



การวัดแบบอิงกลุ่มและการวัดแบบอิงเกณฑ์ : ทฤษฎีและแนวปฏิบัติ

Norm-Referenced and Criterion-Referenced Tests : Theory and Practice

ระพินทร์ โพธิ์ศรี¹, สุกัญญา รุจิเมธามาส² และสุรเชษฐ์ บุญยรักษ์³

Rapin Posrie¹, Sukunya Rujimethabhas² and Surachet Boonyarug³

คณะครุศาสตร์ มหาวิทยาลัยราชภัฏอุตรดิตถ์^{1,2,3}

Faculty of Education Uttaradit Rajabhat University^{1,2,3}

Corresponding author, E-mail: r.sukunya@uru.ac.th²

สาระสังเขป

แบบทดสอบอิงกลุ่มและแบบทดสอบอิงเกณฑ์เป็นแบบทดสอบวัดผลสัมฤทธิ์หรือสมรรถนะทางการเรียนที่ใช้อยู่ทั่วไปขณะนี้ แบบทดสอบแต่ละประเภทมีพื้นฐานทางทฤษฎีที่แตกต่างกัน การสร้างโดยประยุกต์ใช้ทฤษฎีไม่ถูกต้องจะทำให้เกิดการแปลความหมายของคะแนนที่ผิดพลาด นำไปสู่การตัดสินใจดำเนินการต่อเรื่องที่ผิด เกิดผลเสียต่อผู้เรียนและผู้มีส่วนได้เสียเป็นอย่างมาก การวัดแบบอิงกลุ่มและอิงเกณฑ์เป็นแนวคิดที่พัฒนาต่อเนื่องมาเป็นเวลาไม่ต่ำกว่าครึ่งศตวรรษ ในบทความนี้จะเรียกแนวคิดทั้งสองว่า “ทฤษฎี”

ทฤษฎีการวัดแบบอิงกลุ่มอธิบายว่า ในการวัดผลทั่วไปนั้นกรอบสาระหลักมีขนาดใหญ่มาก ไม่เป็นเนื้อเดียวกัน การวัดกรอบเนื้อหาข้างต้นใช้วิธีสุ่มตัวอย่างเนื้อหาบางส่วนมาสร้างเป็นแบบทดสอบ ไม่ได้นำเนื้อหาทั้งหมดมาสร้างเป็นแบบทดสอบ คะแนนที่ได้จากแบบทดสอบจึงไม่ใช่คะแนนจริง และไม่เสมือนจริง ไม่สามารถแปลความหมายของคะแนนโดยคิดเป็นร้อยละได้ ต้องแปลความหมายอิงกลุ่มผู้สอบที่เข้าสอบ ในขณะที่ทฤษฎีการวัดแบบอิงเกณฑ์อธิบายว่าในบางกรณีการวัดผลนั้นกลุ่มสาระหลักมีขนาดเล็กและเป็นเนื้อเดียวกันสุ่มเนื้อหาสอบเพียงนิดเดียวก็สามารถแทนเนื้อหาทั้งหมดได้ เหมือนสุ่มแกงเพียงหนึ่งช้อนจากแกงหนึ่งหม้อ สามารถใช้แกงหนึ่งช้อนประเมินแกงหนึ่งหม้อได้ ในทำนองเดียวกันในการวัดแบบอิงเกณฑ์จะถือว่าคะแนนจากแบบทดสอบเป็นคะแนนเสมือนจริงเป็นตัวประมาณค่าคะแนนจริงที่เชื่อถือได้ ดังนั้นจึงสามารถแปลความหมายคะแนนโดยคิดเป็นร้อยละ แล้วเทียบหรืออิงกับเกณฑ์ ซึ่งเป็นกลุ่มสาระหลักว่าผู้สอบมีความรู้ความสามารถคิดเป็นร้อยละของความรู้ทั้งหมดเท่าไรได้

คำสำคัญ: การวัดแบบอิงกลุ่ม, การวัดแบบอิงเกณฑ์

SUMMARY

Norm-referenced and criterion-referenced tests are both achievement or proficiency tests, each of which has a different theoretical background. Construction of these tests with incorrect applications of theoretical backgrounds will lead to incorrect interpretation of raw score, resulting in huge negative effects on students and all other stakeholders. Both approaches have long been developing for more than half a century, therefore, in this article they will be named “theories”.

Norm-referenced test theory explains that in general, the test domain is large and heterogeneous. A test measuring such a domain is only a sample of tasks. A raw score is not a true score and not dependable as a true score. It cannot be interpreted as a percent correct as a true score. It is, then, interpreted through norm referencing. While the criterion-referenced test theory explains that in some cases, the test domain is small and homogeneous. A small sample of tasks can represent the whole domain, just like a spoon of soup can represent the whole soup in a pot. Therefore, a raw score is a virtual true score and dependable as a true score. It, then, can be interpreted in terms of a percent correct referencing directly to the test domain.

Keywords: Norm-Referenced, Criterion-Referenced



บทนำ

แบบทดสอบวัดผลสัมฤทธิ์หรือสมรรถนะทางการเรียนที่ใช้อยู่ทั่วไปขณะนี้แบ่งออกเป็น 2 ประเภทใหญ่ ๆ คือ แบบทดสอบอิงกลุ่ม (norm-referenced test) และแบบทดสอบอิงเกณฑ์ (criterion-referenced test) แบบทดสอบแต่ละประเภทมีพื้นฐานทางทฤษฎีและการพัฒนาให้เป็นแบบทดสอบมาตรฐานที่แตกต่างกัน การนำทฤษฎีไปประยุกต์ใช้ไม่ถูกต้องจะทำให้เกิดการแปลความหมายของคะแนนที่ผิดพลาด นำไปสู่การตัดสินใจดำเนินการต่อเนื่องที่ผิด เกิดผลเสียต่อผู้เรียนและผู้มีส่วนได้เสียอื่น ๆ ด้วย เช่น การนำผลการวัดผลการเรียนรู้ในชั้นเรียนมาให้ระดับคะแนนหรือตัดเกรด โดยกำหนดล่วงหน้าว่า ควรได้เกรด A B C D F ร้อยละเท่าไร โดยอิงกับการแจกแจงปกติ หรือ นำคะแนนดิบมาแปลงเป็นคะแนนมาตรฐานเช่นคะแนนที่ก่อนแล้วจึงนำมาตัดเกรดแบบอิงกลุ่ม ซึ่งอาจทำให้มีนักเรียนบางคนได้เกรด D หรือ F ทั้ง ๆ ที่เขาอาจจะได้เกรด C ถ้าดูจากร้อยละของคะแนนดิบที่เขาทำได้ หรือในทางตรงกันข้ามกันกรณีกลุ่มนักเรียนเรียนอ่อนมากทำคะแนนได้สูงสุดเพียง 50 ต่ำสุด 20 คะแนนจากคะแนนเต็มร้อยละคะแนนถ้านำคะแนนชุดนี้มาตัดเกรดแบบอิงกลุ่มจะกลายเป็นว่านักเรียนที่ได้คะแนนเพียง 50% อาจได้เกรด A หรือ B ก็ได้ ที่เกิดปัญหาดังกล่าวขึ้นก็เพราะการวัดผลในชั้นเรียนเป็นการวัดตามทฤษฎีการวัดแบบอิงเกณฑ์ คะแนนที่ได้เป็นคะแนนเสมือนจริงสามารถประเมินผลและตัดเกรดจากร้อยละของคะแนนดิบได้เลย ไม่ต้องกังวลว่าคะแนนจะแจกแจงแบบปกติหรือไม่

ความผิดพลาดอีกกรณีหนึ่งคือ การวัดผลเป็นการวัดแบบอิงกลุ่มแต่นำคะแนนจากการวัดผลมาแปลความหมายแบบอิงเกณฑ์ ทำให้แปลความหมายผิดพลาด นักเรียนระดับปานกลางอาจกลายเป็นนักเรียนระดับอ่อน เช่น กรณีการสอบโอเน็ตในระยะแรก ๆ ที่นำคะแนนดิบมาคิดเป็นร้อยละแล้วสังคมทั่วไปแปลความหมายว่าวิชาใดที่นักเรียนได้คะแนนต่ำกว่า 50% ถือว่าเรียนอ่อน ซึ่งอาจเป็นการแปลความหมายที่ผิด เพราะการสอบโอเน็ตแต่ละวิชาที่ สทศ. (สถาบันทดสอบทางการศึกษาแห่งชาติ) ดำเนินการมานั้นมีลักษณะเป็นการวัดแบบอิงกลุ่ม คือ กรอบเนื้อหาสาระสำหรับการสอบมีขนาดใหญ่และไม่ชัดเจน เช่น ข้อสอบวิชาภาษาอังกฤษชั้นมัธยมศึกษาปีที่ 6 จะรวมเนื้อหาสาระชั้น ม.4 ม. 5 และ ม.6

ซึ่งถือว่ามีความจำนวนมาก ในเอกสารหลักสูตรแกนกลางระบุนสาระการเรียนรู้ไว้ในรูปมาตรฐานการเรียนรู้และตัวชี้วัดแบบกว้าง ๆ โรงเรียนแต่ละโรงเรียนก็เลือกใช้แบบเรียนกันเอง ตีความกันเอง เวลา สทศ. จะออกข้อสอบก็ตีความกันเองอีกชั้นหนึ่ง คุณครูผู้สอนไม่สามารถทราบได้ว่ากรอบเนื้อหาสาระที่จะสอบมีอะไรบ้าง จะต้องสอนอะไรเน้นอะไร ข้อสอบโอเน็ตแต่ละวิชาให้เวลานักเรียนแค่ 2 ชั่วโมง เหมือนกับเอาสิ่งที่เรียนมา 3 ปีมาออกข้อสอบเพียง 2 ชั่วโมง คะแนนดิบที่ได้ไม่สามารถใช้ประมาณค่าคะแนนจริงของนักเรียนได้อย่างถูกต้อง ดังนั้นการนำคะแนนดิบมาคิดเป็นร้อยละและแปลความหมายว่านักเรียนมีผลการเรียนเป็นอย่างไร จึงเป็นการแปลความหมายที่ยังไม่ถูกต้อง เพราะคะแนนมีลักษณะเป็นคะแนนตามทฤษฎีการวัดแบบอิงกลุ่ม ต้องแปลความหมายแบบอิงกลุ่ม คือนำคะแนนมาเทียบเกณฑ์ปกติ (norm) หรือเทียบกันเองภายในกลุ่ม เช่นทำเป็นเปอร์เซ็นต์ไทล์ (PR) แล้วแปลงเป็นคะแนนมาตรฐานระบบใดระบบหนึ่งแล้วจึงแปลความหมายออกมาว่าผู้สอบมีความสามารถระดับใด เช่น ถ้าใช้ระบบคะแนนมาตรฐานเก้า (stanine) จะสามารถแปลความหมายได้ดังภาพที่ 1 ดังนี้

PR	Stanine Range	Stanine	ความหมาย	เกรด
96 – 99	Stanine > 8.55	9	เก่งอย่างยิ่ง	A ⁺
89 – 95	7.55 – 8.54	8	เก่งมาก	A ⁻
77 – 88	6.55 – 7.54	7	เก่ง	B ⁺
60 – 76	5.55 – 6.54	6	ค่อนข้างเก่ง	B ⁻
40 – 59	4.55 – 5.54	5	ปานกลาง	C ⁺
23 – 39	3.55 – 4.54	4	ค่อนข้างอ่อน	C ⁻
11 – 22	2.55 – 3.54	3	อ่อน	D ⁺
4 – 10	1.55 – 2.54	2	อ่อนมาก	D ⁻
1 – 3	Stanine < 1.54	1	อ่อนอย่างยิ่ง	F

ภาพที่ 1 การแปลความหมายคะแนนด้วยระบบคะแนนมาตรฐานเก้า (ที่มา: ปรับมาจาก Popham, 1981)

การแปลความหมายคะแนนโอเน็ตระยะต่อมาได้ปรับมาใช้วิธีแบบอิงกลุ่มด้วยคะแนนมาตรฐานที่ (standard T-score) ดังตัวอย่างใบรายงานผลคะแนนในภาพที่ 2 ซึ่งเป็นตัวอย่าง



ใบรายงานผลสอบโอเน็ตของนักเรียนคนหนึ่ง จากโรงเรียนแห่งหนึ่ง จะเห็นได้ว่า มีคะแนนมาตรฐานที่เพิ่มเข้ามาอีกหนึ่งสมมติ ดังภาพที่ 2

ผลการทดสอบทางการศึกษาระดับชาตินิยมพื้นฐาน (O-NET)									
ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558									
1. ข้อมูลผู้เข้าสอบ O-NET ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558									
ชื่อ-สกุล :		เลขที่ประจำตัวสอบ :			เลขประจำตัวประชาชน :				
โรงเรียน :		ขนาดโรงเรียน :			สังกัดโรงเรียน :				
จังหวัด :		ภาค :		สังกัด :					
2. คะแนน O-NET									
วิชา	กลุ่มสาระการเรียนรู้	คะแนนเต็ม	คะแนน O-NET	มาตรฐาน (T-Score)	ร้อยละ (Per Cent)	อันดับ (Ranking)	ค่าเฉลี่ย	ค่าเบี่ยงเบน	ค่ามาตรฐาน
01	ภาษาไทย	100.00	89.00	62.81	55.76	54.22	50.53	49.96	50.10
04	คณิตศาสตร์	100.00	87.50	77.66	28.01	30.40	27.74	26.86	27.04
05	วิทยาศาสตร์	100.00	54.00	74.43	36.40	35.52	33.80	33.56	33.40
02	สังคมศึกษา ศาสนาและวัฒนธรรม	100.00	53.00	66.08	41.44	41.76	40.32	40.00	39.83
03	ภาษาอังกฤษ	100.00	31.00	54.83	26.00	28.76	26.58	24.68	25.15

ภาพที่ 2 ผลคะแนนโอเน็ตของนักเรียนเป็นรายบุคคล
ที่มา Net News. (เมษายน, 2559) <https://www.niets.or.th>

แนวคิดทฤษฎี

ทฤษฎีที่อธิบายเกี่ยวกับคะแนนจากการวัดผลประกอบด้วย ทฤษฎีการวัดผลแบบดั้งเดิม (classical test theory) ทฤษฎีหลักนัยทั่วไป (generalizability theory) ทฤษฎีการวัดแบบอิงกลุ่ม (norm-referenced test theory) และทฤษฎีการวัดแบบอิงเกณฑ์ (criterion-referenced test theory) แต่ละทฤษฎีมีสาระสำคัญดังนี้

ทฤษฎีวัดผลแบบดั้งเดิม ผู้ที่ริเริ่มสร้างทฤษฎีนี้ คือ Spearman (1904) ซึ่งเป็นนักจิตวิทยาชาวอังกฤษ ทฤษฎีนี้อธิบายว่า คะแนนดิบจากการวัดผล (X) ประกอบด้วย คะแนนจริง (True Score : T) และคะแนนความคลาดเคลื่อน (Error Score : E) เขียนเป็นสมการได้ดังนี้

$$X = T + E$$

X คือ คะแนนดิบ

T คือ คะแนนจริง แทนความสามารถจริงของผู้สอบ

E คือ คะแนนความคลาดเคลื่อน อาจเป็นได้ทั้งบวกและลบ

ข้อบกพร่องที่สำคัญของทฤษฎีนี้คือ ทฤษฎีนี้ไม่ได้อธิบายไว้ชัดเจนว่า คะแนนจริง ซึ่งเป็นความสามารถจริงของผู้สอบนั้นจะวัดได้อย่างไร

ทฤษฎีหลักนัยทั่วไป

ทฤษฎีนี้พัฒนาขึ้นโดยนักวิชาการด้านการวัดผล 4 ท่าน คือ Cronbach, Glaser, Nanda and Rajaratnam (1963) ได้เสนอวิธีประมาณค่าความเชื่อมั่นของแบบทดสอบโดยใช้การวิเคราะห์ความแปรปรวน ซึ่งสามารถประมาณค่าได้หลายรูปแบบตามแบบการทดลองใช้แบบทดสอบและการกำหนดให้เกิดความคลาดเคลื่อนในการวัด ที่สำคัญอย่างยิ่งคือ ทฤษฎีนี้อธิบายเพิ่มเติมว่าจะวัดคะแนนจริงได้อย่างไร โดยอธิบายว่า คะแนนจริง (Brennan, 1983, p. 4) คือ ค่าเฉลี่ยของคะแนนจากการวัดผลที่ได้จากการนำแบบทดสอบที่สร้างจากการสุ่มตัวอย่างเนื้อหาในรอบสาระหลัก ที่สามารถสุ่มได้ทั้งหมด (all conditions) ของคุณลักษณะที่ต้องการวัดมาทำการสอบวัด

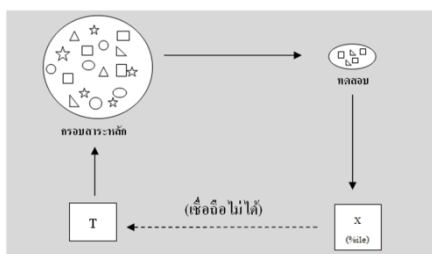
จากคำอธิบายของทฤษฎีหลักนัยทั่วไปทำให้เราทราบว่าการสอบวัดเพื่อให้ได้คะแนนจริงนั้นจะต้องสอบวัดจากกรอบเนื้อหาทั้งหมดของคุณลักษณะที่ต้องการวัด คะแนนจริงคือพารามิเตอร์ที่แสดงถึงคุณลักษณะของประชากร ส่วนคะแนนจากการวัดผลที่ได้จากแบบทดสอบแต่ละฉบับนั้นเป็นเพียงคะแนนดิบหรือค่าสถิติจากกลุ่มตัวอย่างไม่ใช่คะแนนจริง แบบทดสอบต่าง ๆ ไม่ว่าจะเป็นแบบทดสอบที่ใช้ในชั้นเรียนหรือแบบทดสอบมาตรฐานต่าง ๆ เช่น แบบทดสอบ TOEFL (Test of English as a Foreign Language) แบบทดสอบ IELTS (International English Language Testing System) หรือ แบบทดสอบ PISA (Programme for International Student Assessment) ต่างก็เป็นแบบทดสอบที่สุ่มเนื้อหาเพียงบางส่วนจากกรอบเนื้อหาทั้งหมดที่ต้องการวัด (domain) มาทำการสอบวัด

การวัดแบบอิงกลุ่มและอิงเกณฑ์

แนวคิดการวัดแบบอิงกลุ่มและการวัดแบบอิงเกณฑ์เป็นแนวคิดที่พัฒนาต่อเนื่องมาจากทฤษฎีวัดผลแบบดั้งเดิม แนวคิดการวัดแบบอิงกลุ่มถือว่าเป็นทฤษฎีย่อยของทฤษฎีวัดผลแบบดั้งเดิม ส่วนแนวคิดการวัดแบบอิงเกณฑ์เป็นแนวคิดใหม่ที่อธิบายเกี่ยวกับการแปลความหมายคะแนนที่แตกต่างไปจากแนวคิดการวัดแบบอิงกลุ่ม แนวคิดทั้งสองนี้ได้พัฒนามาเป็นเวลานานไม่ต่ำกว่าครึ่งศตวรรษ เป็นหลักการที่สำคัญในการแปลความหมายคะแนนจากแบบทดสอบ ดังนั้นในบทความนี้จะเรียกว่า

ทฤษฎีการวัดแบบอิงกลุ่ม (norm-referenced test theory) และ ทฤษฎีการวัดแบบอิงเกณฑ์ (criterion-referenced test theory)

ทฤษฎีการวัดแบบอิงกลุ่ม ทฤษฎีการวัดแบบอิงกลุ่ม อธิบายเกี่ยวกับการแปลความหมายคะแนนมีสาระสำคัญโดยสรุปได้ว่า (Ebel, 1970, Lyman, 1978, Popham, 1981) ในการวัดผลทั่วไปนั้นกรอบสาระหลักมีขนาดใหญ่มากมีลักษณะเป็นอเนกนัย (heterogeneous) ไม่เป็นเนื้อเดียวกัน เช่น การวัดเชาวน์ปัญญา การวัดสมรรถนะทางภาษา การวัดความถนัดทางการเรียนการวัด กรอบเนื้อหาข้างต้นใช้วิธีสุ่มตัวอย่างเนื้อหาบางส่วนมาสร้างเป็นแบบทดสอบ ไม่ได้นำเนื้อหาทั้งหมดมาสร้างเป็นแบบทดสอบ แล้วจึงนำไปสอบวัดกลุ่มเป้าหมาย ดังนั้นคะแนนที่ได้จากแบบทดสอบจึงไม่ใช่คะแนนจริงเพราะไม่ได้นำเนื้อหาทั้งหมดมาทำการทดสอบ และไม่เสมือนจริง เพราะตัวอย่างเนื้อหาที่สุ่มมาสร้างแบบทดสอบนั้นไม่เหมือนเนื้อหาในกรอบสาระหลักดังภาพที่ 3



ภาพที่ 3 ทฤษฎีการวัดแบบอิงกลุ่ม

จากภาพที่ 3 จะเห็นได้ว่า กรอบสาระหลักประกอบไปด้วย เนื้อหาหลายลักษณะ เช่น การวัดสมรรถนะภาษาอังกฤษ กรอบสาระหลักจะประกอบไปด้วยทักษะการฟัง พูด อ่าน และเขียน ซึ่งแต่ละทักษะประกอบด้วยเนื้อหาเป็นจำนวนมาก ในภาพแสดงไว้ เป็นสัญลักษณ์ลักษณะต่าง ๆ เมื่อทำการสุ่มเนื้อหาสร้างเป็นแบบทดสอบ เนื้อหาที่สุ่มมาจึงไม่เหมือน กับเนื้อหาทั้งหมดในกรอบ สาระหลัก ด้วยเหตุนี้เอง คะแนนดิบ X ที่ได้จากการวัดด้วยแบบทดสอบประเภทนี้จึงเป็นคะแนนที่เชื่อถือไม่ได้ว่าเป็นคะแนนจริง (T) ไม่สามารถนำมาคิดเป็นร้อยละแล้วแปลความหมายว่าผู้สอบ มีความรู้ความสามารถคิดเป็นร้อยละเท่าไรของความรู้ทั้งหมดได้ เพราะถ้านำมาคิดเป็นร้อยละคนทั่วไปจะเข้าใจว่าเป็นคะแนนจริง

แล้วนำไปเทียบกับ 50% ถ้าได้คะแนนต่ำกว่า 50% ถือว่าเรียนอ่อน เหมือนกรณีการสอบโอเน็ตในระยะแรก ๆ ที่กระบวนการวัด มีลักษณะเป็นการวัดแบบ อิงกลุ่มแต่รายงานผลคะแนนสอบ เป็นร้อยละทำให้คนทั่วไปเข้าใจว่าถ้านักเรียนได้คะแนนต่ำกว่า 50% ถือว่าเรียนอ่อนซึ่งความจริงอาจไม่ได้เรียนอ่อนก็ได้ถ้าคะแนนที่ได้ สูงกว่าค่าเฉลี่ย หรือต่ำกว่าค่าเฉลี่ยแต่ต้องไม่เกินผลคูณของ 0.5 กับส่วนเบี่ยงเบนมาตรฐาน ($0.5 \times SD$) จะถือว่าอยู่ระดับปานกลาง สำหรับกรณีที่เป็นกรวัดแบบอิงกลุ่มต้องนำคะแนนดิบมาแปล ความหมายโดยเปรียบเทียบกันเองภายในกลุ่มโดยคิดเป็น เปอร์เซ็นต์ไทล์ (Percentile) แล้วแปลงเป็นคะแนนมาตรฐานระบบใด ระบบหนึ่ง เช่น คะแนนมาตรฐานเก้า (stanine) ดังกรณีการสอบ IELTS ดังแสดงในภาพที่ 4

IELTS Listening Scores

Correct Answers	Band Scores
39 - 40	9
37 - 38	8.5
35 - 36	8
32 - 34	7.5
30 - 31	7
26 - 29	6.5
23 - 25	6
18 - 22	5.5
16 - 17	5
13 - 15	4.5
11 - 12	4

ภาพที่ 4 ระบบคะแนนการสอบ IELTS

ที่มา: IELTS Liz. (2024.). <https://ieltsliz-band-scores/>

จากภาพที่ 4 จะเห็นได้ว่าระบบการสอบ IELTS รายงาน คะแนนผลการสอบด้วยคะแนนมาตรฐานเก้า (band score) ไม่รายงานเป็นร้อยละของคะแนนดิบ เพราะแบบทดสอบ IELTS เป็นแบบทดสอบอิงกลุ่มเช่นเดียวกับคะแนนจากการสอบ PISA ซึ่งรายงานผลด้วยคะแนนแปลงรูป (transformed score) ในลักษณะมาตรวัดคะแนน PISA (PISA score scale) โดยให้มีค่าเฉลี่ยเป็น 500 ส่วนเบี่ยงเบนมาตรฐานเป็น 100 แล้วแปล ความหมายโดยวิธีเรียงลำดับที่เป็นระดับความสามารถ 7 ระดับ ดังกรณีการแปลความหมายคะแนนความรู้ทางคณิตศาสตร์ (mathematical literacy) กำหนดไว้ 7 ระดับ แสดงดังภาพที่ 5 ดังนี้



Level	Score points on the PISA Scale
6	At or above 669.30
5	At or above 606.99 but less than 669.30
4	At or above 544.68 but less than 606.99
3	At or above 482.38 but less than 544.68
2	At or above 420.07 but less than 482.38
1a	At or above 357.77 but less than 420.07
1b	At or above 295.47 but less than 357.77

ภาพที่ 5 ระดับคะแนนความรอบรู้ทางคณิตศาสตร์

ของการสอบ PISA ปี 2022

คะแนนแต่ละระดับจะมีความหมายตามคำนิยามที่อธิบายไว้ว่านักเรียนสามารถทำอะไรได้ทำอะไรเป็นบ้าง เช่น ระดับ 1b หมายความว่า “นักเรียนสามารถแก้ปัญหาโจทย์ปัญหาคณิตศาสตร์พื้น ๆ ที่ง่ายต่อการเข้าใจ โดยโจทย์ปัญหามีสถานการณ์ที่ประกอบกับการแก้ปัญหาได้อย่างชัดเจน (ตัวอย่างเช่น ระบุไว้ในรูปตาราง หรือ ภาพ) แบบกระชับ และใช้ภาษาง่าย ๆ นักเรียนสามารถที่จะแก้ปัญหาได้ถูกต้องโดยดำเนินการตามขั้นตอนที่กำหนดไว้ในโจทย์ปัญหา” ระดับ 6 หมายความว่า “นักเรียนสามารถแก้โจทย์ปัญหาคณิตศาสตร์ที่เป็นนามธรรม สามารถคิดแบบสร้างสรรค์และคิดแบบยืดหยุ่นได้ ตัวอย่างเช่น นักเรียนสามารถ ตีโจทย์ปัญหาที่ซับซ้อนได้ สามารถเข้าใจสิ่งที่ทางคณิตศาสตร์ที่ลุ่มลึก สามารถเชื่อมโยงสารสนเทศเข้าด้วยกัน รวมทั้งการใช้ การจำลองสถานการณ์หรือ ผลการวิเคราะห์ข้อมูลด้วยโปรแกรมคอมพิวเตอร์ ประกอบการแก้โจทย์ปัญหาคณิตศาสตร์ได้ นักเรียนที่มีความสามารถระดับนี้มีความสามารถในการคิดเชิงวิพากษ์ และมีความรอบรู้ในการใช้สัญลักษณ์และกระบวนการดำเนินงานและความสัมพันธ์ทางคณิตศาสตร์ในการสื่อสารแสดงการแก้ปัญหาได้อย่างชัดเจน ตลอดจนสามารถสะท้อนคิดถึงความถูกต้องในการแก้ปัญหาทางคณิตศาสตร์ได้”

ทฤษฎีการวัดแบบอิงกลุ่มมีอิทธิพลอย่างมากต่อการวัดและประเมินผลในชั้นเรียนในทศวรรษที่ 50 และ 60 หรือประมาณปี ค.ศ. 1950-1969 (พ.ศ. 2493-2512) ในยุคนั้นการจัดการเรียนการสอนในวิทยาลัยหรือมหาวิทยาลัยทั่ว ๆ ไปต้องตัดเกรดโดยนำคะแนนดิบมาแปลงเป็นคะแนนมาตรฐานปกติที่เสียก่อนแล้วจึงตัดเกรดแบบอิงกลุ่ม ซึ่งอาจมีปัญหา นักเรียนตกได้เกรด F ทั้ง ๆ ที่ถ้าดูจากคะแนนดิบแล้วเขาไม่ควรสอบตกก็ได้

ทฤษฎีการวัดแบบอิงเกณฑ์ Glaser (1963) ได้เสนอทฤษฎีการวัดผลใหม่ เรียกว่า ทฤษฎีการวัดแบบอิงเกณฑ์ ซึ่งเป็นทฤษฎีการวัดผลกรณีที่กลุ่มสารหลักมีขนาดเล็ก แตกต่างไปจากทฤษฎีการวัดแบบอิงกลุ่มอย่างมาก Glaser ได้แนวคิดการวัดแบบอิงเกณฑ์ในขณะที่เรียนอยู่ที่มหาวิทยาลัยอินเดียนา (Popham, 2514 <https://www.ascd.org/el/articles/criterion-referenced-measurement-half-a-century-wasted>) ในขณะที่ทดลองใช้ชุดกิจกรรมจัดการเรียนการสอน พบว่านักเรียนส่วนใหญ่ทำคะแนนได้สูงและใกล้เคียงกันมาก การแจกแจงของคะแนนเป็นรูปเบ้ซ้าย ไม่ใช้การแจกแจงแบบปกติ การแปลความหมายของคะแนนแบบเดิมที่เคยทำกันมาคือ นำคะแนนมาแปลงเป็นเปอร์เซ็นต์ไทล์และแปลงเป็นคะแนนมาตรฐานนั้นไม่เกิดประโยชน์แต่อย่างใด เพราะกระตือรือร้นนักเรียนได้คะแนนสูงคะแนนติด ๆ กันนั้นการแปลความหมายคะแนนโดยใช้เปอร์เซ็นต์ไทล์นั้นไม่แสดงถึงความสามารถที่แท้จริงของนักเรียน นักเรียนได้คะแนนห่างกันเพียง 1 หรือ 2 คะแนน แต่ตำแหน่งเปอร์เซ็นต์ไทล์จะดูห่างกันมาก ดังนั้นการตีข้อมูลแจกแจงแบบเบ้ซ้ายนักเรียนส่วนใหญ่จะได้คะแนนเต็มหรือเกือบเต็ม การนำเปอร์เซ็นต์ไทล์หรือลำดับที่มาเปรียบเทียบกับจึงไม่มีความหมาย เช่น มีนักเรียนสอบ 10 คน มีได้คะแนน เต็ม 9 คน มีเพียงคนเดียวที่ได้คะแนนเกือบเต็มทำผิดไปข้อเดียว ถ้าแปลความหมายแบบอิงกลุ่มโดยใช้ลำดับที่จะกลายเป็นว่านักเรียนที่ได้คะแนนเกือบเต็มได้ข้อ 10 เหมือนกับว่าเป็นนักเรียนเรียนอ่อน ซึ่งไม่จริง จากแนวคิดนี้ต่อมา Glaser (1963, p. 519) ได้เขียนบทความทางวิชาการเสนอแนวคิดการวัดแบบอิงเกณฑ์ดังกล่าว อธิบายตอนหนึ่งที่มีความดังนี้

“Under the concept of achievement measurement is the notion of a continuum of knowledge acquisition regarding from no proficiency at all to perfect performance. An individual's achievement level falls at some point on the continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion referenced-measures of achievement or proficiency. The standard against which a student's performance is compared when

measured in this manner is the behavior which defined each point along the achievement continuum.

The term “criterion” when used in this way, does not necessarily refer to-final end-of-course behavior. Criterion levels can be established at any-point in instruction as to the adequacy of an individual’s performance. In other words, the specified behaviors implied at each level of proficiency. can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. It is this sense that measures of proficiency-can be criterion-referenced.

Along such a continuum of attainment, a student’s score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement.

Measures which assess student achievement in terms of criterion standard thus provide information as to the degree of competence attained by a particular student is independent of reference to the performance of others.”

ในยุคแรก ๆ ที่แนวคิดนี้เกิดขึ้นใหม่ ๆ นั้นคนทั่วไปเข้าใจว่าการวัดแบบอิงเกณฑ์หรือแบบทดสอบอิงเกณฑ์คือแบบทดสอบที่กำหนดเกณฑ์จุดตัด (cut-score) สำหรับประเมินผลการสอบผ่าน/ตกไว้ ซึ่งเป็นความเข้าใจที่ผิด Popham (1981, pp. 17-18) อธิบายไว้ว่า แบบทดสอบอิงเกณฑ์คือแบบทดสอบที่ใช้ประเมินระดับความสามารถของบุคคลโดยเปรียบเทียบกับเกณฑ์ที่เป็นกรอบพฤติกรรมที่กำหนดไว้อย่างชัดเจน (a defined behavioral domain) ในขณะที่แบบทดสอบอิงกลุ่มคือแบบทดสอบที่มุ่งเปรียบเทียบความสามารถของผู้สอบภายในกลุ่มที่สอบ หากไปตีความ หมายถึงว่าแบบทดสอบอิงเกณฑ์คือแบบทดสอบที่มีการกำหนดคะแนนเกณฑ์จุดตัดไว้เพื่อประเมินการผ่าน/ตก

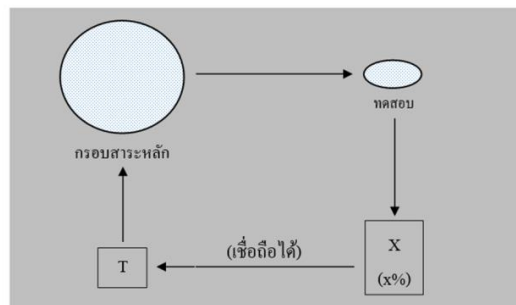
ของผู้สอบ การสร้างแบบทดสอบอิงเกณฑ์ก็ง่ายมาก เพียงแค่เอาแบบทดสอบอะไรก็ได้มากำหนดคะแนนจุดตัด ผ่าน/ตก เท่านั้น ก็จะกลายเป็นแบบทดสอบอิงเกณฑ์ไปหมดซึ่งไม่ถูกต้อง เหตุที่คนทั่วไปเข้าใจผิดก็เพราะคำว่า “เกณฑ์” หรือ criterion ที่เคยใช้กันมาแต่ก่อนนั้นมีความหมายสองนัย คือ หมายถึง ระดับสมรรถนะ (a desired level of proficiency) ซึ่งเป็นคะแนนจุดตัด หรือ หมายถึง กลุ่มพฤติกรรมเป้าหมาย (the target behaviors) ก็ได้ แต่เนื่องจากในบทความของ Glaser ใช้คำว่า “criterion” ทั้งสองความหมาย จึงทำให้เกิดความสับสน บางครั้งก็ใช้ในความหมายว่า หมายถึง “กลุ่มพฤติกรรมเฉพาะเจาะจง” (specific behaviors) ซึ่งหมายถึง การที่บุคคลสามารถทำ หรือ ไม่สามารถทำอะไรได้อะไร เป็นที่สื่อถึงการวัดแบบอิงเกณฑ์อันแสดงถึงการนำคะแนนความสามารถของผู้สอบไปเปรียบเทียบกับเกณฑ์ซึ่งเป็นกรอบพฤติกรรมหลัก (defined behavioral domain) ที่กำหนดไว้สำหรับการสอบวัด ในขณะที่บางตอนของบทความ Glaser อ้างถึง “ความต่อเนื่องของความรู้” (continuum of knowledge) และ “ความรู้ความสามารถที่คาดหวัง ณ.ระดับสมรรถนะ ที่กำหนด” (desired performance at any specified level) กรณีนี้ Glaser ต้องการจะสื่อว่าการวัดแบบอิงเกณฑ์ก็เปรียบเทียบความสามารถของผู้สอบกับเกณฑ์ระดับสมรรถนะด้วยเหมือนกัน (level of proficiency) เพื่อให้เข้าใจแนวคิดของ Glaser จะแปลถอดความข้อเขียนของ Glaser ข้างต้นดังนี้ “หลักคิดสำคัญของการวัดผลการเรียนรู้ คือ ความต่อเนื่องของการเรียนรู้ที่เริ่มจากการไม่มีความรู้เลยไปจนถึงการมีความรู้อย่างสมบูรณ์ ระดับผลสัมฤทธิ์การเรียนรู้ของแต่ละบุคคลจะตกอยู่ ณ.จุดใดจุดหนึ่งบนมาตรวัดนี้ ตามความสามารถของบุคคลที่วัดได้จากแบบทดสอบ คะแนนความสามารถในการเรียนรู้ ณ ระดับใดระดับหนึ่ง ในกรณีนี้สามารถวัดได้จากวิธีการวัดผลการเรียนรู้แบบอิงเกณฑ์ มาตรฐานที่นำผลการเรียนรู้มาเปรียบเทียบกับกรณีการวัดในลักษณะนี้คือ พฤติกรรมการเรียนรู้ซึ่งแจกแจงรายละเอียดการเรียนรู้แต่ละระดับบนมาตรวัดผลการเรียนรู้ คำว่า “เกณฑ์” ในกรณีนี้ไม่จำเป็นต้องหมายถึงพฤติกรรมสุดท้ายของการเรียนรู้เมื่อเรียนจบรายวิชาใดวิชาหนึ่ง (end-of-course behavior) การกำหนดเกณฑ์สามารถที่จะกำหนด ณ ระดับใดระดับหนึ่งตามความเหมาะสม หรือกล่าว



อีกนัยหนึ่ง เกณฑ์ในการนี้ก็คือรายละเอียดของพฤติกรรม การเรียนรู้เฉพาะเรื่อง (specified behaviors) ณ ระดับสมรรถนะ (level of proficiency) แต่ละระดับที่บุคคลสามารถกระทำได้ ก่อนที่จะประเมินได้ว่าเขามีความสามารถบรรลุตามเกณฑ์ที่กำหนด การวัดในลักษณะนี้จึงจะถือว่าการวัดแบบอิงเกณฑ์คะแนนของ นักเรียนที่วัดตามระบบมาตรฐานวัดความรู้ ที่เป็นการวัดแบบอิงเกณฑ์ จะเป็นคะแนนที่แสดงไว้อย่างชัดเจนว่านักเรียนสามารถทำอะไรได้ หรือทำอะไรไม่ได้บ้าง หน่วยวัดจากการวัดแบบอิงเกณฑ์คือ คลังความรู้ที่ประกอบไปด้วยพฤติกรรมการเรียนรู้ต่าง ๆ ซึ่งสอดคล้องกันระหว่าง สิ่งที่บุคคลกระทำได้กับมาตรฐาน ความต่อเนื่องของผลการเรียนรู้ การวัดแบบอิงเกณฑ์เป็นการ วัดผลการเรียนรู้ของบุคคลที่มีความหมายสมบูรณ์ในตัวเอง โดยไม่ต้องเปรียบเทียบกับคนอื่น ๆ ในกลุ่มเดียวกัน

จากแนวคิดการวัดแบบอิงเกณฑ์ข้างต้นจะเห็นได้ว่า Glaser ใช้คำว่า “criterion” ทั้งในความหมายว่า “specified behaviors” ซึ่งหมายถึง รายละเอียดของพฤติกรรมการเรียนรู้เฉพาะเรื่องและ “level of proficiency” ซึ่งหมายถึง ระดับสมรรถนะ หรือ คะแนนจุดตัด จึงเกิดความสับสนเกี่ยวกับการวัดอิงเกณฑ์ (criterion-referenced measurement) หมายความว่าอย่างไรแน่ และเกิดความเข้าใจผิดว่า แบบทดสอบอิงเกณฑ์ก็คือ แบบทดสอบอะไรก็ได้ที่มีการกำหนด เกณฑ์จุดตัดเพื่อการประเมินเข้าไว้ ดังคำอธิบายของ Popham ที่กล่าวมาแล้วข้างต้น อย่างไรก็ตามต่อมา Hambleton, et al. (1978) ได้อธิบายสรุปความหมายของคำว่า “เกณฑ์” ในวลี “การวัดแบบอิงเกณฑ์” ไม่ได้หมายถึงระดับสมรรถนะที่เป็นคะแนนจุดตัด เพื่อการประเมิน ผ่าน/ตก แต่หมายถึง กรอบพฤติกรรมการเรียนรู้ ที่ระบุไว้อย่างชัดเจน (criterion-as-domain) ว่า ผู้เรียนต้องทำอะไรได้ อะไรเป็นบ้าง ซึ่งเป็นความหมายที่เป็นประโยชน์อย่างยิ่งต่อการจัดการเรียนรู้ในชั้นเรียน และแม้ว่า การกำหนดเกณฑ์คะแนน จุดตัด ผ่าน/ตก เป็นเรื่องที่จะต้องทำในการวัดแบบอิงเกณฑ์ แต่จุดประสงค์สำคัญของการวัดแบบอิงเกณฑ์ก็คือ การวัดว่า นักเรียนสามารถทำอะไรได้ทำอะไรเป็นได้เท่าไรจากกรอบ พฤติกรรมที่กำหนดไว้อย่างชัดเจนแล้ว ซึ่งจะทำให้ครูผู้สอนให้ความสำคัญกับกรอบเนื้อหาอย่างจริงจัง มากกว่าการไปกำหนด เกณฑ์คะแนนจุดตัดผ่าน/ตก เพื่อให้เกิดความเข้าใจแนวคิดการวัด

แบบอิงเกณฑ์ของ Glaser ให้ชัดเจนยิ่งขึ้น จะขยายความแนวคิดนี้ต่อไปดังภาพที่ 6



ภาพที่ 6 ทฤษฎีการวัดแบบอิงเกณฑ์

จากภาพที่ 6 จะเห็นได้ว่าการวัดแบบอิงเกณฑ์อธิบาย สถานการณ์ในการวัดกรณีกรอบสาระหลัก (domain) มีขนาดเล็ก เฉพาะเจาะจง มีความเป็นเนื้อเดียวกันเหมือนแกงหนึ่งหม้อ เวลาทำ การสอบวัด สุ่มมาเพียงนิดเดียวก็สามารถแทนทั้งหมดได้ เราสามารถใช้แกงที่ตักมาหนึ่งช้อนประเมินแกงทั้งหมดได้ว่ามี รสชาติเป็นอย่างไร ชาตรสเค็มหรือเปรี้ยว ในทำนองเดียวกันกรณี การวัดแบบอิงเกณฑ์ตัวอย่างเนื้อหาที่สุ่มมาสร้างเป็นแบบทดสอบ ซึ่งเปรียบเสมือนแกงหนึ่งช้อนนั้นเหมือนกับกรอบเนื้อหาทั้งหมด เปรียบเสมือนกับแกงหนึ่งหม้อ ดังนั้นเมื่อนำแบบทดสอบอิงเกณฑ์ มาทำการสอบวัดความสามารถของบุคคล คะแนนที่ได้แม้ว่าจะไม่ใช่ คะแนนจริงเพราะไม่ได้นำเนื้อหาทั้งหมดมาสอบวัดแต่ก็เป็นคะแนน เสมือนจริง ซึ่งเชื่อถือได้ว่าเป็นคะแนนจริง (dependable) เพราะเนื้อหาที่สุ่มมาสร้างเป็นแบบทดสอบนั้นเหมือนกับเนื้อหา ในกลุ่มสาระหลัก เป็นตัวแทนเนื้อหาในกลุ่มสาระหลักได้ ดังนั้น เราจึงสามารถนำคะแนนดิบที่ได้มาคิดเป็นร้อยละแล้วเทียบกับเกณฑ์ ซึ่งได้แก่กลุ่มสาระหลัก ว่านักเรียนมีความรู้ความสามารถคิดเป็น ร้อยละเท่าไรของความรู้ทั้งหมดในกลุ่มสาระหลัก ดังที่ Popham (2014) อธิบายไว้ว่า ในการวัดความสามารถในการเขียนสะกดคำ ที่มักเขียนผิด จำนวน 250 คำ สร้างแบบทดสอบสะกดคำโดยสุ่มคำ ที่มักเขียนผิดมา 30 คำ นักเรียนคนหนึ่งทำแบบทดสอบฉบับนี้ ได้คะแนน 27/30 คิดเป็น 90% สามารถแปลความหมายคะแนนนี้ โดยอิงกับเกณฑ์ซึ่งได้แก่กลุ่มสาระหลักประกอบด้วยคำที่มักเขียน ผิด 250 คำ ไดว่านักเรียนคนนี้มีสามารถในการเขียนสะกดคำ

ได้ถูกต้อง 90% ของคำที่มักเขียนผิดทั้งหมด 250 คำ ตัวอย่างที่เป็นรูปธรรมอีกตัวอย่างหนึ่งของการวัดแบบอิงเกณฑ์คือการสอบใบขับขี่ ดังแสดงในภาพที่ 7

การสอบใบขับขี่

เจ้าหน้าที่กรมการขนส่งทางบก จะมอบหนังสือรวมป้ายจราจร และกฎจราจรเบื้องต้นให้อ่าน ก่อนทำการสอบ บาสที่จะเป็นข้อสอบให้เขียนกัน ซึ่งในข้อสอบจะมีป้ายจราจรและกฎจราจรที่เราอ่านมาเป็นข้อสอบ จำนวน 50 ข้อ เป็นข้อสอบแบบปรนัย มีให้เลือก ก-ง จะผ่านเกณฑ์การสอบที่ 45 คะแนน หรือคิดเป็น 90% ของข้อสอบทั้งหมด



ภาพที่ 7 การสอบใบขับขี่ การวัดแบบอิงเกณฑ์

ที่มา : KAPOOK. (n.d.) <https://car.kapook.com/view173898.html> car.kapook.com/view173898.html

จากภาพที่ 7 จะเห็นได้ว่า ในการสอบใบขับขี่นั้นกรอบสาระหลักคือ ป้ายจราจรและกฎจราจรซึ่งมีลักษณะเฉพาะเจาะจงและเป็นเอกนัย จึงมีลักษณะเป็นการวัดแบบอิงเกณฑ์เมื่อทำการทดสอบด้วยแบบทดสอบที่มี 50 ข้อ ถ้าผู้สอบทำคะแนนได้ 45/50 คิดเป็น 90% จะสามารถแปลความหมายได้ว่าผู้สอบมีความรู้ความสามารถเกี่ยวกับป้ายจราจรและกฎจราจร 90% ของเนื้อหาความรู้ทั้งหมด ไม่จำเป็นต้องนำคะแนนมาจัดลำดับที่หรือทำเป็นคะแนนมาตรฐานแต่อย่างใด เพราะคะแนนดิบเป็นคะแนนเสมือนจริง มีความสมบูรณ์ในตัวเอง เป็นตัวประมาณค่าคะแนนจริงที่เชื่อถือได้ สามารถคิดเป็นร้อยละและเทียบกับเกณฑ์ความรู้ความสามารถทั้งหมดในกรอบสาระหลักได้เลย

สรุปการประยุกต์ใช้ทฤษฎีการวัดแบบอิงกลุ่มและอิงเกณฑ์

การประยุกต์ใช้ทฤษฎีวัดแบบอิงกลุ่มและอิงเกณฑ์ในการสร้างแบบทดสอบและแปลความหมายคะแนนมีความแตกต่างกันดังนี้

ทฤษฎีการวัดแบบอิงกลุ่ม ควรนำทฤษฎีการวัดแบบอิงกลุ่มไปประยุกต์ใช้ในการสร้างแบบทดสอบและแปลความหมายคะแนน

กรณีต้องการวัดคุณลักษณะที่ค่อนข้างเป็นนามธรรม กรอบสาระหลักมีขนาดใหญ่ประกอบด้วยทักษะและพฤติกรรมการเรียนรู้ที่หลากหลายไม่เป็นเนื้อเดียวกันไม่สามารถแจออกมาได้ทั้งหมด เช่น การวัดความถนัดทางการเรียน การวัดสมรรถนะทางภาษา การวัดเชาวน์ปัญญา การวัดเจตคติ ตัวแปรเหล่านี้โดยทั่วไปจะถือว่าเป็นตัวแปรโครงสร้าง (construct variable) ต้องวิเคราะห์ความตรงเชิงโครงสร้างและวิเคราะห์ความเชื่อมั่นด้วยสูตรสัมประสิทธิ์แอลฟาของ Cronbach (1951)

ทฤษฎีการวัดแบบอิงเกณฑ์ ควรนำทฤษฎีการวัดแบบอิงเกณฑ์ไปประยุกต์ใช้ในการสร้างแบบทดสอบและแปลความหมายคะแนนกรณีต้องการวัดคุณลักษณะที่ค่อนข้างเป็นรูปธรรม กรอบสาระหลักมีขนาดเล็กประกอบด้วยทักษะ และพฤติกรรม การเรียนรู้ที่ไม่หลากหลายเป็นเนื้อเดียวกันสามารถแจออกมาได้ทั้งหมด เช่น การวัดผลการเรียนรู้จากการทดลองใช้ชุดกิจกรรมแต่ละบท หรือการวัดผลการเรียนรู้ในชั้นเรียนในการจัดการเรียนการสอนทั่วไป ตัวแปรประเภทนี้ถือว่าเป็นตัวแปรสังเกต (observe variable) ต้องวิเคราะห์ความแม่นยำตรงเชิงเนื้อหา และวิเคราะห์ความเชื่อมั่นด้วยสูตรของ Brennan and Kane (1977)

เอกสารอ้างอิง

- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277-289.
- Brennan, R. L. (1983). *Elements of Generalizability Theory*. Iowa: ACT Publications.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 292-334.
- Ebel, R. L. (1970). *Essentials of Educational Measurement*. New Jersey: Prentice-Hall.
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 48(1), 1-47.



- IELTS Liz. (2024.). *IELTS band score*. <https://ieltsliz-band-scores/>
- Glaser, R. (1963). Instructional technology, and the measurement of learning outcomes, Some questions, *American Psychologist*, 18, 519-521
- KAPOOK. (n.d.). *สอบใบขับขี่รถยนต์*.
<https://car.kapook.com/view173898.html>.
- Lyman, H. B. (1978). *Test scores and what they mean* (3rd Ed.). Englewood Cliffs.
NJ: Prentice Hall.
- Net News. (2559). *ใบรายงานผลโอเน็ตมีกี่แบบ*.
<https://www.niets.or.th>
- PISA Technical Report. (2022). *Proficiency Scale Construction for the Core Domains*.
<https://www.oecd.org/pisa/data/pisa2022technicalreport/>.
- Popham, W. J. (1981). *Modern Educational Measurement*. Englewood Cliffs. NJ: Prentice Hall.
- Popham, W. J. (2014). *Criterion-Referenced Measurement: Half a Century Wasted*. <https://www.ascd.org/el/articles/criterion-referenced-measurement-half-century-waste>.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.