $\bar{x}$ (เอ็กซ์บาร์) คือ ค่าเฉลี่ยเลขคณิต

$\sum x$ คือ ผลบวกของข้อมูลทุกค่า

n คือ จำนวนข้อมูลทั้งหมด

$$S.D. = \sqrt{\frac{\sum(x - \bar{x})^2}{N}} \qquad (3)$$

เมื่อ *S.D.* คือ ส่วนเบี่ยงเบนมาตรฐาน

x คือ ข้อมูล (ตัวที่ 1,2,3...,n)

$\overline{x}$ คือ ค่าเฉลี่ยเลขคณิต

n คือ จำนวนข้อมูลทั้งหมด

แต่ทุกรูปแบบจะตัดสินคะแนนสมมูลจากค่าคะแนนมาตรฐานเดียวกันซึ่งในงานวิจัยนี้คือ

$$\frac{Y - M_y}{S_y} = \frac{X - M_x}{S_x} \qquad (4)$$

เมื่อ *X* คือ คะแนนจากแบบสอบฉบับ X

Y คือ คะแนนจากแบบสอบฉบับ Y

M_x คือ ค่าเฉลี่ยจากคะแนนแบบสอบฉบับ X โดยใช้สมการค่าเฉลี่ยตามสมการที่ (2)

M_y คือ ค่าเฉลี่ยจากคะแนนแบบสอบฉบับ Y โดยใช้สมการค่าเฉลี่ยตามสมการที่ (2)

S_x คือ ค่าส่วนเบี่ยงเบนมาตรฐานของคะแนนแบบสอบฉบับ X โดยใช้สมการค่าส่วนเบี่ยงเบนมาตรฐานตามสมการที่ (3)

S_y คือ ค่าส่วนเบี่ยงเบนมาตรฐานของคะแนนแบบสอบฉบับ Yโดยใช้สมการค่าส่วนเบี่ยงเบนมาตรฐานตามสมการที่ (3)

ในแต่ละรูปแบบการรวบรวมข้อมูลของแองกอฟฟ์ ได้เสนอรายละเอียดแยกระหว่างกรณีที่แบบสอบเทียบมาตรามีความเที่ยงเท่ากัน และเมื่อแบบสอบเทียบมาตรามีความเที่ยงไม่เท่ากัน

และรูปแบบที่ทางผู้จัดทำเลือกใช้คือรูปแบบที่4 กลุ่มที่ไม่ได้มาจากการสุ่มสองกลุ่มแต่ละกลุ่มทำแบบสอบเพียงฉบับเดียว และทำแบบสอบร่วมที่เหมือนกันอีกส่วนหนึ่ง กลุ่มผู้สอบที่ไม่ได้มาจากการสุ่มเป็นกลุ่มที่เกิดขึ้นในสถานการณ์จริง ซึ่งจะต้องสอบแบบสอบต่างชุดในเวลาต่างกัน ผู้สอบจึงไม่ได้ถูกสุ่มมาจากประชากรเดียวกัน

2.3. วิธีแบบภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation)

วิธีการประมาณค่าความเป็นไปได้สูงสุดตามทฤษฎีการตอบสนองต่อข้อสอบที่นำค่าความสามารถของผู้สอบที่มีการประเมินค่าความสามารถของผู้สอบในทุกๆครั้งที่ผู้สอบทำการตอบข้อสอบโดยใช้สมการความเป็นไปได้สูงสุดเพื่อประมาณความสามารถและหาค่าความคลาดเคลื่อนมาตรฐาน จากนั้นก็ประมาณค่าความสามารถของผู้สอบโดยใช้สมการความเป็นไปได้สูงสุด จนกว่าการทดสอบจะสิ้นสุดตามเกณฑ์ที่กำหนดไว้ตามสมการประเมินระดับความสามารถ

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} a_i[u_i - P(\hat{\theta}_s)]}{\sum_{i=1}^{N} a_i^2 P_i(\hat{\theta}_s) Q(\hat{\theta}_s)} \qquad (5)$$

โดย θ_s คือค่าระดับความสามารถของผู้สอบ

a_i คือค่าอำนาจจำแนกของข้อสอบข้อที่ *i* , *i*=1, 2, ... , N

u_i คือผลการตอบข้อสอบข้อที่ *i*

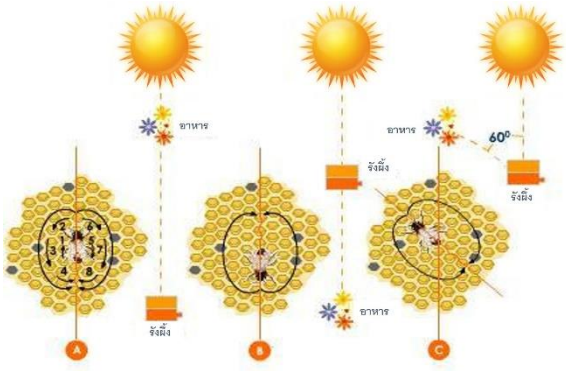
$P(\theta_i)$ คือค่าความน่าจะเป็นในการตอบข้อสอบข้อที่i ถูกต้องเมื่อผู้สอบมีระดับความามารถที่θ

$Q(\theta_s) = 1 - P_i(\theta_s)$ คือค่าความน่าจะเป็นในการตอบข้อสอบข้อที่iไม่ถูกต้อง เมื่อผู้สอบมีระดับความสามารถที่θ

หมายเหตุ u_i จะมีค่าเท่ากับ1เมื่อตอบข้อสอบถูก และ u_i จะมีค่าเท่ากับ0เมื่อตอบข้อสอบผิด

2.4. อัลกอริทึมอาณานิคมผึ้งเทียม (Bee Algorithm)

อัลกอริทึมอาณานิคมผึ้งเทียมเป็นวิธีการแก้ไขปัญหาแบบการหาคำตอบให้ใกล้เคียงเป้าหมายเรื่อย ๆ จนให้ได้ค่าที่ดีที่สุด (Optimization Algorithm) ซึ่งเลียนแบบมาจากพฤติกรรมการออกหาอาหารของผึ้งตามธรรมชาติ โดยอาณานิคมผึ้งจะประกอบไปด้วยหน้าที่ 3 กลุ่ม คือ ผึ้งงาน (Employed bees) ผึ้งผู้สังเกตการณ์ (Onlooker bees) และผึ้งสำรวจ (Scout bees)



รูปที่ 2 การทำงานของอาณานิคมผึ้งเทียม

จากรูปที่ 2 ผึ้งสำรวจจะออกเดินทางหาอาหารทุกทิศทางรอบๆรัง เมื่อพบแหล่งอาหารผึ้งสำรวจก็จะกลับมาที่รัง และไปที่ฟลอร์เต้นรำ (Dance Floor) แล้วทำการเต้นรำ (Waggle Dance) เพื่อรายงานให้ผึ้งสังเกตการณ์ทราบถึงเส้นทาง ระยะทางและปริมาณของอาหาร ซึ่งระยะเวลาในการเต้นรำจะสัมพันธ์กับระยะทางและปริมาณของอาหาร หลังจากที่เต้นรำเสร็จผึ้งสำรวจก็เปลี่ยนหน้าที่เป็นผึ้งงานเพื่อออกเก็บอาหารเพิ่มที่แหล่งเดิม ตราบใดที่แหล่งอาหารนั้นยังคงอุดมสมบูรณ์ เส้นทางนั้นก็จะถูกบอกต่อให้กับผึ้งสังเกตการณ์ตัวอื่น ๆ ให้ออกมาเก็บอาหารในเส้นทางนั้นมากขึ้น จนกระทั่งแหล่งอาหารนั้นหมดไปผึ้งสำรวจก็จะออกเดินทางหาแหล่งอาหารใหม่อีกครั้ง

ในที่นี้ระบบจัดสร้างข้อสอบนั้นได้ใช้ Bee Algorithm เป็นกระบวนการแก้ไขปัญหา โดยให้ Master Node เปรียบเสมือนรังผึ้งและ Slave Node เปรียบเสมือนผึ้งงาน เมื่อได้ Constraints ของชุดข้อสอบที่ผู้ใช้ต้องการแล้ว Master ก็จะนำ Constraints นี้ ไปสร้างเป็นชุดข้อสอบต้นแบบ โดยวิธีสุ่มเป็นจำนวนหนึ่ง หลังจากนั้น Master จะส่งชุดต้นแบบเหล่านั้นไปยัง Slave Node เพื่อทำการแก้ไขชุดข้อสอบต้นแบบเหล่านั้นให้ได้ค่าความยากใกล้เคียงกับค่าความยากที่ต้องการ มากยิ่งขึ้น หลังจากที่แก้ไขเสร็จครบตามจำนวนชุดที่ได้รับมา ก็จะส่งชุดที่ทำการแก้ไขแล้วกลับไปที่ Master Node เพื่อไปจัดลำดับค่าความยากที่มีความใกล้เคียงกับค่าความยากที่ต้องการมากที่สุด แล้วนำชุดเหล่านั้นส่งไปให้ Slave Node คำนวณหาค่าที่ดีที่สุดต่อไปจนกว่าจะได้ชุดข้อสอบที่มีความยากและจำนวนข้อสอบตามที่ผู้ใช้กำหนดไว้

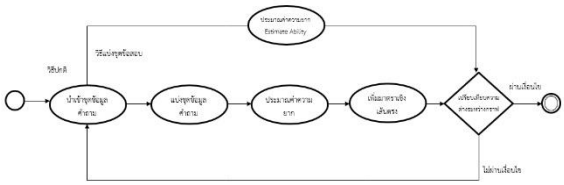
2.5. วิธีการทดลอง

2.5.1. การสร้างชุดข้อมูล

1. ใช้ผู้ใช้งานในการกำหนดเงื่อนไขที่ต้องการ เช่น จำนวนชุดข้อสอบ จำนวนข้อสอบ และจำนวนผู้เข้าสอบที่ต้องการ

2.ใช้การจำลองข้อมูลการทำข้อสอบในการทดลอง

2.5.2. สถาปัตยกรรมของระบบ



รูปที่ 3 สถาปัตยกรรมของระบบ

จากรูปที่ 3 เป็นโครงสร้างการทำงานของระบบโดยที่ระบบมีการทำงานดังนี้

1.นำข้อสอบที่ประกอบด้วย คำถาม เฉลย เข้าสู่ระบบฐานข้อมูล

2.แบ่งวิธีการประมาณค่าความยากขอข้อสอบออกเป็น 2 วิธี

2.1. แบบปกติ

- ประมาณค่าความยากข้อสอบจากการทำข้อสอบแบบปกติ

2.2. แบบแบ่งชุดข้อสอบ

- แบ่งชุดคำถาม โดยที่แต่ละชุดมีข้อทับซ้อนทุกชุด
- นำคำถามแต่ละชุดมาทำการประมาณค่าความยาก
- นำค่าความยากที่ได้จากข้อทับซ้อนมาทำการเทียบมาตราเชิงเส้นตรงเพื่อปรับค่าความยากข้อที่เหลือในชุดเพื่อให้ได้ค่าความยากของข้อสอบทุกข้อ ที่อยู่บนมาตรฐานเดียวกัน

3.นำค่าความยากของข้อสอบทั้งแบบปกติ และแบบแบ่งชุดข้อสอบมาเปรียบเทียบกัน เพื่อหาค่าความผิดพลาด

4.หากมีค่าความผิดพลาดสูงกว่า 0.5 ให้กลับไปทำขั้นตอนที่ 2 ใหม่อีกครั้ง

5.หากมีค่าความผิดพลาดน้อยกว่า 0.5 และมีข้อทับซ้อนที่เหมาะสมกับจำนวนชุดให้จบการทำงาน

3. ผลการทดลอง

ตารางที่ 1 การแจกแจงของพารามิเตอร์

พารามิเตอร์	คุณสมบัติ	ค่าจำลองเริ่มต้น	ผลลัพธ์
a	ช่วง	0 ~ 1	0 ~ 1
	ค่าเฉลี่ย	0.502	0.518
	ค่าเบี่ยงเบนมาตรฐาน	0.287	0.246
b	ช่วง	-3 ~ 3	-3 ~ 3
	ค่าเฉลี่ย	-0.301	-0.397
	ค่าเบี่ยงเบนมาตรฐาน	1.221	0.981

ตารางที่ 2 รายละเอียดข้อจำกัดสำหรับสร้างชุดคำถาม

จำนวนคำถามในแต่ละชุดคำถาม	จำนวนผู้ทำการสอบ
น้อยกว่า100ข้อ	มากกว่าจำนวนของคำถาม

ในการทดลองครั้งนี้ทางผู้จัดทำได้ทำการสร้างชุดคำถามจำลองตามตารางที่ 1 และมีเงื่อนไขของการสร้างชุดคำถามตามตารางที่ 2 หลังจากนั้นได้ทำการทดลองตามหัวข้อย่อย 2.5

ทางผู้จัดทำได้ทำการประมาณค่าพารามิเตอร์(ผลลัพธ์) หลังจากที่รวมชุดคำถามด้วยอัลกอริทึมอาณานิคมผึ้งเทียมและคำนวณค่าความผิดพลาดระหว่างค่าผลจำลองเบื้องต้นและค่าผลลัพทธ์ด้วย Root Mean Square Error (RMSE) จากการทดลอง จากตารางที่3จะสามารถเห็นได้ว่ามีความแตกต่างที่เห็นได้ชัดที่ระหว่างการใช้ชุดคำถามร่วมจำนวน 10 และ 20 เปอเซนต์

ตารางที่ 3 ผลการทดลอง

จำนวนข้อของการใช้ชุดคำถามร่วม (เปอเซนต์)	RMSE
10	1.568
20	0.898
30	0.914
40	0.753
50	0.698
60	0.685
70	0.434
80	0.269
90	0.258

4. อภิปรายผลและสรุป

จากการที่ได้ศึกษาและทดลองทำได้พบว่าการกำหนดค่าความยากของข้อสอบนั้นหากมีชุดข้อมูลของผู้สอบที่น้อยเกินไป จะทำให้การประเมินค่าความยากของข้อสอบนั้นมีการเบี่ยงเบนตามความสามารถของผู้สอบ เช่น ข้อสอบที่มีความยากปานกลาง หากให้กลุ่มผู้สอบที่เก่งทำจะทำให้ข้อสอบข้อนั้นกลายเป็นข้อที่ยาก และหากให้แต่กลุ่มผู้สอบที่อ่อนทำจะทำให้ข้อสอบข้อนั้นเป็นข้อสอบที่ยากจึงต้องให้ผู้สอบจำนวนเยอะเพื่อให้ได้ค่าความยากของข้อสอบที่มีคุณภาพ

5. องค์ความรู้ใหม่

ระบบจัดหาค่าความยากของข้อสอบโดยทฤษฎีการตอบสนองต่อข้อสอบ เป็นระบบที่ประเมินค่าความยากข้อสอบจำนวนมากผ่านทฤษฎีการตอบสนองข้อสอบ(Item Response Theory) ซึ่งจะต้องใช้ข้อสอบที่ผ่านการทดสอบและใช้กลุ่มคนทำข้อสอบเป็นกลุ่มคนทำเดียวกันเพื่อเป็นมาตรฐานความยากของข้อสอบ แต่เพื่อลดปัญหาของจำนวนของข้อสอบที่มากเกินไปจึงทำการจัดข้อสอบแบ่งออกเป็นหลายๆชุด และในแต่ละชุดจะมีข้อสอบที่ทับซ้อนกัน และนำค่าความยากของข้อสอบในแต่ละชุดมาปรับให้

เป็นชุดเดียว โดยการใช้การปรับเปลี่ยนคะแนนเชิงเส้นตรง (Linear Equating)

ทางคณะผู้จัดทำคาดหวังว่าระบบจะสามารถประเมินค่าความยากของข้อสอบทั้งหมดได้อยู่บนมาตรฐานเดียวกัน โดยที่มีจำนวนข้อสอบซ้อนทับที่น้อยที่สุด แต่ยังสามารถให้ผลลัพธ์ค่าความยากได้อย่างมีคุณภาพ

6. เอกสารอ้างอิง

[1] Lord, F.M. (1980). Applications of item response theory to practical testing problems. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

[2] Baker, Frank B., and Seock-Ho Kim, eds. Item response theory: Parameter estimation techniques. CRC Press, 2004.

[3] Songmuang, P., & Ueno, M. (2010). Bees algorithm for construction of multiple test forms in e-testing. IEEE Transactions on Learning Technologies, 4(3), 209-221

[4] จิระนาฏ ฉวีพัฒน์ และคณะ. (2017). การเปรียบเทียบคุณภาพของการปรับเทียบคะแนนตามแนวตั้ง โดยการใช้แบบทดสอบร่วมภายในระหว่างวิธีเชิงเส้นตรง กับวิธีทฤษฎีการตอบสนองข้อสอบแบบสามพารามิเตอร์. Available: https://www.tci-thaijo.org/index.php/RMCS/article/view/45707

[5] Haward Wainer and Robert J. Misley “Item Response Theory, Item Calibration and Proficiency Estimation”, New Jersey: Lawrence Erlbaum Associates, 2000

[6] Nathan Thompson, PhD, “what is item response theory” [Online]. Available: http://www.assess.com/what-is-item-response-theory/?nabe=5447296772997120:0

[7] van ver Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for IRT-based test design with practical constraints. Psychometrika, 54(2), 237-247. W.J.

[8] Armstrong, R. D., Jones, D. H., & Wu, L. (1992). An automated test development of parallel tests from a seed test. Psychometrika, 57(2), 271-288.

[9] G.-J. Hwang, P.-Y. Yin, and S.-H. Yeh, “A Tabu Search Approach to Generating Test Sheets for Multiple Assessment Criteria,” IEEE Trans. Education, vol. 49, no. 1, pp. 88-97, Sept. 2006.

[10] Songmuang, P., & Ueno, M. (2011). Bees algorithm for construction of multiple test forms in e-testing. IEEE Transactions on Learning Technologies, 4(3), 209-221.

[11] Chen, P. H., Chang, H. H., & Wu, H. (2012). Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach. Educational and Psychological Measurement, 72(6), 933-953.

[12] Chang, T. Y., & Shiu, Y. F. (2012). Simultaneously construct IRT-based parallel tests based on an adapted CLONALG algorithm. Applied Intelligence, 36(4), 979-994.

[13] Porter, A., Polikoff, M. S., Barghaus, K. M., & Yang, R. (2013). Constructing aligned assessments using automated test construction. Educational Researcher, 42(8), 415-423.

[14] Ishii, T., Songmuang, P., & Ueno, M. (2014). Maximum clique algorithm and its approximation for uniform test form assembly. IEEE Transactions on Learning Technologies, 7(1), 83-95.

[15] Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques. CRC Press.

[16] Pham, D. T., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., & Zaidi, M. (2005). The bees algorithm. Technical Note, Manufacturing Engineering Centre, Cardiff University, UK.