การพัฒนาแบบสอบผลสัมฤทธิ์ทางการเรียนวิชา การวัดและประเมินผลในชั้นเรียนโดยการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค

A Development of Measurement and Assessment in the Classroom Achievement Test : using the Bookmark standard setting procedure

ดร.สุรีพร อนุศาสนนันท์*
E-mail : sirimal@buu.ac.th

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) สร้างแบบสอบผลสัมฤทธิ์ทางการเรียนวิชา การวัดและประเมินผลในชั้นเรียน 2) ตรวจสอบความยาก อำนาจจำแนก ความตรง และความเที่ยงของแบบสอบผลสัมฤทธิ์ทางการเรียนวิชา การวัดและประเมินผลในชั้นเรียนที่สร้างขึ้น และ 3) หาคะแนนจุดตัดของแบบสอบที่สร้างขึ้นโดยการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค กลุ่มตัวอย่าง มี 2 กลุ่มคือ 1) ผู้ตัดสิน คือ อาจารย์ที่สอนรายวิชา การวัดและประเมินผลในชั้นเรียน จำนวน 6 คน 2) ผู้สอบ คือ นิสิตชั้นปีที่ 3 คณะศึกษาศาสตร์ ที่ลงเรียนรายวิชา 400204 การวัดและประเมินในชั้นเรียน ปีการศึกษา 2553 ใช้วิธีการสุ่มแบบยกกลุ่ม (cluster samping) จำนวน 667 คน เครื่องมือที่ใช้ในการวิจัยได้แก่ 1) แบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาการวัดและประเมินผลในชั้นเรียน แบบเลือกตอบ 4 ตัวเลือก และแบบอัตนัย จำนวน 2 ฉบับ ผลการสอบนำมาวิเคราะห์ค่าความยาก อำนาจจำแนก ความเที่ยงโดยใช้ทฤษฎีตอบสนองข้อสอบแบบ 3 พารามิเตอร์ และ Partial-Credit Model (PCM) วิเคราะห์ความตรงโดย หาสหสัมพันธ์ 2) คู่มือการจัดเรียงข้อสอบ (ordered item booklet : OIB) เพื่อใช้กำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค

ผลการวิจัยสรุปได้ดังนี้

1) แบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาการวัดและประเมินในชั้นเรียนประเภทเลือกตอบที่พัฒนาแล้ว ฉบับที่ 1 มีค่าความยากระหว่าง-2.50 ถึง 3.00 ค่าอำนาจจำแนกระหว่าง .49 ถึง .88 และค่าการเดาระหว่าง .11 ถึง .29 และประเภทอัตนัย ค่าอำนาจจำแนกเท่ากับ .98 และค่าความยากมีค่าระหว่าง -2.17 ถึง .47 ส่วนฉบับที่ 2 ค่าความยากระหว่าง -1.39 ถึง 3.00 ค่าอำนาจจำแนกระหว่าง .50 ถึง .90 และการเดาระหว่าง .11 ถึง .29 และประเภทอัตนัย ค่าอำนาจจำแนกเท่ากับ .47 และค่าความยากระหว่าง -0.98 ถึง .59

2) ค่าสารสนเทศของแบบสอบฉบับที่ 1 อยู่ในช่วงประมาณ 3.9 ถึง 5.0 และมีค่าสูงสุดอยู่ที่ระดับความสามารถ (*0*) ประมาณ -1.0 ฉบับที่ 2 ค่าสารสนเทศของแบบสอบอยู่ในช่วงประมาณ 3.9 ถึง 12.0 และมีค่าสูงสุดอยู่ที่ระดับความสามารถ (*0*) ประมาณ -0.5

* อาจารย์ประจำภาควิชาวิจัย และจิตวิทยาประยุกต์ คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

3) คะแนนจุดตัดที่อยู่ในรูปคะแนนดิบ และคะแนนความสามารถ จากแบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาการวัดและประเมินในชั้นเรียน ซึ่งกำหนดคะแนนจุดตัดด้วยวิธีบุ๊คมาร์ค 7 ระดับ ดังนี้ ระดับดีเยี่ยม (A) เท่ากับ 109 (1.453) ระดับดีมาก (B+) เท่ากับ 91 (1.293) ระดับดี (B) เท่ากับ 83 (1.203) ระดับดีพอใช้ (C+) เท่ากับ 65 (1.133) ระดับพอใช้ (C) เท่ากับ 49 (1.093) ระดับอ่อน (D+)เท่ากับ 30 (1.0130) และระดับอ่อนมาก (D) เท่ากับ 13 (.933)

คำสำคัญ : การกำหนดมาตรฐาน คะแนนจุดตัด การกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค การพัฒนาแบบสอบ

Abstract

The objectives of this study are 1) to develop achievement tests for the educational measurement and evaluation of bachelor degrees at Burapha University, 2) to investigate the item difficulty indices, the item discriminating indices, and validity and reliability in achievement tests, and 3) to investigate the cut scores on the bookmark standard setting method. The sample was divided into two groups: 1) 6 Educational Measurement and Evaluation in the Classroom lecturers assessors (acting as assessors) from the Education faculty of Burapha University, and 2) 677 undergraduate students who studied Educational Measurement and Evaluation in the Classroom, in 2010. The research instruments included: 1) both of midterm and final in educational measurement and evaluation achievement tests: multiple choice items and essay items. Their scores were analyzed to find difficulty indices ,discriminating indices ,the reliability—by using the IRT model; 3 PL model and the Partial-Credit Model (PCM) and the validity – by finding the correlation. 2) the ordered item booklet: the OIB used in Bookmark standard setting method.

The following are the research findings:

1. The first test, multiple choice items, indicated difficulty index were 2.50 to 3.00 , the discriminating index were .49 to .88 and the guessing index were .11 to .29 . The eassy items indicated the discrimination index was 0.98 and the item difficulty indices were -2.17 to .47. The second test, multiple choice items, the difficulty index were -1.39 to 3.00, the discriminating index were .50 to .90 and the guessing were .11 to .29. The eassy items indicated the discriminating index was 0.47 and the item difficulty indices were -.98 to .59.

2. The test information function of the first test was between 3.9 and 5.0, and the highest ability level (*0*) was -1.0. For the second test, the test information function was between 3.9 to 12 and the highest ability level (*0*) was -0.5

3. The cut scores in raw scores and examinee's ability (*0*) of the achievement tests were divided into 7 levels ,using the Bookmark method, as follows: excellent (A) was 109

(1.453= *0*) , very good (B+) was 91 (1.293= *0*), good (B) was 83 (1.203= *0*), rather good (C+) was 65 (1.133= *0*), fair (C) was 49.5 (1.093= *0*), poor (D+) was 30 (1.013= *0*), and very poor (D) was 13 (0.933 = *0*).

Keywords : standard setting ; cut off scores ; Bookmark standard setting ; development of test

บทนำ

การจัดการเรียนการสอนของคณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา ได้บรรจุรายวิชาการวัดและประเมินผลทางการศึกษาเป็นรายวิชาบังคับ ในหลักสูตรปริญญาตรีทุกสาขาวิชาเพื่อให้สอดคล้องกับมาตรฐานวิชาชีพครู โดยใช้ชื่อวิชา 400204 การวัดและประเมินในชั้นเรียน ซึ่งในแต่ละปี คณะศึกษาศาสตร์ต้องเปิดกลุ่มรายวิชาการวัดและประเมินผลการศึกษาประมาณ 15 - 20 กลุ่ม เพื่อให้การเรียนการสอนและการวัดและประเมินไปในแนวทางเดียวกัน จึงจำเป็นต้องหาวิธีการวัดและประเมินผลที่เป็นมาตรฐาน ผู้วิจัยซึ่งเป็นหนึ่งในฐานะผู้สอนในรายวิชานี้ เล็งเห็นวิธีการหนึ่งคือ การมีแบบสอบวัดผลสัมฤทธิ์ทางการเรียนที่เป็นมาตรฐานในรายวิชาการวัดและประเมินผลการศึกษา ผู้วิจัยจึงมีความสนใจพัฒนาแบบสอบผลสัมฤทธิ์ทางการเรียนวิชาการวัดและประเมินผลในชั้นเรียน เพื่อให้ได้แบบสอบมาตรฐานที่สามารถวัดนิสิตได้ใกล้เคียงกับความสามารถที่แท้จริง นอกจากนี้แล้วควรมีการกำหนดมาตรฐาน หรือคะแนนจุดตัดของแบบสอบที่มีคุณภาพเช่นเดียวกัน

การกำหนดมาตรฐานมีมากมายหลายวิธี วิธีที่ได้รับความนิยมใช้กันอย่างแพร่หลายคือ วิธีการกำหนดมาตรฐานด้วยวิธีแองกอฟ แต่วิธีแองกอฟเหมาะสำหรับข้อสอบที่มีระบบการให้คะแนนแบบ 0, 1 คือ ถูกให้ 1 คะแนน ผิดให้ 0 คะแนน และเมื่อกำหนดคะแนนที่หลากหลายระดับจะเกิดความยุ่งยาก และเสียเวลาในการกำหนดคะแนนจุดตัด ปัจจุบันมีการกำหนดมาตรฐานอีกวิธีหนึ่งคือ วิธีการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค (Bookmark standard setting) วิธีการบุ๊คมาร์คยังไม่เป็นวิธีที่แพร่หลายในประเทศไทย แต่ในต่างประเทศวิธีนี้ได้รับการนิยมอย่างแพร่หลาย เช่น ในสหรัฐอเมริกาวิธีนี้ได้รับความนิยมใช้กันอย่างกว้างขวางใน 28 มลรัฐ (Egan, 2001, cited in Beretvas, 2004) จุดเด่นของวิธีนี้คือ ช่วยจัดระบบการคิดแก่ผู้ตัดสินทำให้ผู้ตัดสินตัดสินง่ายขึ้น เนื่องจากมีการจัดเรียงข้อสอบที่เป็นระบบในรูปของคู่มือจัดเรียงข้อสอบ วิธีนี้ใช้ได้กับแบบสอบที่มีข้อสอบที่ให้คะแนนมากกว่า 2 ค่า (แบบอัตนัย) และการให้คะแนนแบบ 1 ค่า (แบบเลือกตอบ) และมีการกำหนดคะแนนจุดตัดที่หลากหลายระดับ ซึ่งสอดคล้องกับลักษณะแบบสอบที่ผู้วิจัยสร้างขึ้นประกอบด้วย แบบเลือกตอบ และแบบอัตนัย และสอดคล้องกับแนวทางการประเมินผลในรายวิชานี้ที่กำหนดเกรด 8 เกรด โดยมีคะแนนจุดตัด 7 ระดับ คือ ระดับดีเยี่ยม (A) ระดับดีมาก (B+) ระดับดี (B) ระดับดีพอใช้ (C+) ระดับพอใช้ (C) ระดับอ่อน (D+) และ ระดับอ่อนมาก (D) ดังนั้น ผู้วิจัยจึงได้นำวิธีการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์คมาหาคะแนนจุดตัด 7 ระดับในแบบสอบที่ผู้วิจัยสร้างขึ้น ผลการวิจัยนี้ ได้แบบสอบวิชาการวัดและประเมินในชั้นเรียนและคะแนนจุดตัดที่มีคุณภาพ ตลอดจนได้ข้อความรู้เกี่ยวกับวิธีการกำหนดมาตรฐานแนวใหม่ คือวิธีการบุ๊คมาร์คซึ่งนำไปใช้ในการสอบที่มีผลกระทบสูง

วัตถุประสงค์การวิจัย

1. เพื่อสร้างแบบสอบผลสัมฤทธิ์ทางการเรียนวิชา การวัดและประเมินผลในชั้นเรียน ระดับปริญญาตรี มหาวิทยาลัยบูรพา

2. เพื่อตรวจสอบความยาก อำนาจจำแนก ความตรง และความเที่ยงของแบบสอบผลสัมฤทธิ์ทางการเรียนวิชา การวัดและประเมินผลในชั้นเรียนที่สร้างขึ้น

3. เพื่อหาคะแนนจุดตัดของแบบสอบที่สร้างขึ้นโดยการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค

นิยามคำศัพท์

วิธีการกำหนดมาตรฐาน (Standard Setting) หมายถึง กระบวนการหามาตรฐานเพื่อแบ่งความสามารถของนักเรียนออกเป็น 8 ระดับ คือ ระดับดีเยี่ยม (A) ระดับดีมาก (B+) ระดับดี (B) ระดับดีพอใช้ (C+) ระดับพอใช้ (C) ระดับอ่อน (D+) ระดับอ่อนมาก (D) และระดับตก (F) ในการวิจัยครั้งนี้ใช้วิธีการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค

วิธีการกำหนดมาตรฐานบุ๊คมาร์ค (Bookmark Method) หมายถึง วิธีการหามาตรฐาน หรือคะแนนจุดตัด โดยจัดให้มีคู่มือการเรียงข้อสอบจากข้อง่ายสุดไปข้อยากสุดข้อละหนึ่งหน้า มีการวิเคราะห์ข้อสอบโดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) คู่มือที่ได้จะนำมาให้ผู้ตัดสินพิจารณาหาคะแนนจุดตัด โดยนำที่คั่นหนังสือคั่นในหน้าที่ผู้ตัดสินพิจารณาว่าเป็นคะแนนจุดตัดทั้งหมด 7 ระดับ

วิธีการวิจัย

1. ประชากรและกลุ่มตัวอย่าง แบ่งออกเป็น 2 กลุ่มคือ 1) ประชากรผู้ตัดสิน คือ อาจารย์มหาวิทยาลัยบูรพาที่ทำการสอนวิชาการวัดและประเมินผลระดับปริญญาตรี จำนวน 7 คน กลุ่มตัวอย่างผู้ตัดสินจำนวน 6 คนโดยวิธีการสุ่มตัวอย่างอย่างง่าย 2) ประชากรผู้สอบ คือ นิสิตชั้นปีที่ 3 คณะศึกษาศาสตร์ ที่ลงเรียนรายวิชา 400204 การวัดและประเมินในชั้นเรียน ปีการศึกษา 2553 จำนวนทั้งหมด 990 คน การกำหนดขนาดกลุ่มตัวอย่างใช้สูตรของ โคชแรน(Cochran) ใช้เทคนิคการสุ่มแบบยกกลุ่ม (Cluster random samping) ได้กลุ่มตัวอย่างผู้สอบจำนวน 667 คน

2. เครื่องมือที่ใช้ในการวิจัย คือ แบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาการวัดและประเมินผลในชั้นเรียน การพัฒนาแบบสอบดำเนินการดังนี้

2.1 วิเคราะห์คำอธิบายรายวิชา สร้างตารางโครงสร้างเนื้อหา นำตารางโครงสร้างเนื้อหาไปให้ผู้เชี่ยวชาญจำนวน 3 คน พิจารณากำหนดน้ำหนักเนื้อหา และความครอบคลุมของเนื้อหา ผู้เชี่ยวชาญแล้วดำเนินการตัดสินความสอดคล้องระหว่างข้อสอบและวัตถุประสงค์การเรียนรู้พบว่า ค่าดัชนีความสอดคล้อง (IOC) อยู่ในช่วง 0.6 – 1.0 แสดงว่าผู้เชี่ยวชาญมีความคิดเห็นและพิจารณาตัดสินแล้วว่าข้อสอบทั้งหมดนั้นวัดได้ตรงตามวัตถุประสงค์

2.2 จากนั้นผู้วิจัยสร้างแบบสอบวัดผลสัมฤทธิ์ทางการเรียน คือ ฉบับที่ 1 เป็นแบบสอบกลางภาค ประกอบด้วยแบบเลือกตอบจำนวน 90 ข้อ อัตนัย 1 ข้อ มีเนื้อหา 6 เรื่องคือ หลักการวัดและประเมินการศึกษา แนวทางการวัดและประเมินผลการเรียนรู้ การวางแผนประเมินทางการศึกษา การสร้างแบบสอบผลสัมฤทธิ์ทางการเรียน การประเมินทางจิตพิสัย การประเมินตามสภาพจริง และการประเมินการปฏิบัติ ฉบับที่ 2 แบบสอบปลายภาค ประกอบด้วยแบบเลือกตอบจำนวน 70 ข้อ อัตนัย 1 ข้อ มีเนื้อหา 4 เรื่องคือ สถิติเบื้องต้นเกี่ยวกับการวัดและประเมินผล การวิเคราะห์คุณภาพข้อสอบรายข้อ การวิเคราะห์คุณภาพแบบสอบทั้งฉบับ และการรายงานผลการเรียน

3 การเก็บรวบรวมข้อมูล และวิเคราะห์ข้อมูล

3.1 เป็นการทดลองใช้แบบสอบครั้งที่ 1 จำนวนนิสิต 5 คน เพื่อตรวจสอบความเป็นปรนัยของข้อสอบโดยการสัมภาษณ์ ผลการทดลองใช้ ปรากฏว่านิสิตมีความเข้าใจในคำชี้แจง และข้อคำถามแต่ละข้อดี

จากนั้นทดลองใช้แบบสอบครั้งที่ 2 กับนิสิตจำนวน 47 คน เพื่อปรับปรุงข้อสอบวิเคราะห์ด้วยโปรแกรม EXCEL และ SPSS เพื่อหาค่าความยาก อำนาจจำแนก โดยใช้สูตรในทฤษฎีการทดสอบแบบดั้งเดิม พบว่า ข้อสอบแบบเลือกตอบฉบับที่ 1 และฉบับที่ 2 มีค่าความยากง่ายเฉลี่ยปานกลาง (.513 และ .412 ตามลำดับ) แต่ค่าอำนาจจำแนกเฉลี่ยฉบับที่ 1 และฉบับที่ 2 ต่ำ (.175 และ .175 ตามลำดับ) ส่วนข้อสอบอัตนัยฉบับที่ 1 มีความยากง่ายปานกลาง (.548) ฉบับที่ 2 ค่อนข้างยาก (.366) ส่วนอำนาจจำแนกของข้อสอบอัตนัยฉบับที่ 1 และฉบับที่ 2 ดีมาก(.470 และ .591 ตามลำดับ) สามารถจำแนกเด็กเก่ง และเด็กอ่อนออกจากกันได้ สำหรับค่าความเที่ยงของแบบสอบทั้งสองฉบับอยู่ในระดับปานกลาง คือ 0.795 และ 0.749 ตามลำดับ จากนั้นผู้วิจัยตัดข้อคำถามที่ไม่ได้คุณภาพได้แบบสอบฉบับที่ 1 แบบเลือกตอบ 60 ข้อ แบบอัตนัย 1 ข้อ แบบสอบฉบับที่ 2 แบบเลือกตอบ 50 ข้อ แบบอัตนัย 1 ข้อ เพื่อนำไปเก็บข้อมูลจริงต่อไป

3.2 ดำเนินการเก็บข้อมูลจริงกับนิสิตจำนวน 667 คน นำข้อมูลมาวิเคราะห์หาค่าพารามิเตอร์ของข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ (Item response theory) ด้วยโปรแกรม Xcalibre และ Mulitilog คือ ค่าความยาก ค่าอำนาจจำแนก ค่าการเดา ฟังก์ชันสารสนเทศของแบบสอบ และหาค่าความสามารถของผู้สอบ (*θ*) ในแบบสอบประเภทเลือกตอบและอัตนัยที่ระดับความน่าจะเป็นในการตอบถูก .67 โดยโปรแกรม EXCEL ค่าความสามารถ (*θ*) ในงานวิจัยนี้ มาจาก Wright และ Stone.(1979, cited in Cizek และ Bunch, 2007) ได้เสนอสูตรดั้งเดิมแบบโมเดล Rasch สำหรับข้อสอบแบบเลือกตอบ ที่มีระบบการให้คะแนนเป็น 1, 0 โดยกำหนดความน่าจะเป็นในการตอบถูก P (X=1) เมื่อกำหนดค่าความยาก β_j และความสามารถของผู้สอบ (θ_i) ดังนี้

$$P(x = 1/\theta_i, \beta_j) = \exp(\theta_i - \beta_j)/[1 + \exp(\theta_i - \beta_j)] \quad (1)$$

เมื่อกำหนด θ_i = ค่าความยากของข้อสอบ

θ_i = ความสามารถของผู้สอบ

p = ความน่าจะเป็นของการตอบถูก

exp = natural logarithm e (2.71828...)

เมื่อ แทนค่า p = 2/3 แล้วแก้สมการหา θ_i ดังนี้

$$\exp(\theta_i - \beta_j)/[1 + \exp(\theta_i - \beta_j)] = 2/3 \quad (2)$$

$$\exp(\theta_i - \beta_j) = 2/3 * [1 + \exp(\theta_i - \beta_j)] \quad (3)$$

$$\exp(\theta_i - \beta_j) = 2/3 + 2/3 * \exp(\theta_i - \beta_j) \quad (4)$$

$$\exp(\theta_i - \beta_j) - 2/3 * \exp(\theta_i - \beta_j) = 2/3 \quad (5)$$

$$1/3 * \exp((\theta_i - \beta_j) = 2/3 \quad (6)$$

$$\theta_i = \beta_j + .693 \quad (7)$$

สำหรับสูตรการหาค่า θ_i จาก**ข้อสอบแบบให้คะแนนหลายค่า** เช่น ข้อสอบอัตนัย มาจาก Wright และ Masters (1982, cited in Cizek และ Bunch, 2007) ได้เสนอสูตรตามกรอบแนวคิดของ Rasch Model และ Partial- Credit Model (PCM) ซึ่งยกตัวอย่างข้อสอบที่ให้คะแนนแบบ 6 ค่า (0, 1, 2, 3, 4, 5) ซึ่งการคำนวณจากสูตรข้างล่างนี้ ผู้วิจัยใช้โปรแกรม EXCEL ช่วยในการคำนวณ

กำหนด $\pi_{nix} = \frac{\exp\sum\theta_n - \delta_{ij})}{\sum\exp\sum(\theta_n - \delta_{ij})}$ (8)

เมื่อกำหนด π_{nix} = likelihood ของผู้สอบแต่ละคน

θ_n = ความสามารถของผู้สอบ

δ_{ij} = ค่าความยากของข้อสอบข้อที่ i ณ คะแนน j

i = ข้อสอบ

j = คะแนน

เมื่อคะแนนเท่ากับ 0 ค่า $\delta_{io} \equiv 0$ แทนค่าในสมการ 11 ได้ดังนี้

$\sum(\theta_n - \delta_{ij}) = 0$ และ $\exp\sum(\theta_n - \delta_{ij}) = 1$ (9)

ขั้นที่ 1 $\sum(\theta_n - \delta_{ij}) = \sum(\theta_n - \delta_{io}) + \theta_n - \delta_{i1}$

$= 0 + \theta_n - \delta_{i1}$

$= \theta_n - \delta_{i1}$ (10)

ขั้น 2 – 4 ทำคล้ายขั้น 1 ได้สมการดังนี้

ขั้นที่ 2 $\sum(\theta_n - \delta_{ij}) = 2\theta_n - \delta_{i1} - \delta_{i2}$ (11)

ขั้นที่ 3 $\sum(\theta_n - \delta_{ij}) = 3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3}$ (12)

ขั้นที่ 4 $\sum(\theta_n - \delta_{ij}) = 4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4}$ (13)

ขั้นที่ 5 $\sum(\theta_n - \delta_{ij}) = 4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4} - \delta_{i5}$ (14)

นำค่า exponential มาคูณค่า summation ของสมการข้างต้นได้ดังนี้

ขั้นที่ 1 คะแนนเท่ากับ 0 $\exp(0)$ (15)

ขั้นที่ 2 คะแนนเท่ากับ 1 $\exp(\theta_n - \delta_{i1})$ (16)

ขั้นที่ 3 คะแนนเท่ากับ 2 $\exp(2\theta_n - \delta_{i1} - \delta_{i2})$ (17)

ขั้นที่ 4 คะแนนเท่ากับ 3 $\exp(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})$ (18)

ขั้นที่ 5 คะแนนเท่ากับ 4 $\exp(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})$ (19)

ขั้นที่ 6 คะแนนเท่ากับ 5 $\exp(5\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4} - \delta_{i5})$ (20)

3.3 นำค่าความยากมาเป็นข้อมูลเพื่อสร้างคู่มือการจัดเรียงข้อสอบ (Ordered Item Booklet: OIB) สำหรับผู้ตัดสินพิจารณาในการกำหนดคะแนนจุดตัด

4) ผู้ตัดสิน 6 คน ดำเนินการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค ขั้นตอนการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค มีดังนี้ (Buckendahl, 2002; Wang, 2003; Beretvas, 2004; Lewis, Green, Mitzel, Baum &

Patz, 1998; Kiplinger, 1997; Dawber & Lewis, 2005; Cizek, Bunch & Koons, 2004; Ferrara, Johnson & Chen, 2005)

1. สร้างคู่มือการจัดเรียงข้อสอบ (Ordered Item Booklet: OIB) โดยจัดเรียงข้อสอบจากข้อสอบง่ายที่สุดไปยังข้อสอบยากที่สุด มีรูปแบบการจัดเรียงข้อสอบ 1 ข้อ ต่อ 1 หน้า แล้วให้ ผู้ตัดสินทั้ง 6 คน ร่วมสร้างนิยามความสามารถของนักเรียน 8 ระดับ ว่านักเรียนแต่ละระดับควรมีความรอบรู้อย่างไร

2. ผู้ตัดสินพิจารณากำหนดมาตรฐาน โดยดำเนินการพิจารณา 3 รอบ ดังนี้

รอบที่ 1 ผู้ตัดสินแต่ละคนทำการกำหนดคะแนนจุดตัดที่ระดับอ่อนมากเป็นจุดแรก โดยให้ผู้ตัดสินแต่ละคนพิจารณาข้อสอบทีละข้อในคู่มือจัดเรียงข้อสอบ ผู้ตัดสินทำการกำหนดบุ๊คมาร์คหรือกำหนดคะแนนจุดตัด (คั่นหนังสือ) บนหน้าที่ผู้ตัดสินคิดว่า *"ผู้ที่คาบเส้นระดับอ่อนมากจะมีโอกาสตอบข้อสอบข้อนั้นถูก 67%"* หรือมีผู้คาบเส้นระดับอ่อนมาก 100 คน จะมีผู้ตอบถูก 67 คน จากนั้นผู้ตัดสินดำเนินการหาคะแนนจุดตัดระดับอ่อน พอใช้ ดีพอใช้ ระดับดี มาก ระดับดีเยี่ยม ตามลำดับเช่นเดียวกับข้างต้น **รอบที่ 2** จัดให้มีการอภิปรายภายในกลุ่ม ก่อนการอภิปรายผู้ตัดสินได้รับข้อมูลเลขที่หน้าในคู่มือจัดเรียงข้อสอบที่ผู้ตัดสินคนอื่นกำหนดบุ๊คมาร์ค (คั่นหนังสือ) ผู้ตัดสินร่วมกันอภิปรายถึงเหตุผล และความเหมาะสมของคะแนนจุดตัดที่ผู้ตัดสินแต่ละคนตัดสินไว้ หลังจากนั้นเปิดโอกาสให้ผู้ตัดสินพิจารณาคู่มือจัดเรียงข้อสอบอีกครั้งหนึ่ง เพื่อหาคะแนนจุดตัดในรอบที่ 2 เมื่อได้เลขที่หน้าของผู้ตัดสินแต่ละคน นำเลขที่หน้ามาหาค่าเฉลี่ยเพื่อหาคะแนนจุดตัดที่เป็นคะแนนดิบ **รอบที่ 3** ผู้อำนวยความสะดวกนำข้อมูลเลขที่หน้าในคู่มือจัดเรียงข้อสอบที่ผู้ตัดสินแต่ละคนกำหนดบุ๊คมาร์ค (คั่นหนังสือ) ในรอบที่ 2 มาให้ผู้ตัดสินพิจารณา จากนั้นผู้ตัดสินอภิปรายร่วมกัน แล้วเปิดโอกาสให้ผู้ตัดสินทำการกำหนดคะแนนจุดตัดอีกครั้ง เลขที่หน้าของผู้ตัดสินถูกแปลงเป็นค่าความสามารถ (*θ*) เลขที่หน้าที่ทำการคั่นคู่มือจัดเรียงข้อสอบไว้นี้ตรงกับค่า *θ* เท่าไร *θ* ก็นำค่าเหล่านี้มาหาค่าเฉลี่ยเพื่อนำมาประมาณความสามารถขั้นต่ำของนักเรียน นอกจากนี้หาคะแนนจุดตัดที่เป็นคะแนนดิบ โดยนำเลขที่หน้า ที่ผู้ตัดสินแต่ละคนกำหนด มาหาค่าเฉลี่ยในแต่ละระดับ

3. เมื่อได้คะแนนจุดตัดทั้ง 7 ระดับแล้ว จากนั้นผู้ตัดสินระดมพลังสมองเพื่อเขียนบรรยายระดับการปฏิบัติของนิสิตในประเด็นความรู้ ทักษะ และความสามารถของนิสิต ณ ระดับมาตรฐานของแต่ละระดับ รวมทั้งสิ้น 8 ระดับ

ผลการวิจัย

1. ผลการวิเคราะห์ข้อมูลเบื้องต้น

แบบสอบฉบับที่ 1 นิสิตส่วนใหญ่ทำคะแนนได้เกินครึ่งของคะแนนเต็มเล็กน้อย ส่วนฉบับที่ 2 นิสิตส่วนใหญ่ได้คะแนนเกินครึ่งของคะแนนเต็มค่อนข้างมาก และเมื่อพิจารณาค่าความเบ้ พบว่า ฉบับที่ 1และ 2 ติดค่าลบ หรือโค้งเบ้ซ้าย แสดงว่าแบบสอบค่อนข้างง่าย ดังตารางที่ 1

ตารางที่ 1 ค่าสถิติพื้นฐานของแบบสอบผลสัมฤทธิ์ทางการเรียนฉบับที่ 1 และฉบับที่ 2

ค่าสถิติพื้นฐาน (N=664)	*ฉบับที่ 1 แบบสอบกลางภาคเรียน*		*ฉบับที่ 2 แบบสอบปลายภาคเรียน*	
	แบบเลือกตอบ	*แบบอัตนัย*	*แบบเลือกตอบ*	*แบบอัตนัย*
คะแนนเต็ม	60	5	50	5
คะแนนเฉลี่ย	32.60	3.15	29.11	3.18
มัธยฐาน	33.00	3.00	30.00	4.00
ส่วนเบี่ยงเบนมาตรฐาน	6.84	1.38	8.67	1.62
ความเบ้	-.17	-.249	-.30	-.54
ความโด่ง	.06	-.90	-.673	-.86
คะแนนต่ำสุด	9	0	9	0
คะแนนสูงสุด	51	5	48	5

2. ผลการพัฒนาแบบสอบผลสัมฤทธิ์ทางการเรียนวิชาการวัดและประเมินผลในชั้นเรียน

แบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาการวัดและประเมินในชั้นเรียนประเภทเลือกตอบ ฉบับที่ 1 มีค่าความยากอยู่ในช่วง -2.50 ถึง 3.00 คิดเป็นค่าความยากเฉลี่ยเท่ากับ 0.60 ค่าอำนาจจำแนกอยู่ในช่วง .49 ถึง .88 คิดเป็นค่าอำนาจจำแนกเฉลี่ยเท่ากับ .57 และค่าการเดาอยู่ในช่วง .11 ถึง .29 คิดเป็นค่าเฉลี่ยการเดาเท่ากับ .24 **ส่วนฉบับที่ 2** มีค่าความยากอยู่ในช่วง -1.39 ถึง 3.00 คิดเป็นค่าความยากเฉลี่ยเท่ากับ .36 ค่าอำนาจจำแนกอยู่ในช่วง .50 ถึง .90 คิดเป็นค่าอำนาจจำแนกเฉลี่ยเท่ากับ .67 และค่าการเดาอยู่ในช่วง .11 ถึง .29 คิดเป็นค่าเฉลี่ยการเดาเท่ากับ .24 จะเห็นว่า แบบสอบทั้ง 2 ฉบับนั้น ข้อสอบส่วนใหญ่ค่อนข้างยาก และสามารถจำแนกผู้สอบที่มีความสามารถแตกต่างกันได้ดี ส่วนข้อสอบอัตนัยกลางภาค และปลายภาคฉบับละข้อ สามารถจำแนกผู้สอบที่มีความสามารถแตกต่างกันได้ดี ส่วนค่าความยากทั้งสองข้อ แต่ละระดับคะแนน มีค่า $\delta_{01} < \delta_{12} < \delta_{23} < \delta_{34} < \delta_{45}$ ดังนั้นค่าความยากอยู่ในเกณฑ์ใช้ได้ดังตารางที่ 2 และ 3

ตารางที่ 2 ค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าการเดา (c) ของแบบสอบผลสัมฤทธิ์ทางการเรียนประเภทเลือกตอบ

ค่าพารามิเตอร์	*ฉบับที่ 1 กลางภาคเรียน*	*ฉบับที่ 2 ปลายภาคเรียน*	*รวม 2 ฉบับ*
ค่าความยาก	-2.50 ถึง 3.00	-1.39 ถึง 3.00	-2.50 ถึง 3.00
ค่าความยากเฉลี่ย	0.61	0.36	.495
ส่วนเบี่ยงเบนมาตรฐานความยากเฉลี่ย	1.60	1.05	1.42
ค่าอำนาจจำแนก	0.49 ถึง 0.88	0.50 ถึง 0.90	.49 ถึง 0.90
ค่าอำนาจจำแนกเฉลี่ย	0.57	0.67	.61
ส่วนเบี่ยงเบนมาตรฐานอำนาจจำแนกเฉลี่ย	0.080	0.12	.11
ค่าการเดา	0.11 ถึง 0.29	0.11 ถึง 0.29	.11 ถึง .29
ค่าการเดาเฉลี่ย	0.24	0.23	.24
ส่วนเบี่ยงเบนมาตรฐานการเดาเฉลี่ย	0.03	0.03	.28

ตารางที่ 3 ค่าความยาก (δ_{ij}) ค่าอำนาจจำแนก (α) ของแบบสอบผลสัมฤทธิ์ทางการเรียนประเภทอัตนัย

แบบสอบ/ข้อที่	α	δ_{01}	δ_{12}	δ_{23}	δ_{34}	δ_{45}
แบบสอบกลางภาค ข้อ 1	0.98	-2.17	-1.44	-0.25	0.24	0.47
แบบสอบปลายภาค ข้อ 1	0.47	-0.98	-1.41	0.22	0.27	0.59

สำหรับค่าสารสนเทศหรือค่าความเที่ยง ของแบบสอบฉบับที่ 1 แบบสอบกลางภาคอยู่ในช่วงประมาณ 3.9 ถึง 5.0 และมีค่าสูงสุดอยู่ที่ระดับความสามารถ (*0*) ประมาณ -1.0 ฉบับที่ 2 แบบสอบปลายภาค ค่าสารสนเทศของแบบสอบอยู่ในช่วงประมาณ 3.9 ถึง 12.0 และมีค่าสูงสุดอยู่ที่ระดับความสามารถ (*0*) ประมาณ -0.5 และเมื่อรวมข้อสอบทั้ง 2 ฉบับ 100 ข้อ ค่าสารสนเทศของแบบสอบอยู่ในช่วงประมาณ 2.0 ถึง 39.0 และมีค่าสูงสุดอยู่ที่ระดับความสามารถ (*0*) ประมาณ 1.5 จะเห็นว่าแบบสอบทั้ง 2 ฉบับ เหมาะสำหรับนิสิตค่อนข้างเก่ง ส่วนความตรงตามสภาพของแบบสอบ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างระดับผลการเรียนเฉลี่ย (GPAX) กับคะแนนสอบวิชาวัดและประเมินผลในชั้นเรียนกลางภาคเท่ากับ 0.576 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างระดับผลการเรียนเฉลี่ยกับคะแนนสอบ วิชาวัดและประเมินผลในชั้นเรียนปลายภาคเท่ากับ 0.422 มีระดับนัยสำคัญที่ระดับ .01

และดาวเบอร์ และคณะ(Dawber et al., 2002) ว่า วิธีการกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค เป็นวิธีการที่นิยมใช้กันอย่างกว้างขวาง เนื่องจากวิธีการนี้ช่วยผู้ตัดสินในการคิดกำหนดคะแนนจุดตัดให้ง่ายในทางปฏิบัติ และมีหลักฐานสำคัญที่ช่วยผู้ตัดสินเข้าใจ และเชื่อมั่นกระบวนการในการใช้

เมื่อผู้ตัดสินกำหนดคะแนนจุดตัดจากคะแนนเต็ม 120 คะแนน พบว่าคะแนนจุดตัด 7 ระดับ มีดังนี้คือ คะแนนจุดตัดระดับดีเยี่ยม (A) 109 คะแนน คะแนนจุดตัดระดับดีมาก (B+) คือ 91 คะแนน คะแนนจุดตัดระดับดี (B) คือ 83 คะแนนจุดตัดระดับดีพอใช้ (C+) คือ 65 คะแนน คะแนนจุดตัดระดับพอใช้ (C) คือ 49.5 คะแนน คะแนนจุดตัดระดับอ่อน (D+)คือ 30 คะแนน คะแนนจุดตัดระดับอ่อนมาก (D) คือ 13 คะแนน จากที่กล่าวมา เป็นที่น่าสังเกตว่าการหาคะแนนจุดตัดด้วยบุ๊คมาร์คจะให้คะแนนจุดตัดระดับอ่อนมากต่ำมากคือ 13 คะแนน และคะแนนระดับดีเยี่ยมมีคะแนนจุดตัดสูงมากคือ 109 คะแนน จากผลการวิจัยดังกล่าวสอดคล้องกับการตั้งสมมติฐานของ กรีนและคณะ (Green et al., 2003); ยินและชูลส์ (Yin & Schulz, 2005) กล่าวว่า วิธีบุ๊คมาร์คน่าจะให้คะแนนจุดตัดต่ำที่สุดเมื่อเทียบกับวิธีการอื่น

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

การกำหนดมาตรฐานด้วยวิธีการบุ๊คมาร์คเหมาะสำหรับการกำหนดมาตรฐานที่ผู้สอบต้องนำผลการสอบไปตัดสินอนาคตของผู้สอบ เช่น การทดสอบระดับชาติ การสอบใบประกอบวิชาชีพต่าง ๆ เนื่องจากการสอบเหล่านี้ต้องการคะแนนจุดตัดที่เป็นมาตรฐานเดียว และต้องการคะแนนจุดตัดที่มีคุณภาพสูง และวิธีนี้เหมาะสำหรับแบบสอบที่มีข้อสอบจำนวนมาก มีลักษณะการให้คะแนนแบบหลายค่า หรือการให้คะแนนแบบสองค่า หรือทั้งสองแบบอยู่ในฉบับเดียวกัน และวิธีนี้ยังเหมาะสำหรับคะแนนจุดตัดหลายระดับ

ข้อเสนอแนะในการวิจัย

1. เนื่องจากงานวิจัยนี้ยังไม่มีการตรวจสอบคุณภาพมาตรฐานหรือคะแนนจุดตัดว่ามีความตรงหรือไม่ จึงควรมีการเปรียบเทียบคุณภาพของการกำหนดคะแนนจุดตัดด้วยวิธีบุ๊คมาร์ค (Bookmark Method) กับวิธีอื่นเช่น Single-Passage Bookmark Method, Contrasting Groups, Jaeger – Mills ซึ่งเป็นการหาความตรงภายนอก นั่นคือเปรียบเทียบคะแนนจุดตัดวิธีบุ๊คมาร์คกับวิธีอื่นว่ามีความสอดคล้องของคะแนนจุดตัดมากน้อยเพียงใด ส่วนความตรงภายในสามารถศึกษาจาก generalizability theory เพื่อประมาณค่า variance components

2. วิธีการกำหนดมาตรฐานหรือคะแนนจุดตัดแนวใหม่ ที่มีลักษณะแบบสอบเป็นการให้คะแนนแบบสองค่า (เช่น แบบเลือกตอบ) และมากกว่าสองค่า (เช่นอัตนัย) ในฉบับเดียวกัน หรือเป็นแบบสอบแบบใดแบบหนึ่ง เช่น วิธีแองกอฟแบบปรับขยาย (Extented Angoff) (Hambleton & Plake, 1995; Impara & Plake, 1997; Brandon, 2004) วิธีการเส้นภาพเด่น (Dominant Profile) (Putnam, Pence & Jaeger, 1995; Plake, Hambleton & Jaeger, 1997) วิธีการใช้การตัดสินเชิงนโยบาย (Judgmental Policy Capturing) (Jaeger, 1995) ซึ่งยังไม่ได้ศึกษาวิธีการดังกล่าว จึงควรทำวิจัยเกี่ยวกับการกำหนดคะแนนจุดตัดด้วยวิธีดังกล่าว

3. งานวิจัยนี้ยังไม่ได้ศึกษาถึงความคิดเห็นของผู้ตัดสินในการใช้การกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค จึงควรทำวิจัยเชิงคุณภาพกับผู้ตัดสินหลังจากกำหนดมาตรฐานด้วยวิธีบุ๊คมาร์ค กับวิธีอื่น ๆ เพื่อตรวจสอบประสบการณ์ และกระบวนการคิดในการกำหนดคะแนนจุดตัด ความเหมาะสมและความเป็นไปได้ของวิธีการดังกล่าว เพื่อนำผลที่ได้ไปใช้ปรับปรุงกระบวนการกำหนดคะแนนจุดตัดต่อไป

เอกสารอ้างอิง

Beretvas, S. N. (2004). Comparison of bookmark difficulty locations under different item response models. *Applied Psychological Measurement, 28*(1), 25-47.

Berk, R.A. (1995). Something old, something new, something borrowed, a lot to do. Applied Measurement in Education, 8(1), 99-109.

Brandon, P. R. (2004). Conclusions about frequently studied modified angoff standard- setting topics. *Applied Measurement in Education, 17*(1), 59-88.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of angoff and bookmark standard setting methods. *Journal of Educational Measurement, 33* (3), 253-263.

Cizek, G. J. ; Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance standards on tests*. California: Sage Publications, Inc.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Dawber, T., Lewis, D. M., & Rogers, W. T. (2002). The cognitive experience of bookmark standard setting participants. *Paper presented at the annual meeting of the American EducationalResearch Association*, New Orleans, LA.

Ferrara, S., Johnson, E., & Chen, L. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education, 18*(1), 35-59.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three Different standard setting procedures. *Educational Measurement: Issuers and Practice, 22*(1), 22-32.

Hambleton, R. K., Plake, B. S. (1995). Using an extended angoff procedure to set standards on complex performant assessments. *Applied Measurement in Education, 8*(1), 41-55.

Imprara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 69-81.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*(1), 15-40.

Kiplinger, V. L. (1997). *Standard-setting procedures for the specification of performance levels on a standards-based assessment*. Retrieved 5/7/2004 from http://www.cde.state.co.us/cdeassess/csap/asperf.htm

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). The bookmark standard setting procedure: Methodology and Recent Implementations. *Paper presented at the National Council for Mearsurement in Education annual meeting*, San Diego, CA.

Nunnally, J. C. and Berrnstein, I. H. (1994). *Psychometric Theory*. (3rded.) New York: McGraw-Hill,Inc.

Plake, B. S., Hambleton, R. K., Jaeger, R. M. (1997). A new standard-setting method for performance assessment: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57* (3), 355-366.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8* (1), 57-83.

Wang, N. (2003). Use of the rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement, 40*(3), 231-253.

Yin, P., & Schulz, E. M. (2005, April). A comparison of cul scores and cul score variability from Angoff-based and Bookmark-based procedures in standard setting. *Paper presented at the annual meeting of the National Council on Measurement in Education*, Montreal, Canada.