การกำหนดมาตรฐานการประเมินการศึกษา: แนวคิด วิธีการ และการปรับใช้

Standard Setting for Educational Evaluation: Concepts, Methods and Applications

ศิริพันธ์ ติยะวงศ์สุวรรณ* ณัฏฐภรณ์ หลาวทอง และศิริชัย กาญจนวาสี

Siripan Tiyawongsuwan*, Nuttaporn Lawthong and Sirichai Kanjanawasee

ABSTRACT

A standard setting method is used to determine levels of student performance in educational programs. A cutscore is the score that serves to classify students into the various performance levels. The results of the Ordinary National Education Test in Thailand must formally adopt achievement descriptors for outcomes of educational processes; consequently, the levels should also be determined so that students can be classified into various performance categories. Levels should be named using neutral labels to avoid the excessive meanings that are often attached to more descriptive labels. The next task is to develop performance descriptors that state in words what students know and can do, and to choose an appropriate method for setting cutscores. The literature in this field indicates there is a preference by some to apply resources to one well-selected method and implement it as best they can, while other studies suggest using several different methods of setting cutscores and combining their results. There are several standard setting methods currently in use. The two most popular procedures are the Angoff method and the Bookmark method. Under the Angoff Method, panelists are asked to estimate the probability that the "minimally competent candidate" at each achievement level would answer each item correctly. Under the Bookmark Method, panelists examine a booklet of test items arranged in order of difficulty. Each panelist places a bookmark between the items at the place where the panelist believes minimally competent candidates should have mastered the items below the bookmark and not have mastered the items above the bookmark.

Keywords: standard setting, cutscores, criterion-referenced test

ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ 10330.

Department of Educational Research and Psychology, Chulalongkorn University, Bangkok 10330, Thailand.

* Corresponding author, e-mail: siripantiya@gmail.com

บทคัดย่อ

การกำหนดมาตรฐานเป็นการดำเนินการเพื่อกำหนดเกณฑ์ในการตัดสินเกี่ยวกับระดับความสามารถของผู้สอบสำหรับการประเมินผลแบบอิงเกณฑ์ โดยจะกำหนดเป็น “คะแนนจุดตัด” ซึ่งเป็นคะแนนที่ใช้ในการแบ่งคะแนนเป็น 2 กลุ่มหรือมากกว่าสองกลุ่มที่สะท้อนถึงระดับความสามารถของผู้สอบที่ต่างกัน สำหรับการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานของประเทศไทยมีเป้าหมายสำคัญเพื่อตรวจสอบผลการจัดการศึกษา ดังนั้นผลการสอบควรจะสามารถใช้เพื่ออธิบายระดับความสามารถของผู้สอบอันเป็นการสะท้อนถึงระดับคุณภาพของการจัดการศึกษา จึงควรกำหนดระดับมาตรฐานมากกว่า 2 ระดับ เช่น 3, 4 หรือ 5 ระดับ เป็นต้น สำหรับชื่อของแต่ละระดับมาตรฐานควรใช้คำกลาง ๆ เพื่อหลีกเลี่ยงการแปลความหมายตามคำที่ใช้จนเกินที่ต้องการสื่อจริง จากนั้นจึงเขียนคำอธิบายเกี่ยวกับความรู้ และทักษะของผู้ที่อยู่ในแต่ละระดับมาตรฐานควรจะมี และกำหนดคะแนนจุดตัดด้วยวิธีการที่เหมาะสม ซึ่งการเลือกวิธีการกำหนดคะแนนจุดตัดอาจเลือกวิธีเดียวแล้วดำเนินการตามวิธีนั้นอย่างเคร่งครัด หรืออาจเลือกหลาย ๆ วิธี แล้วสรุปผลจากแต่ละวิธีเพื่อให้ได้คะแนนจุดตัดที่เหมาะสมที่สุด ซึ่งวิธีการกำหนดคะแนนจุดตัดที่นิยมใช้มากที่สุดคือ วิธี Angoff และวิธี Bookmark โดยวิธี Angoff จะให้ผู้เชี่ยวชาญพิจารณาความน่าจะเป็นที่ผู้มีความสามารถขั้นต่ำจะสามารถตอบข้อสอบได้ถูกต้อง ส่วนวิธี Bookmark จะให้ผู้เชี่ยวชาญพิจารณาคู่มือเรียงข้อสอบ ซึ่งจัดทำโดยการเรียงข้อสอบจากข้อง่ายไปข้อยาก และพิจารณาเลือกคั่นระหว่างข้อสอบที่ความเหมาะสมที่จะเป็นจุดตัด โดยข้อที่อยู่ก่อนหน้าที่คั่นหนังสือจะเป็นข้อที่ผู้ที่มีความสามารถขั้นต่ำของระดับความสามารถที่กำลังพิจารณาอยู่ มีความน่าจะเป็นที่จะตอบข้อสอบได้ถูกต้อง

คำสำคัญ: การกำหนดมาตรฐาน คะแนนจุดตัด การประเมินผลแบบอิงเกณฑ์

บทนำ

ในปัจจุบันการประเมินผลการศึกษานิยมประเมินผลแบบอิงเกณฑ์ กล่าวคือคะแนนจากการสอบจะถูกตีความโดยอิงกับเกณฑ์หรือมาตรฐานที่กำหนดไว้อย่างเหมาะสม (Berk, 1980; Cizek & Bunch, 2007; Zieky, Perie & Livingston, 2008) ดังนั้น การรายงานผลการจัดการศึกษาจะเป็นการนำเสนอผลการศึกษาเปรียบเทียบกับมาตรฐานที่ตั้งไว้ ซึ่งการกำหนดมาตรฐานสำหรับการประเมินการศึกษาได้แพร่หลายในหลายประเทศ อาทิ สหรัฐอเมริกา อังกฤษ ออสเตรเลีย เป็นต้น สำหรับการกำหนดมาตรฐานในประเทศสหรัฐอเมริกาเป็นการดำเนินการเพื่อกำหนดเกณฑ์เพื่อประเมินผลพัฒนาการทางการเรียนเป็นรายปี แต่ละรัฐจะต้องกำหนดมาตรฐานอย่างน้อย 3 ระดับ ได้แก่ ระดับชำนาญการบางส่วน (partially proficient) ระดับชำนาญการ (proficient) และระดับชำนาญการขั้นสูง (advanced proficient) จากการศึกษา พบว่า ระดับที่นิยมกำหนดมากที่สุดคือ 4 ระดับ ประกอบด้วย ต่ำกว่าพื้นฐาน (below basic), พื้นฐาน (basic), ชำนาญการ (proficient) และระดับเชี่ยวชาญ (advanced) ผลการสอบจะถูกตีความโดยอิงกับมาตรฐานที่กำหนดขึ้น และรายงานผลการสอบต่อผู้ปกครองทุกปี โดยนักเรียน ผู้ปกครอง และผู้ที่เกี่ยวข้องจะสามารถดูรายละเอียดเกี่ยวกับความรู้และทักษะของนักเรียนที่อยู่ในแต่ละระดับได้จากคำอธิบายระดับความสามารถ (Performance Level Descriptor: PLD) อันจะทำให้การตีความหมายผลการสอบเป็นไปอย่างมีมาตรฐาน (Perie, 2008)

นอกจากนี้ยังมีการสอบสำคัญจำนวนมากที่นิยมตีความคะแนนสอบโดยการเปรียบเทียบกับมาตรฐานที่กำหนดไว้ ส่งผลให้นักวิชาการหลายท่านพยายามศึกษาเพื่อกำหนดมาตรฐานที่มีความเหมาะสมและน่าเชื่อถือขึ้น เพื่อช่วยให้การตีความคะแนนมีความแม่นยำ อาทิ Gomez (2007) ได้ศึกษาเพื่อกำหนดระดับมาตรฐานความสามารถทางการ

อ่านภาษาอังกฤษ สำหรับการสอบวัดระดับความรู้ความสามารถทางการอ่านภาษาอังกฤษผ่านทางอินเทอร์เน็ต (TOEFL iBT reading test) 3 ระดับ ได้แก่ ระดับต่ำ (low), ระดับปานกลาง (intermediate) และระดับสูง (high) นอกจากนี้ Watermann and Klieme (as cited in Gomez, 2007) ศึกษาเพื่อกำหนดมาตรฐาน 4 ระดับ สำหรับการทดสอบเพื่อการประเมินผลการศึกษาวิทยาศาสตร์และคณิตศาสตร์ (Third International Mathematics Science Study: TIMSS) รวมถึงพัฒนาคำอธิบายระดับความสามารถที่เป็นที่ยอมรับ ซึ่งในรายงานผลสอบจะไม่นำเสนอเพียงคะแนนสอบเท่านั้นแต่จะอธิบายถึงความรู้และทักษะของผู้สอบด้วย

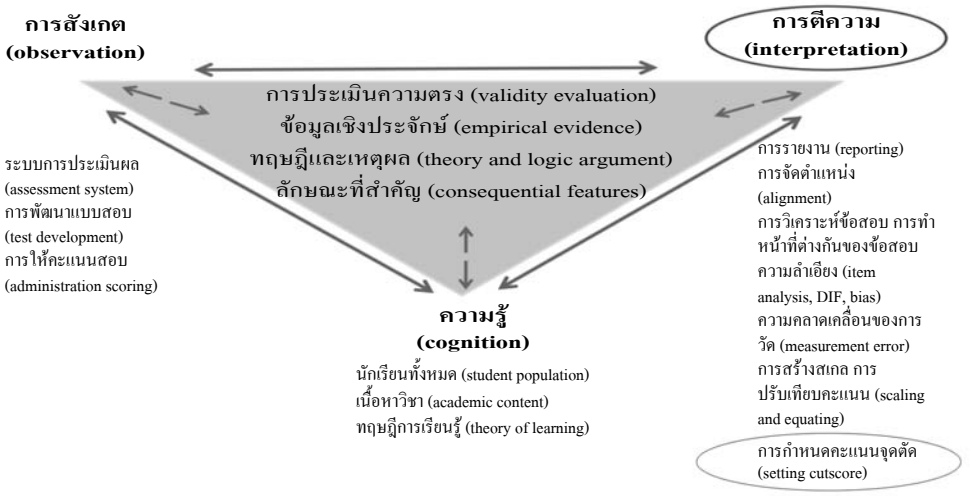
มาตรฐานที่กำหนดขึ้นส่วนใหญ่มิได้เป็นเพียงเกณฑ์สำหรับการตีความคะแนนสอบเท่านั้น แต่ยังเป็นเสมือนเป้าหมายสำหรับท้าทายให้ไปให้ถึงอีกด้วย ดังปรากฏในเป้าหมายของประเทศสหรัฐอเมริกาที่ตั้งเป้าเกี่ยวกับคุณภาพผู้เรียนไว้ว่า ในปี ค.ศ. 2014 ผู้เรียนจะต้องมีความสามารถในระดับชำนาญการ (proficient) ทุกคน ทุกรายวิชา โดยจะอาศัยสารสนเทศจากการทดสอบเพื่อวัดพัฒนาการทางการเรียนเป็นรายปีประกอบการวางแผนเพื่อดำเนินการไปสู่เป้าหมายที่คาดหวัง นอกจากที่กล่าวมาข้างต้นแล้ว มาตรฐานยังสามารถนำไปใช้ในหลายๆ บทบาท (Green, 1995) อาทิ ใช้เพื่อการรับรอง (certifiers) เช่น ใช้เพื่อการรับรองการจบการศึกษาและรับรองความสามารถในการทำงาน ใช้เพื่อทำนาย (predictors) เช่น ใช้เพื่อคัดเลือกบุคคลเข้าศึกษาต่อในมหาวิทยาลัย ใช้เพื่อบรรยาย (descriptors) เช่น ใช้แบ่งกลุ่มผู้สอบตามระดับความสามารถ และใช้เพื่อกระตุ้น (motivators) เช่น ใช้เป็นคะแนนที่เป็นเป้าหมายของการจัดการศึกษา หรือการเรียนการสอน เป็นต้น

จากบทบาทและความสำคัญของการกำหนดมาตรฐานที่กล่าวมา ผู้วิจัยจึงสนใจศึกษาแนวคิด และวิธีการของการกำหนดมาตรฐาน จากเอกสารและงานวิจัยต่าง ๆ ที่เกี่ยวข้อง เพื่อให้เกิดความเข้าใจอันจะนำไปสู่การสร้างข้อเสนอแนะในการนำการกำหนดมาตรฐานมาปรับใช้เพื่อการพัฒนาคุณภาพการศึกษาไทยต่อไป

แนวคิดเรื่องการกำหนดมาตรฐาน

การกำหนดมาตรฐาน (standard setting) มีวิวัฒนาการมากว่า 60 ปี ในปี ค.ศ.1970 วิธีการกำหนดมาตรฐานถูกใช้อย่างแพร่หลายสืบเนื่องมาจากความนิยมการทดสอบแบบอิงเกณฑ์ (criterion referenced testing) การกำหนดมาตรฐานเป็นการดำเนินการเพื่อกำหนดคะแนนที่มีความเหมาะสมและน่าเชื่อถือสำหรับใช้เป็นเกณฑ์หรือขีดจำกัด (threshold) ในการแบ่งคะแนนเป็น 2 กลุ่มหรือมากกว่าสองกลุ่มที่สะท้อนถึงระดับความสามารถที่ต่างกัน (Berk, 1980; Cizek & Bunch, 2007; Zieky et al., 2008; Towles-Reeves, 2008) จึงมีผู้เชี่ยวชาญบางท่านเรียกคะแนนที่ได้จากการกำหนดมาตรฐานว่า “คะแนนจุดตัด (cutscores)” (Zieky et al., 2008, p. 1) สำหรับบทความฉบับนี้จะใช้คำนี้เช่นเดียวกัน

ในปี ค.ศ. 2001 คะแนนจุดตัดได้ถูกใช้อย่างกว้างขวางมากในสหรัฐอเมริกาเนื่องจากรัฐบาลกลางของสหรัฐอเมริกาออกกฎหมายปฏิรูปการศึกษาที่ชื่อว่า “No Child Left Behind” ที่มีผลบังคับทั้งโรงเรียนประถมศึกษาและมัธยมศึกษาทั่วสหรัฐอเมริกา โดยมีสาระสำคัญว่านักเรียน เกรด 3–8 จะต้องมีพัฒนาการทางการเรียนด้านการอ่านและการคำนวณ และมีการจัดการทดสอบเพื่อวัดผลของพัฒนาการทางการเรียนเป็นรายปี เรียกว่า “Adequate Yearly Progress (AYP)” โดยได้กำหนดคะแนนจุดตัดเพื่อใช้ตัดสินระดับความสามารถทางด้านการอ่านและการคำนวณของนักเรียน (Zieky & Perie, 2006)



ภาพที่ 1 บทบาทของการกำหนดคะแนนจุดตัดในกระบวนการวัดและประเมินผลการเรียนรู้
ที่มา : ปรับปรุงจาก Marion และ Pellegrino (2006)

จากภาพที่ 1 การกำหนดคะแนนจุดตัด มีบทบาทเพื่อการตีความหมายคะแนนสอบในกระบวนการวัดและประเมินผลการเรียนรู้ โดยใช้เป็นขีดจำกัดในการแบ่งคะแนนสอบเป็นกลุ่มๆ ซึ่งคะแนนแต่ละกลุ่มก็จะสามารถอธิบายความหมายตามเป้าหมายของการกำหนดคะแนนจุดตัด

ผู้ที่จะนำคะแนนจุดตัดมาใช้ควรทำความเข้าใจเบื้องต้นเกี่ยวกับคะแนนจุดตัดก่อนที่จะนำไปใช้ในการตัดสินใจ เพื่อให้สามารถใช้ได้อย่างถูกต้องและเหมาะสม ซึ่งผู้วิจัยได้รวบรวมไว้ดังนี้ (Kane, 1994; Norcini, 2003; Zieky et al., 2008)

1. คะแนนจุดตัดมีบทบาทเพื่อการตัดสินใจ ซึ่งการตัดสินโดยการใช้คะแนนจุดตัด จะทำให้เกิดลักษณะของการตัดสินใจ 4 ลักษณะดังนี้คือ

1.1 การตัดสินว่าเป็นผู้ไม่ผ่านเกณฑ์ แต่ในสภาพความเป็นจริงแล้วเป็นผู้ที่มีความรู้พอที่จะผ่านเกณฑ์ได้ ซึ่งถือว่าเป็นการตัดสินใจคลาดเคลื่อนแบบไม่ยอมรับ (error of rejection) หรือเรียกว่า ความผิดพลาดแบบลบ (false negative)

1.2 การตัดสินว่าเป็นผู้ผ่านเกณฑ์ แต่ในสภาพความเป็นจริงแล้วเป็นผู้ที่ไม่มีความรู้พอที่จะผ่านเกณฑ์ ซึ่งถือว่าเป็นการตัดสินใจคลาดเคลื่อนแบบยอมรับ (error of acceptance) หรือเรียกว่า ความผิดพลาดแบบบวก (false positive)

1.3 การตัดสินว่าผ่านเกณฑ์และในสภาพความเป็นจริงก็เป็นผู้ที่มีความรู้พอที่จะผ่านเกณฑ์ ถือว่าเป็นการตัดสินที่ถูกต้องไม่มีความคลาดเคลื่อนเกิดขึ้น

1.4 การตัดสินว่าไม่ผ่านเกณฑ์และในสภาพความเป็นจริงก็เป็นผู้ที่ไม่มีความรู้พอที่จะผ่านเกณฑ์ ถือว่าเป็นการตัดสินที่ถูกต้องไม่มีความคลาดเคลื่อนเกิดขึ้น

2. คะแนนจุดตัดเป็นคะแนนซึ่งขึ้นอยู่กับการตัดสินของผู้เชี่ยวชาญ กล่าวคือคะแนนจุดตัดเป็นคะแนนที่ได้จากการตัดสินของผู้ตัดสินหรือผู้เชี่ยวชาญซึ่งจะตัดสินโดยอาศัยความรู้และประสบการณ์ในขณะนั้นเท่านั้น ดังนั้นการเลือกผู้เชี่ยวชาญจึงเป็นกระบวนการที่สำคัญมาก โดยต้องเลือกผู้ที่มีความเชี่ยวชาญในเรื่องนั้น ๆ และต้องมีเอกสารประกอบการตัดสินใจที่เพียงพอเพื่อให้คะแนนจุดตัดที่ได้มีความแม่นยำ

3. ไม่มีคะแนนจุดตัดจริง กล่าวคือคะแนน

จุดตัดที่กำหนดขึ้นเป็นเพียงคะแนนที่ได้จากการตัดสินของผู้เชี่ยวชาญกลุ่มนั้นๆ เท่านั้นและใช้กับการตัดสินใจหนึ่งๆ เฉพาะเท่านั้น

วิธีการกำหนดมาตรฐาน

จากบทบาทที่สำคัญของคะแนนจุดตัดที่กล่าวมา ได้มีนักวิชาการพยายามพัฒนาวิธีการกำหนดคะแนนจุดตัดขึ้นหลากหลายวิธี เพื่อให้ได้คะแนนจุดตัดที่มีความแม่นยำในการจำแนกผู้สอบตามระดับความสามารถและมีความน่าเชื่อถือ ซึ่งสามารถแบ่งวิธีการกำหนดมาตรฐานได้ 2 กลุ่ม คือ วิธีแบบสอบเป็นศูนย์กลาง (test-centered methods) และวิธีผู้สอบเป็นศูนย์กลาง (Examinee-centered methods) (Berk, 1996; Plake, 2008; Lee & Lewis, 2008; Näsström & Nyström, 2008) ผู้วิจัยได้สังเคราะห์ข้อมูลและสรุปรายละเอียดเกี่ยวกับวิธีกำหนดมาตรฐาน ดังแสดงในตารางที่ 1

ตารางที่ 1 สรุปรายละเอียดเกี่ยวกับวิธีการกำหนดมาตรฐาน

วิธี	การพิจารณา	ข้อดี	ข้อด้อย
1. วิธีแบบสอบเป็นศูนย์กลาง (test-centered methods)			
1.1 วิธี Angoff (Angoff method)	ผู้ตัดสินพิจารณาความน่าจะเป็นที่ผู้มีความสามารถขั้นต่ำ จะสามารถตอบข้อสอบแต่ละข้อได้ถูกต้อง	1. เป็นวิธีที่ง่ายต่อการนำไปปฏิบัติ 2. ใช้สถิติในการวิเคราะห์ไม่ยาก	การพิจารณากำหนดคะแนนมาตรฐานต้องใช้เวลามากในการพิจารณา
1.2 วิธี Bookmark (Bookmark method)	ผู้ตัดสินพิจารณากำหนดคะแนนมาตรฐานโดยพิจารณาจากคู่มือเรียงข้อสอบ ซึ่งจัดทำโดยการเรียงข้อสอบจากข้อง่ายไปข้อยาก	1. กำหนดคะแนนมาตรฐานได้หลายค่า 2. ใช้ได้ทั้งกับข้อสอบที่ให้คะแนน 2 ค่า และแบบให้คะแนนมากกว่า 2 ค่า 3. วิธีนี้มีการคำนวณค่าความยากของข้อสอบเพื่อประกอบการพิจารณาทำให้การตัดสินทำได้ง่ายขึ้น และส่งผลให้คะแนนมาตรฐานมีความน่าเชื่อถือมากขึ้น	1. ต้องใช้ข้อสอบจำนวนมากมิฉะนั้นอาจเกิด floor and ceiling effect ได้ 2. สถิติที่ใช้ในการวิเคราะห์ข้อมูลค่อนข้างซับซ้อน
1.3 วิธี Ebel (Ebel's method)	ผู้ตัดสินกำหนดร้อยละที่คาดหวังที่ผู้ที่มีความสามารถขั้นต่ำจะสามารถผ่านได้ถือว่าเป็นปริมาณความคาดหวังของผู้ที่มี ความสามารถขั้นต่ำที่จะสอบผ่าน	ผู้ตัดสินพิจารณาเพื่อกำหนดคะแนนมาตรฐานเพียง 12 ลักษณะเท่านั้นคือ พิจารณาตามลักษณะที่แบ่งไว้ทั้ง 12 ลักษณะซึ่งถือว่าเป็นการตัดสินที่ทำได้ง่ายและใช้เวลาไม่มาก	การคำนวณคะแนนมาตรฐานที่ต้องนำจำนวนข้อสอบในแต่ละลักษณะไปคูณกับค่าร้อยละที่คาดหวังนั้นเป็นเรื่องที่ค่อนข้างยุ่งยากและเสียเวลามาก
1.4 วิธี Nedelsky (Nedelsky method)	ผู้ตัดสินพิจารณาตัดตัวเลือกที่ผู้สอบที่มีความสามารถขั้นต่ำน่าจะตอบผิด ระดับการผ่านขั้นต่ำสำหรับข้อสอบถูกคำนวณจากจำนวนตัวลวงที่ยังคงเหลืออยู่	มีประโยชน์มากในศาสตร์ที่ตัวเลือกที่ผิดมีความสำคัญ เพราะทำให้เข้าใจว่าผู้สอบมีแนวคิดเกี่ยวกับเรื่องที่สอบผิดในเรื่องใด	ใช้เวลามากในการกำหนดคะแนนมาตรฐาน
2. วิธีผู้สอบเป็นศูนย์กลาง (Examinee-centered methods)			
2.1 วิธี Borderline-group (Borderline-group method)	ผู้ตัดสินจะต้องพิจารณาเพื่อระบุผู้สอบที่มีความสามารถระดับคาบเส้นซึ่งผู้สอบที่มี ความสามารถคาบเส้นคือผู้ที่มีความสามารถในระดับที่ผ่านพอดี	ง่ายในการนำไปใช้และง่ายในการทำความเข้าใจ	การจะตัดสินว่าผู้สอบคนใดมีความสามารถในระดับคาบเส้นนั้นทำได้ยากหากผู้ตัดสินไม่ได้คลุกคลี และสังเกตผู้สอบอย่างจริงจัง
2.2 วิธี Contrasting-group (Contrasting-group method)	ผู้ตัดสินควรจะเป็นครู หรือผู้ที่มีประสบการณ์ในการให้คะแนนและจัดอันดับการปฏิบัติงาน เพื่อพิจารณาแบ่งผู้สอบออกเป็น 2 กลุ่ม คือ กลุ่มที่มีระดับการปฏิบัติงานสูงกว่ามาตรฐาน และกลุ่มที่มีระดับการ	เป็นวิธีที่ง่ายต่อการอธิบายให้ผู้ตัดสินเข้าใจวิธีการกำหนดคะแนนมาตรฐาน	หากกำหนดคะแนนมาตรฐานหลายค่าผู้ตัดสินก็ต้องกำหนดกลุ่มเป็นหลาย ๆ กลุ่มระดับความสามารถซึ่งต้องใช้เวลาในการพิจารณามาก

ตารางที่ 1 (ต่อ)

วิธี	การพิจารณา	ข้อดี	ข้อด้อย
	ปฏิบัติงานต่ำกว่ามาตรฐาน หลังจากนั้นก็แจกแจงคะแนนสำหรับสองกลุ่ม เพื่อใช้ตัดสินและกำหนดคะแนนมาตรฐาน		
2.3 วิธีเลือกงานผู้สอบ (Examinee paper selections method)	ผู้ตัดสินจะเลือกพิจารณาชิ้นงานที่เป็นตัวแทนของงานทั้งหมด คะแนนเฉลี่ยจากชิ้นงานที่ถูกเลือกในแต่ละระดับจะถูกใช้เป็นค่าการผ่านขั้นต่ำ (minimum passing value) และจะถูกรวบรวมเพื่อใช้ตัดสินมาตรฐานการปฏิบัติแต่ละมาตรฐาน	เป็นการกำหนดคะแนนมาตรฐานโดยการใช้ผลงานประกอบการพิจารณา	วิธีนี้เหมาะสำหรับการตัดสินกับกลุ่มเล็ก ๆ เท่านั้น และการเลือกชิ้นงานเพื่อเป็นตัวแทนในการพิจารณานั้นทำได้ยาก
2.4 วิธีองค์รวม (Holistic method)	ผู้ตัดสินจะกำหนดคะแนนมาตรฐานโดยตัดสินจากผลงานที่สมบูรณ์ และตัดสินว่าชิ้นงานชิ้นใดเป็นตัวแทนของผลงานของกลุ่มผู้สอบคาบเส้น	ผู้ตัดสินกำหนดคะแนนมาตรฐานจากชิ้นงานของผู้สอบจริงๆ	เหมาะสำหรับการตัดสินกับกลุ่มเล็ก ๆ และการเลือกชิ้นงานเพื่อเป็นตัวแทนในการพิจารณานั้นทำได้ยาก
2.5 การวิเคราะห์กลุ่ม (Cluster analysis)	วิเคราะห์เพื่อกำหนดกลุ่มจากตัวแปรหรือลักษณะที่สนใจ โดยพยายามให้กลุ่มเดียวกันมีลักษณะที่สนใจเหมือนหรือคล้ายกันมากที่สุดและหน่วยที่อยู่ต่างกลุ่มกันจะต้องมีลักษณะที่สนใจต่างกันให้มากที่สุด	เป็นการกำหนดคะแนนมาตรฐานโดยปราศจากความลำเอียงใดๆ	เนื่องจากเป็นการวิเคราะห์เพื่อแบ่งกลุ่มจากตัวแปรที่สนใจ อันได้แก่ ผลการตอบข้อสอบ ดังนั้นจึงไม่สามารถกำหนดคะแนนมาตรฐานที่สูงกว่าหรือต่ำกว่าผลการตอบข้อสอบของนักเรียนกลุ่มที่เลือกมาวิเคราะห์

จากการศึกษาสรุปว่า วิธีแบบสอบเป็นศูนย์กลางเป็นวิธีที่ใช้ง่าย เหมาะกับแบบสอบปรนัยหรือแบบตอบสั้น ส่วนวิธีผู้สอบเป็นศูนย์กลางเหมาะสำหรับการสอบแบบปฏิบัติ

ในปัจจุบันวิธีกำหนดคะแนนจุดตัดที่ได้รับความนิยมมากที่สุดคือ วิธี Angoff และวิธี Bookmark (Zieky et al., 2008) บทความฉบับนี้จะนำเสนอรายละเอียดวิธีการกำหนดมาตรฐานสองวิธีดังกล่าว ซึ่งมีรายละเอียดต่อไปนี้

วิธี Angoff

วิธี Angoff ถูกคิดขึ้นโดย William H. Angoff ในปี ค.ศ.1970 เพื่อกำหนดมาตรฐานสำหรับข้อสอบแบบเลือกตอบ (multiple choice tests) วิธีนี้ได้รับความนิยมอย่างกว้างขวางโดยเฉพาะการทดสอบที่มีความสำคัญ (high stakes testing) (Angoff, as cited in Kane, 1994; Zieky et al., 2008) ทั้งนี้เนื่องจากว่าวิธีนี้มีขั้นตอนการดำเนินการที่ง่าย และเน้นการพิจารณาข้อสอบเป็นหลัก กล่าวได้ว่าวิธีการนี้สามารถดำเนินการได้ทั้งก่อนและหลังการสอบ วิธีนี้จะให้ผู้เชี่ยวชาญเป็นผู้พิจารณากำหนดมาตรฐาน โดยมีหลักในการพิจารณาความน่าจะเป็นที่ผู้มีความสามารถขั้นต่ำ (minimally competent candidate) จะสามารถตอบข้อสอบได้ถูกต้อง โดยคะแนนจุดตัดจะได้มาจากผลรวมของค่าความน่าจะเป็นของข้อสอบทุกข้อนั่นเอง

วิธี Angoff ได้รับการปรับปรุง และพัฒนาอย่างต่อเนื่องโดยนักวิชาการหลายท่าน เป้าหมายเพื่อให้คะแนนจุดตัดที่กำหนดขึ้นมีคุณภาพเป็นที่ยอมรับ กล่าวคือมีความเที่ยงภายในผู้ตัดสิน (intra-rater reliability) มีความเที่ยงระหว่างผู้ตัดสิน (inter-rater reliability) มากขึ้น ลดความคลาดเคลื่อนมาตรฐาน (standard error) ให้น้อยลง เป็นต้น วิธี Angoff ถูกปรับปรุงหลายลักษณะภายใต้ชื่อ “Modified Angoff” “Extended Angoff” และ “Yes/No Angoff” ซึ่งลักษณะการปรับปรุงรวมทั้งจุดเด่นและจุดด้อยของแต่ละลักษณะ ผู้วิจัยได้สังเคราะห์ข้อมูลและสรุปไว้ในตารางที่ 2

ตารางที่ 2 สรุปการปรับปรุงวิธี Angoff ข้อดีและข้อเสีย

การปรับปรุงวิธี Angoff	นักวิชาการ	ข้อดี	ข้อเสีย
1. การเพิ่มจำนวนรอบของการพิจารณาตัดสิน เช่น 2, 3 หรือ 4 รอบ	Ferdous and Plake (2005); Giraud, Impara and Plake (2005); Hess, Subhiyah and Giordano (2007); Fowell, Fewtrell and McLaughlin (2008); Hurtz, Jones, and Jones, (2008); Näsström and Nyström (2008); MacCann (2008); Clauser et al. (2009)	- เป็นการเปิดโอกาสให้ผู้ตัดสินได้ทบทวนผลการกำหนดคะแนนจุดตัด - ผู้ตัดสินได้อภิปรายและแลกเปลี่ยนความคิดเห็นกันเกี่ยวกับคะแนนจุดตัดทำให้การกำหนดคะแนนมีความสอดคล้องกันมากขึ้น	ใช้เวลาในการกำหนดคะแนนจุดตัดมาก
2. การนำเสนอข้อมูลสารสนเทศที่สำคัญในการตัดสิน เช่น ความยากของข้อสอบ สัดส่วนของผู้ที่ตกกับผู้ที่ผ่าน	Ferdous and Plake (2005); Hess, et al. (2007); Näsström and Nyström (2008); Fowell et al. (2008); Clauser et al. (2008)	- ช่วยให้ผู้ตัดสินมีข้อมูลสนับสนุนคะแนนจุดตัดที่กำหนด - ช่วยให้ผู้ตัดสินกำหนดคะแนนจุดตัดได้ง่ายขึ้น	หากให้ข้อมูลสารสนเทศมากเกินไปอาจส่งผลในด้านลบต่อการกำหนดคะแนนจุดตัดเนื่องจากเป็นการชี้นำการตัดสินของผู้ตัดสิน
3. การเปลี่ยนวิธีการพิจารณาให้ง่ายขึ้น ได้แก่ วิธี Yes/No Angoff	Yudkowsky (2008); Ferdous and Plake (2008)	ง่ายต่อการพิจารณาตัดสิน ทำให้กำหนดคะแนนจุดตัดได้ในกรณีที่ข้อสอบให้คะแนนไม่เท่ากัน	เป็นการพิจารณาตัดสินที่หยาบเกินไป
4. การให้ค่าน้ำหนักแก่ข้อสอบที่วัดแต่ละมิติ	Hambleton and Plake (1995)	ใช้ในการตัดสินสำหรับข้อสอบที่วัดหลายมิติได้	ใช้เวลาค่อนข้างมาก
5. การนำทฤษฎีการตอบสนองข้อสอบมาประยุกต์ใช้	Hurtz et al. (2008); Ferdous and Plake (2008); Clauser et al. (2009)	- ทำให้ผลการตัดสินน่าเชื่อถือมากขึ้น - ช่วยให้ตัดสินใจได้ง่ายขึ้น	- ต้องมีความรู้ในเรื่องสถิติที่ต้องใช้ - ต้องใช้เวลาในการวิเคราะห์และจัดเตรียมข้อมูลมาก

วิธี Angoff และวิธี Modified Angoff มีข้อดีคือเป็นวิธีที่ง่ายต่อการนำไปปฏิบัติ กล่าวคือผู้เชี่ยวชาญหรือผู้ที่เกี่ยวข้องสามารถปฏิบัติได้โดยไม่ต้องใช้เวลาในการเรียนรู้มาก รวมทั้งใช้สถิติในการวิเคราะห์ที่ไม่ซับซ้อน ส่วนข้อบกพร่อง คือ ต้องใช้เวลามากในการพิจารณาโดยเฉพาะในกรณีที่มีข้อสอบจำนวนข้อมากๆ ซึ่งหากสามารถปรับปรุงวิธี Angoff เพื่อลดข้อบกพร่องนี้ เช่น การสุ่มข้อสอบเพียงบางส่วนมาพิจารณาโดยที่ยังคงกำหนดคะแนนจุดตัดได้เท่ากับการกำหนดจากแบบสอบฉบับเต็มจะเป็นประโยชน์มาก

วิธี Bookmark

นับตั้งแต่ Lewis and Green (as cited in Zieky et al., 2008) ได้เสนอวิธี Bookmark ครั้งแรกในปี ค.ศ. 1996 ก็ได้รับความนิยมอย่างรวดเร็ว โดยเฉพาะในปี ค.ศ. 2000 วิธี Bookmark ถูกใช้อย่างแพร่หลายในสหรัฐอเมริกา จากการสำรวจพบว่าวิธี Bookmark ถูกนำไปใช้ใน 18 รัฐ ซึ่งถือว่าถูกใช้มากที่สุดและวิธี Angoff ถูกนำไปใช้ 6 รัฐ (Nellhaus, 2000) ต่อมาในปี ค.ศ. 2001 วิธี Bookmark ก็ได้ถูกนำมาใช้เพื่อกำหนดมาตรฐานมากกว่า 28 รัฐ (Egan, as cited in

Lin, 2008) และปี ค.ศ. 2005 มี 31 รัฐ ที่ใช้วิธี Bookmark (Perie, 2005 cited in Lee & Lewis, 2008) Karantonis and Sireci (2006) กล่าวว่า ในแวดวงของการวัดผลการศึกษานั้น วิธี Bookmark เป็นวิธีที่นิยมใช้มากที่สุด

วิธี Bookmark สามารถใช้ได้ทั้งข้อสอบที่เป็นแบบเลือกตอบที่ให้คะแนน 2 ค่า และข้อสอบแบบให้คะแนนมากกว่า 2 ค่า วิธีนี้มีจุดเด่นอยู่ที่การจัดทำคู่มือเรียงข้อสอบ (Ordered Item Booklet: OIB) ซึ่งจัดทำขึ้นโดยการเรียงข้อสอบจากข้อง่ายไปข้อยาก โดยใช้ทฤษฎีการตอบสนองข้อสอบ (Item response theory: IRT) ช่วยในการวิเคราะห์หาค่าความยากของข้อสอบ ซึ่งเป็นการทำให้ความสามารถของนักเรียนอยู่ในระดับเดียวกัน (same continuum)

วิธี Bookmark จะดำเนินการกำหนดคะแนนจุดตัด 3 รอบ ดังนี้

รอบที่ 1 ผู้เชี่ยวชาญแต่ละท่านพิจารณาคู่มือเรียงข้อสอบในแต่ละหน้า และพิจารณาในประเด็นต่อไปนี้ 1) ผู้สอบต้องมีความรู้ความสามารถในเรื่องใดจึงจะสามารถตอบข้อสอบแต่ละข้อได้ถูกต้อง และ 2) เหตุใดข้อสอบข้อนี้จึงยากกว่าข้อก่อนๆ จากนั้นผู้เชี่ยวชาญกำหนดคะแนนจุดตัดอย่างอิสระโดยนำที่คั่นหนังสือมาคั่นที่ข้อสอบซึ่งผู้เชี่ยวชาญคิดว่ามีความเหมาะสมที่จะเป็นจุดตัด โดยข้อที่อยู่ก่อนหน้าที่คั่นหนังสือจะเป็นข้อที่ผู้ที่มีความสามารถขั้นต่ำของระดับความสามารถที่กำลังพิจารณาอยู่ มีความน่าจะเป็น 0.67 ที่จะตอบข้อสอบได้ถูกต้อง โดยผู้เชี่ยวชาญจะพิจารณาจากระดับความสามารถต่ำที่สุดไปจนถึงระดับความสามารถที่สูงที่สุด

รอบที่ 2 ผู้เชี่ยวชาญร่วมกันอภิปรายกลุ่มย่อยเกี่ยวกับคะแนนจุดตัดที่กำหนดในรอบที่ 1 และกำหนดคะแนนจุดตัดอีกครั้ง โดยอาศัยหลักการเดียวกับรอบที่ 1

รอบที่ 3 ผู้เชี่ยวชาญร่วมกันอภิปรายกลุ่มใหญ่เกี่ยวกับคะแนนจุดตัดที่กำหนดในรอบที่ 2 และกำหนดคะแนนจุดตัดอีกครั้ง โดยอาศัยหลักการเดียวกับรอบที่ 1

จากการศึกษางานวิจัยเกี่ยวกับการกำหนดมาตรฐานด้วยวิธี Bookmark พบว่า มีการกำหนดความน่าจะเป็นในการตอบข้อสอบถูก (RP) และการเลือกใช้โมเดลการตอบสนองข้อสอบที่หลากหลาย ซึ่งผู้วิจัยได้สังเคราะห์ข้อมูลและสรุปรายละเอียดไว้ดังแสดงในตารางที่ 3

จากการศึกษาสรุปได้ว่า ในการวิเคราะห์ข้อสอบแบบให้คะแนน 2 ค่า ควรใช้โมเดลการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ (three parameter model) (Lewis as cited in Lee & Lewis, 2008) ส่วนการวิเคราะห์ข้อสอบแบบให้คะแนนคำตอบถูกบางส่วนนั้นควรวิเคราะห์ด้วยโมเดลการตอบสนองข้อสอบแบบ Partial credit model 2 พารามิเตอร์ นอกจากนี้ Cizek, Bunch and Koons (2004) เสนอว่า การวิเคราะห์ข้อสอบด้วยโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ ก็สามารถใช้ในการวิเคราะห์ข้อสอบทั้งที่เป็นแบบให้คะแนน 2 ค่าและให้คะแนนมากกว่า 2 ค่าได้ดี ส่วนการกำหนดความน่าจะเป็นของการตอบถูก (RP)นั้น ส่วนใหญ่จะกำหนดเป็น 0.67 หรือ 2 ใน 3 ของผู้ที่มีความสามารถขั้นต่ำจะสามารถตอบข้อสอบได้ถูก โดยมีเหตุผลของการกำหนดความน่าจะเป็นค่านี้ด้วย 2 เหตุผล คือ 1) สำหรับโมเดลการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ฟังก์ชั่นสารสนเทศของข้อสอบจะมีค่าสูงสุดที่ สำหรับ เมื่อไม่พิจารณาโอกาสการเดาถูก และ 2) การกำหนดความน่าจะเป็นของการตอบถูกไว้ที่ 0.67 นั้นทำความเข้าใจและปฏิบัติได้ง่าย (Williams & Schulz (as cited in Yin & Sconing, 2008) ส่วน Wang (as cited in Beretvas, 2004) เสนอว่า หากนำโมเดลการวิเคราะห์ข้อสอบแบบ 1 พารามิเตอร์ (Rash model) มาใช้ในการวิเคราะห์ข้อสอบ ฟังก์ชั่นสารสนเทศของข้อสอบจะมีค่าสูงสุดเมื่อโอกาสการตอบถูกมีค่าเท่ากับ 0.50

วิธี Bookmark มีจุดเด่น คือ สามารถกำหนดมาตรฐานได้ทั้งกับข้อสอบที่ให้คะแนน 2 ค่าและ

ตารางที่ 3 สรุปการศึกษาวิจัยเกี่ยวกับการกำหนดคะแนนมาตรฐานโดยวิธี Bookmark

นักวิชาการ	วิชา	ระดับความสามารถ	ข้อสอบ	RP	IRT
Buckendahl, Smith, Impara, and Plak (2002)	คณิตศาสตร์	2 ระดับ	ข้อสอบแบบเลือกตอบ	0.67	three-parameter logistic
Beretvas (2004)	คณิตศาสตร์	4 ระดับ	- ข้อสอบแบบเลือกตอบ - คำถามปลายเปิด	0.50 และ 0.80	- Rash model - Partial credit - Generalize Partial Credit - Graded Response
Lewis (2006)	คณิตศาสตร์และภาษาอังกฤษ	2 ระดับ	ข้อสอบแบบเลือกตอบ	0.67	three-parameter logistic
Lee and Lewis (2008)	คณิตศาสตร์	4 ระดับ	ข้อสอบแบบเลือกตอบ	0.67	three-parameter logistic
Yin and Sconing (2008)	คณิตศาสตร์	3 ระดับ	ข้อสอบแบบเลือกตอบ	0.67	three-parameter logistic
MacCann (2008)	คณิตศาสตร์และภาษาอังกฤษ	2 ระดับ	ข้อสอบแบบเลือกตอบ	0.67	three-parameter logistic
Lissitz and Kroopnick (2008)	คณิตศาสตร์และภาษาอังกฤษ	2 ระดับ	ข้อสอบแบบเลือกตอบ	0.50	Rash model

แบบให้คะแนนมากกว่า 2 ค่า และข้อสอบถูกเรียงไว้ตามความยากซึ่งทำให้การพิจารณาทำได้ง่ายขึ้น ส่วนข้อด้อย คือ ต้องใช้ข้อสอบจำนวนมากมิฉะนั้นอาจเกิด floor and ceiling effect

การกำหนดมาตรฐานเพื่อการประเมินทางการศึกษาในประเทศไทย

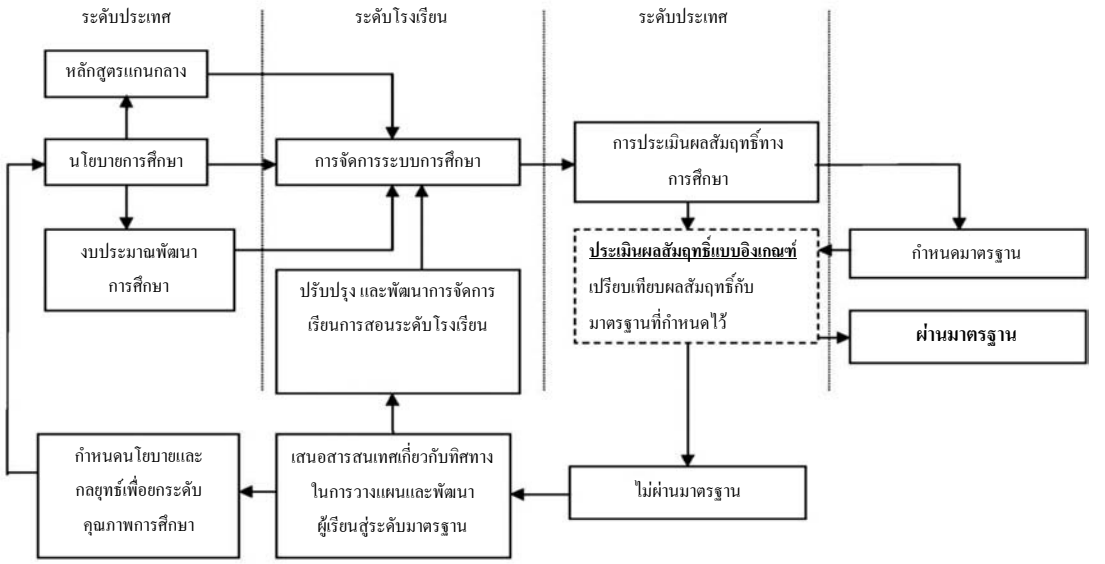
นับจากการปฏิรูปการศึกษาทั้งระบบตามความมุ่งหมายในการจัดการศึกษาที่ประกาศไว้ในพระราชบัญญัติการศึกษาแห่งชาติ พ.ศ. 2542 และที่แก้ไขเพิ่มเติม (ฉบับที่ 2) พ.ศ. 2545 คุณภาพผู้เรียนยังคงเป็นปัญหาที่ต้องเร่งพัฒนาปรับปรุง เนื่องจากผู้เรียนมีผลสัมฤทธิ์ทางการศึกษาต่ำ ขาดคุณลักษณะที่พึงประสงค์ทั้งการคิด วิเคราะห์ ใฝ่เรียนรู้และแสวงหาความรู้อย่างต่อเนื่อง (สำนักงานเลขาธิการสภาการศึกษา, 2552) เห็นได้จากผลการทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน ปี 2548-2552 พบว่า นักเรียนชั้นมัธยมศึกษาปีที่ 6 มีคะแนนเฉลี่ยต่ำกว่าร้อยละ 50 ใน 5 วิชาหลัก คือ ภาษาไทย สังคมศึกษา คณิตศาสตร์ วิทยาศาสตร์และภาษาอังกฤษ มาโดยตลอด (สำนักงานเลขาธิการสภาการศึกษา, 2551ก) นอกจากนี้ผลการประเมินในระดับ

นานาชาติ วิชาคณิตศาสตร์และวิทยาศาสตร์ ในโครงการประเมินผลการศึกษานักเรียนนานาชาติ พ.ศ. 2549 (Program for International Student Assessment : PISA 2006) และโครงการศึกษาแนวโน้มการจัดการศึกษาคณิตศาสตร์และวิทยาศาสตร์ พ.ศ. 2550 (Third International Mathematics Science Study: TIMSS) ประเทศไทยได้คะแนนเฉลี่ยต่ำกว่าหลาย ๆ ประเทศ ยกเว้นอินโดนีเซียเพียงประเทศเดียว (สำนักงานเลขาธิการสภาการศึกษา, 2551ข) ซึ่งแสดงให้เห็นว่าการดำเนินการต่างๆ เพื่อเพิ่มผลสัมฤทธิ์ทางการเรียนยังไม่ได้ผล

การตรวจสอบและพัฒนาคุณภาพผู้เรียนตลอดจนการปรับปรุงคุณภาพการศึกษาถือเป็นภารกิจและนโยบายสำคัญที่ผู้มีส่วนเกี่ยวข้องต่างเร่งดำเนินการและพยายามที่จะพัฒนามาโดยลำดับ หนึ่งในวิธีการคือการจัดให้มีการทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ เรียกว่า การทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน (Ordinary National Education Testing: O-NET) ซึ่งเป็นแบบทดสอบมาตรฐานสำหรับประเมินคุณภาพผู้เรียนในระดับชาติตามมาตรฐานการเรียนรู้ตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน โดยสถานศึกษาต้องจัดให้ผู้เรียนทุกคนที่เรียนในชั้นประถมศึกษาปีที่ 3 ชั้นประถมศึกษาปีที่ 6 ชั้นมัธยมศึกษาปีที่ 3 และชั้นมัธยมศึกษาปีที่ 6 เข้ารับการประเมินทุกคน เป้าหมายสำคัญคือเพื่อตรวจสอบผลการจัดการศึกษา และใช้เป็นข้อมูลสำหรับการวางแผนเพื่อดำเนินการปรับปรุงหลักสูตร วิธีการสอน และส่งเสริมให้เกิดความเสมอภาค โดยสารสนเทศที่ได้จากการสอบจะเป็นประโยชน์หรือไม่ก็ขึ้นอยู่กับวิธีการตีความหมายคะแนนสอบ สำหรับวิธีการแปลความหมายคะแนนสอบแบบอิงกลุ่ม ซึ่งเป็นการแปลความหมายคะแนนโดยเปรียบเทียบคะแนนของผู้สอบด้วยกัน ประกอบด้วย ค่าเฉลี่ย ค่าสูงสุด ค่าต่ำสุด ค่าเปอร์เซ็นต์ไทล์ ค่าการกระจาย เป็นต้น ถือว่าเป็นสารสนเทศที่ยังมีรายละเอียดไม่เพียงพอต่อการนำใช้ กล่าวคือยังขาดความชัดเจนเกี่ยวกับความรู้และทักษะที่ผู้สอบแต่ละคนมี (Green, 1995) ซึ่งการแปลผลเพื่อให้ได้สารสนเทศเกี่ยวกับความรู้และทักษะของผู้สอบแต่ละคน ควรแปลความหมายของคะแนนสอบแบบอิงเกณฑ์ กล่าวคือคะแนนสอบจะถูกแปลความหมายโดยอิงกับเกณฑ์หรือมาตรฐานที่กำหนดไว้ (Cizek & Bunch, 2007; Zieky et al., 2008) ซึ่งสารสนเทศที่ได้จะสามารถนำไปพิจารณาเพื่อหาปัญหาและจุดอ่อนของผู้สอบสำหรับแสวงหาข้อเสนอแนะที่เป็นประโยชน์แก่ครูผู้สอนในแง่วิธีการสอน ความครอบคลุมของหลักสูตรและจุดเน้นในกระบวนการเรียนการสอน รวมถึงใช้เพื่อเป็นข้อมูลในการเทียบเคียงคุณภาพการศึกษาในระดับต่าง ๆ และนำไปใช้ในการวางแผนยกระดับคุณภาพการจัดการศึกษา ตลอดจนเป็นข้อมูลสนับสนุนการตัดสินใจในระดับนโยบายของประเทศ

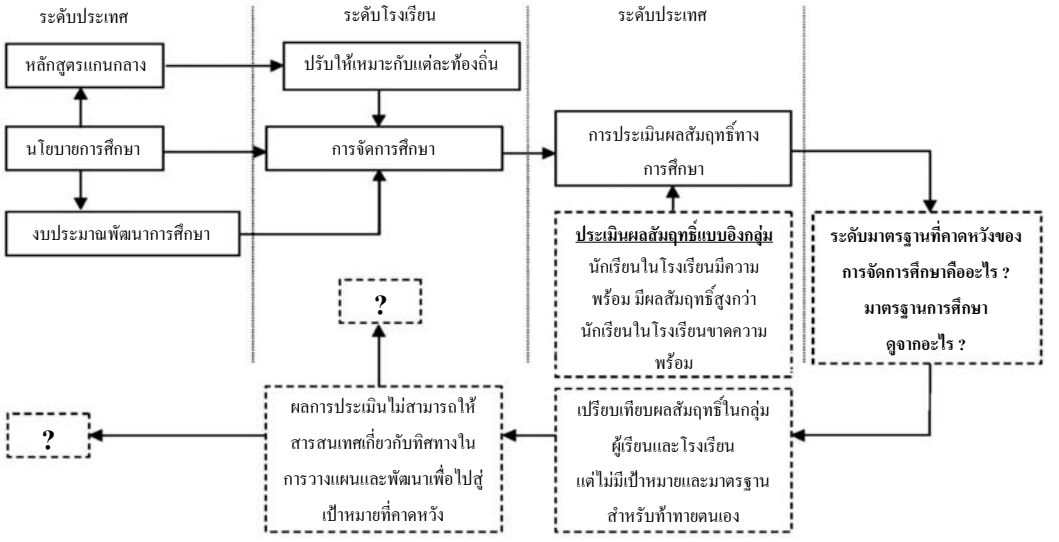
จากการศึกษา พบว่า การทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐานยังไม่มีการกำหนดมาตรฐานสำหรับระดับมัธยมศึกษาปีที่ 3 และ 6 กล่าวคือยังไม่มีการกำหนดเป้าหมายสำหรับท้าทายตัวเองเพื่อให้ไปถึง ทำให้ผลการประเมินไม่สามารถให้สารสนเทศเกี่ยวกับทิศทางในการวางแผนและพัฒนาเพื่อไปสู่ป้าหมายที่คาดหวัง กล่าวคือการนำเสนอผลสอบยังคงแปลความหมายคะแนนสอบแบบอิงกลุ่ม ซึ่งเป็นการแปลความหมายคะแนนโดยการเปรียบเทียบคะแนนของนักเรียนด้วยกันเอง (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2550) ซึ่งจะทำให้ได้สารสนเทศที่มีรายละเอียดไม่เพียงพอต่อการนำใช้ปรับปรุงและพัฒนาผู้เรียน กล่าวคือยังขาดความชัดเจนเกี่ยวกับความรู้และทักษะที่ผู้สอบแต่ละคนมี เพื่อแสดงให้เห็นความแตกต่างของบทบาทของการทดสอบทางการศึกษาและการแปลความหมายคะแนนสอบที่มีต่อการวางแผนและพัฒนาการศึกษาระหว่างต่างประเทศกับประเทศไทย ผู้วิจัยขอนำเสนอเป็นแผนภาพ ดังแสดงในภาพที่ 2 และ 3 ตามลำดับ

ต่างประเทศ



ภาพที่ 2 บทบาทของการทดสอบทางการศึกษาและการแปลความหมายคะแนนสอบที่มีต่อการวางแผนและพัฒนาการศึกษาของต่างประเทศ

ประเทศไทย



ภาพที่ 3 บทบาทของการทดสอบทางการศึกษาและการแปลความหมายคะแนนสอบที่มีต่อการวางแผนและพัฒนาการศึกษาของประเทศไทย

จากที่กล่าวมา การทดสอบทางการศึกษาขั้นพื้นฐานควรมีการกำหนดมาตรฐานเพื่อใช้เป็นเกณฑ์ในการประเมินผลการจัดการเรียนการสอน โดยควรเริ่มต้นจากการกำหนดจำนวนของระดับมาตรฐานและชื่อของระดับที่เหมาะสมก่อน จากการทบทวนเอกสารและงานวิจัยเกี่ยวกับการกำหนดจำนวนของระดับมาตรฐาน พบว่า การกำหนดจำนวนของระดับมาตรฐานจะกำหนดโดยพิจารณาจากเป้าหมายของการนำผลไปใช้ หากเป้าหมายเพื่อต้องการนำผลไปใช้ในการรับรอง หรือ อนุมัติ (licensure) จะกำหนดระดับมาตรฐานเพียง 2 ระดับ คือ ผ่านและตก (pass / fail) ก็เพียงพอ แต่หากมีเป้าหมายเพื่อทำนาย บรรยาย หรือเพื่อกระตุ้น ก็ควรกำหนดจำนวนระดับมาตรฐานเป็นหลายระดับเพื่อใช้จำแนก (classified) ผู้สอบตามระดับมาตรฐาน กล่าวคือควรกำหนดระดับมาตรฐานมากกว่า 2 ระดับ เช่น 3, 4 หรือ 5 ระดับ เป็นต้น (Woehr, Arthur, & Fehrmann, 1991, Giraud et al., 2005, Perie, 2008)

จากที่กล่าวมาข้างต้นพอสรุปได้ว่า การทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน เป็นการทดสอบที่มีเป้าหมายสำคัญเพื่อตรวจสอบผลการจัดการศึกษา ดังนั้นผลการสอบควรจะสามารถใช้เพื่ออธิบายระดับความสามารถของผู้สอบอันเป็นการสะท้อนถึงระดับคุณภาพของการจัดการศึกษา ดังนั้นการจะกำหนดระดับมาตรฐานควรกำหนดมากกว่า 2 ระดับจึงจะเหมาะสม

สำหรับแนวทางการกำหนดชื่อของระดับมาตรฐานนั้น จากการศึกษาเอกสารและงานวิจัยเกี่ยวกับการกำหนดชื่อของระดับมาตรฐาน พบว่า ชื่อของระดับต้องมีความชัดเจน (clear) เหมาะสม (reasonable) ใช้ภาษาที่ไม่ขัดแย้งกันเอง และไม่เป็นชื่อที่คลุมเครือ (nebulous) โดยควรกำหนดให้สอดคล้องกับวัตถุประสงค์ของการวัด (purpose) โครงสร้างของการวัด (construct) ควรเป็นชื่อที่สนับสนุนการสรุปผล (inferences) และการกำหนดกลุ่ม (classifications) (Cizek and Bunch, 2007; Perie, 2008) นอกจากนี้ควรใช้คำกลาง ๆ ในการตั้งชื่อระดับ เช่น ความสามารถระดับที่ 1, ระดับที่ 2 และระดับที่ 3 เป็นต้น เพื่อหลีกเลี่ยงการแปลความหมายตามคำที่ใช้จนเกินที่ต้องการสื่อจริง (Zieky et al., 2008)

เมื่อกำหนดจำนวนของระดับมาตรฐานและชื่อที่เหมาะสมแล้วควรกำหนดคำอธิบายสำหรับแต่ละระดับมาตรฐาน เรียกว่า คำอธิบายระดับความสามารถ (Performance Level Descriptor: PLD) จากการศึกษาเอกสารและงานวิจัยเกี่ยวกับการกำหนดคำอธิบายระดับความสามารถ พบว่า คำอธิบายระดับความสามารถเป็นการเขียนบรรยายเกี่ยวกับความรู้และทักษะที่ผู้สอบในแต่ละระดับมาตรฐานควรจะมีหากเป็นไปได้ควรเขียนอธิบายระดับความสามารถให้เสร็จก่อนที่จะดำเนินการสร้างข้อสอบ ทั้งนี้เนื่องจากว่าหากผู้ที่สร้างข้อสอบทราบว่าระดับมาตรฐานแต่ละระดับจะต้องสามารถทำอะไรหรือรู้อะไรบ้าง ก็จะสามารถสร้างข้อคำถามเพื่อวัดความสามารถในระดับต่างๆ ได้อย่างถูกต้องและครอบคลุมทุกระดับมาตรฐานที่กำหนด (Perie, 2008; Zieky et al., 2008) การเขียนคำอธิบายระดับความสามารถต้องเขียนให้มีความชัดเจน เนื่องจากคำอธิบายระดับความสามารถถือเป็นปัจจัยสำคัญปัจจัยหนึ่งที่ส่งผลต่อคุณภาพของคะแนนจุดตัด กล่าวคือหากเขียนไม่ชัดเจนย่อมส่งผลให้การกำหนดคะแนนจุดตัดผิดพลาดได้ (Ferdous and Plake, 2005; Perie, 2006) เมื่อได้จำนวนของระดับมาตรฐาน ชื่อของระดับมาตรฐาน และมีคำอธิบายแต่ละมาตรฐานที่เหมาะสมแล้ว ขั้นต่อไปก็จะเป็นการกำหนดคะแนนจุดตัด ซึ่งต้องเลือกวิธีการที่เหมาะสม การเลือกวิธีการกำหนดคะแนนจุดตัดมี 2 แนวคิด กล่าวไว้โดย Zieky et al. (2008) แนวคิดแรกเสนอให้เลือกวิธีการกำหนดคะแนนจุดตัดเพียงวิธีเดียวแล้วดำเนินการตามวิธีนั้นอย่างเคร่งครัดและดีที่สุดเท่าที่จะทำได้ ส่วนแนวคิดที่สอง เสนอให้เลือกหลาย ๆ วิธี หรืออย่างน้อย 2 วิธี เพื่อให้ได้คะแนนจุดตัดที่มีความถูกต้องและเหมาะสมที่สุด และหากเกิดกรณีที่คะแนนจุดตัดที่

กำหนดด้วยวิธีที่ต่างกันมีค่าแตกต่างกัน Green, Trimble, and Lewis (2003) ก็ได้เสนอแนวทางในการแก้ปัญหาดังกล่าวไว้ 3 แนวทาง โดยแนวทางแรกให้พิจารณาเลือกคะแนนจุดตัดที่กำหนดจากวิธีที่ผู้วิจัยมั่นใจว่าได้มีการดำเนินการตามขั้นตอนอย่างเคร่งครัดและถูกต้องที่สุด ซึ่งอาจพิจารณาจากผลการตรวจสอบคุณภาพของคะแนนจุดตัดเป็นต้น แนวทางที่สอง คือการหาค่าเฉลี่ยของคะแนนจุดตัดจากทุก ๆ วิธีที่กำหนดได้ เพื่อเป็นตัวแทนของทุกวิธี และแนวทางที่สาม คือการให้ผู้เชี่ยวชาญที่ไม่ได้เข้าร่วมกำหนดคะแนนจุดตัดวิธีใดวิธีหนึ่งที่กำหนดไว้พิจารณาคะแนนจุดตัดทั้งหมดและให้ตัดสินใจเลือกคะแนนจุดตัดที่เหมาะสมที่สุด

สรุปและข้อเสนอแนะ

การกำหนดมาตรฐานถือว่ามีความสำคัญอย่างมากต่อการประเมินผลการศึกษา กล่าวคือมาตรฐานที่กำหนดขึ้นสามารถนำมาใช้เป็นเกณฑ์สำหรับตีความคะแนนจากการสอบและเป็นเป้าหมายสำหรับท้าทายให้ไปให้ถึง ในการกำหนดมาตรฐานโดยเฉพาะการกำหนดมาตรฐานสำหรับการสอบที่มีความสำคัญและมีผลกระทบสูง อย่างเช่น การทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานของประเทศไทยควรกำหนดระดับมาตรฐานมากกว่า 2 ระดับเพื่อใช้ในการอธิบายระดับความสามารถของผู้สอบอันเป็นการสะท้อนถึงระดับคุณภาพของการจัดการศึกษา ในการตั้งชื่อระดับควรใช้คำกลาง ๆ เช่น ความสามารถระดับที่ 1, ระดับที่ 2 และระดับที่ 3 เป็นต้น เพื่อหลีกเลี่ยงการแปลความหมายตามคำที่ใช้จนเกินที่ต้องการสื่อจริง นอกจากนี้ควรเขียนอธิบายระดับความสามารถให้เสร็จก่อนที่จะดำเนินการสร้างข้อสอบ เพื่อให้สามารถสร้างข้อคำถามเพื่อวัดความสามารถในระดับต่างๆ ได้อย่างถูกต้องและครอบคลุมทุกระดับมาตรฐานที่กำหนดไว้ การเลือกใช้วิธีกำหนดคะแนนจุดตัดเสนอให้เลือกหลาย ๆ วิธี หรืออย่างน้อย 2 วิธี เพื่อให้ได้คะแนนจุดตัดที่มีความถูกต้องและเหมาะสมที่สุด ซึ่งจะทำให้คะแนนจุดตัดที่ได้เป็นที่ยอมรับสำหรับผู้ที่เกี่ยวข้อง และได้คะแนนจุดตัดสำหรับการแปลความหมายคะแนนสอบที่เหมาะสม รวมทั้งมีเป้าหมายที่ท้าทายสำหรับการศึกษาไทย และได้สารสนเทศที่เป็นประโยชน์ต่อการวางแผนเพื่อยกระดับการศึกษาไทยให้มีมาตรฐานทัดเทียมต่างประเทศต่อไป

กิตติกรรมประกาศ

บทความนี้เป็นส่วนหนึ่งของงานวิจัยที่ได้รับทุนสนับสนุนจาก “ทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย” กองทุน รัชดาภิเษก สมโภช

เอกสารอ้างอิง

สำนักงานเลขาธิการสภาการศึกษา. (2551ก). *สภาวะการศึกษาไทยปี 2550/2551 “ปัญหาความเสมอภาคและคุณภาพของการศึกษาไทย”*. กรุงเทพฯ: สกศ.

สำนักงานเลขาธิการสภาการศึกษา. (2551ข). *สมรรถนะการศึกษาไทยในเวทีโลก พ.ศ. 2551*. กรุงเทพฯ: สกศ.

สำนักงานเลขาธิการสภาการศึกษา. (2552). *ข้อเสนอการปฏิรูปการศึกษาในทศวรรษที่สอง (พ.ศ. 2552-2561)*. กรุงเทพฯ: สกศ.

สถาบันทดสอบทางการศึกษาแห่งชาติ. (2550). *คู่มือสำหรับผู้บริหารในการนำผลการสอบ O-NET ไปวางแผนปรับปรุงการจัดการเรียนการสอนและพัฒนาครู-นักเรียน*. กรุงเทพฯ: สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน).

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

Berk, R. A. (1980). *Criterion-referenced testing: State of the art*. Baltimore: John Hopkins University Press.

Berk, R. A. (1996). Standard setting: The next generation (Where few psychometricians have gone before!). *Applied Measurement in Education*, *9*(3), 215–235.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plak, B. S. (2002). A Comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253–263.

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

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education*, *22*, 1–21.

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18*(3), 257–267.

Ferdous, A. A., & Plake, B. S. (2008). Item response theory–based approaches for computing minimum passing scores from an Angoff-based standard - setting study. *Educational and Psychological Measurement, 68*(5), 778–796.

Fowell, S. L., Fewtrell, R., & McLaughlin, P. J. (2008). Estimating the minimum number of judges required for test-centred standard setting on written assessments: Do discussion and iteration have an influence? *Advances in Health Sciences Education, 13,* 11–24.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18*(3), 223–232.

Gomez, P. G. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*(3): 417–444.

Green, B. F. (1995). S*etting performance standard: content, goals, individual differences*. Princeton, NJ: Educational Testing Service.

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

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

Hess, B., Subhiyah, R. G., & Giordano, C. (2007). Convergence between cluster analysis and the Angoff method for setting minimum passing scores on credentialing examinations. *Evaluation & the Health Professions, 30*(4), 362–375.

Hurtz, G. M., Jones, J. P., & Jones, C. N. (2008). Conversion of proportion-correct standard-setting judgments to cutoff scores on the item response theory θ scale. *Applied Psychological Measurement, 32*(5), 385–406.

Kane, M. (1994). Validating the performance standards associated with passing score. *Review of Educational Research, 64*(3), 425–461.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12.

Lee, G., & Lewis, D. M. (2008). A generalizability theory approach to standard error estimates for bookmark standard settings. *Educational and Psychological Measurement, 68*(4), 603–620.

Lewis, D. M. (2006). *Standard setting for grades 3-8 mathematics and English language arts*. Retrieved from http://www.nysut.org/files/k12_071210_assessments03_standards.pdf.

Lin, J. (2008). *The bookmark standard setting procedure: Strengths and weaknesses*. Retrieved from http://www.education.ualberta.ca/educ/psych/crame/files/standard_setting.pdf.

Lissitz, R. W., & Kroopnick, M. H. (2008). *Running ahead: An adaptive procedure for standard setting*. Retrieved from http://www.education.umd.edu/EDMS/MARCES/presentation/Lissitz_KroopnickFNLNCME07.pdf.

MacCann, R .G. (2008). A modification to Angoff and Bookmarking cutscores to account for the imperfect reliability of test scores. *Educational and Psychological Measurement, 68*, 197–214.

Marion, S., & Pellegrino, J. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*(4), 47–57.

Näsström, G., & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research & Evaluation, 13*(9), 1–12.

Nellhaus, J. M. (2000). *States with NAEP like performance standard*. Retrieved from http://www.nagb.org/pubs/studentperfstandard.pdf.

Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education, 37*(5), 464– 469.

Perie, M. (2006). Standard setting. Presentation at the inclusive assessment seminar. Denver CO / Alexandria, VA [Online]. 2008. Retrieved from http://www.naacpartners.org/products/workshops/presentations/16210.pdf.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15–29.

Plake, B. S. (2008). Standard setters: Stand up and take a stand!. *Education Measurement: Issues and Practice, 27*(1), 3–9.

Towles-Reeves, E. (2008). *Standard setting approaches for alternate assessment*. National Alternate Assessment Center. Retrieved from http://www.ccsso.org/content/PDFs/190_Towles-Reeves.pdf.

Woehr, D. J., Arthur, W., & Fehrmann, M. L. (1991). An empirical comparison of cutoff score methods for content-related and criterion-related validity settings. *Educational and Psychological Measurement, 51,* 1029–1039.

Yin, P., & Sconing, J. (2008). Estimating standard errors of cutscores for item rating and mapmark procedures a generalizability theory approach. *Educational and Psychological Measurement, 68*, 25–41.

Yudkowsky, R. (2008). Simpler standards for local performance examinations: The yes/no Angoff and whole-test Ebel. *Teaching and Learning in Medicine, 20*(3), 212–217.

Zieky, M. J., & Perie, M. (2006). *A primer on setting cutscores on tests of educational achievement*. Princeton, NJ: Educational Testing Service.

Zieky, M. J., Perie, M., & Livingston, S. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

TRANSLATED THAI REFERENCES

Office of the Education Council (ONEC). (2008a). *The stage of Thai education 2007-2008 "equality issues and quality of Thai education"*. Bangkok: Office of Educational Council. [in Thai]

Office of the Education Council (ONEC). (2008b). *The competencies of Thai education on the world class 2008*. Bangkok: Office of Educational Council. [in Thai]

Office of the Education Council (ONEC). (2009). *The education reform proposal in the second decade (2008-2018)*. Bangkok: Office of Educational Council. [in Thai]

National Institute of Educational Testing Service. (2007). *A guide for administrators to use the O-NET test results for the plan to improve teaching and student-teacher*. Bangkok: National Institute of Educational Testing Service. [in Thai]