หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 2

วรรณคดีที่เกี่ยวข้อง

การสอนในโรงเรียนมักเป็นกระบวนการที่เป็นลำดับขั้น นั่นคือสิ่งที่นักเรียนเรียนในขั้นสองจะต้องอาศัยพื้นฐานที่เรียนมาจากขั้นหนึ่ง เรียกว่าเป็นกระบวนการที่ต่อเนื่องกัน กระบวนการดังนี้จะมีประสิทธิภาพดีก็เมื่อสิ่งที่นำมาให้นักเรียนเรียนควรจัดลำดับต่อเนื่องกันอย่างได้ความหมาย ซึ่งเป็นสิ่งที่จำเป็นทั้งในด้านการเตรียมการสอนและการตัดสินความก้าวหน้าของนักเรียนในการเรียนลำดับต่อไป[1]

วิธีการที่จะประเมินความก้าวหน้านี้ได้คือ การทดสอบอิงเกณฑ์ เครื่องมือที่ใช้คือแบบสอบอิงเกณฑ์ (Criterion-or Domain-Referenced Test) ซึ่ง ไฮฟ์ลี่(Hively) วิจารณ์ไว้ว่า

> คำว่า "เกณฑ์" ใน "แบบสอบอิงเกณฑ์" นี้ นำไปสู่การแปลความหมายของข้อมูลอย่างผิด ๆ และ เพื่อหลีกเลี่ยงความเข้าใจผิดดังกล่าวจึงได้เสนอให้ใช้คำว่า "โดเมน" (Domain) แทน โดยที่ในโดเมนหนึ่ง ๆ จะประกอบด้วยชุดย่อย ๆ ของความรู้ ทักษะ ความเข้าใจ หรือทัศนคติซึ่งจะบรรยายให้เห็นคุณสมบัติที่จำเป็นของเนื้อหาที่คาดว่าจะเกิดขึ้นในตัวของนักเรียน และบรรยายให้เห็นวัตถุประสงค์ซึ่งหวังว่านักเรียนจะแสดงออกมาว่าได้รับอะไรไปบ้าง[2]

---

[1] C.M. Lindvall and A.J. Nitko, Measuring Pupil Achievement and Aptitude (New York : Harcourt Brace Jovanovich, 1975), p.14.

[2] Wells Hively et al., "Criterion - Referenced Testing," Educational Technology 14 (1974) : 10.

ได้มีนักวัดและประเมินผลหลายท่านให้คำนิยามของแบบสอบอิงเกณฑ์หรือแบบสอบอิงโดเมนไว้ เช่น มิลแมน[1] ซึ่งใช้คำว่าโดเมนแทนเกณฑ์ กล่าวว่า "การสอนอิงโดเมนเป็นการวัดสภาพที่เป็นปัจจุบันของนักเรียนคนใดคนหนึ่ง โดยอ้างอิงถึงชุดของงานที่ทำที่ได้บรรยายไว้อย่างดีแล้ว เรียกว่า โดเมน แบบสอบอิงโดเมนคือ กลุ่มตัวอย่างของข้อสอบจากโดเมนซึ่งอาจได้มาจากการสุ่ม หรือการสุ่มแยกประเภท" ในขณะเดียวกันเกลเซอร์และนิทโก้[2] นิยามว่า "แบบสอบอิงเกณฑ์เป็นแบบสอบที่สร้างขึ้นโดยมีเจตนาให้เป็นการวัดที่สามารถแปลความหมายออกมาได้ตรง ๆ ในรูปของมาตรฐานการกระทำที่เฉพาะ ซึ่งในที่นี้บ่งชี้ได้โดยให้นิยามโดเมนของงานที่นักเรียนหรือผู้สอบควรทำได้ จากนั้นนำตัวอย่างที่เป็นตัวแทนของงานจากโดเมนนี้มาจัดเป็นแบบสอบ ผลจากการวัดจะแสดงถึงพฤติกรรมของแต่ละคนเพื่อเทียบกับโดเมนที่นิยามไว้"

การวัดอิงเกณฑ์เป็นวิธีการวัดที่รู้จักกันมานานแล้ว แต่ไม่มีการศึกษาถึงการสร้างหรือพัฒนาเครื่องมือที่ใช้ในการวัด จนกระทั่งเมื่อไม่กี่ปีนี้จึงได้มีนักวัดและประเมินผลบางท่านที่พยายามจะศึกษาเรื่องนี้ขึ้น โดยมีแรงจูงใจจากแนวโน้ม 4 ประการ อันเนื่องมาจากการใช้แบบสอบอิงกลุ่มในการวัดและประเมินผลการศึกษาดังที่ แอราเซี่ยน และมาดอสกล่าวไว้คือ

1. การวิพากษ์วิจารณ์ถึงเรื่องความหมายของการวัด รวมทั้งการใช้แบบสอบมาตรฐานที่ให้ข้อมูลไม่เพียงพอและไม่เหมาะสมกับเนื้อหาบางประการที่ต้องการวัด

---

[1]Jason Millman, "Criterion-Referenced Measurement," in Evaluation in Education, ed. W.J. Popham (Berkeley : McCutchan Publishing Corparation, 1974), p.327.

[2]R. Glaser and A.J. Nitko, "Measuring in Learning and Instruction," in Educational Measurement, ed. R.L. Thorndike (Washington : American Council on Education, 1971), p.653.

2. การวิพากษ์วิจารณ์เกี่ยวกับระดับคะแนน (Grade) ว่า ทำให้จุดหมายของการศึกษาเปลี่ยนไป กลายเป็นการแข่งขันกันเพื่อจะให้ได้ระดับคะแนนสูงๆ และกลายเป็นสิ่งต่อรองในการเข้าศึกษา การสมัครงาน ฯลฯ จบลงด้วยว่า ระดับคะแนนไม่ให้ข้อมูลว่า เด็กทำอะไรได้บ้าง บอกแต่ว่า เขาดีกว่าหรือด้อยกว่าผู้อื่นในกลุ่มของเขาเท่านั้น

3. จากวิวัฒนาการการเคลื่อนไหวทางเทคนิคการสอนต่าง ๆ ที่ทำให้พบว่า แบบสอบอิงกลุ่มไม่เพียงพอในการที่จะใช้ประเมินพฤติกรรมเป็นรายบุคคล หรือแม้แต่วิธีการสอนที่ใช้ ว่าดี หรือไม่ เพียงใด

4. ความเชื่อของนักการศึกษาส่วนมากที่ว่า คนทุกคนหรือคนส่วนมากจะสามารถเรียนรู้หรือสัมฤทธิ์ในวิชาต่าง ๆ ได้ ถ้ามีวิธีการที่ดีกว่า จึงทำให้การทดสอบเปลี่ยนจากการเปรียบเทียบในกลุ่มเป็นการตรวจสอบ และให้รางวัลการเรียนในรูปของการกระทำ (เช่น ถ้าทุกคนทำได้ถึงเกณฑ์ ทุกคนอาจได้คะแนนระดับ "ก")[1]

กล่าวโดยสรุปคือ การวัดอิงกลุ่มนั้นไม่เหมาะจะนำมาใช้วัดสัมฤทธิผลทางการศึกษาบางอย่าง ทั้งนี้เนื่องจากว่า บอกเราได้ไม่ครบเท่าที่เราต้องการ นั่นคือใช้แบบสอบไม่ตรงกับจุดประสงค์ของการทดสอบที่เราต้องการกระทำ

กรอนลันด์ (Gronlund) ได้เสนอการตัดสินใจในเรื่องการใช้แบบสอบ โดยพิจารณาจากวัตถุประสงค์ของการสอบ และลักษณะของแบบสอบวัดสัมฤทธิผล 4 ชนิด ดังตาราง[2]

---

[1] Airasian, "Criterion-Referenced Testing in the Classroom," pp. 76-77.

[2] Norman E. Gronlund, Constructing Achievement Test (N.J.: Prentice-Hall, 1977), p.19.

ตารางที่ 1 ลักษณะของแบบสอบวัดสัมฤทธิ์ผล 4 ชนิด

ชนิดของแบบสอบ	หน้าที่ของแบบสอบ	การเลือกตัวอย่างข้อสอบ	ลักษณะข้อสอบ
วางตัวบุคคล (Placement)	วัดทักษะเบื้องต้น ตัดสินการกระทำเริ่มแรกเกี่ยวกับวัตถุประสงค์ของวิชา	รวมไว้ด้วยพฤติกรรมที่ต้องการวัดเบื้องต้นก่อนเรียน เลือกตัวอย่างที่เป็นตัวแทนของวัตถุประสงค์ของวิชา	โดยทั่วไปจะง่ายและแปลผลแบบอิงเกณฑ์ มีความยากง่ายต่าง ๆ (มีพิสัยของความยากกว้าง) และอิงกลุ่ม
วัดความก้าวหน้า (Formative)	ให้ข้อมูลย้อนกลับแก่ผู้เรียนและผู้สอนในด้านความก้าวหน้าในการเรียน	รวมไว้ด้วยทุกหน่วยของวัตถุประสงค์ถ้าเป็นไปได้(หรือวัดสิ่งที่จำเป็นที่สุด)	ข้อสอบยากหรือง่ายแล้วแต่วัตถุประสงค์ของแต่ละหน่วย และอิงเกณฑ์
วัดเพื่อวินิจฉัย (Diagnostic)	เพื่อพิจารณาสาเหตุของความขัดข้องทางการเรียนที่เกิดขึ้นซ้ำๆ	รวมไว้ด้วยตัวอย่างของงาน ซึ่งขึ้นอยู่กับแหล่งของความผิดพลาดในการเรียนที่มีร่วมกันอยู่	โดยทั่วไปจะง่ายและใช้เพื่อชี้จุดบกพร่องของนักเรียน และอิงเกณฑ์
จัดลำดับความรู้รวบยอด (Summative)	ให้ระดับคะแนนหรือรับรองความรอบรู้ท้ายหน่วยการสอน	เลือกตัวอย่างที่เป็นตัวแทนของวัตถุประสงค์	มีพิสัยความยากกว้างและอิงกลุ่ม

จะเห็นว่า กรอนลันด์มีความเห็นว่า วัตถุประสงค์ในการสอบต่างกัน แบบสอบที่ใช้ก็ต่างกันไปด้วย ในขณะที่แบบสอบที่ใช้ในโครงการ I P I (Individually Prescribed Instruction)[1] ซึ่งเป็นแบบสอบที่ใช้วัดพัฒนาการของนักเรียนในโปรแกรมการสอนรายบุคคลวิชาคณิตศาสตร์ และวิชาอื่น ๆ ในชั้นประถมที่มีการจัดลำดับแบบสอบตามวัตถุประสงค์การสอนและตามหน่วยการสอนได้ใช้แบบสอบอิงเกณฑ์เพียงแบบเดียว (หลายฉบับ) เพื่อวัตถุประสงค์ในการวัด 4 ประการ คือ

1. การวางตัวบุคคล (Placement Testing) ใช้แบบสอบเพื่อบอกนักเรียนว่า เขาควรจะเริ่มเรียนชั้นใด ระดับใด ของหลักสูตรโดยให้นักเรียนทำแบบสอบในตอนเริ่มปีการศึกษา

2. สอบเพื่อวินิจฉัย (Diagnostic Testing) ใช้แบบสอบก่อนเรียน (Pre-test) ที่เตรียมไว้ให้นักเรียนทำก่อนที่จะเริ่มเรียนหน่วยเรียนใหม่ เนื้อหาเป็นเนื้อหาที่นักเรียนถูกตัดสินให้ผ่านมาแล้ว

3. สอบหลังจากเรียน (Post-testing) นักเรียนจะเลือกได้ด้วยตนเองว่าจะเรียนหน่วยใดต่อไป เมื่อผ่านแต่ละหน่วยแล้ว เช่น เมื่อจบหน่วย D อาจเรียนหน่วย E หรือ F ต่อก็ได้ วิธีการก็คือ จะให้นักเรียนทำแบบสอบอีกครั้งหลังจากสอบไปแล้วหลายสัปดาห์ เพื่อดูว่ายังมีความรู้เหลืออยู่หรือไม่ (อาจสุ่มตัวอย่างมาให้ทำเป็นบางข้อ)

4. สอบเพื่อสำรวจความก้าวหน้า (Monitoring Progress) เพื่อตอบคำถามให้แก่นักเรียนที่ว่า "ฉันรอบรู้ทักษะนี้แล้วจริงหรือ?" จะใช้แบบสอบ C E T (Curriculum Embedded Test) ให้ข้อมูลเพิ่มเติมเกี่ยวกับความรอบรู้ของนักเรียน

ในขณะที่กรอนลันด์เสนอว่า ควรใช้แบบสอบอิงกลุ่มวัดความรู้รวบยอดนั้น โครงการ IPI ใช้แบบสอบอิงเกณฑ์วัดความรู้หลังจากเรียน (ซึ่งเป็นความรู้รวบยอดเช่นเดียวกัน) นั่นคือ ใช้แบบสอบอิงเกณฑ์วัดได้ทุกวัตถุประสงค์ที่ต้องการจะวัด

002615

A

[1]Lindvall, Measuring Pupil Schievement..., p. 198.

แฮมเบิลตันและเอ็คเน่อร์ (Hambleton and Eignor)[1] ได้เสนอเกณฑ์ที่จะใช้ในการตัดสินคุณภาพของแบบสอบอิงเกณฑ์ เมื่อต้องการที่จะซื้อแบบสอบอิงเกณฑ์มาใช้ โดยมีแนวทางในการถามตามหัวข้อต่อไปนี้

วัตถุประสงค์

1. ได้มีการแถลงจุดมุ่งหมายของแบบสอบชัดแจ้งหรือไม่
2. แต่ละวัตถุประสงค์ที่แถลงไว้ชัดเจนพอที่จะชี้ให้เห็นถึงจักรวาลของข้อสอบหรือไม่
3. จุดประสงค์แถลงไว้ชัดเจนหรือไม่ว่า แบบสอบนั้นวัดอะไร
4. มีเหตุผลเหมาะสมหรือไม่ ในการบรรจุวัตถุประสงค์แต่ละข้อลงในแบบสอบ
5. ผู้นำแบบสอบไปใช้ที่มีศักยภาพสามารถปรับปรุงแบบสอบให้เหมาะกับความต้องการของท้องถิ่นหรือไม่ โดยเพียงแต่พิจารณาตัดสินว่าวัตถุประสงค์ใดบ้างในแบบสอบที่ควรจะวัดด้วยแบบสอบนั้น
6. มีการระบุเนื้อเรื่องที่จะวัดด้วยแบบสอบนั้น ๆ. และสถานภาพที่จะใช้แบบสอบฉบับนั้นไว้คู่กันหรือไม่
7. ระบุชื่อผู้ที่เขียนวัตถุประสงค์ในแบบสอบหรือไม่
8. ชุดของวัตถุประสงค์ที่จะวัดโดยแบบสอบนั้น เป็นตัวแทนของชุดจากโดเมนเนื้อหาที่เราต้องการหรือไม่

---

[1]Ronald K. Hambleton and Daniel R. Eignor "Guidelines for Evaluating Criterion-Referenced and Test Manuals," Journal of Educational Measurement 15 (1978) : 322-324.

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

ข้อสอบ

1. ได้บรรยายถึงกระบวนการสร้างข้อสอบหรือไม่
2. ข้อสอบมีความตรงตามวัตถุประสงค์ที่จะวัดหรือไม่
3. เป็นตัวแทนของประชากรของข้อสอบที่วัดวัดถุประสงค์นั้นหรือไม่
4. มีข้อบกพร่องทางเทคนิคหรือไม่
5. รูปแบบเหมาะกับวัตถุประสงค์ที่จะวัดหรือไม่
6. มีอคติหรือไม่ (เช่น ต่อเพศ เชื้อชาติ)
7. ใช้กลุ่มตัวอย่างที่เป็นวิวิธพันธ์ในการศึกษานำหรือไม่
8. การวิเคราะห์ข้อมูลเป็นไปเพียงเพื่อหาข้อสอบที่บกพร่องหรือไม่

การบริหารการสอบ

1. ในคำสั่งมีข่าวสารเกี่ยวกับจุดประสงค์ของการสอบ เวลาที่ใช้ ตัวอย่างคำถาม กระดาษคำตอบ และการให้คะแนนหรือไม่
2. คำสั่งในแบบสอบชัดเจนหรือไม่
3. สะดวกในการตรวจให้คะแนนหรือไม่
4. คู่มือแบบสอบชี้ถึงบทบาทและความรับผิดชอบของผู้คุมสอบหรือไม่

การวางรูปแบบของแบบสอบ

1. ตัวแบบสอบมีการวางรูปแบบให้ชวนสนใจหรือไม่
2. ให้ความสะดวกต่อผู้สอบหรือไม่

ความเที่ยง

1. ชนิดของความเที่ยงที่เสนอไว้เหมาะกับการนำคะแนนจากแบบสอบไปใช้หรือไม่

2. กลุ่มตัวอย่างที่นำมาศึกษาเรื่องความเที่ยง มีขนาดและเป็นตัวแทนประชากรที่เหมาะสมกับผู้ที่จะนำแบบสอบไปวัดหรือไม่
3. ความยาวของแบบสอบเหมาะสมที่จะให้แบบสอบที่มีระดับความเที่ยงของคะแนนพอหรือไม่
4. เสนอข่าวสารเกี่ยวกับความเที่ยงในคู่มือแบบสอบสำหรับวัตถุประสงค์ในการนำแบบสอบไปใช้หรือไม่

<u>คะแนนเกณฑ์</u>

1. แถลงเหตุผล เหมาะสมในการเลือกวิธีการในการตัดสินใช้คะแนนเกณฑ์ตัดสินหรือไม่
2. อธิบายถึงวิธีการที่ใช้เลือกเกณฑ์ตัดสินหรือไม่ วิธีที่ใช้นั้นเหมาะสมหรือไม่
3. แสดงหลักฐานด้านความตรงของคะแนนเกณฑ์ที่เลือกใช้หรือไม่

<u>ความตรง</u>

1. ความตรงที่เสนอไว้เหมาะสมกับการนำไปใช้หรือไม่
2. ได้วิจารณ์อย่างมีเหตุผลในองค์ประกอบที่มีผลต่อความตรงของคะแนนในคู่มือหรือไม่

<u>ปกติวิสัย</u>

1. เสนอปกติวิสัยในรูปที่เหมาะสมหรือไม่
2. บรรยายถึงกลุ่มตัวอย่างที่นำมาใช้เป็นกลุ่มปกติวิสัยหรือไม่
3. มีคำเตือนที่มีเหตุผลในการแปลความหมายคะแนนอย่างถูกต้องหรือไม่

<u>การรายงานเกี่ยวกับคะแนนจากแบบสอบ</u>

1. มีการรายงานคะแนนของผู้สอบแต่ละวัตถุประสงค์หรือไม่

2. มีวิธีการในการรายงานผลการสอบเสนอไว้หลายวิธีพอที่ผู้ใช้จะเลือกนำไปใช้หรือไม่ (เช่น รายงานตามชั้นและระดับชั้นภายในโรงเรียน)
3. มีวิธีการที่สะดวกในการตรวจให้คะแนนด้วยมือ และมีแบบฟอร์มในการรายงานคะแนนที่ได้จากแบบสอบหรือไม่

การแปลความหมายคะแนน

1. มีคำเตือนในการแปลความหมายคะแนนของแต่ละคนหรือกลุ่มไว้ในคู่มือหรือไม่
2. ได้มีการเสนอแนวทางที่เหมาะสมในการนำคะแนนจากแบบสอบไปใช้บรรยายถึงสถานภาพ การตัดสินในการสอน การประเมินโปรแกรมการสอน หรือเพื่อการอื่น ๆ หรือไม่

การประมาณค่าคะแนนโดเมน

คะแนนโดเมนหรือเรียกว่า คะแนนจริงในแบบสอบอิงกลุ่มมีวิธีการในการประมาณค่าต่างจากการประมาณค่าคะแนนจริงในแบบสอบอิงกลุ่มที่เห็นได้ชัดประการหนึ่ง คือ จะต้องได้มาจากแบบสอบที่เป็นเอกพันธ์ในด้านเนื้อหา หรือพูดให้เข้าใจง่าย ๆ คือ ได้จากแบบสอบที่วัดวัตถุประสงค์ข้อเดียว ถ้าแบบสอบทั้งฉบับวัดวัตถุประสงค์หลายข้อก็จะต้องแบ่งออกเป็นแบบสอบย่อย ๆ ตามวัตถุประสงค์ที่วัด แล้วประมาณค่าคะแนนโดเมนตามแบบสอบย่อยในขณะที่การประมาณค่าคะแนนจริงในแบบสอบอิงกลุ่มจะไม่คำนึงถึงความเป็นเอกพันธ์ในด้านเนื้อหาของแบบสอบเลย การประมาณค่าคะแนนโดเมนมีหลายวิธี

แต่ละวิธีมีข้อจำกัดและข้อตกลงแตกต่างกันไปดังนี้[1]

1. วิธีหาสัดส่วนความถูกต้อง (Proportion-Correct Score Estimate) ถ้าแบบสอบอิงเกณฑ์นั้นสร้างอย่างดี ข้อสอบทุกข้อมีความเป็นตัวแทนของข้อสอบทั้งหมดเท่าที่เป็นไปได้ในโดเมนนั้นจริง และมีความยาวพอเพียง ค่าที่ได้จากการประมาณวิธีนี้จะเป็นค่าประมาณที่ไม่ลำเอียงของคะแนนโดเมน แต่จะมีความเที่ยงต่ำมากถ้าจำนวนข้อสอบที่วัดแต่ละวัตถุประสงค์มีจำนวนน้อย

$$(\Phi_i) = \frac{X_i}{n}$$

X_i คือ คะแนนที่แต่ละคนได้จากแบบสอบ

n คือ จำนวนข้อสอบ

Φ_i คือ คะแนนโดเมน

2. วิธีคลาสลิคแบบที่ 2 (Classical Model II) เป็นการประมาณค่าคะแนนโดเมนที่ใช้ข้อมูลของกลุ่มเข้ามาคิดด้วย ผู้ที่คิดขึ้นคือ เคลลี่ (Kelly, 1972) เป็นวิธีประมาณค่าการกระถดถอยของคะแนนโดเมนโดยเป็นผลรวมของน้ำหนักองค์ประกอบสองอย่าง อย่างหนึ่งขึ้นอยู่กับคะแนนของผู้สอบแต่ละคน อีกอย่างขึ้นกับค่ามัชฌิมเลขคณิตของกลุ่ม แจ็คสัน (Jackson, 1972) ได้แก้ไขวิธีการของเคลลี่มาใช้กับข้อมูลแบบชาคทอนแบ่งสอง(Bianary Data) โดยใช้สมการของฟรีแมน-ทิวกี้ (Freeman-Tukey Transformation) ดังนี้

---

[1]Ronald K. Hambleton et al., "Criterion-Referenced Testing and Measurement : A Review of Technical Issues and Developments," Review of Educational Research 48 (1978) : 5 - 9.

แปลงคะแนนดิบให้อยู่ในรูป g_i

$$g_i = \frac{1}{2}\left\{ \sin^{-1}\sqrt{\frac{X_i}{n+1}} + \sin^{-1}\sqrt{\frac{X_i+1}{n+1}} \right\} \quad \ldots (1)$$

แปลงคะแนนโดเมน (π_i) เป็น y_i

$$y_i = \sin^{-1}\sqrt{\pi_i} \quad \ldots (2)$$

เมื่อ π_i มีค่าอยู่ระหว่าง .15 ถึง .85 และจำนวนข้อสอบไม่ต่ำกว่า 8 ข้อ การแจกแจงของ g_i จะเป็นปกติโดยประมาณ โดยมีค่ามัชฌิมเลขคณิตเท่ากับ y_i โดยประมาณ และค่าความแปรปรวน (ν) เท่ากับ $(4n + 2)^{-1}$

และจะได้ค่าประมาณของ y_i จากสูตร

$$\hat{y}_i = g_i\hat{\phi} + (4n+2)^{-1}g_\cdot \; / [\hat{\phi} + (4n+2)^{-1}] \quad \ldots (3)$$

เมื่อ $g_\cdot$ คือ ค่ามัชฌิมเลขคณิตของกลุ่มตัวอย่างหาได้จากสูตร

$$g_\cdot = N^{-1} \sum_{i=1}^{N} g_i \quad \ldots (4)$$

เมื่อ N คือจำนวนผู้สอบ

$\hat{\phi}$ คือค่าความแปรปรวนของกลุ่มตัวอย่างของ y_i หาได้จากสูตร

$$\hat{\phi} = (N-1)^{-1} \sum_{i=1}^{N} (g_i - g_\cdot)^2 - (4n+2)^{-1} \quad \ldots (5)$$

เมื่อหาค่า $\hat{y}_i$ ได้ ก็จะประมาณค่าคะแนนโดเมนได้จากสูตร

$$\hat{\pi}_i = (\pi + .5/n) \sin^2 \hat{y}_i - .25/n \quad \ldots (6)$$

3. วิธีประมาณค่าของเบส์แบบที่ 2 (Bayesian Model II Estimate) การประมาณค่าแบบข้อ 2 อาจไม่เป็นไปตามอุดมคติเพราะไม่ได้พิจารณาถึงข้อมูลเดิม (Prior Data) ของผู้สอบด้วยและค่าพี ($\hat{\phi}$) ที่ประมาณได้อาจให้ค่าเป็นลบได้ เพื่อหลีกเลี่ยงการได้ค่าพีเป็นลบ เบส์ (Bayes) ได้เสนอวิธีประมาณค่าคะแนนโดเมนโดยใช้ข้อมูลสามอย่างคือ คะแนนที่ผู้สอบทำได้จากแบบสอบ คะแนนของกลุ่ม และข้อมูลของผู้สอบจากการสอบครั้งก่อน ๆ หรือจากการสอบที่วัดวัตถุประสงค์อื่น ๆ (Direct, Callateral and Prior Informations)

จากการประมาณค่าด้วยวิธีในข้อ 2 การแจกแจงของ g_i นั้น เราทราบค่าความแปรปรวน แต่ไม่ทราบค่ามัชฌิมเลขคณิต (y_i) การแจกแจงของ g_i จึงเป็นการแจกแจงที่มีเงื่อนไข

$$(g_i / y_i) \sim N(y_i, \nu) \quad \ldots (7)$$

เมื่อ $N(y_i, \nu)$ เป็นการแจกแจงปกติที่มีค่ามัชฌิมเลขคณิตเท่ากับ y_i และความแปรปรวนเท่ากับ ν การประมาณค่าแบบนี้สรุปออกมาในรูปการแจกแจงภายหลังของค่าพารามิเตอร์

เพื่อที่จะได้การแจกแจงภายหลังของ y_i จึงต้องใช้ข้อตกลงว่า y_i คือตัวอย่างที่ได้จากการสุ่มจากการแจกแจงอันใดอันหนึ่ง (จากมวลประชากรเดียว) หรือการแจกแจงเดิมของ y_i เป็นปกติด้วยค่ามัชฌิมเลขคณิตที่ยังไม่ทราบค่าคือ แอลฟ่า(α) และความแปรปรวนที่ไม่ทราบค่าคือ ฟี (ϕ) หมายความว่า การแจกแจงเดิมของ y_i ขึ้นอยู่กับค่าของแอลฟ่า และฟี

โนวิคและคณะ (1973) ได้ทำการศึกษาและเสนอว่า เป็นไปได้ที่จะทราบค่าฟี โดยใช้การแจกแจงแบบอินเวอร์ส ไคสแควร์ (Inverse-Chi-Square) ที่มีค่าชั้นของความเป็นอิสระ ν และสเกลพารามิเตอร์ λ เมื่อทราบค่าของ ν และ λ ก็จะทราบรายละเอียดของฟี และจากผลการศึกษาพบว่า ควรกำหนดค่า $\nu = 8$ ซึ่งจะได้ค่า $\bar{\phi} = \lambda/(\nu - 2)$ โดยที่ β คือค่ามัชฌิมเลขคณิตของการแจกแจงแบบอินเวอร์สไคสแควร์ ดังนั้นค่า $\lambda = (\nu-2)\bar{\phi}$ โดยที่ค่า $\bar{\phi}$ มีค่าโดยประมาณ $= (4n + 2)^{-1}$

และจะได้ค่าประมาณร่วม (Joint Estimate) ของ y_i จากสูตร

$$y_i = \frac{g_i\left[\frac{\lambda}{n + \nu - 1}\right] + y_\cdot\left[\frac{1}{4n + 2}\right]}{\left[\frac{\lambda + \sum (y_i - y_\cdot)^2}{n + \nu - 1}\right] + \left[\frac{1}{4n + 2}\right]}$$

เมื่อ $y_\cdot = N^{-1} \sum_{i=1}^{N} y_i$

4. การประมาณค่ามัชฌิมโดยส่วนรวม (Marginal Mean Estimate) การประเมินแบบข้อ 3 เป็นประโยชน์ในการตัดสินผู้สอบทั้งกลุ่ม แต่การวัดแบบอิงเกณฑ์เราควรต้องมีการตัดสินแยกแต่ละบุคคลด้วย โดยใช้การประมาณค่าคะแนนโดเมนอีกอย่างเรียกว่า Separate (หรือ Marginal) Estimate of Domain Score)

เลวิส แวง และโนวิค (1973) ได้เสนอการประมาณค่าแบบนี้จากสูตร

$$y_i = g_. + \rho^* (g_i - g_.) \qquad \ldots (9)$$

ค่า ρ^* ขึ้นอยู่กับค่า ν และค่า λ ถ้าได้ค่า 2 ค่านี้ จะอ่านค่า ρ^* ได้จากตารางของ แวง (1973) และได้ค่าประมาณคะแนนโดเมนจากสูตร

$$\hat{\pi}_i = (1 + {}^{.5}/_{n}) \sin^2 y_i - {}^{.25}/_{n} \qquad \ldots (10)$$

และถ้าตั้งข้อตกลงว่า α มีการแจกแจงเป็นปกติ มีค่ามัชฌิมเลขคณิตเท่ากับ θ และความแปรปรวนเท่ากับ ϕ/n แล้ว

$$y_i = \rho g_i + (\tau - \rho) g_. + (1 - \tau) \theta \qquad \ldots (11)$$

และประมาณค่า $\hat{\pi}_i$ ได้จากสูตรที่ (10)

5. การประมาณแบบกึ่งวิธีการของเบส์ ("Quasi" Bayesian Estimate) จากข้อตกลงที่ว่า ความเชื่อเดิมเกี่ยวกับค่า α และ ϕ สามารถแสดงได้ในรูปของการแจกแจงนั้น อาจมีข้อเบี่ยงเบนไปได้หลายทางคือ ถ้าไม่สามารถระบุออกมาได้ในรูปของการแจกแจงก็ต้องสามารถระบุได้ว่า

1) สามารถแสดงค่า α และ ϕ เดิมในรูปของการแจกแจงรูปสี่เหลี่ยมพื้นผ้า (A Uniform Distribution) และค่า ϕ เป็นค่าที่ระบุได้แน่นอน

2) ทั้งค่า α และ ϕ สามารถระบุได้แน่นอนทั้งสองค่า

ในกรณี 1) จะได้

$$y_i = \frac{g_i \phi + (4n + 2)^{-1} g_.}{\phi + (4n + 2)^{-1}} \qquad \ldots (12)$$

ในกรณี 2) จะได้

$$y_i = \frac{g_i \phi + (4n + 2)^{-1}}{\phi + (4n - 2)^{-1}} \quad \ldots (13)$$

จากการศึกษาเปรียบเทียบของ แฮมเบิลตั้น แฮทตั้น และสแวมมินนาธาน ในวิธีการทั้งห้า โดยดูองค์ประกอบด้านขนาดของกลุ่มตัวอย่าง ความยาวของแบบสอบ ความเป็นเอกพันธ์ของการกระจายของคะแนนโดเมน คะแนนเกณฑ์ และรายละเอียดเกี่ยวกับข้อมูลเดิมแล้ว ถ้าได้ข้อมูลแน่นอนมาใช้ เช่น ทราบค่า α และ ϕ การประเมินที่ให้ค่าคลาดเคลื่อนสัมบูรณ์น้อยที่สุดคือวิธีที่ 5 โดยค่า y_i ได้จากสมการที่ 13 โดยจะมีค่าความคลาดเคลื่อน (ϵ) เป็น

$$\epsilon = \sum_{i=1}^{N} | y_j - y_i |$$

ถ้าบอกค่าได้ไม่แน่นอน วิธีการที่ให้ค่าความคลาดเคลื่อนน้อยที่สุดคือ วิธีที่ 5 ที่ได้จากสมการที่ 12 วิธีที่ให้ค่าความคลาดเคลื่อนรองลงไปคือ วิธีที่ได้จากสมการที่(10) (3),(8) และวิธีที่ 1 แต่เนื่องจากในทางปฏิบัติเราไม่ทราบค่า α และ ϕ ดังนั้นวิธีที่ใช้ได้จริง ๆ คือวิธีของเบธ ซึ่งมิลแมน[1]ได้แสดงลำดับขั้นในการประมาณดังต่อไปนี้

1. แปลงคะแนนดิบของผู้สอบแต่ละคนให้อยู่ในรูปสัดส่วน โดยหารด้วยจำนวนข้อสอบ $P_j = X_j / n$

2. แปลงคะแนน P_j เป็น g_j จากสูตร

$$g_j = \text{arc sin } \sqrt{(X_j + 3/8) / (n+3/4)}$$

เมื่อ n คือ จำนวนข้อสอบ

---

[1]Millman, "Criterion-Referenced Measurement," pp.348-349.

3. หาค่ามัชฌิมเลขคณิตของคะแนนแปลง g_j

$$g_. = \sum_{j=1}^{m} g_j/m$$

เมื่อ m คือจำนวนผู้สอบ

4. หาค่าความแปรปรวนของคะแนนแปลง g_j

$$s_g^2 = \frac{\sum (g_j - g_.)^2}{m}$$

5. กำหนดค่า ν ซึ่งจากการศึกษาของเลวิสและคณะในปี 1973 พบว่าค่าที่เหมาะสมคือ 8

6. กำหนดค่า t ซึ่งเป็นจำนวนข้อสอบที่เราคิดว่าเป็นจำนวนที่น่าพอใจในการที่จะใช้วัดความรู้ในวัตถุประสงค์นั้นหรือโดเมนนั้น

7. คำนวณหาค่า λ จากสูตร

$$\lambda = \frac{\nu - 2}{4(t + 1)}$$

8. ใช้ค่า s_g^2 , ν , λ , n และ m หาค่า ρ^* จากตารางที่แวง (Wang, 1973) ได้สร้างขึ้น

9. คำนวณหาค่า μ_j จากสูตร

$$\mu_j = \rho^* g_j + (1 - \rho^*) g_.$$

10. ใช้ตาราง Arc Sin Transformation[1] เปลี่ยนค่า μ_j เป็น $\hat{\pi}_j$ โดยที่ค่า $\mu_j = \theta = \arcsin \sqrt{\hat{\pi}_j}$

---

[1] Ibid., p.349.

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

การวิเคราะห์รายข้อ

ซู (Hsu, 1971) กล่าวว่า ข้อสอบที่ดีตามแนวอิงเกณฑ์ควรมีอำนาจจำแนกดีโดยที่ผู้ที่เรียนรู้แล้วควรทำได้เป็นจำนวนมากกว่าผู้ที่ยังไม่ได้เรียน ค่าอำนาจจำแนกที่ซูเสนอคือ D%[1] ซึ่งเป็นความแตกต่างระหว่างระดับความยากของข้อสอบก่อนและหลังเรียนที่ได้มาจากการวิเคราะห์แบบอิงกลุ่ม และด้วยแนวความคิดคล้ายคลึงกัน คริสพินและเฟลฮูสัน[2] (Kryspin and Feldhusen, 1974) ได้เสนอดัชนีเอส (Sensitivity Index - S) โดยมีสูตรว่า

$$S = \frac{R_{pos} - R_{pre}}{T}$$

R_{pos} คือ จำนวนผู้ที่ตอบถูกหลังจากเรียน

R_{pre} คือ จำนวนผู้ที่ตอบถูกก่อนเรียน

T คือ จำนวนผู้เข้าสอบ

---

[1]Thomas M. Haladyna, "Effects of Different Samples on Item and Test Characteristics of Criterion-Referenced Test," Journal of Educational Measurement 11 (1974) : 94. Citing T. Hsu, Empirical Data on Criterion-Referenced Tests. Paper Presented at the Annual Meeting of the American Educational Research Association, New York : 1971.

[2]สมศักดิ์ สินธุระเวชญ์, "การประเมินผลแบบอิงกลุ่ม และอิงเกณฑ์," วารสารวิจัยการศึกษา 8 (มิถุนายน 2521) : 84 อ้างจาก William J. Kryspin and John F. Feldhusen, Developing Classroom Test (Minneapolis, Minnesota : Burgess Publishing Company, 1974).

ถ้ามองในแง่ของการวัดอิงกลุ่มแล้ว สูตรนี้ก็เปรียบได้กับค่าอำนาจจำแนกนั่นเอง แต่เปลี่ยนจากกลุ่มสูงเป็นกลุ่มหลังเรียน และจากกลุ่มต่ำเป็นกลุ่มก่อนเรียน ข้อสอบที่ต้องการคือ ข้อสอบที่มีค่าดัชนีเอสสูงและเป็นบวก (ระหว่าง 0.00 – 1.00) ซึ่งแสดงว่าข้อสอบนั้นสามารถจำแนกผู้ที่เรียนแล้วและผู้ที่ยังไม่ได้เรียนออกจากกัน

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

ความเที่ยงของแบบสอบอิงเกณฑ์

ในการตัดสินเกี่ยวกับความเที่ยงของแบบสอบอิงเกณฑ์ ได้มีผู้เสนอวิธีการในการประมาณค่าสัมประสิทธิ์ความเที่ยงไว้หลายวิธี แบ่งออกตามแนวความคิดเป็นสองฝ่าย ฝ่ายแรกประยุกต์วิธีคลาสสิคมาใช้ โดยมองความแปรปรวนของการวัดอิงเกณฑ์เป็นความเบี่ยงเบนของคะแนนผู้สอบแต่ละคนจากคะแนนเกณฑ์ ยิ่งความเบี่ยงเบนนี้มีค่ามากเท่าไรความถูกต้องของการตัดสินให้ผู้สอบได้หรือตกก็ยิ่งมากขึ้นเท่านั้น ผู้ที่เสนอสูตรนี้มาเป็นที่ยอมรับกันในฝ่ายที่สนับสนุนความคิดนี้คือ ลีฟวิงสตัน เสนอวิธีการประมาณค่าความเที่ยง โดยมีข้อตกลงเบื้องต้นว่า[1] "วัตถุประสงค์ของการวัดอิงเกณฑ์ก็เพื่อจะแยกคะแนน (โดเมน) ของแต่ละคนออกจากคะแนนเกณฑ์" จากสูตร[2]

$$r_{cc} = \frac{r_{xx}s^2_x + (\bar{X} - C)^2}{s^2_x + (\bar{X} - C)^2}$$

r_{xx} คือ ความเที่ยงของแบบสอบอิงกลุ่ม

s^2_x คือ ความแปรปรวนของคะแนน

$\bar{X}$ คือ ค่ามัชฌิมเลขคณิต

C คือ คะแนนเกณฑ์

r_{cc} คือ ค่าความเที่ยงอิงเกณฑ์

---

[1] Hambleton, "Criterion-Referenced Testing...," p.16.

[2] William A. Mehrens and Irvin J. Lehmann, Measurement and Evaluation in Psychology (New York : Holt, Rinehart and Winston, 1973), p.121.

หรือ[1]

$$K^2(X_i, T_X) = \frac{\rho^2(X, T_X)\sigma^2_{(X)} + (\mu_X - C_X)^2}{\sigma^2_{(X)} + (\mu_X - C_X)^2} \quad \ldots\ldots (b)$$

$\sigma^2_{(x)}$ คือ ค่าความแปรปรวนของคะแนนดิบ

μ_x คือ ค่ามัชฌิมเลขคณิต

C_x คือ คะแนนเกณฑ์

$\rho^2 (X, T_X)$ คือ ความเที่ยงที่ได้จากอิงกลุ่ม

หรือ[2]

$$K^2(\hat{\pi}, \pi) = \frac{\sigma^2(\pi) + (\bar{\pi} - \pi_o)^2}{\sigma^2(\hat{\pi}) + (\bar{\pi} - \pi_o)^2} \quad \ldots\ldots (c)$$

$\hat{\pi}$ คือ ค่าประมาณคะแนนโดเมน

π คือ คะแนนโดเมน

π_o คือ คะแนนเกณฑ์ (สัดส่วน)

$\bar{\pi}$ คือ ค่ามัชฌิมเลขคณิตของคะแนนโดเมน

สูตรทั้งสาม[(a),(b) และ(c)] ความจริงคือสูตรเดียวกัน แต่สัญญลักษณ์และชนิดของคะแนนที่ใช้ต่างกัน เมื่อลองเปรียบเทียบสูตรการหาความเที่ยงของลีฟวิงสตันกับความเที่ยงแบบคลาสสิคแล้ว พบว่าค่าประมาณความเที่ยงตามสูตรของลีฟวิงสตันจะสูงกว่าแบบคลาสสิค และแม้ว่าความแปรปรวนของคะแนนโดเมนจะเป็นศูนย์

---

[1]Samuel A. Livingston, "Criterion-Referenced Application of Classical Test Theory," Journal of Educational Measurement 9 (1972) : 17 - 18.

[2]Hambleton, Criterion-Referenced Testing...,", p.16.

[3]Ibid.

ค่าประมาณความเที่ยงที่ได้ก็จะไม่เป็นศูนย์เหมือนวิธีคลาสสิค ถ้าความแปรปรวนของคะแนนจากแบบสอบเป็นศูนย์ คือทุกคนทำคะแนนได้เท่ากัน ค่าความเที่ยงจะเป็น 1.00 และถ้าคะแนนเฉลี่ยของกลุ่มมีค่าเท่ากับคะแนนเกณฑ์แล้ว ค่าความเที่ยงของแบบสอบจะมีค่าเท่ากับความเที่ยงอิงกลุ่ม ลีฟวิงสตันกล่าวว่า[1] "ถ้าคะแนนของคนสองกลุ่ม มีค่าความแปรปรวน และความเที่ยงอิงกลุ่มเท่ากันแล้ว กลุ่มที่มีคะแนนมัชฌิมเลขคณิตเบี่ยงเบนจากคะแนนเกณฑ์มากกว่า จะมีค่าความเที่ยงอิงเกณฑ์สูงกว่า" และ "เป็นไปได้ที่จะกล่าวว่า ค่าความเที่ยงอิงกลุ่มเป็นกรณีพิเศษกรณีหนึ่งของความเที่ยงอิงเกณฑ์คือ กรณีที่ค่าความเที่ยงต่ำสุด"

เชฟเวลสัน, บลอค และเรวิทช์ (Shavelson, Block and Ravitch, 1972) แนะว่า การตัดสินให้ผู้สอบแต่ละคนได้หรือตกนั้นควรพิจารณาความคลาดเคลื่อนมาตรฐานในการวัดด้วย โดยเฉพาะพวกที่มีคะแนนอยู่ใกล้จุดตัดมาก ๆ ควรพิจารณาองค์ประกอบอย่างอื่นด้วย นอกจากนั้นได้วิจารณ์การหาความเที่ยงของลีฟวิงสตันว่า ควรจะแบ่งแบบสอบเป็นแบบสอบฉบับย่อย ๆ แล้วให้คะแนนและหาความเที่ยงตามแบบสอบย่อย

แฮริส (1972) ซึ่งเห็นด้วยกับเชฟเวลสัน บลอค และเรวิทช์ กล่าวว่า ในทางพีชคณิตแล้ว สูตรที่ลีฟวิงสตันพัฒนาขึ้นมาใช้นั้นไม่มีที่ผิดพลาด แต่จุดสำคัญอยู่ที่ว่า ค่าความแปรปรวนของความคลาดเคลื่อนในการวัดน่าจะเป็นดัชนีที่ดีกว่าในการที่จะให้ความมั่นใจในการประเมินผลว่าแต่ละคนได้คะแนนสูงหรือต่ำกว่าเกณฑ์ โดยให้เหตุผลว่า แม้ว่าดัชนีของลีฟวิงสตันจะสูงกว่าแบบคลาสสิคก็ไม่ได้มีความหมายว่า แบบสอบนั้นจะมีความเที่ยงสูงกว่าจริง ๆ เพราะค่าความคลาดเคลื่อนมาตรฐานในการวัดก็ยังมีค่าเท่ากัน นอกจากนั้นแล้ว ลีฟวิงสตันยังไม่ได้คำนึงถึงคะแนนจริงของผู้สอบว่าอยู่ทั้งสูงกว่า และต่ำกว่าเกณฑ์ซึ่งทำให้ดูเหมือนเป็นค่าดัชนีของประชากรสองกลุ่มที่มีค่ามัชฌิมเลขคณิตอยู่ห่างจากคะแนนเกณฑ์เท่ากันทั้งต่ำกว่าและสูงกว่า

---

[1]Livingston, "Criterion-Referenced Application...," p.18.

ในกรณีที่ความยาวของข้อสอบเพิ่มขึ้นและต้องการดูว่า ค่าความเที่ยงของแบบสอบจะเพิ่มขึ้นเป็นเท่าไร ลีฟวิงสตันได้เสนอให้ใช้สูตรของ สเปียร์แมน–บราวน์ คือ[1]

$$\frac{n\, K^2(X, T_X)}{1 + (n-1)\, K^2(X, T_X)}$$

เมื่อ n คือ จำนวนเท่าของข้อสอบที่เพิ่มขึ้นจากเดิม

ในกรณีที่ข้อมูล ที่ได้ไม่ใช่ข้อมูลแบบแบ่งสอง (Binary Data) ลีฟวิงสตันเสนอให้ใช้สัมประสิทธิ์ แอลฟ่า[2] (α)

$$\alpha = \frac{\frac{I}{(I-1)}\, \sigma^2(X) - \Sigma\sigma^2(i) + (\mu_X - C_X)^2}{\sigma^2(X) + (\mu_X - C_X)^2}$$

เมื่อ I คือ จำนวนข้อสอบ และ $\sigma^2(i)$ คือความแปรปรวนของข้อสอบแต่ละข้อ

ความเที่ยงของแบบสอบอิงเกณฑ์ในทัศนะของอีกฝ่ายหนึ่ง มองในแง่ความคงที่ของการตัดสิน เช่น ถ้าให้ผู้สอบสองกลุ่มทำแบบสอบเดียวกัน (หรือให้ผู้สอบกลุ่มเดียวทำแบบสอบคู่ขนาน) แล้ว แบบสอบที่มีความเที่ยงน่าจะเป็นแบบสอบที่สามารถแบ่งผู้เรียนออกเป็นพวกรอบรู้ หรือไม่รอบรู้เป็นร้อยละเท่ากันทั้งสองกลุ่ม (หรือทั้งสองแบบสอบ) โดยที่ผู้สอบสองกลุ่มนี้ต้องเป็นกลุ่มที่เท่าเทียมกันด้วย ค่าความเที่ยงอยู่ที่ความใกล้เคียงกันของอัตราร้อยละที่แบ่งผู้เรียนออกเป็นผู้รอบรู้หรือไม่รอบรู้ทั้งสองครั้ง ได้มีผู้เสนอสูตรตามแนวความคิดนี้ไว้หลายสูตรดังนี้

คาร์เวอร์[3] (Carver, 1970) ได้เสนอการประมาณค่าความเที่ยงโดยการเปรียบเทียบสัดส่วนของความคงที่ในการตัดสินได้ตก โดยใช้กลุ่มตัวอย่างกลุ่มเดียว

---

[1] Ibid., p. 19.

[2] Ibid.

[3] Crehen, "Item Analysis,...," p. 256. Citing R.P. Carver, "Special Problems in Measuring Change with Psychometric Devices," in Evaluation Research : Strategies and Methods (Pittsburgh : American Institutes for Research, 1970).

ทำแบบสอบคู่ขนานหลังจากเรียนผลออกมาเป็นรูปของความถี่ในการตัดสินแต่ละครั้ง ดังตาราง

ฉบับ ก. / ฉบับ ข.	ตก	ได้
ได้	b	a
ตก	c	d

ค่าความเที่ยงได้จาก $(a + c)/N$ เมื่อ $N = a + b + c + d$

สแวมมินนาธาน แฮมเบิลตัน และอัลกิน่า[1] ได้เสนอค่าสัมประสิทธิ์แคปป้า (Kappa-K) ของโคเฮนเป็นดัชนีความเที่ยงโดยใช้สูตร

$$K = (P - P_c)/(1 - Pc)$$

เมื่อ P คือ สัดส่วนที่ตัดสินให้ผู้เรียนได้และตกซ้ำกันสองครั้ง

P_c คือ สัดส่วนในการตัดสินที่คาดหวัง

ฮวน[2] (Huynh,1976) แนะว่าเมื่อคะแนนเกณฑ์มีค่าสูงใกล้ค่า n (จำนวนข้อ) ควรใช้สูตรต่อไปนี้

$$K = (P_{11}-P_1^2) / (P_1 - P_1^2)$$

และเมื่อคะแนนเกณฑ์มีค่าต่ำห่างจาก n มาก

---

[1] Huynh Huynh, "On the Reliability of Decisions in Domain-Referenced Testing," Journal of Educational Measurement, 13 (1976), p.253.

[2] Ibid., p.256.

$$K = (P_{oo} - P_o^2) \quad (P_o - P_o^2)$$

เมื่อ $P_{11} = \sum_{x,y=c}^{n} f(x,y)$

$$P_1 = \sum_{x=c}^{n} f(x)$$

$$P_{oo} = \sum_{x,y=o}^{c-1} f(x,y)$$

$$P_o = \sum_{x=o}^{c-1} f(x)$$

แฮมเบิลตันและโนวิค[1] (1973) เสนอสูตรเมื่อต้องการจำแนกผู้เรียนให้อยู่ในระดับการรอบรู้ m ระดับดังนี้

$$P_o = \sum_{k=1}^{m} P_{kk}$$

เมื่อ P_o คือ ดัชนีความเที่ยง

P_{kk} คือ สัดส่วนของผู้สอบที่ถูกตัดสินให้อยู่ระดับรอบรู้ที่ k

ค่า P_o ของแฮมเบิลตันและโนวิคนี้ สแวมมินนาธาน แฮมเบิลตันและอัลกินา[2] (1974) แย้งว่าอาจเป็นค่าที่มากเกินไป เพราะรวมการตัดสินตรงกันซึ่งเกิดโดยบังเอิญไว้ด้วย เขาแนะนำให้ใช้ค่าแก้ จากสูตร

$$K = (P_o - P_c) / (1 - P_c)$$

เมื่อ $P_c = \sum_{k=1}^{m} P_{k.} \cdot P_{.k}$

โดยที่ $P_{k.}$ และ $P_{.k}$ คือ สัดส่วนของผู้รอบรู้ที่ถูกตัดสินให้อยู่ในระดับรอบรู้ที่ k ในการสอบครั้งที่ 1 และ 2 ตามลำดับ

---

[1] Hambleton, "Criterion-Referenced Testing...," p.21.

[2] Ibid.

ดัชนีความเที่ยงของฮวนเป็นวิธีเดียวที่ได้จากการบริหารการสอบครั้งเดียว นอกนั้นต้องบริหารการสอบสองครั้ง แต่ใช้แบบสอบคู่ขนานเช่นเดียวกัน แต่วิธีของฮวนเป็นวิธีที่ขึ้นอยู่กับความรู้ทางสถิติในด้านการแจกแจงมาก และมีข้อจำกัดว่าถ้าจำนวนข้อสอบน้อยค่าที่ได้จะมีความเที่ยงไม่เท่าที่ควร วิธีอีกวิธีที่มีความยุ่งยากและข้อจำกัดน้อยกว่าคือ วิธีของซับโคเวียค ซึ่งมีสูตรว่า[1]

$$P_c = n^{-1} \sum_{i=1}^{N} P_c^{(i)}$$

เมื่อ $P_c^{(i)} = \text{Prob}\ (X_i \geq c).\text{Prob}(X_i' \geq c)+\text{Prob}(X_i < c).\text{Prof}(X_i' < c)$

x_i คือ คะแนนของผู้สอบคนที่ i จากแบบสอบฉบับที่ 1

X_i' คือ คะแนนของผู้สอบคนที่ i จากแบบสอบฉบับที่ 2

c คือ คะแนนเกณฑ์

ข้อตกลงในการประมาณค่าวิธีนี้คือ[2]

1. คะแนนจากแบบสอบคู่ขนานของผู้สอบแต่ละคนต้องเป็นอิสระต่อกัน และมีการแจกแจงแบบเดียวกัน

2. คะแนนจากแบบสอบทั้งสองแบบของผู้สอบ ต้องมีการแจกแจงแบบไบโนเมียล

การตั้งข้อตกลงดังกล่าวก่อให้เกิดปัญหาแม้ว่าธรรมดาการให้คะแนนจะเป็นแบบ 0–1 และการทำแบบสอบจะเป็นอิสระต่อกันก็ตาม เพราะจะต้องยอมรับด้วยว่าข้อกระทงในแบบสอบแต่ละข้อมีความยากง่ายเท่ากัน ภายหลังจึงใช้แบบจำลองทวินามผสม (Compound Binomial Model) แทน

---

[1] Michael J. Subkoviah, "Estimating Reliability from a Single Administration of A Criterion-Referenced Test," Journal of Educational Measurement 13 (1976) : 267.

[2] Hambleton, "Criterion-Referenced Testing," p.22.

จากข้อตกลงสองข้อนี้จึงสามารถแสดงได้ว่า[1]

$$P\ (X_i \geq c) = \sum_{x_i=c}^{n} \binom{n}{x_i} \pi_i^{x_i} (1 - \pi_i)^{n-x_i} \qquad \ldots (1)$$

$$P_c^{(i)} = \left[P\ (X_i \geq c)\right]^2 + \left[1 - P(X_i \geq c)\right] \qquad \ldots (2)$$

เมื่อประมาณค่า $\hat{\pi}_i$ ได้ ก็นำมาแทนค่า π_i ในสูตร (1)

$P_c^{(i)}$ คือ ความน่าจะเป็นของการคงที่ในการตัดสินผู้สอบคนที่ i

ค่าประมาณของสัดส่วนของการตัดสินตรงกันของกลุ่มคือ $N^{-1} \sum_{i=1}^{N} P_c^{(i)}$

ค่าประมาณของซับโคเวียคมีจุดอ่อนตรงที่ตั้งข้อตกลงเบื้องต้นว่า ข้อสอบทุกข้อมีความยากเท่ากัน และวิธีประมาณค่าแบบนี้อาจให้ค่าความเที่ยงสูงไปเนื่องมาจากรวมการตัดสินตรงกันโดยบังเอิญไว้ด้วย

นอกจากนี้ แฮริส[2] (1972 a) ได้เสนอค่าสัมประสิทธิ์ในกรณีที่กำหนดสภาพการรอบรู้เพียงสองระดับ (k = 2) เป็นค่าสหสัมพันธ์กำลังสองระหว่างระดับการรอบรู้กับคะแนนรวม (โดยให้ 0 เมื่อผู้สอบสอบตก และให้ 1 เมื่อผู้สอบสอบได้)

$$\mathcal{M}_c^2 = \frac{SS_B}{SS_B + SS_W}$$

เมื่อ SS_B = ผลบวกกำลังสองระหว่างกลุ่ม

SS_W = ผลบวกกำลังสองภายในกลุ่ม

---

[1] Ibid., p.23.

[2] Subkoviah, Estimating Reliability from...," pp.265-266. Citing C.W. Harris, "An Index of Efficiency for Fixed-Length Mastery Tests," Paper Presented at the Annual Meeting of the American Educational Research Association (Chicago : 1972. (ab).

มาร์แชล และเฮอร์ทัล (Marshall and Heartel, 1975) ได้เสนอการหา P_0 เฉลี่ยโดยบริหารการสอบ 1 ครั้ง แบ่งแบบสอบเป็นแบบสอบคู่ขนานแล้ว ใช้สูตรของสเพียร์แมน-บราวน์ หาค่าสัมประสิทธิ์ความเที่ยงของแบบสอบทั้งฉบับ[1]

ความตรง

ความตรงของแบบสอบอิงเกณฑ์มีรากฐานอยู่บนมโนมติ เกี่ยวกับทฤษฎีของการวัดอิงเกณฑ์ที่ว่าเครื่องมือที่ใช้ควรเป็นเครื่องมือที่สามารถจะแยกผู้ที่เรียนแล้ว และผู้ที่ยังไม่ได้เรียนออกจากกันได้ เช่น ถ้าเครื่องมือที่ใช้คือแบบสอบ เมื่อนำแบบสอบไปทดสอบคนสองกลุ่ม คือ กลุ่มผู้ที่ยังไม่เคยเรียนเนื้อหาในแบบสอบ และกลุ่มผู้ที่เคยเรียนมาแล้ว ผู้สอบกลุ่มหลังควรจะทำข้อสอบนั้นได้มากกว่าผู้สอบกลุ่มแรก จากความคิดนี้ ความตรงในความหมายของการวัดอิงเกณฑ์ก็คือ ความตรงตามทฤษฎี หรือความตรงในการตัดสินผู้สอบนั่นเอง คุณภาพของแบบสอบอิงเกณฑ์ในด้านนี้มีผู้ศึกษาไม่มากนัก จึงไม่ยุ่งยากซับซ้อนหรือขึ้นอยู่กับค่าสถิติเหมือนความเที่ยง ที่เหตุผลส่วนหนึ่งอาจเนื่องมาจากดัชนีที่ใช้เลือกข้อสอบในการวิเคราะห์รายข้อก็เป็นตัวชี้ถึงความตรงโดยส่วนรวมของข้อสอบเป็นข้อ ๆ อยู่แล้วก็อาจเป็นได้ ดัชนีแสดงความตรงในการตัดสินตัวแรกเสนอโดย คาร์เวอร์[2]

---

[1]Ibid., p.266. Citing J.L. Marshall and E.H. Haertel, "A Single-Administration Reliability Index for Criterion-Referenced Tests: The Mean Split-Half Coefficient of Agreement," Paper Presented at the Annual Meeting of the American Educational Research Association (Washington D.C. : 1975).

[2]K.D. Crehen, "Item Analysis for...," p.256. Citing R.P. Carber, "Special Problems in Measuring Change with Psychometric Devices," in Evaluation Research : Strategies and Methods," (Pittburgh : American Institutes for Research, 1970).

ได้ค่าความตรงจากข้อมูลที่ได้จากแบบสอบของกลุ่มผู้เรียนแล้ว และยังไม่เคยเรียนดังตาราง

	ยังไม่เรียน	เรียนแล้ว
ผ่าน	b a	a b
ไม่ผ่าน	c	d

ค่าสัมประสิทธิ์ความตรงได้จาก (a+c) / (a+b+c+d)

สัมประสิทธิ์ความตรงของคาร์เว่อร์จะมากหรือน้อยขึ้นอยู่กับจำนวนผู้สอบที่สอบไม่ผ่านก่อนเรียน และที่สอบผ่านหลังจากเรียน ดังนั้น องค์ประกอบที่ควรจะมีผลต่อความตรงของคาร์เว่อร์ ก็น่าจะเป็นความตรงตามเนื้อหาของข้อสอบแต่ละข้อและประสิทธิภาพในการสอน

ดัชนีความตรงอีกตัวได้มาจากค่าสัมประสิทธิ์สหสัมพันธ์แบบพี[1] ซึ่งมีลักษณะรวมทั้งองค์ประกอบที่มีผลต่อค่าที่ได้คล้ายกับของคาร์เว่อร์ มีสูตรว่า

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

เมื่อ a คือ จำนวนผู้ที่สอบผ่านหลังจากเรียน

b คือ จำนวนผู้ที่สอบผ่านก่อนเรียน

c คือ จำนวนผู้ที่สอบไม่ผ่านหลังจากเรียน

d คือ จำนวนผู้ที่สอบไม่ผ่านก่อนเรียน

---

[1] Millman, "Criterion-Referenced Measurement," pp.370-371.

ความยาวของแบบสอบ การกำหนดคะแนนเกณฑ์ และการตัดสินผู้สอบ

การพิจารณาในเรื่องเหล่านี้มีองค์ประกอบอย่างหนึ่งที่เกี่ยวข้องอยู่ด้วยคือ อัตราความสูญเสีย (Loss Ratio) ซึ่งเกิดจากความคลาดเคลื่อนสองชนิดคือ ความคลาดเคลื่อนอันเกิดจากการยอมรับที่ผิด และความคลาดเคลื่อนอันเกิดจากการปฏิเสธที่ผิด การพิจารณาในเรื่องความยาวของแบบสอบ คะแนนเกณฑ์ และการตัดสินผู้สอบที่เหมาะสม จึงทำโดยพยายามไม่ให้มีความคลาดเคลื่อนสองประการนี้เกิดขึ้นได้ แต่ในทางปฏิบัติแล้วเป็นไปไม่ได้ที่จะป้องกันไม่ให้เกิดขึ้นเลยได้ จึงมีความพยายามในอันที่จะทำให้เกิดน้อยที่สุดเท่าที่จะเป็นไปได้ (Threshold Loss)

ในด้านความยาวของแบบสอบ มิลแมน[1] ได้สร้างตารางเพื่อหาความยาวที่เหมาะสม โดยเปรียบเทียบระหว่างค่าคะแนนโดเมนที่ประมาณไว้ กับคะแนนตัดสินระดับความรอบรู้ (Advancement Score) โดยใช้การทดสอบทวินาม (Binomial Test) เราสามารถหาความน่าจะเป็น หรือโอกาสของการตัดสินที่ผิดพลาดได้ (ตามน้ำหนักที่เราจะให้แก่ความคลาดเคลื่อนแต่ละแบบ) โดยดูจากตารางซึ่งประกอบด้วยคะแนนเกณฑ์ คะแนนตัดสินระดับความรอบรู้ คะแนนประมาณของคะแนนโดเมน และจำนวนข้อสอบ เราสามารถเลือกความยาวของข้อสอบและคะแนนตัดสินระดับความรอบรู้ได้ โดยที่ดูจากค่าความสูญเปล่าที่เรากำหนดไว้

ตารางของมิลแมนมีปัญหาอยู่บ้าง ประการแรกคือค่าที่ได้ขึ้นอยู่กับความถูกต้องของค่าประมาณคะแนนโดเมน และอีกปัญหาคือการพิจารณาคะแนนตัดสินระดับความรอบรู้ร่วมกับ

---

[1] Hambleton, "Criterion-Referenced Testing...," p. 24. Citing J. Millman, "Criterion-Referenced Measurement," in Evauation in Education : Current Practices (Berkeley, California : McCutchan Publisher, 1974).

ความยาวของแบบสอบเพื่อลดโอกาสของความคลาดเคลื่อนในการตัดสินผู้เรียนที่อยู่เหนือจุดตัด จะไปเพิ่มโอกาสของความคลาดเคลื่อนของการตัดสินของผู้ที่อยู่ต่ำกว่าจุดตัด

ในเรื่องเกี่ยวกับการตั้งเกณฑ์ตัดสินผู้สอบ มิลแมน (1973) กล่าวว่าการพิจารณาเรื่องนี้ควรพิจารณาหลาย ๆ ด้านประกอบกัน เช่น

1. คะแนนของคนอื่น ๆ (Performance of Others) เช่น กำหนดจากที่กลุ่มก่อน ๆ ทำได้ หรืออาจกำหนดจำนวนนักเรียนที่จะให้ผ่านไว้ (แต่วิธีหลังนี้คงไม่ดีนัก เพราะจะทำให้ทฤษฎีการวัดอิงเกณฑ์เปลี่ยนไป)

2. เนื้อหาของข้อสอบ (Item Content) พิจารณาข้อสอบแต่ละข้อแล้วนำมารวมกัน เช่นให้ผู้เชี่ยวชาญตัดสินว่าข้อสอบแต่ละข้อควรมีส่วนสัดของผู้สอบทำได้เท่าไร แล้วนำสัดส่วนทุกข้อมาบวกกันจะได้เป็นคะแนนเกณฑ์

3. ผลทางการศึกษาที่ตามมา (Educational Consequences) คำถามที่เกี่ยวข้องคือ คะแนนเกณฑ์เท่าใดจึงจะทำให้เกิดประโยชน์ทางการศึกษามากที่สุด

4. คุณค่าทางจิตวิทยา และทางเศรษฐกิจที่มีต่อผู้สอนและผู้เรียน เช่น ถ้าค่าใช้จ่ายในการสอนซ่อมเสริมสูงมากต้องกำหนดเกณฑ์ให้ต่ำลง เป็นต้น ทั้งนี้ต้องคิดถึงความสูญเปล่าทางอื่นด้วย

5. ความคลาดเคลื่อนจากการเดา และการเลือกตัวอย่างข้อสอบ ถ้ามีปัญหาเรื่องนี้เกิดมาก ควรกำหนดเกณฑ์สูงขึ้น

ปัญหาอันเกิดจากการเดาในข้อ 5 นี้ กรอนลันด์[1] (Gronlund) กล่าวว่า "การที่เราบอกให้นักเรียนตอบข้อสอบให้ครบทุกข้อเป็นความพยายามที่จะทำให้ความแปรปรวนระหว่างผู้สอบในแนวโน้มของการเดามีเท่า ๆ กัน จึงไม่จำเป็นจะต้องมีการแก้การเดา" และ "ยังมีข้อขัดแย้งอยู่เหมือนกันในเรื่องการที่จะแก้การเดาหรือไม่ แต่โดยมากจะเป็นเรื่อง

---

[1] Norman E. Gronlund, Constructing Achievement Test (Englewood Cliffs, N.J. : Prentice-Hall, 1977), p.107.

ของแบบสอบมาตรฐานอาจมีข้อกระทงบางข้อไม่เหมาะสมกับกลุ่มผู้สอบบางกลุ่ม และผู้สอบอาจไม่ได้รับโอกาสที่จะทำแบบสอบได้ทุกข้อ... โดยทั่วไปการวัดสัมฤทธิผลที่ไม่ได้ใช้แบบสอบมาตรฐานไม่ควรมีการแก้การเดาเพราะเนื้อหาในแบบสอบมีความสัมพันธ์กันมากกับประสบการณ์การเรียนรู้ของผู้สอบ และเวลาที่กำหนดให้ก็มากพอที่จะให้ผู้สอบมีเวลาทำทุกข้อ"

จากการศึกษาของบล็อก[1](Block, 1972) พบว่า การกำหนดเกณฑ์ตัดสินมีผลต่อสัมฤทธิผลทางการศึกษา ถ้าตั้งเกณฑ์ตัดสินสูงสัมฤทธิผลก็สูง ถ้าตั้งเกณฑ์ต่ำสัมฤทธิผลก็ต่ำ

เบอร์ก[2](Berk, 1976) เสนอว่าควรพิจารณากำหนดเกณฑ์จากผู้ทำแบบสอบเดียวกันสองกลุ่ม คือกลุ่มผู้ที่เรียนมาแล้วและกลุ่มผู้ที่ยังไม่ได้เรียน ข้อเสนอของเขาอยู่ที่การที่เขาสังเกตว่าคุณภาพของแบบสอบอิงเกณฑ์ขึ้นอยู่กับอำนาจของมันในอันที่ให้ข้อมูลที่เพียงพอในการที่จะตัดสินแบ่งผู้เรียนได้ถูกต้อง ภายใต้ข้อตกลงว่าผู้สอบกลุ่มที่เรียนมาแล้วเป็นผู้รอบรู้ กลุ่มผู้ไม่ได้เรียนเป็นผู้ไม่รอบรู้ เขาก็ตั้งเกณฑ์ตัดสินผู้สอบออกเป็นผู้รอบรู้และผู้ไม่รอบรู้อย่างเป็นอิสระต่อกัน ผู้สอบกลุ่มแรกจะถูกตัดสินแบ่งเป็นสองพวกคือ ผู้รอบรู้จริง(True Master) และพวกเรียนแล้วแต่ไม่รอบรู้(False Nonmaster) ผู้สอบกลุ่มที่สองจะถูกตัดสินแบ่งเป็นสองพวก พวกรอบรู้ไม่จริง(False Master) และพวกไม่รอบรู้จริง(True Nonmaster) การกระจายของคนสองกลุ่มนี้จะเป็นตัวกำหนดขอบเขตตัวแรกว่าผู้สอบจะเป็นพวกรอบรู้จริงหรือไม่รอบรู้จริง เกณฑ์จะได้จากการตรวจสอบความน่าจะเป็นของเฉลี่ยในคะแนน

---

[1]Hambleton, "Criterion-Referenced Testing...," p.29, Citing J.H. Block, "Student Learning and the Setting of Mastery Performance Standard," Educational Horizons, (1972):50, 183-190.

[2]Hambleton, "Criterion-Referenced Testing...," p.31, Citing R.A. Berk, "Determination of Optional Cutting Scores in Criterion-Referenced Measurement," Journal of Experimental Education, (1976): 45. 4-9.

จุดตัดแต่ละประเภท คะแนนจุดตัดที่เหมาะสมคือ ค่าที่ให้สัดส่วนในการตัดสินผิดพลาดน้อยที่สุด ค่าที่ให้สัดส่วนในการตัดสินได้ถูกต้องในการใช้จุดตัดใดๆคือสัดส่วนของผู้สอบที่เรียนมาแล้ว และถูกตัดสินให้เป็นพวกรอบรู้ รวมกับสัดส่วนของผู้ที่ไม่ได้เรียน และถูกตัดสินให้เป็นพวกไม่รอบรู้

วิธีของเบอร์ค คล้ายกับวิธีการหาความตรงของ คาร์เว่อร์ และจัดได้ว่าเป็นวิธีที่ให้ประโยชน์มาก รวมทั้งสามารถนำไปใช้ได้ไม่ยุ่งยากในทางปฏิบัติ โดยเฉพาะในการเรียนที่เรียงหน่วยการเรียนการสอนต่อเนื่องกันไป เกณฑ์ของแต่ละหน่วยก็ตัดสินได้โดยพิจารณาถึงสัมฤทธิผลที่ต้องการนำไปใช้ในหน่วยต่อ ๆ ไปด้วย วิธีของเบอร์คนี้ สมหวัง พิธิยานุวัฒน์ ได้นำไปศึกษาโดยให้กลุ่มผู้สอบกลุ่มเดียวทำแบบสอบก่อนเรียนและหลังเรียน เพื่อหาเกณฑ์ที่ทำให้เกิดความสูญเสียน้อยที่สุด

รายงานการวิจัยที่เกี่ยวข้อง

เอ็มริค[1] ได้ทำการศึกษาถึงความสูญเสียสองแบบ คือ แอลฟ่า และ เบต้า โดยใช้ข้อมูลจากแบบสอบ IPI Math Placement Test "Numerical D" ซึ่งมีความยาว 16 ข้อ คำนวณหาเกณฑ์ที่เหมาะสม เมื่อกำหนดให้ $\beta = 3\alpha$, RR = 10 จากสูตร

$$K = \frac{\log \frac{\beta}{1-\alpha} + n^{-1}(\log RR)}{\log \frac{\alpha\beta}{(1-\alpha)(1-\beta)}}$$

---

[1] John A. Emrick, "An Evaluation Model for Mastery Testing," Journal of Educational Measurement 8 (1971) : 321-326.

เมื่อ K คือ เกณฑ์ที่อยู่ในรูปของร้อยละ

n คือ ความยาวของแบบสอบ

β คือ ค่าประมาณความน่าจะเป็นของความผิดพลาดในการยอมรับที่ผิด

α คือ ค่าประมาณความน่าจะเป็นของความผิดพลาดในการปฏิเสธที่ผิด

ค่า α และ β หาได้จากสูตร

$$\alpha + \beta = 1 - \sqrt{\bar{r}_i}$$

เมื่อ $\bar{r}_i$ คือ ค่าความเที่ยงของข้อกระทง คำนวณได้โดยใช้สูตรของสเปียร์แมน-บราวน์

RR คือ อัตราส่วนของความรุนแรงของความคลาดเคลื่อน β ต่อ α (Ratio of Regret of β to α)

ได้ค่า K = 56.25% นั่นคือจะต้องทำข้อสอบได้อย่างน้อย 9 ข้อ จาก 16 ข้อ จึงจะถือว่าผ่าน

สแวมมินาธาน, แฮมเบิลตัน และ อัลกินา[1] ได้แสดงการประมาณค่าความสูญเสียที่น้อยที่สุด (Threshold Loss) ในการตัดสินด้วยวิธีของเบส์ โดยกำหนดค่า α = 1 หน่วย และ β = 2 หน่วย พบว่าเมื่อกำหนดเกณฑ์เป็น .8 ความยาวของแบบสอบเป็น 10 และจำนวนผู้สอบ 25 คนค่าความสูญเสียจะน้อยที่สุดเมื่อให้ผู้ที่ทำข้อสอบได้ 9 ข้อขึ้นไปสอบได้

---

[1] Hariharan Swaminathan; Ronald K. Hambleton; and James Algina, "A Bayesian Decision-Theoretic Procedure for Use with Criterion-Referenced Test," Journal of Educational Measurement, 12 (1975):87-98.

และผู้ที่ได้ต่ำกว่า 9 ข้อ สอบตก และเมื่อกำหนดจุดตัดใหม่เป็น .6 และ .8 โดยแบ่งระดับการรอบรู้เป็น 3 ระดับ คือรอบรู้ทั้งหมด รอบรู้เป็นบางส่วน และไม่รอบรู้ ปรากฏว่าค่าความสูญเสียจะต่ำสุด เมื่อตัดสินผู้ที่ทำข้อสอบได้ 10 ข้อขึ้นไปผ่าน ผู้ที่ทำได้ 7 ถึง 9 ข้อต้องสอนทบทวนอีกเล็กน้อย และผู้ที่ทำได้ 6 ข้อลงไปต้องเรียนใหม่

** เครเฮน[1] ได้ศึกษาเปรียบเทียบผลของความเที่ยงและความตรงของแบบสอบอิงเกณฑ์ โดยใช้แบบสอบคู่ขนานซึ่งใช้วิธีการเลือกข้อกระทงเพื่อแบ่งเป็นแบบสอบคู่ขนาน 6 วิธีคือ

1. วิธีของค็อกซ์-แวกาส (Cox-Vagas)
2. วิธีของเบรนนั่น (Brennan)
3. เรียงลำดับตามสัดส่วนในการตอบของผู้สอบหลังจากเรียน
4. เรียงตามที่ครูผู้สร้างแบบสอบเลือก
5. เรียงตามค่าสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล
6. เรียงตามลำดับการสุ่ม

พบว่าไม่มีวิธีใดให้ผลต่อความเที่ยงต่างจากวิธีอื่น ๆ อย่างเห็นได้ชัดนัก ในด้านความตรงได้เลือกเกณฑ์ตัดสินที่ให้ค่าความประมาณความตรงสูงสุด โดยใช้จุดเดียวกันทั้งสองฉบับ ปรากฏว่าวิธีของค็อกซ์-แวกาส และ เบรนนั่น ให้ความตรงสูงสุด ในปีเดียวกันนี้ ฮาลาไดน่า[2] (Haladyna) ได้ศึกษาผลของการใช้กลุ่มตัวอย่างต่างกันที่มีต่อลักษณะของข้อกระทงและแบบสอบ กลุ่มตัวอย่างที่ใช้เป็นนิสิตเรียนทางการศึกษาเรียนรายวิชาการวัดและประเมินผลการศึกษาระดับปริญญาตรี 189 คน ซึ่งเป็นตัวอย่างที่เทียบเท่ากับประชากรในระดับอุดมศึกษาทั่ว ๆ ไป ทั้งเกรดเฉลี่ยและคะแนนจาก American College Test ให้กลุ่มตัวอย่างเรียนเรื่อง Mastery Learning ของบลูม (Bloom, 1968) ซึ่งมีอยู่สองหน่วย ใช้เวลาสอน

---

[1]Crehen, "Item Analysis...," pp.255-262.

[2]Haladyna, "Effects of Different Samples...," pp.93-99.

7 ชั่วโมง แล้วให้ทำแบบสอบคู่ขนาน 3 แบบ ซึ่งสร้างให้ข้อสอบในแบบสอบวัดจุดมุ่งหมายที่กำหนดไว้เพื่อจำแนกระหว่างผู้รอบรู้และไม่รอบรู้ แล้วสุ่มข้อกระทงใส่ในแบบสอบ นำแบบสอบไปวัดก่อนและหลังเรียน ใช้เกณฑ์ตัดสินร้อยละ 70 หลังจากนั้นกำหนดค่าสถิติจากกลุ่มตัวอย่างสามกลุ่มคือ กลุ่มก่อนเรียน กลุ่มหลังเรียนและกลุ่มผสมจากสองกลุ่ม ใช้สูตรของคูเดอร์-ริชาร์ดสัน สูตรที่ 20 คำนวณความเป็นเอกพันธ์ภายในของแบบสอบ พบว่า ค่ามัชฌิมเลขคณิตของกลุ่มหลังเรียนสูงกว่ากลุ่มก่อนเรียนมาก ($p < .01$) ความแปรปรวนของคะแนนจะเพิ่มขึ้นตามประสิทธิภาพของการสอน และจากการรวมคะแนนก่อนและหลังเรียนเข้าด้วยกัน ความเป็นเอกพันธ์ดูจะเป็นมโนมติที่ใช้การได้เมื่อใช้กลุ่มผสม และมีจำนวนหัวข้อที่จะวัดในแต่ละหน่วยต่ำ ดัชนีค่าอำนาจจำแนกแบบ พอยท์ ไบซีเรียล ให้ค่าใกล้เคียงกับ D% ของซู (Hsu) ในขณะที่ค่า D% ต้องใช้การคำนวณสองครั้ง ค่าที่ได้จากพอยท์ ไบซีเรียล ใช้การคำนวณเพียงครั้งเดียว ฮาลาได้นำ กล่าวว่า ในแง่ของการสอนแล้ว D% ตรงกับความหมายของแบบสอบอิงเกณฑ์มากกว่า ในขณะที่ค่าที่ได้จากพอยท์ ไบซีเรียล มีประสิทธิภาพที่สุดในการให้ข้อมูลเกี่ยวกับความครบถ้วนถูกต้อง (Adequacy) ของข้อกระทงในแบบสอบอิงเกณฑ์

ฮวน[1] ได้ใช้ข้อมูลจากแบบสอบเลขคณิตซึ่งมีความยาว 5 ข้อ กลุ่มตัวอย่างที่ใช้ 91 คน คำนวณหาค่าแคปป้า (K) ได้ข้อสรุปว่า ทั้งความเป็นวิวิธพันธ์ของคะแนนและคะแนนเกณฑ์ที่ต่างกันมีผลต่อค่าแคปป้า ความแปรปรวนของคะแนนมีสหสัมพันธ์ทางบวกกับค่าความเที่ยง และค่าความเที่ยงจะเพิ่มขึ้นตามความยาวของแบบสอบ อัตราการเพิ่มจะลดลงเมื่อแบบสอบยาวมาก ๆ ในปีเดียวกันนี้ ซับโคเวียค[2] ได้ทำการประมาณค่าความเที่ยงโดยบริหารการสอบเพียง 1 ครั้ง ใช้สูตรในการประมาณค่าคะแนนจริงสามวิธีคือ ใช้ค่า P_o ของ

---

[1] Huynh, "On the Reliability of...," pp. 253-264.

[2] Subkoviak, "Estimating Reliability...," pp.265-276.

แฮมเบิลตั้น และ โนวิค (1973) ค่า P_c ของซับโคเวียค และ $\hat{P}_i$ ซึ่งใช้สูตรของคูเดอร์-ริชาร์ดสัน สูตรที่ 21 คือ

$$\hat{P}_i = \alpha_{21/x}\left(\frac{x_i}{n}\right) + (1 - \alpha_{21/x})\left(\frac{M_X}{n}\right)$$

เมื่อ $$\alpha_{21/x} = \frac{n}{n-1}\left[1 - \frac{M_X(n-M_X)}{n\sigma_X^2}\right]$$

เกณฑ์ที่ใช้มีตั้งแต่ 12 ถึง 24 พบว่าค่าความเที่ยงที่ได้จาก P_o มากกว่า $\hat{P}_i$ และ $\hat{P}_i$ มากกว่า P_c ผลที่เกิดจากจุดตัดที่ต่างกันต่อค่าสัมประสิทธิ์ความเที่ยงทั้งสามแบบคล้ายคลึงกัน กล่าวคือมีค่าสูงสุดเมื่อเกณฑ์มีค่าต่ำสุดและสูงสุด และมีค่าต่ำสุดเมื่อเกณฑ์อยู่ตรงกลาง ๆ

ซับโคเวียค[1] ได้ศึกษาเปรียบเทียบค่าดัชนีความเที่ยงเมื่อคำนวณด้วยวิธีต่าง ๆ 4 วิธีคือ

1. บริหารการสอบสองครั้ง คือวิธีของสแวมมินนาธาน, แฮมเบิลตั้น และอัลกิน่า (1974, 1975)

2. บริหารการสอบหนึ่งครั้ง คือวิธีของ

2.1 ฮวน (Huynh, 1976)

2.2 มาร์แชล และ แฮร์เทล (Marshall and Haertel, 1976)

2.3 ซับโคเวียค (Subkoviak, 1976)

เปรียบเทียบค่าความเที่ยง 4 วิธีที่ได้จากตัวอย่างกับที่ได้จากประชากรโดยตรง โดยใช้นักเรียนจำนวน 1586 คน เป็นประชากร แล้วทำการสุ่มแบบแทนที่เป็นกลุ่มตัวอย่างกลุ่มละ 30 คน รวม 50 กลุ่ม ให้ทุกคนทำแบบสอบวัดความถนัดทางวิชาการ (Scholastic Aptitude Test) ที่เป็นแบบสอบคู่ขนานมีความยาว 10, 30 และ 50 ข้อ ใช้เกณฑ์ตัดสิน

---

[1]Michael J. Subkoviak, "Empirical Investigation of Procedures for Estimating Reliability for Mastery Test," Journal of Educational Measurement 15 (1978) : 111 - 116.

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

4 เกณฑ์ คือ ร้อยละ 50, 60, 70 และ 80 พบว่า ถ้าแบบสอบยาวขึ้น ค่าความคลาดเคลื่อนมาตรฐานจะลดลง และค่าประมาณที่ได้จะคงที่และถูกต้องมากขึ้นเมื่อค่าความเที่ยงที่ได้จากประชากรมีค่าสูงมากหรือต่ำมาก และพบว่าแม้ว่าจะใช้ข้อมูลจากแบบสอบที่เป็นวิวิธพันธ์ แต่ไม่มีข้อจำกัดที่สำคัญมากนัก และที่น่าสนใจคือ ค่าประมาณของมาร์แชล, ซับโคเวียค และฮวน กลับให้ค่าประมาณที่ถูกต้อง เมื่อเปรียบเทียบวิธีการทั้งสี่แล้ว วิธีของฮวนให้ค่าประมาณที่ถูกต้องที่สุด เมื่อพิจารณาจากหลาย ๆ ด้านรวมกัน ในปีเดียวกัน อัลกิน่า และ โน[1] ได้ศึกษาถึงความเชื่อถือได้ของความเที่ยงในการตัดสินเมื่อใช้ค่าประมาณคะแนนโดเมน ความเที่ยงของแบบสอบ จำนวนผู้สอบและเกณฑ์ตัดสินต่าง ๆ กัน สูตรที่ใช้คำนวณค่าประมาณคะแนนโดเมน คือ สัดส่วนความถูกต้อง (P_i) และ $\hat{T}_i$ ซึ่งคำนวณจากสูตร

$$\hat{T}_i = \hat{\beta} P_i + (1 - \hat{\beta}) \hat{\mu}_P \quad (i = 1, 2, \ldots N),$$

เมื่อ

P_i คือ สัดส่วนความถูกต้อง

$\hat{\beta}$ คือ ค่าสัมประสิทธิ์จากสูตรคูเดอร์-ริชาร์ดสัน สูตรที่ 20

$\hat{\mu}_P$ คือ ค่ามัชฌิมเลขคณิตในรูปของสัดส่วน

ปรากฏข้อค้นพบสรุปได้ดังนี้

1. เมื่อคะแนนเกณฑ์มีค่าใกล้มัชฌิมเลขคณิตของคะแนนดิบ ค่าสัมบูรณ์ของความคลาดเคลื่อนในการประมาณจะมีค่าสูงสุด (มีค่าเป็นลบเมื่อใช้ $\hat{T}_i$ และเป็นบวกเมื่อใช้ P_i) และจะลดลงเมื่อคะแนนเกณฑ์หางออกมาทั้งสองข้าง จนกระทั่งเครื่องหมายเปลี่ยนไป

---

[1] James Algina and Michael J. Noe, "A Study of The Accuracy of Subkoviak's Single - Administration Estimate of the Coefficient of Agreement Using Two True - Score Estimates," Journal of Educational Measurement 15 (1978) : 101 - 110.

แล้วจึงเพิ่มขึ้นเรื่อย ๆ แล้วลดลงอีก ความเปลี่ยนแปลงนี้เห็นได้ชัดในแบบสอบที่มีความยาวที่สุด

2. ค่าความเที่ยงของแบบสอบต่างกันจะให้ผลต่อความเที่ยงในการตัดสินต่างกัน และเมื่อใช้ค่าประมาณคะแนนจริงต่างกันก็ให้ผลเช่นเดียวกัน เมื่อค่าความเที่ยงของแบบสอบเกิน .50 ควรใช้ P_i ประมาณค่าคะแนนจริง และถ้าใช้ $\hat{T}_i$ ประมาณค่าคะแนนจริงแล้วทำให้ค่าความเที่ยงในการตัดสินสูงมากไป วิธีที่ดีคือใช้ค่าเฉลี่ยของสองแบบรวมกัน

3. ความแปรปรวนจากตัวประมาณค่าทั้งสองตัวเพิ่มขึ้นเมื่อลดจำนวนผู้สอบแต่ไม่มากนัก

สมหวัง พิธิยานุวัฒน์[1] ได้ศึกษาเรื่องจุดแบ่งที่พอดีของแบบสอบอิงเกณฑ์ทางสถิติศาสตร์ เพื่อระบุหาจุดแบ่งที่พอดีของแบบสอบอิงเกณฑ์ทางสถิติศาสตร์โดยเริ่มต้นหาจุดแบ่งขั้นต่ำซึ่งเป็นจุดตัดระหว่างการแจกแจงของคะแนนกลุ่มที่มีความรู้พื้นฐานทางสถิติอย่างนำไปใช้การได้กับกลุ่มที่ขาดความรู้พื้นฐานทางสถิติ และหาจุดแบ่งขั้นสูงซึ่งประมวลจากความเห็นของบลูม (Bloom) และ บล็อค (Block) จากจุดแบ่งขั้นต่ำและจุดแบ่งขั้นสูงดังกล่าวนำมาระบุหาจุดแบ่งที่พอดีของแบบสอบอิงเกณฑ์ทางสถิติศาสตร์ประจำหน่วยบทเรียนที่สี่ คิดเป็นคะแนนร้อยละ 70

** จากงานค้นคว้าและวิจัยที่เกี่ยวข้อง พบว่า ในขณะที่การศึกษาเกี่ยวกับแบบสอบอิงเกณฑ์ในประเทศไทยยังอยู่ในระยะเริ่มต้นนั้น นักวัดและประเมินผลในสหรัฐอเมริกากำลังให้ความสนใจศึกษาในเรื่องนี้กันอย่างกว้างขวางและนำไปใช้ในการทดสอบหลายสาขา และเพื่อจุดประสงค์หลายอย่าง มิใช่เพียงในด้านการวัดสัมฤทธิผลทางการศึกษาเท่านั้น ดังที่ แฮมเบิลตั้น และ เอคเนอร์[2] กล่าวว่า "ในสองสามปีที่ผ่านมานี้ในประเทศสหรัฐอเมริกาได้

---

[1]สมหวัง พิธิยานุวัฒน์, "จุดแบ่งที่พอดีของแบบสอบอิงเกณฑ์ทางสถิติศาสตร์ : การศึกษาเบื้องต้น," ใน <u>เอกสารประกอบการประชุมซิมโปเซี่ยม</u>, (กรุงเทพมหานคร : สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, 2520), หน้า 13.

[2]Hambleton, "Guidelines for Evaluating...," p.321.

มีการสร้างแบบสอบอิงเกณฑ์ขึ้นมาใช้ทั่วไป ไม่ว่าในระดับรัฐ ท้องถิ่น โรงเรียนและหน่วยงานต่าง ๆ ที่มีหน้าที่เกี่ยวกับการทดสอบหรือการแนะแนว การออกแบบก็เพื่อสำรวจความก้าวหน้าในโปรแกรมการสอนในโรงเรียนเพื่อวินิจฉัยหาข้อบกพร่องของนักเรียน เพื่อรายงานความก้าวหน้าของนักเรียนแก่ผู้ปกครอง เพื่อประเมินโปรแกรมการศึกษาต่าง ๆ หรือเพื่อให้การรับรอง หรือให้อนุญาตในการประกอบอาชีพในสาขาต่าง ๆ มากมาย" พร้อมกันนี้ก็ได้เสนอรายชื่อแบบสอบอิงเกณฑ์ที่มีอยู่ในปัจจุบันจำนวนหนึ่งดังต่อไปนี้[1]

---

[1] Hambleton, "Guidelines for Evaluating...," p.324.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 2 รายชื่อแบบสอบอิงเกณฑ์ที่พัฒนาขึ้นใช้ในประเทศสหรัฐอเมริกา

รหัส (Code)	ชื่อแบบสอบ (Name of Test)	ระดับชั้น (Grades)	ระดับ (Levels)	ฟอร์ม (Forms)	ปีที่พิมพ์ (Publication Date)	ผู้พิมพ์ (Publisher)
1	1976 Stanford Diagnostic Mathematics Test	1-12	4	2	1976	Harcourt Brace Jovanovich
2	1976 Stanford Diagnostic Reading Test	1-12	4	2	1976	Harcourt Brace Jovanovich
3	Skills Monotoring	3-5	3	1	1975	Harcourt Brace Jovanovich
4	Individual Pupil Monitoring System-Mathematics	1-6	6	2	1974	Houghton-Mifflin
5	Individual Pupil Monitoring System-Mathematics	1-8	8	2	1974	Houghton-Mifflin
6	Diagnostic Mathematics Inventory	1.5-7,5	7	1	1977	CTB/McGraw-Hill
7	Prescriptive Reading Inventory	K-6.5	6	1	1977	CTB/McGraw-Hill
8	Diagnosis : An Instructional Aid-Mathematics and Reading	1-6	2	2	1974	Science Research Associates
9	Mastery : An Evaluation Tool-SOBAR Reading	K-9	10	2	1975	Science Research Associates
10	Mastery : An Evaluation Tool-Mathematics	K-8	9	2	1974	Science Research Associates
11	Fountain Valley Support System in Mathematics	K-8	9	1	1974	Richard L. Zweig Associates