ผลของการกำหนดน้ำหนักคะแนนต่อคุณภาพของแบบสอบ

สุพัฒน์ สุกมลสันต์

แนวคิดเชิงทฤษฎี

เป็นที่ทราบกันดีในหมู่นักทดสอบและประเมินผลว่า คุณภาพที่สำคัญของแบบสอบมีหลายอย่าง แต่ที่จำเป็นที่แบบสอบต้องมี คือ ความตรง (validity) และความเที่ยง (reliability) ซึ่งลักษณะทั้ง 2 ประการนี้ มีความสัมพันธ์กัน เช่น แบบสอบจะมีค่าความตรงเชิงพยากรณ์สูงสุดได้ไม่เกิน $\sqrt{r_{tt}}$ เป็นต้น (Magnuson, 1966 : 149) ค่าความเที่ยงและความตรงของแบบสอบจะมีค่ามากหรือน้อยขึ้นอยู่กับการกระจายของคะแนนและความยาวของแบบสอบ เช่น ในกรณีของความเที่ยงนั้น ถ้าการกระจายของคะแนนของแบบสอบยิ่งมีมาก แบบสอบก็ยิ่งมีค่าความเที่ยงมาก และถ้าแบบสอบมีความยาวเพิ่มขึ้น ค่าความเที่ยงก็จะเพิ่มขึ้นเช่นเดียวกัน (Magnuson, 1966 : 66)

สำหรับการวัดและประเมินผลนั้น นักทดสอบต้องการวัดสิ่งที่มุ่งวัดให้ได้มากที่สุด และต้องการให้ผลการวัดมีความคงที่มากที่สุดเท่าที่จะสามารถทำได้ ดังนั้น นักทดสอบจึงพยายามทำให้แบบสอบมีความตรงและความเที่ยงให้สูงที่สุดเท่าที่จะทำได้ (Ebel, 1974 : 408; Lord, 1980 : 5) แต่ว่าในขณะเดียวกัน ก็ต้องการทำให้แบบสอบมีจำนวนข้อที่เหมาะสม เพราะหากว่า แบบสอบมีความยาวมากเกินไป จะใช้เวลาสอบนานเกินไป จะทำให้มีผลเสียทางจิตวิทยา เช่น ความเมื่อยล้า เบื่อหน่าย และหงุดหงิด เป็นต้น ซึ่งผลดังกล่าวนี้มีผลทำให้ผลการสอบคลาดเคลื่อนได้

ดังนั้น ในปัจจุบันนี้ นักทดสอบจึงพยายามหาทางทำให้แบบมีความสั้นลง (Ebel, 1972 : 107) เช่น ทำให้แบบสอบมีน้อยข้อ และใช้เวลาสอบน้อยที่สุด (Stecklein and Others, 1981 : 208) กำหนดความยาวที่เหมาะสมของแบบสอบ (Lord, 1980 : 173) กำหนดจำนวนตัวเลือกที่เหมาะสม (Lord, 1977 : 33) และการสอบที่เหมาะสมเป็นราย

บุคคล (tailored testing : Urry, 1977 : 181) เป็นต้น แต่ว่าวิธีการต่าง ๆ เหล่านี้ ผู้ใช้แบบสอบก็ยังต้องการให้แบบสอบมีความตรงและความเที่ยงของแบบสอบมีมากที่สุดเท่าที่จะสามารถทำได้ แต่ให้มีข้อสอบจำนวนน้อย ซึ่งนอกจากวิธีต่าง ๆ ดังกล่าวแล้วข้างต้น วิธีหนึ่งที่นักทดสอบได้ทำการศึกษากันมานานแล้ว ก็คือ การให้คะแนนที่เหมาะสมแก่ข้อสอบแต่ละข้อหรือแต่ละตัวเลือก เพราะเชื่อว่า วิธีการดังกล่าวนี้จะสามารถทำให้แบบสอบมีการกระจายของคะแนนมากเพิ่มมากกว่าการให้คะแนนแบบประเพณีนิยม ซึ่งให้ข้อที่ถูก = 1 คะแนน และข้อที่ผิด = 0 คะแนน เท่านั้น

แต่อย่างใดก็ตาม แม้ว่านักทดสอบจะได้พยายามให้คะแนนที่เหมาะสมแก่ข้อสอบรายข้อ หรือรายตัวเลือกแตกต่างกันหลาย ๆ วิธี เช่น 1) ใช้คะแนนโดยอาศัยความเห็นของผู้เชี่ยวชาญ 2) ใช้ค่ามาตรวัดจากวิธีของ Guttman 3) ใช้ค่ามาตรวัดจากวิธีการเปรียบเทียบรายคู่ของ Thurstone 4) ใช้คะแนนเมื่อหักค่าการเดาแล้ว 5) ใช้คะแนนตามน้ำหนักความเชื่อมั่นในการตอบ (Confidence weighting) และวิธีอื่น ๆ อีกมาก แต่ว่าวิธีการต่าง ๆ เหล่านี้แม้จะมีผู้ทำการศึกษาติดต่อกันมาเรื่อย ๆ ประมาณ 30 ปีแล้ว (อ้างจาก Wang and Stanley, 1970 : 691) แต่ว่าผลของการศึกษามักไม่คงที่ กล่าวคือ บางครั้งบางวิธีทำให้แบบสอบมีค่าความเที่ยงสูง และความตรงเชิงพยากรณ์สูง แต่บางครั้งทำให้ค่าทั้งสองต่ำ หรือทำให้บางค่าสูง แต่บางค่าต่ำกว่าค่าความเที่ยงและความตรงที่คิดจากคะแนนที่ให้โดยวิธีประเพณีนิยม

แต่ในปัจจุบันนี้ นักทดสอบมีความหวังว่า 1) การให้คะแนนโดยอาศัยแนวคิดจากรูปแบบของความสามารถแฝง (Latent Trait Models) และ 2) การให้คะแนนโดยอาศัยแนวคิดจากการถดถอยพหุคูณ น่าจะเป็นวิธีที่จะทำให้การให้คะแนนข้อสอบแต่ละข้อหรือแต่ละตัวเลือกมีค่าน้ำหนักคะแนนแตกต่างกัน จะทำให้แบบสอบมีค่าความเที่ยงและความตรงเพิ่มขึ้นได้ (Wang and Stanley, 1970 : 670) ดังนั้น ตั้งแต่ปี 1970 เป็นต้นมา การวิจัยที่เกี่ยวข้องกับการให้คะแนนข้อสอบหรือตัวเลือกแตกต่างกัน จึงได้รับความสนใจจากนักทดสอบมาก ผู้วิจัยเองก็มีความสนใจในเรื่องดังกล่าว จึงได้ทำการวิจัยเกี่ยวกับผลของการกำหนดน้ำหนักคะแนน ให้ข้อสอบแต่ละข้อ 6 วิธี เพื่อเปรียบเทียบผลที่เกิดขึ้นว่าจะมีผลต่อคุณภาพของข้อสอบมากน้อยเพียงใด วิธีการให้คะแนนข้อสอบดังกล่าวแล้ว คือ

1. วิธีใช้ค่าเดลต้า (Δ)
2. วิธีใช้ค่าเบต้าแปลง (β_t)
3. วิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w)
4. วิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2)
5. วิธีใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) และ
6. วิธีประเพณีนิยม

วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบว่า วิธีการให้คะแนนข้อสอบแต่ละข้อ 6 วิธีดังกล่าวแล้ว ทำให้แบบสอบการอ่านเข้าใจความภาษาอังกฤษมีค่าความตรงเชิงพยากรณ์แตกต่างกันอย่างมีนัยสำคัญหรือไม่ ในกรณีที่เป็นแบบสอบอิงกลุ่ม

2. เพื่อเปรียบเทียบว่า วิธีการให้คะแนนข้อสอบแต่ละข้อ 6 วิธีดังกล่าวแล้ว ทำให้แบบสอบการอ่านเข้าใจความภาษาอังกฤษมีค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญหรือไม่ ในกรณีที่เป็นแบบสอบอิงกลุ่ม

3. เพื่อเปรียบเทียบว่า วิธีการให้คะแนนทั้ง 6 วิธี ทำให้ค่าความเที่ยงของแบบสอบอิงเกณฑ์ (ค่าดัชนีความเชื่อถือ : Dependability Index) แตกต่างกันหรือไม่

4. เพื่อศึกษาว่า วิธีการให้คะแนนข้อสอบแต่ละข้อต่างกัน 6 วิธีนั้น ทำให้อันดับที่ของคะแนนการสอบคงที่หรือไม่ และมากน้อยเพียงใด เมื่อเปรียบเทียบกับอันดับที่ของคะแนนจากการให้คะแนนวิธีประเพณีนิยม

5. เพื่อศึกษาว่า คะแนนที่ได้จากวิธีการให้คะแนนทั้ง 6 วิธี มีความสัมพันธ์ซึ่งกันและกันมากน้อยเพียงใด

สมมุติฐานในการวิจัย

1. ในกรณีการสอบแบบอิงกลุ่ม แบบสอบที่ให้คะแนนรายข้อโดยวิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w) มีค่าความตรงเชิงพยากรณ์สูงที่สุด

2. ในกรณีการสอบแบบอิงกลุ่ม แบบสอบที่ให้คะแนนรายข้อโดยวิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w) มีค่าความเที่ยงสูงที่สุด

3. ในกรณีการสอบแบบอิงเกณฑ์ แบบสอบที่ให้คะแนนรายข้อโดยวิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w) มีค่าความเที่ยง (ค่าดัชนีความเชื่อถือ) สูงที่สุด

4. การให้คะแนนต่างกันแต่ละวิธี ทำให้อันดับที่ของคะแนนการสอบการอ่านเข้าใจความภาษาอังกฤษเปลี่ยนไปจากอันดับที่ของคะแนนที่ให้โดยวิธีประเพณีนิยม

5. คะแนนที่ได้จากวิธีการให้คะแนนรวมข้อทั้ง 6 วิธี ต่างมีความสัมพันธ์ซึ่งกันและกันอย่างมีนัยสำคัญ

ขอบเขตการวิจัย

1. ข้อความเพื่อการอ่านเข้าใจความภาษาอังกฤษมีความยาวประมาณ 300 คำ ที่มีความยากง่าย (readability) ในระดับเหมาะสมกับนิสิตชั้นปีที่ 1

2. การอ่านเข้าใจความมุ่งเน้นที่การอ่านไม่ออกเสียง

3. พลวิจัย (Subjects) เพื่อการวิจัย ได้แก่ นิสิตชั้นปีที่ 1 ของจุฬาลงกรณ์มหาวิทยาลัย ที่ลงทะเบียนเรียนวิชาภาษาอังกฤษพื้นฐานเฉพาะการอ่าน II

ข้อตกลงเบื้องต้น

1. ความสามารถในการอ่านเข้าใจความภาษาอังกฤษมีลำดับขั้น ผู้ที่อ่านข้อความได้เข้าใจมาก ย่อมมีและใช้ความสามารถเพื่อทำความเข้าใจเรื่องที่อ่านมาก ส่วนผู้ที่อ่านเรื่องเดียวกันแล้ว เข้าใจได้น้อยกว่า แสดงว่า มีและ/หรือใช้ความสามารถเพื่อการอ่านน้อยกว่า

ดังนั้น ข้อสอบข้อที่ง่ายหรือข้อที่วัดลำดับขั้นความเข้าใจในการอ่านในลำดับต่ำจึงย่อมมีค่าน้ำหนักคะแนนน้อยกว่าข้อที่ยากกว่า หรือข้อที่วัดลำดับขั้นความเข้าใจในการอ่านในลำดับที่สูงกว่า (สำหรับการให้คะแนนวิธีที่ 1, 2 และ 5)

2. ระดับความสัมพันธ์ของข้อสอบกับสิ่งที่มุ่งทดสอบ เป็นปัจจัยหนึ่งที่แสดงถึงระดับความเข้าใจเนื้อหาสาระของสิ่งที่อ่าน

ดังนั้น ข้อสอบข้อที่มีระดับความสัมพันธ์กับสิ่งที่มุ่งทดสอบต่ำ จึงย่อมมีค่าน้ำหนักคะแนนน้อยกว่าข้อที่มีค่าดังกล่าวสูงกว่า (สำหรับการให้คะแนนวิธีที่ 4)

3. คุณลักษณะที่สำคัญของข้อสอบแต่ละข้อ คือ ระดับความยากง่าย อำนาจจำแนกและค่าการเดา ค่าดังกล่าวนี้มีความสัมพันธ์ซึ่งกันและกันในการวัดความสามารถในการอ่านเข้าใจความ

ดังนั้น ข้อสอบที่มีความสัมพันธ์กันของค่าทั้งสามต่ำ ย่อมมีค่าน้ำหนักคะแนนน้อยกว่าข้อที่มีค่าดังกล่าวสูงกว่า (สำหรับการให้คะแนนวิธีที่ 3)

วิธีการดำเนินการวิจัย

ประชากรและพลวิจัย

ประชากรสำหรับการวิจัยครั้งนี้ ได้แก่ นิสิตชั้นปีที่ 1 ของจุฬาลงกรณ์มหาวิทยาลัยที่เรียนภาษาอังกฤษพื้นฐานเฉพาะการอ่าน 2 ประจำภาคปลายปีการศึกษา 2527 ซึ่งมีจำนวน 2,794 คน ส่วนพลวิจัยครั้งนี้ได้จากการสุ่มตัวอย่าง โดยการจับสลากห้องเรียน จาก 97 ห้องเรียนได้ตัวอย่าง 75 ห้องเรียน จาก 14 คณะ และมีพลวิจัย 1,872 คน หรือประมาณร้อยละ 67% จากประชากรทั้งหมด

เครื่องมือการวิจัย

เครื่องมือการวิจัย ได้แก่ แบบทดสอบแบบเลือกตอบ เพื่อการทดสอบการอ่านเข้าใจความ แบบทดสอบนี้ประกอบด้วยเรื่อง 2 เรื่อง แต่ละเรื่องมีความยาวประมาณ 300 คำ และมีระดับความยากง่ายที่วัดด้วยวิธีการโคลซ (cloze procedure) แล้วว่า อยู่ในระดับที่เหมะสมกับนิสิตชั้นปีที่ 1 แบบทดสอบนี้มีความยาวเที่ยง $(KR_{20}) = 0.882$ $(SE_m = 2.970)$ และมีความยาว 50 ข้อ ใช้เวลาสอบ 1 ชั่วโมง แบบทดสอบนี้ได้ทำการวิเคราะห์โดยการแยกองค์ประกอบแล้วพบว่า มีเพียงมิติเดียว (ตามวิธีของ Lord, 1980:21) สร้างให้มีลักษณะเป็นแบบสอบอิงกลุ่ม (Norm-referenced test) และแบบทดสอบอิงเกณฑ์

(Criterion–referenced test) โดยยึดวิธีการแปลความหมาย (interpretation) ของคะแนนเป็นเกณฑ์ในการแบ่งตามแนวคิดของนักทดสอบหลายท่าน (Hively, 1974 : 143; Hambleton and Others, 1978 : 3, Brennan, 1984 : 329; Block and Dockrell, 1984 : 36)

การรวบรวมข้อมูล

จัดทำแบบสอบให้มีลักษณะแตกต่างกัน 2 ฉบับ แต่ละฉบับเรียงลำดับเรื่องสำหรับอ่านแตกต่างกัน แล้วติดต่อให้อาจารย์ประจำรายวิชา นำแบบสอบไปสอบกับนิสิตภายในชั่วโมง การทดสอบนี้กระทำพร้อมกันภายในหนึ่งสัปดาห์

การวิเคราะห์ข้อมูล

1. ทำแบบทดสอบมาวิเคราะห์ข้อสอบรายข้อ 3 วิธี คือ 1) แบบประเพณีนิยม (เทคนิค 27%) เพื่อหาค่าเดลต้า ($\triangle$) 2) ตามรูปแบบราสช์ (Rasch Model) เพื่อหาค่าเบต้า (β) และ 3) ตามรูปแบบใช้ค่าพารามิเตอร์ 3 ตัว (Three–parameter Model) เพื่อหาค่าอำนาจจำแนก (a) ค่าความยากง่าย (b) และค่าการเดา (c)

2. นำค่าต่าง ๆ ที่ได้จากการวิเคราะห์ข้อสอบมาแปลง หรือคำนวณหาค่าน้ำหนักคะแนนรายข้อของข้อสอบ 3 ค่า คือ ค่าเดลต้า ($\triangle$) ค่าเบลต้าแปลง (β_t) และค่าน้ำหนักคะแนนที่เหมาะสม (w) ตามวิธีของ Lord (Lord, 1980 : 76)

3. นำแบบทดสอบไปวิเคราะห์ตัวประกอบ (factors) เพื่อหาค่าผลบวกความแปรปรวนร่วม (h^2) ของข้อสอบแต่ละข้อ เพื่อใช้เป็นค่าน้ำหนักคะแนนรายข้อของแบบสอบ

4. นำค่าน้ำหนักคะแนนที่คำนวณได้แต่ละวิธีไปแทนค่าน้ำหนักของข้อสอบรายข้อ คือ 1) ค่าเดลต้า ($\triangle$) 2) ค่าเบต้าแปลง (β_t) 3) ค่าน้ำหนักคะแนนที่เหมาะสม (w) 4) ค่าผลบวกความแปรปรวนร่วม (h^2) 5) ค่าลำดับขั้นความเข้าใจในการอ่าน (h) ที่ข้อสอบทำการทดสอบ (Valette and Disick, 1972 : 161 และ 6) ค่าน้ำหนักตามแบบประเพณีนิยม คือ ข้อที่ถูก = 1 คะแนน ข้อที่ผิด = 0 คะแนน

5. คำนวนหาค่าความตรงเชิงพยากรณ์และความเที่ยงแบบ Cronbach's α ของแบบสอบอิงกลุ่มที่ให้คะแนนแต่ละวิธี โดยใช้คะแนนรวมของการสอบการอ่านเข้าใจความกลางภาคและปลายภาคของนิสิตเป็นเกณฑ์

6. คำนวณค่าความเที่ยง (ค่าดัชนีความเชื่อถือ) ของแบบสอบอิงเกณฑ์ที่ให้คะแนแต่ละวิธีตามวิธีของ Brennan และ Kane, (Brennan and Kane,1977 : 277–283)

7. ทดสอบความแตกต่างของค่าความตรงเชิงพยากรณ์ของแบบสอบตามวิธีของ Olkin (อ้างจาก Tatnaik and Tranb, 1973 : 284) และทดสอบความแตกต่างของค่าความเที่ยงตามวิธีของ Kristof (Kristof, 1964 : 105)

8. นำค่าความเที่ยง (ค่าดัชนีความเชื่อถือ) ของจุดตัด 0%–100% ของแบบสอบที่คิดคะแนนวิธีต่าง ๆ มาเขียนกราฟเปรียบเทียบกัน

9. คำนวณหาความสัมพันธ์ของอันดับที่ของคะแนนที่ได้จากการให้คะแนนวิธีต่าง ๆ ด้วยวิธี Kendall's Tou และหาค่าสหสัมพันธ์ซึ่งกันและกันของคะแนนที่ได้จากวิธีการให้คะแนนทั้ง 6 วิธี

ผลของการวิจัย

ก. ความแตกต่างของค่าความตรงเชิงพยากรณ์

ตารางที่ 1 ค่าความตรงเชิงพยากรณ์ของแบบสอบเมื่อให้น้ำหนักคะแนนวิธีต่าง ๆ

ที่	วิธีให้คะแนน	r_{xy}
1.	วิธีใช้ค่า h	0.372**
2	วิธีประเพณีนิยม	0.370**
3	วิธีใช้ค่า h^2	0.368**
4	วิธีใช้ค่า △	0.366**
5	วิธีใช้ค่า t	0.361**
6	วิธีใช้ค่า w	0.357*

*$p < .05$

**$p > .05$

จากตารางที่ 1 แสดงให้เห็นว่า การให้คะแนนแบบสอบโดยวิธีใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) ทำให้แบบสอบมีค่าความตรงเชิงพยากรณ์สูงที่สุด คือ 0.372 และเมื่อทำการทดสอบความแตกต่างด้วยวิธีของ Olkin แล้ว พบว่า ค่าความตรงดังกล่าวมีขนาดแตกต่างจากค่าความตรงที่คิดจากวิธีประเพณีนิยม วิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) วิธีใช้ค่าเบต้า $(\triangle)$ และวิธีใช้ค่าเบต้าแปลง (β_t) อย่างไม่มีนัยสำคัญ แต่ความตรงดังกล่าวเหล่านี้ต่างมีค่าสูงกว่าค่าความตรงของแบบสอบที่ให้คะแนนโดยวิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w) ซึ่งมีค่า 0.357 อย่างมีนัยสำคัญ

ดังนั้น ผลการวิจัยพบว่า ขัดแย้งกับสมมติฐานในการวิจัยข้อที่ 1

ข. ความแตกต่างของค่าความเที่ยงของแบบสอบอิงกลุ่ม

ตารางที่ 2 ค่าความเที่ยงของแบบสอบอิงกลุ่มและค่าความคลาดเคลื่อนมาตรฐานในการวัด

ที่	วิธีให้คะแนน	r_{tt}	SE_m
1	วิธีใช้ค่า h	0.885*	1.702
2	วิธีประเพณีนิยม	0.859*	2.726
3	วิธีใช้ค่า h^2	0.843**	1.268
4	วิธีใช้ค่า $\triangle$	0.842**	5.301
5	วิธีใช้ค่า β_t	0.830**	34.793
6	วิธีใช้ค่า w	0.801*	40.072

*$p < .05$

**$p > .05$

จากตารางที่ 2 แสดงให้เห็นว่า ความเที่ยงของแบบสอบที่ให้คะแนนโดยวิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w) มีค่าสูงที่สุดคือ 0.885 และการให้คะแนนโดยวิธีใช้ค่าเบต้าแปลง (β_t) ทำให้แบบสอบมีค่าความเที่ยงต่ำที่สุด คือ 0.801 และเมื่อทำการทดสอบความแตกต่างของค่าความเที่ยงโดยวิธีของ Kristof แล้วพบว่า ค่าความเที่ยงของแบบสอบที่ให้คะแนนโดยวิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) วิธีใช้ค่าลำดับขั้นความเข้าใจใน

การอ่าน (h) และวิธีใช้ค่าเดลต้า (△) แตกต่างกันอย่างไม่มีนัยสำคัญ แต่มีค่าน้อยกว่าค่าความเที่ยงที่คิดจากคะแนนโดยวิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w) และวิธีประเพณีนิยม และสูงกว่าค่าความเที่ยงที่คิดจากวิธีใช้ค่าเบต้าแปลง (β_t)

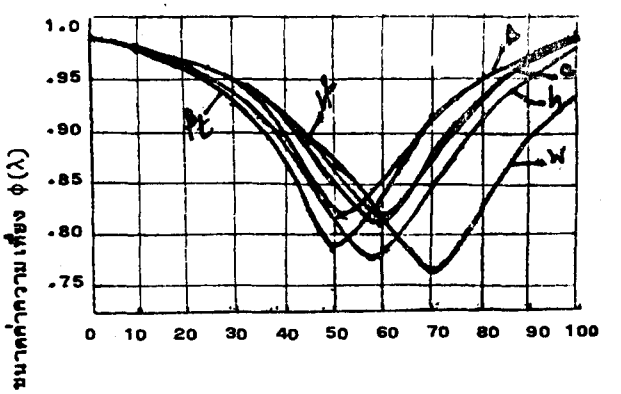
ดังนั้น ผลการวิจัยพบว่า สอดคล้องกับสมมุติฐานในการวิจัยข้อที่ 2

ค. ความแตกต่างของค่าความเที่ยง (ค่าความเชื่อถือได้) ของแบบสอบอิงเกณฑ์

ตารางที่ 3 และแผนภาพที่ 1 ค่าความเที่ยง (ค่าดัชนีความเชื่อถือ) ของแบบสอบเมื่อให้คะแนนโดยวิธีต่าง ๆ

	วิธีให้คะแนน					
λ	c*	Δ	β_t	h	h^2	w
0%	.988	.987	.985	.984	.988	.982
10%	.983	.982	.978	.978	.982	.976
20%	.975	.971	.964	.968	.973	.967
30%	.959	.950	.937	.948	.956	.952
40%	.927	.906	.875	.905	.920	.923
50%	.865	.822	.795	.825	.852	.870
60%	.825	.835	.842	.780	.820	.842
70%	.882	.910	.919	.855	.885	.778
80%	.940	.952	.957	.922	.937	.845
90%	.964	.972	.974	.955	.965	.906
100%	.977	.982	.988	.972	.977	.942
$\bar{\lambda}$	.935	.933	.928	.917	.932	.916

*c = วิธีประเพณีนิยม



จากตารางที่ 3 และแผนภาพที่ 1 จะสังเกตได้ว่า เมื่อจุดตัด (λ) = 0% ค่าความเที่ยงของแบบสอบอิงเกณฑ์ (ค่าดัชนีความเชื่อถือ) จะมีค่าใกล้เคียงกันมาก คือ มีค่าระหว่าง 0.982–0.988 และมีค่าความเที่ยงโดยเฉลี่ยแล้วใกล้เคียงกัน คือ ระหว่าง 0.916–0.935 แต่เมื่อจุดตัดเพิ่มขึ้นค่าความเที่ยงนั้นจะค่อย ๆ ลดลง และจะมีค่าต่ำที่สุดเมื่อจุดตัดมีค่าระหว่าง 50%–70% เป็นที่น่าสังเกตว่าค่าความเที่ยงของแบบสอบที่ให้คะแนนโดยวิธีใช้ค่าเดลต้า (△)

และใช้ค่าเบต้าแปลง (β_t) มีค่าต่ำที่สุดเมื่อจุดตัดเท่ากับ 50% วิธีใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) วิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) และวิธีประเพณีนิยม ต่างมีค่าต่ำที่สุดเมื่อจุดตัดเท่ากับ 60% และวิธีใช้ค่าคะแนนที่เหมาะสม (w) มีค่าความเที่ยงต่ำที่สุดเมื่อจุดตัดเท่ากับ 70% ดังนั้นจะเห็นได้ว่า กระสวน (pattern) ของการลดและเพิ่มของค่าความเที่ยงจากการให้คะแนนโดยวิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) และวิธีใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) มีความคล้ายคลึงกับค่าความเที่ยงของการให้คะแนนโดยวิธีประเพณีนิยมมากที่สุด

ดังนั้น เมื่อพิจารณาถึงขนาดของค่าความเที่ยงแล้ว ผลการวิจัยปรากฏว่า ขัดแย้งกับสมมุติฐานการวิจัยข้อที่ 3

ง. ความคงที่ของอันดับที่ของคะแนนสอบ

ตารางที่ 4 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างอันดับที่ของคะแนนสอบที่ให้คะแนนวิธีต่างกันกับอันดับที่ของวิธีประเพณีนิยม และร้อยละของความคงที่

ที่	วิธีให้คะแนน	Kendall's Tou	ร้อยและความคงที่
1	วิธีใช้ค่า h^2	0.948*	89.87
2	วิธีใช้ค่า △	0.935*	87.42
3	วิธีใช้ค่า h	0.934*	87.23
4	วิธีใช้ค่า β_t	0.904*	81.72
5	วิธีใช้ค่า w	0.902*	81.36

*$p < .05$

จากตารางที่ 4 แสดงให้เห็นว่า อันดับที่ของคะแนนสอบของแบบสอบที่คิดคะแนนโดยวิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) มีความคงที่มากที่สุด คือ ประมาณร้อยละ 90 ส่วนการให้คะแนนโดยวิธีใช้ค่าคะแนนที่เหมาะสม (w) ทำให้อันดับที่ของคะแนนมีความคงที่น้อยที่สุด คือ ประมาณร้อยละ 81 เท่านั้น ส่วนวิธีให้คะแนนวิธีอื่น ๆ ทำให้แบบสอบมีความคงที่ประมาณร้อยละ 82–87 หรืออาจกล่าวได้ว่า การให้คะแนนโดยวิธี

ต่าง ๆ ทำให้อันดับที่ของคะแนนสอบแตกต่างไปจากอันดับที่เดิมที่ได้จากการให้คะแนนโดยวิธีประเพณีนิยม ประมาณร้อยละ 10–20

ดังนั้น ผลการศึกษาครั้งนี้สอดคล้องกับสมมุติฐานของการวิจัยข้อที่ 4

จ. ความสัมพันธ์ของคะแนนสอบ

ตารางที่ 5 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคะแนนสอบของแบบสอบที่ให้คะแนนโดยวิธีต่างกัน

ที่	วิธีให้คะแนน	ใช้ค่า h^2	ใช้ค่า $\triangle$	ใช้ค่า h	ใช้ค่า β_t	ใช้ค่า w
1	วิธีประเพณีนิยม	0.997*	0.995*	0.992*	0.986*	0.980*
2	วิธีใช้ค่า h^2	–	0.996*	0.989*	0.989*	0.977*
3	วิธีใช้ค่า $\triangle$		–	0.996*	0.998*	0.954*
4	วิธีใช้ค่า h			–	0.979*	0.970*
5	วิธีใช้ค่า β_t				–	0.923*
6	วิธีใช้ค่า w					–

*$p < .05$

จากตารางที่ 5 แสดงว่า โดยทั่วไปแล้วคะแนนของแบบสอบที่ให้คะแนนโดยวิธีต่างกันทั้ง 6 วิธี ต่างมีความสัมพันธ์ซึ่งกันและกันในระดับที่สูงมาก คือ ระหว่าง 0.923–0.997 โดยเฉพาะคะแนนที่ให้โดยวิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) ใช้ค่าเดลต้า $(\triangle)$ ใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) และใช้ค่าเบต้าแปลง (β_t) และใช้ค่าคะแนนที่เหมาะสม (w) มีความสัมพันธ์กับคะแนนที่ให้โดยวิธีประเพณีนิยมสูงมาก คือ ระหว่าง 0.980–0.997

ผลการวิจัยครั้งนี้พบว่า สอดคล้องกับสมมุติฐานในการวิจัยข้อที่ 5

สรุปผลการวิจัย

1. ในกรณีของแบบสอบอิงกลุ่ม การให้คะแนนโดยวิธีใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) วิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) วิธีใช้ค่าเบต้าแปลง (B_t) วิธีใช้ค่าเดลต้า ($\triangle$) และวิธีประเพณีนิยม ต่างก็ทำให้แบบสอบมีค่าความตรงเชิงพยากรณ์แตกต่างกันอย่างไม่มีนัยสำคัญ แต่ทว่าวิธีต่าง ๆ เหล่านี้ทำให้แบบสอบมีค่าความตรงสูงกว่าค่าที่ได้จากการให้คะแนนโดยวิธีใช้ค่าคะแนนที่เหมาะสม (w) อย่างมีนัยสำคัญ

2. ในกรณีของแบบสอบอิงกลุ่ม วิธีให้คะแนนโดยใช้ค่าคะแนนที่เหมาะสม (w) ทำให้แบบสอบมีค่าความเที่ยงสูงที่สุด และสูงกว่าค่าความเที่ยงที่ได้จากการให้คะแนนโดยวิธีอื่น ๆ อีก 5 วิธี อย่างมีนัยสำคัญ และการให้คะแนนโดยวิธีใช้ค่าเบต้าแปลง (B_t) ทำให้แบบสอบมีค่าต่ำที่สุด แต่วิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) วิธีใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) และวิธีใช้ค่าเดลต้า ($\triangle$) ต่างมีค่าความเที่ยงแตกต่างกันอย่างไม่มีนัยสำคัญ

3. ในกรณีของแบบสอบอิงเกณฑ์ การให้คะแนนโดยวิธีประเพณีนิยม ทำให้แบบสอบมีค่าความเที่ยง (ค่าดัชนีความเชื่อถือ) โดยเฉลี่ยสูงที่สุด และค่าดังกล่าวมีกระสวนการลดและเพิ่มคล้ายกับค่าความเที่ยงของการให้คะแนนโดยวิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) และวิธีใช้ค่าลำดับขั้นความเข้าใจในการอ่าน (h) มาก แต่การให้คะแนนโดยวิธีใช้ค่าคะแนนที่เหมาะสม (w) ทำให้แบบสอบมีค่าความเที่ยงต่ำที่สุด แต่ว่าโดยเฉลี่ยแล้วยังมีค่าในระดับสูงมาก (คือ 0.916)

4. อันดับที่ของคะแนนสอบของผู้สอบเมื่อแบบสอบให้คะแนนโดยวิธีต่าง ๆ จะแตกต่างไปจากอันดับที่เดิม เมื่อคิดคะแนนโดยวิธีประเพณีนิยมประมาณ 10 %–20 % การให้คะแนนโดยวิธีใช้ค่าผลบวกความแปรปรวนร่วม (h^2) ทำให้อันดับที่เปลี่ยนไปน้อยที่สุด คือประมาณ 10% แต่การให้คะแนนโดยวิธีใช้ค่าคะแนนที่เหมาะสม (w) ทำให้อันดับที่เปลี่ยนไปมากที่สุด คือประมาณ 20%

5. คะแนนสอบของแบบสอบเมื่อให้คะแนนโดยวิธีต่าง ๆ มีความสัมพันธ์ซึ่งกันและกันสูงมาก คือระหว่าง 0.923–0.997

ข้อเสนอแนะ

ผลจากการวิจัยครั้งนี้เป็นประโยชน์ในการนำไปใช้ดังนี้

1. ในกรณีของการสอบแบบอิงกลุ่ม ถ้าต้องการทำให้แบบสอบมีความเที่ยงสูงที่สุด ผู้สร้างแบบสอบควรให้คะแนนโดยวิธีใช้ค่าคะแนนที่เหมาะสม (w) แต่ถ้าต้องการให้แบบสอบมีค่าความตรงเชิงพยากรณ์สูงที่สุด ก็ควรให้คะแนนโดยวิธีประเพณีนิยม

2. ในกรณีของการสอบแบบอิงเกณฑ์ ถ้าต้องการให้แบบสอบมีความเที่ยง (ค่าดัชนีความเชื่อถือ) สูงที่สุด ก็ควรให้คะแนนโดยวิธีประเพณีนิยม หรือวิธีอื่นอีก 4 วิธี ที่คิดว่าสะดวก แต่ไม่ควรให้คะแนนโดยวิธีใช้ค่าน้ำหนักคะแนนที่เหมาะสม (w) เว้นแต่ว่าต้องการใช้แบบสอบนั้นเป็นแบบสอบอิงกลุ่ม และแบบสอบอิงเกณฑ์ด้วย ก็ควรใช้วิธีดังกล่าวนี้

บรรณานุกรม

สุพัฒน์ สุกมลสันต์ **ผลของการกำหนดน้ำหนักคะแนนต่อคุณภาพของแบบสอบ** สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย, 2529.

Brennan, R.L. "Estimating the Dependability of tee Scores." In *A Quide to Criterion-Referenced Test Construction,* pp. 292–333 : Edited by R.A. Berk. Baltimore : the John Hopkins University Press, 1984.

Brennan, R.L. and Kane, M.T. "An Index of Dependability for Mastery Tests." *Journal of Educational Measurement* 14 (1977) : 227-283.

Block, H.D. and Dockrell, W.B. *Criterion-Referenced Assessment in the Classroom.* Edinburgh : the Scottish Council for Research in Education, 1984.

Ebel, R.L. *Essentials of Educational Measurement..* New Jersey : Prentice–Hall, Inc., 1972.

Hambleton, R.K. and Others. "Criterion–Referenced Testing and Measurement : Review of Technical Issues and Developments." *Review of Educatianal Research* 45 (Winter 1978) : 1–47.

Hively, W. "Somt Comments." In *Domain-Referenced Testing*, p.p. 137 – 149. Edited by W. Hively. New Jercey : Educational Technology Pablication, 1974.

Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*. New Jersey : Lawrence Erlbaum, 1980.

"Optimal Number of Cnoices per Item : A Comparison of Four Approaches." *Journal of Educational Measurement* 14 (1977 b) : 33–38.

Magnusson, D. *Test Theory*. Stockholm : Stockholm Universitp Press, 1966.

Patnaik, D. and Traub, R.E. "Differential Weihting by Judged Degree of Correctness." *Journal of Educational Measurement* 10 (Winter 1973) : 281–286.

Stecklein, J.E. and Others. "Optimized Testing : An Empirical Study of Classroom Testing Based on Sepuential Analytic Theory." In *Selected Readings in Educational Measurement and Evaluation*, pp. 208– 240. Edited by S. Pitiyanuwat. Faculty of Education, Chulalongkorn University, 1981.

Urry, V.W. "Tailored Testing : A Successful Applicatioh of Latent Trait Theory." *Journal of Educational Measurement* 14 (1977) : 181–196.

Valette, R.M. and Disick, R.S. *Modern Language Performance Objectives and Individualization : A Handbook*, New York : Harcourt Brance, 1972.

Wang, M.W. and Stanley, J.C. "Differential Weighting : A Review of Methods and Empirical Studies." *Review of Educationat Research* 40 (1960) : 663–705.