เอกสารประมวลสาระรายวิชา

233545 การทดสอบ การประเมินผล และการวัดผลการเรียนภาษา

Testing, Evaluation, and Assessment in Language Learning

เอกสารประมวลสาระรายวิชา

233545 การทดสอบ การประเมินผล และการวัดผลการเรียนภาษา

Testing, Evaluation, and Assessment in Language Learning

ผู้จัดทำ

ผู้ช่วยศาสตราจารย์ ดร.กุลภักดิ์ กองสุวรรณกุล

สาขาวิชาภาษาต่างประเทศ

สำนักวิชาเทคโนโลยีสังคม

มหาวิทยาลัยเทคโนโลยีสุรนารี

ธันวาคม 2563

คำนำ

ผู้จัดทำขอกราบขอบพระคุณสาขาวิชาภาษาต่างประเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี ที่ได้มอบหมายให้ผู้เรียบเรียง–รวบรวมรับผิดชอบรายวิชา 233545 การทดสอบ การประเมินผล และการวัดผลการเรียนภาษา ทั้งยังเล็งเห็นความสำคัญของการจัดทำเอกสารประกอบการเรียนการสอน เพื่อจัดระบบองค์ความรู้ เอกสารประมวลสาระรายวิชาที่จัดทำขึ้นนี้นับเป็นการผลักดันความก้าวหน้าทางวิชาการสำหรับผู้วิจัยอย่างเป็นรูปธรรม

ผู้จัดทำขอขอบคุณหนังสือการวัดผลทางภาษาของ Brown & Abeywickrama (2010) ที่ผู้จัดทำใช้เป็นแกนหลักในการรวบรวมนำเนื้อหามาเรียบเรียงให้เป็นเอกสารประมวลฯ ที่มีความกระชับมากขึ้น นอกจากนี้ ผู้จัดทำขอขอบคุณ Brown (2016) ที่ผู้จัดทำใช้เป็นเนื้อหาหลักในบทที่ 6

ทั้งนี้เอกสารประมวลฯ ได้จัดทำขึ้นมาในฉบับสองภาษา (bilingual) ด้วยเหตุผลสองประการ ประการแรกคือ หลักสูตรมหาบัณฑิตของสาขาวิชาภาษาต่างประเทศมีนักศึกษาต่างชาติเข้ามาศึกษาทุกปี หากเอกสารประมวลฯ จัดทำในรูปแบบภาษาไทยแต่เพียงอย่างเดียวก็อาจจะไม่เป็นประโยชน์มากนักสำหรับนักศึกษากลุ่มนี้ จึงเห็นควรว่า เนื้อหาหลักและแบบฝึกหัดท้ายบทควรจะต้องเป็นภาษาอังกฤษ เหตุผลประการที่สองสำหรับการจัดทำเอกสารประมวลฯ เป็นฉบับสองภาษา ก็เพื่อผู้สนใจชาวไทยโดยทั่วไป เอกสารประมวลฯ มีรูปแบบการเผยแพร่ผ่านคลังปัญญาของมทส. ที่มีเป้าหมายเพื่อให้ข้อมูลในฐานะคลังความรู้ของแผ่นดิน การสืบค้นผ่านอินเตอร์เน็ตก็จะสามารถนำผู้สนใจชาวไทยมาสู่ตัวเอกสารประมวลฯ ได้ การออกแบบให้มีภาษาไทยด้วยจึงน่าจะเป็นประโยชน์ต่อผู้สนใจเหล่านั้น

ผู้วิจัยขอขอบคุณบุคลากรและเจ้าหน้าที่ที่เกี่ยวข้องทุกท่าน ที่ช่วยอำนวยความสะดวกเพื่อให้การเผยแพร่เอกสารประมวลฯ ผ่านคลังปัญญามทส.สำเร็จลุล่วงไปได้

และผู้วิจัยขอขอบคุณทุกๆ คนในครอบครัว A strong family support just makes it even better!

คำอธิบายรายวิชา (Course Description)

หลักการ แนวคิด ทฤษฎีของการวัดและประเมินผล ประเภทของแบบทดสอบ การออกแบบและการสร้างแบบทดสอบภาษา การวิเคราะห์ข้อสอบเป็นรายข้อ การวัดผลแบบทางเลือก แนวโน้มและประเด็นที่เกี่ยวข้องกับการทดสอบ

Principles and theories in testing, assessment and evaluation; types of tests; language test design and construction; item analysis; alternative assessment; trends and issues in language testing

วัตถุประสงค์รายวิชา (Course Objectives)

1. นักศึกษามีความรู้ความเข้าใจในแนวคิดและทฤษฎีของการวัดและประเมินผลทางการเรียนการสอนภาษาอังกฤษ
2. นักศึกษาสามารถวิเคราะห์และวิพากษ์ข้อสอบ การวัดผล และประเด็นที่เกี่ยวข้องกับการทดสอบด้านภาษาอย่างมีระบบ
3. นักศึกษาสามารถสร้างแบบทดสอบภาษาโดยใช้กรอบแนวคิด หลักการ และทฤษฎีของการวัดและประเมินผลได้

1. The students understand the principles and theories in testing and assessing the achievements in language learning.
2. The students can analyze and systematically develop a critique of tests, assessments, and issues relating to language testing.
3. The students can construct a language test using the principles and theories in language testing and assessment.

ผลการเรียนรู้ระดับรายวิชา (Course Learning Outcomes)

1. สามารถสร้างแบบสอบที่มีความยากง่าย และความเที่ยงเหมาะสม
2. มีความรู้ความเข้าใจแนวคิดพื้นฐานเรื่องการทดสอบ และการวัดผลทางภาษา
3. สามารถใช้ภาษาอังกฤษเพื่อวัตถุประสงค์เชิงวิชาการและวิชาชีพได้ดี
4. แสดงความเข้าใจทฤษฎีที่สำคัญและงานวิจัยที่เกี่ยวกับการทดสอบและการประเมินผลภาษาอังกฤษได้
5. ประยุกต์ความรู้เพื่อใช้พัฒนาเครื่องมือวัดและแบบทดสอบภาษาอังกฤษที่มีประสิทธิภาพได้
6. บูรณาการความรู้ที่ศึกษาเพื่อออกแบบแบบสอบทางภาษาได้
7. วิเคราะห์เนื้อหาของโจทย์ข้อสอบและด้านที่วัด ตลอดจนคำนวณสถิติข้อสอบและแบบสอบได้
8. ระบุและสังเคราะห์ประเด็นวิจัยร่วมสมัยได้จากวรรณกรรมงานวิชาการด้านการทดสอบทางภาษา
9. บูรณาการความรู้ที่ศึกษาเพื่อศึกษาวิจัยด้านการทดสอบทางภาษาได้
10. สืบค้นแบบสอบมาตรฐานที่มีในปัจจุบันและข้อมูลที่เกี่ยวข้องได้
11. ดำเนินการเก็บข้อมูลการตอบสนองต่อข้อสอบที่พัฒนาขึ้นได้ด้วยตนเอง
12. คำนวณค่าสถิติข้อสอบและแบบสอบเพื่อตอบโจทย์วิจัยได้
13. นำเสนอบทวิพากษ์บทความและการศึกษาค้นคว้าผ่านสื่อดิจิตัลได้

1. Generate a test that is of an appropriate difficulty level and is reliable.
2. Demonstrate knowledge and understanding in basic concepts of language testing and assessment.
3. Demonstrate a good command of the English language for academic and professional purposes.
4. Demonstrate understanding of key theories and research related to English language testing, assessment, and evaluation.
5. Apply the acquired knowledge to develop effective English assessment tools.

6. Integrate the knowledge acquired for language test design.
7. Analyze test-task content and the construct competence measured by the test created, and calculate test and item statistics.
8. Identify and synthesize contemporary issues for language-testing research from the literature.
9. Integrate the knowledge acquired for research on language testing and assessment.
10. Search for a modern standardized test and its details.
11. Administer a test engineered and collect its item responses responsibly.
12. Perform test and item analyses for research purposes.
13. Present an evaluative discussion of the articles read and information searched for with a digital aid.

แผนการสอน (12-Week Term Plan)

สัปดาห์ที่ (Week)	หัวข้อ/รายละเอียด (Content)	ผลการเรียนรู้ระดับรายวิชาที่ (CLO Number)	แนวทางการสอนและการเรียนรู้ (Teaching and Learning Approaches)	กิจกรรมการประเมิน (Assessment Activities)
1	บทนำรายวิชา Course introduction	2, 4	บรรยาย lecture, อภิปรายกลุ่มย่อย small-group discussion	ประเมินภาพรวม holistic observation, มอบหมายหัวข้อให้นำเสนอในชั้นเรียน working on a selected topic for presentation
2	หลักการของการวัดและประเมินผล Principles of measurement and evaluation	2, 3, 4	บรรยาย lecture, อภิปรายกลุ่มย่อย small-group discussion	ประเมินภาพรวม holistic observation
3	แนวคิด ทฤษฎีของการวัดและประเมินผลทางภาษา Basic concepts in language testing and assessment	2, 3, 4, 5, 8, 13	บรรยาย lecture, การเรียนรู้ โดยใช้ปัญหา problem-based teaching	นำเสนอในชั้นเรียน in-class oral presentation, วิพากษ์งานเกี่ยวกับความตรง (validity) critiquing validity studies
4	วัตถุประสงค์และวิธีทดสอบของการวัดและประเมินผล Assessment purposes and approaches	3, 4, 5, 10	บรรยาย lecture, การเรียนรู้ โดยใช้ปัญหา problem-based teaching	มอบหมายงานสำรวจแบบทดสอบมาตรฐาน surveying a standardized test
5	ชนิดของแบบทดสอบและแบบทดสอบมาตรฐาน Test types and standardized tests	3, 10, 13	บรรยาย lecture, อภิปรายกลุ่มย่อย small-group discussion, การเรียนรู้ โดยใช้ปัญหา problem-based teaching	นำเสนอแบบทดสอบมาตรฐาน presenting the standardized test, มอบหมายงานเขียนพิมพ์เขียวแบบสอบ drafting a test blueprint
6	การออกแบบและการสร้างแบบทดสอบ Designing and writing tests	3, 5, 6, 7	บรรยาย lecture, การเรียนรู้ โดยใช้ปัญหา problem-based teaching, การนำเสนอในชั้นเรียน oral presentation, การเรียนรู้แบบค้นพบ discovery teaching	นำเสนอพิมพ์เขียนแบบสอบ presenting the test blueprint

7	สอบกลางภาค Midterm examination	2, 3, 4	สอบข้อเขียน Written exams: Essay	ประเมินความครบถ้วนและความเหมาะสมของเนื้อหาที่ตอบคำถามในการสอบ evaluating completeness and appropriateness of the answers
8	การวัดทักษะทางภาษาทั้งสี่ Assessing four language skills	1, 3	การบรรยาย lecture,	มอบหมายงานเขียนข้อสอบฉบับต้นแบบ writing a pilot test
9	เทคนิคและขั้นตอนการทดสอบ Testing techniques and procedures	4, 6	การเรียนรู้จากโจทย์ปัญหา problem-based teaching, การอภิปรายกลุ่ม small-group discussion	นำเสนอและวิพากษ์ข้อสอบฉบับต้นแบบ presenting and critiquing the pilot test, มอบหมายงานเก็บข้อมูลด้วยแบบสอบที่พัฒนาขึ้น collecting item responses with the developed test
10	การประเมินผลการเรียนรู้แบบทางเลือก Alternative Assessments	1, 3, 6	การบรรยาย lecture การอภิปรายกลุ่มจากตัวอย่างผลการประเมินแบบทดสอบ Group discussion based on real examples	ประเมินจากการอภิปรายและวิพากษ์ evaluating discussion and critiques
11	การวิเคราะห์แบบทดสอบ Item Analysis สถิติด้านการวัดและประเมินผล Test Statistics	7, 8, 12	การวิเคราะห์ผลการประเมินแบบทดสอบ hands-on calculation	มอบหมายงานสังเคราะห์หัวข้อในวรรณกรรม synthesizing insights on a topic
12	แนวโน้มการวัดและประเมินผลทางภาษาอังกฤษ Current trends and issues in language testing การนำเสนอโครงงานการพัฒนาและประเมินแบบทดสอบภาษาอังกฤษ Presenting a test development project	3, 6, 8, 9, 11, 12, 13	การอภิปรายกลุ่ม การนำเสนอปากเปล่า Group discussion Oral presentation	ประเมินการนำเสนอปากเปล่า evaluating oral presentation

การประเมินผลการเรียนรู้ (Assessment Activities)

กิจกรรมการประเมิน (เช่น การเขียนรายงาน โครงงาน การสอบย่อย การสอบกลางภาค การสอบปลายภาค) (Assessment Activity)	ผลการเรียนรู้ (CLO)	วิธีการประเมินผล (Assessment Method)	สัดส่วนของการประเมินผล (Assessment Proportion)	
			ผลการเรียนรู้ (Outcome)	กิจกรรมการประเมิน (Activity)
การมีส่วนร่วม กิจกรรม และงานที่มอบหมายทั่วไป Overall class participation and responsibility	2, 3, 4	สังเกตพฤติกรรม Observing behaviors in class	5%	5%
การนำเสนอบทความ Presenting an academic paper or chapter	2, 3, 13	ประเมินความเข้าใจขณะนำเสนอและประเมินการตั้งคำถามต่อการนำเสนอ Evaluating understanding displayed in the presentation, and the questions posed	10%	10%
การนำเสนอข้อสอบมาตรฐานจากการสำรวจ Presenting a standardized test from a survey	3, 10, 13	ประเมินความสมบูรณ์ของรายละเอียดข้อสอบและการใช้ความรู้ที่เรียนวิเคราะห์แบบสอบมาตรฐานและองค์ประกอบที่ได้จากการสำรวจ Evaluating completeness of the details about the test obtained from the survey	10%	10%
การนำเสนอพิมพ์เขียวข้อสอบต้นแบบ Presenting a test blueprint	3, 5, 6	ประเมินการนำเสนอในชั้นเรียน ประเมินการทำโครงงาน Evaluating oral presentation in class and the quality of the test blueprint	5%	5%
การอภิปรายประเด็นร่วมสมัย Discussing contemporary issues in language testing and assessment	8, 9, 10	ประเมินการอภิปราย Evaluating the discussion and active participation	10%	10%
การนำเสนอปากเปล่า: โครงงานพัฒนาแบบสอบ Presentation: Test development project	3, 6, 8, 9, 11, 12, 13	ประเมินการนำเสนอปากเปล่า และรายงานโครงงานพัฒนาแบบสอบ Evaluating an oral presentation and its report	40%	40%
การสอบกลางภาค Midterm Examination	2, 3, 4	ประเมินคำตอบข้อสอบอัตนัย	20%	20%

กิจกรรมการประเมิน (เช่น การเขียนรายงาน โครงงาน การสอบย่อย การสอบกลางภาค การสอบปลายภาค) (Assessment Activity)	ผลการเรียนรู้ (CLO)	วิธีการประเมินผล (Assessment Method)	สัดส่วนของการประเมินผล (Assessment Proportion)	
			ผลการเรียนรู้ (Outcome)	กิจกรรมการประเมิน (Activity)
		Assessing answers to open-ended questions		
รวม (Total)			100%	100%

นักศึกษาสามารถตรวจสอบผลคะแนนของงานที่มอบหมายภายใน 2 สัปดาห์หลังการส่งงาน ตรวจสอบคะแนนหลังการสอบย่อยหลังการสอบแต่ละครั้ง และตรวจสอบคะแนนเก็บรวมกับอาจารย์ผู้สอนก่อนการส่งงานที่ได้รับมอบหมายชิ้นงานสุดท้าย

Students can get feedback about their assignments and check the scores of their assignments within two weeks after submission. Their formative scores will be posted for checking before the submission of the final assignment.

การประเมินและปรับปรุงรายวิชา (Course Evaluation and Improvement)

1. การประเมินประสิทธิผลของรายวิชา ได้ดำเนินการโดยให้นักศึกษาประเมินทุกภาคการศึกษาผ่านระบบการประเมินการสอนผ่านอินเทอร์เน็ต และให้คณะกรรมการในระดับสาขาวิชา และระดับสำนักวิชา ตรวจสอบผลการประเมินการเรียนรู้ของนักศึกษา
2. กระบวนการปรับปรุงรายวิชา ได้ดำเนินการปรับปรุงทุกภาคการศึกษาตามผลการประเมินของนักศึกษา และปรับปรุงทุก 5 ปี ตามข้อเสนอแนะของผู้มีส่วนได้ส่วนเสียของหลักสูตร

นักศึกษาสามารถอุทธรณ์ หรือร้องเรียนผลการประเมินได้ที่ email: complaintist@sut.ac.th

1. Evaluation Strategies on course effectiveness by students

   Students evaluate the course through the university's online evaluation system.

2. Course review and improvement plan

   The course is reviewed every trimester in light of the results of evaluation from the students and improved every 5 years according to the feedback and suggestions of the program's stakeholders.

Students can appeal and complain about unsatisfactory matters by sending an e-mail message to: complaintist@sut.ac.th

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1 แนวคิดของการวัดและประเมินผลทางภาษา

Chapter 1 Concepts in Language Testing and Assessment

ในเอกสารประมวลสาระรายวิชานี้ จะนำเสนอเนื้อหาเกี่ยวกับแนวคิดและหลักการของการวัดและประเมินผล ประเภทของแบบทดสอบ การออกแบบและการสร้างแบบทดสอบทางภาษา การวิเคราะห์ข้อสอบเป็นรายฉบับและรายข้อ การประเมินผลทางเลือก ตลอดจนแนวโน้มและประเด็นที่เกี่ยวข้องกับการทดสอบ ในบทนี้จะกล่าวถึงแนวคิดของการวัดและประเมินผล โดยเฉพาะที่เกี่ยวข้องกับการทดสอบทางภาษา เนื้อหาในบทนี้ประกอบด้วย

1.1 ความแตกต่างระหว่างการทดสอบ การวัด และการประเมินค่า

1.2 ความสมเหตุสมผลของการตีความความหมายคะแนน

1.3 ความสำคัญของการทดสอบทางภาษา

1.4 Exercise (แบบฝึกหัดท้ายบท)

In this course document, contents will be presented as to concepts and principles of language testing and assessment, types of test, designs and construction of language tests, test and item analyses, alternative assessment, and trends as well as issues in language testing and assessment. In this chapter, some basic concepts in language testing and assessment will be covered. The organization of this chapter is as follows:

1.1 Distinctions among testing, measurement, and evaluation

1.2 Construct validity

1.3 Importance of language testing

1.4 Exercise

1.1 ความแตกต่างระหว่างการทดสอบ การวัด และการประเมินค่า (Distinctions among Testing, Measurement, and Evaluation)

คำว่า *การทดสอบ การวัด*และ*การประเมินค่า* มักใช้ในความหมายที่ใกล้เคียงกัน และใช้อยู่ในบริบทที่ไม่ต่างกันมากนัก เช่น เราจัดการทดสอบเพื่อวัดและประเมินสมิทธิภาพภาษาอังกฤษของผู้เรียน เป็นต้น ทั้งนี้คำว่า *การทดสอบ* (testing) มักหมายถึง การนำแบบสอบมาใช้วัดหรือเก็บข้อมูล แบบสอบโดยปกติหมายถึงมาตรวัดทางจิตวิทยาหรือทางการศึกษา ที่ออกแบบมาเพื่อดึงเอาพฤติกรรมการแสดงออกออกมา เช่น การเลือกคำตอบจากตัวเลือกที่มีให้ เพื่อให้สามารถอนุมานเกี่ยวกับลักษณะบางประการในบุคคลหนึ่งๆ ได้ เช่น การมีสมิทธิภาพ (proficiency) ภาษาอังกฤษสูงหรือต่ำ เป็นต้น

The terms *testing*, *measurement,* and *evaluation* have close meanings. They are also often used in contexts that are similar. For example, we may administer a test, in order to measure and evaluate the students' English proficiency. The term *testing* usually refers to when a psychological or educational test is used for eliciting some behaviors (such as selecting a choice) from an individual. The aim is to draw an inference about certain characteristics of that individual (e.g., having high or low proficiency in English).

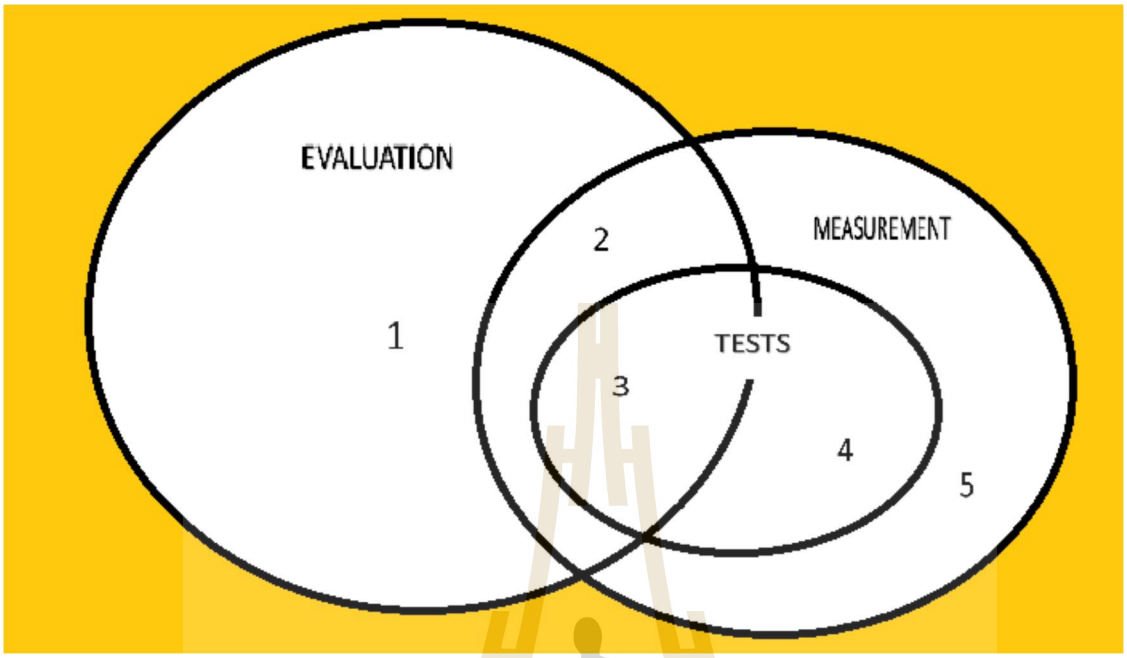
คำว่า *การวัด* (measurement) หมายถึง กระบวนการที่ทำให้ลักษณะที่เก็บได้จากตัวบุคคลที่วัดได้สังเกตได้กลายมาเป็นจำนวนที่นับได้ เช่นการวัดส่วนสูง การชั่งน้ำหนัก การให้ตอบข้อสอบเป็นต้น ลักษณะที่เก็บได้จากตัวบุคคล และที่วัดได้สังเกตได้นี้ เป็นได้ทั้งคุณลักษณะภายในที่เกี่ยวข้องกับสติปัญญาและความรู้ เช่น ความถนัด (aptitude) เชาว์ปัญญา (intelligence) ทัศนคติ (attitude) เป็นต้น หรืออาจจะเป็นความสามารถที่เกี่ยวข้องกับสติปัญญาและความรู้ เช่น ความสามารถในการพูดภาษาอังกฤษ เป็นต้น จุดที่น่าสังเกตในประเด็นนี้คือ คุณลักษณะหรือความสามารถเหล่านี้เป็นคนละสิ่งกับตัวบุคคลที่เราไปเก็บข้อมูลที่วัดได้สังเกตได้มา โดยทั่วไปเราจึงมิได้วัดตัวบุคคล หากแต่เราวัดคุณลักษณะบางประการของตัวบุคคล

The term *measurement* normally refers to the process in which we quantify certain characteristics from individuals, e.g., height, weight, answers in a test. The characteristics could be mental attributes such as aptitude, intelligence, and attitude, or mental abilities such as ability to speak English. What is important is that these attributes

and abilities are not the same as the individuals who have the attributes and abilities. We are thus measuring the attributes or abilities, not the individuals themselves.

คำว่า การประเมินผลหรือการประเมินค่า (evaluation) มีรากศัพท์มาจากคำว่า คุณค่า (value) การประเมินค่าจึงเป็นการตัดสินใจเกี่ยวกับคุณค่าของสิ่งใดสิ่งหนึ่ง มักเกี่ยวข้องกับความสามารถของผู้ตัดสินใจ และคุณภาพของข้อมูลที่ใช้เพื่อตัดสินใจ ตัวอย่างเช่น หากเราตัดสินใจบนพื้นฐานของข้อมูลที่เป็นข่าวลือ ก็อาจกล่าวได้ว่า ข้อมูลที่ใช้ในการตัดสินใจมีคุณภาพต่ำ คำว่าการประเมินค่า ในภาษาไทยมีความใกล้เคียงกับคำว่า การวัดและประเมินผล (assessment) และมักจะใช้ทดแทนกันได้ในหลายสถานการณ์ อย่างไรก็ดี จุดเน้นของคำว่า การประเมินค่าอยู่ที่การตัดสินค่าหรือการตัดสินใจให้ได้ผลลัพธ์อย่างหนึ่งอย่างใดออกมา เช่น ตัวเลข หรือเกรดที่เป็นตัวอักษร ส่วนการวัดและประเมินผล จุดเน้นจะยังคงอยู่ที่การที่ต้องมีการวัด เพื่อความสะดวกในเอกสารประมวลฯ นี้ จะใช้คำว่า การประเมิน การประเมินค่า และการวัดและประเมินผลทดแทนกันได้

The next term is *evaluation*. The root of *evaluation* is the word *value*. The term *evaluation* thus entails a decision regarding the value of something. The term is often involved with the ability of the person who makes a decision, and the quality of the information used for decision making. For example, if we evaluate the students' performance based on some rumor, then it might be said that the quality of the information we use for evaluation is low. In Thai, the word *evaluation* is close to the notion *assessment* in terms of meaning and may be used interchangeably in certain contexts. However, the key semantic component of the word *evaluation* is in the judgment and decision that is made and may result in such outcomes as scores and letter grades. By contrast, the key semantic component of the word *assessment* is in the measurement of something—a quantification of something before making any decision. For convenience, the terms *evaluation* and *assessment* will be used interchangeably when the contexts allow.



รูปที่ 1.1 ความสัมพันธ์ระหว่างการวัด แบบสอบ และการประเมินค่า (relationship among *measurement*, *tests*, and *evaluation*) (Minh 2015)

ในรูปที่ 1.1 แสดงความสัมพันธ์ระหว่างการวัด การทดสอบ และการประเมินค่า พื้นที่หมายเลข 1 หมายถึงการประเมิน โดยที่ไม่มีแบบสอบและเครื่องมือที่ใช้วัดอย่างเป็นกิจจะลักษณะ (no measures) ตัวอย่างเช่น บทบรรยายเชิงคุณภาพถึงการแสดงออกของผู้เรียน (qualitative descriptions of student performance) พื้นที่หมายเลข 2 หมายถึง การใช้เครื่องมือวัดที่ไม่ใช่แบบสอบเพื่อการประเมิน ตัวอย่างเช่น การที่ครูผู้สอนทำการจัดอันดับผู้เรียน (ranking) เพื่อที่จะตัดเกรดผลการเรียน พื้นที่หมายเลข 3 หมายถึง การใช้แบบทดสอบเพื่อการประเมิน เช่น การใช้แบบวัดผลสัมฤทธิ์ทางการเรียนรู้ (achievement test) เพื่อดูความก้าวหน้าของผู้เรียน พื้นที่หมายเลข 4 หมายถึง การใช้แบบสอบที่ไม่ได้ทำเพื่อการประเมิน เช่น การใช้แบบวัดสมิทธิภาพ (proficiency test) ในฐานะเครื่องมือเกณฑ์ (criterion measure) ของการศึกษาวิจัย และพื้นที่หมายเลข 5 หมายถึง การใช้เครื่องมือวัดที่ไม่ใช่แบบสอบและไม่ได้ใช้เพื่อการประเมิน ตัวอย่างเช่น การกำหนดรหัสรายวิชาสำหรับวิชาต่างๆ ในโรงเรียน เป็นต้น

In Figure 1.1, a relationship among measurement, tests, and evaluation is displayed. The no. 1 area represents an evaluation without any test and measures. Examples include qualitative descriptions of student performance. The no. 2 area in the figure represents the use of a non-test measure for evaluation. Examples include teachers'

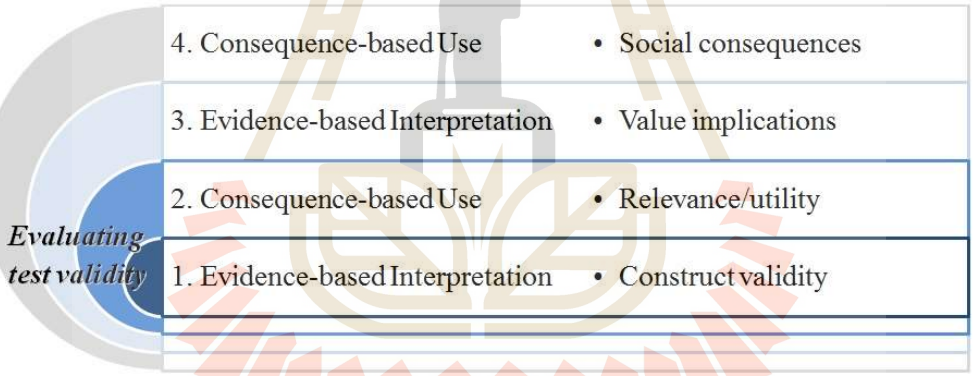
ranking of the students used for assigning grades. The no. 3 area represents the use of a test for the purpose of evaluation. Examples include use of an achievement test to determine students' progress. The no. 4 area represents the use of a test, not for evaluation. Examples include use of a proficiency test as a criterion measure for research purposes. The no. 5 area represents use of a non-test measure, not for evaluation. Examples include when a program administrator assigns code numbers to the subjects in a school.

1.2 ความสมเหตุสมผลของการตีความความหมายคะแนน (Construct Validity)

หนึ่งในทฤษฎีที่ใช้กันอย่างแพร่หลายในแวดวงการทดสอบและการประเมินผลทางภาษา ได้แก่ โมเดลด้านความตรง (validity model) ของ Messick (1989) สาเหตุเพราะทฤษฎีนี้นับเป็นหนึ่งในทฤษฎีเกี่ยวกับความตรงที่สมบูรณ์ครอบคลุมมากที่สุดเท่าที่มีอยู่ในปัจจุบัน (Bachman 2000; Brown 2000; Kane 2006; McNamara 2006; Fulcher & Davidson 2007; Moss 2007; Rigney et al. 2008; Kane 2012) ทฤษฎีนี้ว่าด้วยการพิจารณาความสมเหตุสมผลแบบองค์รวมของความเหมาะสมและความเพียงพอของหลักฐานเชิงประจักษ์และเหตุผลเชิงทฤษฎี ที่สามารถรองรับการตีความและการใช้ผลคะแนนหรือข้อมูลอื่นๆ จากการวัดและการทดสอบได้ (construct validity) ดังแสดงในรูปที่ 1.2 (หน้า 6) ทฤษฎีนี้แบ่งเป็นด้าน (1) ความหมายของคะแนน (construct) (2) ความเกี่ยวข้องและประโยชน์ใช้สอย (relevance and utility) (3) คุณค่าแฝง (value implications) และ (4) ผลกระทบทางสังคม (social consequences) โดยมีความหมายคะแนนเป็นแกนกลางของการพิจารณาความสมเหตุสมผลในทุกๆ ด้าน (progressive matrix with the construct being central in all validity investigations) ตัวอย่างเช่น หากเราจะพิจารณาคุณค่าแฝงของข้อสอบฉบับหนึ่ง (หมายเลข 3 คุณค่าแฝง) เราจำต้องทราบและพิจารณาความหมายของคะแนนจากข้อสอบฉบับนั้นก่อน เช่นว่าคะแนนที่ผู้เรียนได้จากข้อสอบฉบับนี้บ่งชี้ถึงความสามารถด้านใด (หมายเลข 1 ความหมายคะแนน) จากนั้นจึงจะพิจารณาได้ว่า ความสามารถด้านนั้นๆ ที่วัดในข้อสอบมีความเกี่ยวข้องกับหลักสูตรที่ใช้อยู่หรือไม่ และอย่างไร (หมายเลข 2 ความเกี่ยวข้องและประโยชน์ใช้สอย) ก่อนที่เราจะพิจารณาถึงคุณค่าแฝงของข้อสอบนั้นได้ เป็นต้น

One of the theoretical models that are used widely in the field of language testing and assessment is the validity model by Messick (1989). A reason for its wide use lies in the comprehensiveness of the model (Bachman 2000; Brown 2000; Kane 2006; McNamara 2006; Fulcher & Davidson 2007; Moss 2007; Rigney et al. 2008; Kane 2012). In the model, validity is concerned with an evaluative judgment of the appropriateness and

adequacy with which empirical data and theoretical rationales could support inferences and use of test scores and other modes of assessment. This is called construct validity. Illustrated in Figure 1.2, Messick's model covers (1) the construct aspect, also known as the aspect of score meaning, (2) relevance and utility, (3) value implications, and (4) social consequences. In Messick's model, which is a progressive matrix, the construct is indispensable, making it central to all other validity efforts. For example, we want to consider the value implications of a test (no. 3, value implications in Figure 1.2). We would want to know first what the test measures, especially what aspect of human competence a score from the test would represent (no. 1, construct validity). Then we would need to know if and in what way the aspect of competence assessed in the test is related to the curriculum or program in question (no. 2, relevance and utility). Only then can we consider the value implications of the test meaningfully.



รูปที่ 1.2 โมเดลด้านความตรงของ Messick (1989: 20, adapted) (Messick's (1989, adapted) model of construct validity)

รูปที่ 1.2 แสดงให้เห็นว่า โมเดลทฤษฎีของ Messick (1989) ครอบคลุมทั้งการตีความความหมายคะแนนและการนำคะแนนไปใช้ (test-score interpretation and use) ความตรงหรือ validity ในคำนิยามของ Messick (1989: 13) คือ การพิจารณาตัดสินความสมเหตุสมผลแบบองค์รวมของการตีความ การแปลความหมาย ตลอดจนการใช้ผลการวัดและการทดสอบรูปแบบต่างๆ กล่าวได้ว่า ความตรงของความหมายคะแนนเกิดขึ้น เมื่อความหมายที่สร้างขึ้นมาจากคะแนนหรือจากค่าในมาตรวัดใดๆ สามารถสะท้อนทักษะ ความสามารถ หรือองค์ประกอบทางสติปัญญาหนึ่งๆ ของมนุษย์ (human

competence – สามัตถิยะ) ได้อย่างสมเหตุสมผลนั่นเอง ทั้งนี้ จุดเน้นที่ Messick กล่าวไว้เกี่ยวกับความสมเหตุสมผลคือ เมื่อหลักฐานและทฤษฎีมีความเหมาะสมและความเพียงพอเป็นจุดสำคัญ

Figure 1.2 shows that the theoretical model by Messick (1989) covers both test-score interpretation and test-score use. That is, validity in Messick's (1989: 13) definition entails an evaluative, holistic judgment of interpretation and use of the test scores, and outcomes of other forms of assessment. It could be said that validity emerges when the score meaning constructed could reflect the skills, abilities, or other aspects of human competence aptly. A key point in Messick's rationality is when the empirical evidence and theoretical rationales are suitable and sufficient.

สิ่งที่วัดได้สังเกตได้ในการวัดผลหรือการทดสอบแต่ละครั้ง เช่น การที่ตัวเลือกตัวหนึ่ง (option) ถูกเลือกมาจาก 4 ตัวเลือกในข้อสอบปรนัยข้อหนึ่งๆ นำมาผ่านการแปลความหมายเป็นการตอบผิดหรือตอบถูก ด้วยการให้ค่าเป็นคะแนน 0 คือตอบผิด หรือ 1 คือตอบถูก (0 or 1 in a dichotomous test item) เป็นต้น ตามโมเดลของ Messick (1989: 30) คำตอบแต่ละข้อคือการปรากฏรูป (manifestation) ของการที่ด้านใดด้านหนึ่งหรือหลายด้านรวมกันของสามัตถิยะของมนุษย์มีปฏิสัมพันธ์กับชิ้นงาน (competence interacting with a test task) เกิดเป็นพฤติกรรมที่วัดได้สังเกตได้หนึ่งๆ สามัตถิยะเหล่านี้มีอยู่จริง ไม่ว่าการวัดจะเกิดขึ้นหรือไม่ก็ตาม หลักคิดนี้จัดอยู่ในแนวปรัชญาเรียลลิสต์ (realism) (Fulcher 2014) ส่วนการวัดหรือการทดสอบโดยทั่วไปเป็นไปตามแนวปรัชญาคอนสตรัคทิวิสต์ (constructivism) (Messick 1989: 30) เพราะนอกจากโดยทั่วไปจะหมายถึงการที่ผู้เรียนสร้างการเรียนรู้ด้วยตนเองได้แล้ว ยังหมายรวมไปถึงการสร้างบทตีความของสิ่งที่วัดได้และของทฤษฎีที่รองรับการตีความนั้นๆ อย่างเป็นระบบและสมเหตุสมผล

What is measured in each assessment—for example, when one option is selected out of four choices in a selected-response, multiple-choice question—and then is scored in the form of a right or wrong answer, e.g., 0 or 1 in a dichotomous test item, is meaningful for validity investigation. According to Messick (1989), test and item responses are manifestations of a particular aspect of competence in interaction of the test tasks. The result of the interaction is the behavior that is measurable or observable, for example, selecting a choice, saying something in an interview etc. Whether or not there is the assessment, the competence exists. This concept is in accordance with realism (Fulcher

2014). By contrast, the act of measurement or testing is usually in accordance with constructivism (Messick 1989: 30). This is so because, apart from referring to when the learners construct their own learning, constructivism in this context is when we build up or 'construct' a validity argument—the argument that has to systematically and meaningfully deal with (a) the interpretation of what it is that is being measured and (b) the theoretical rationale behind the interpretation.

1.3 ความสำคัญของการทดสอบทางภาษา (Importance of Language Testing)

Heaton (1988: 5) ได้กล่าวไว้ว่า การทดสอบและการสอนมีความเกี่ยวพันกันอย่างแนบแน่น ถึงขนาดที่ว่า เป็นไปไม่ได้เลยที่จะทำงานในด้านหนึ่งโดยที่ไม่ต้องเกี่ยวข้องกับอีกด้านหนึ่งอยู่เนืองๆ จากข้อความนี้อาจพออนุมานได้ว่า การทดสอบมีความเกี่ยวข้องกับการสอนอย่างลึกซึ้ง หากจะแยกแยะว่าการทดสอบแบบใดที่ดีแบบใดที่ไม่ดี ก็คงจะไม่ผิดนักหากจะลองมองย้อนไปที่การสอนว่าแบบใดที่ดีและแบบใดที่ไม่ดี ตัวอย่างเช่น แนวโน้มการสอนภาษาอังกฤษในปัจจุบันนิยมเน้นที่การสื่อสาร แบบทดสอบที่มุ่งเน้นแต่เรื่องไวยากรณ์ การแปล หรือการแต่งประโยคให้เป็นตามข้อความที่ต้องการโดยไม่สนใจสถานการณ์การสื่อสาร ก็อาจถูกมองได้ว่า ไม่สะท้อนการสอนภาษาอังกฤษที่เน้นการสื่อสาร และอาจจะมิใช่แบบทดสอบที่ดีที่ยึดโยงกับการสอนมากนัก เป็นต้น

"Testing and teaching are [so] closely interrelated that it is impossible to work in either field without being constantly concerned with the other" (Heaton 1988: 5). From the statement, we may make an interpretation that testing is deeply related to teaching, to the extent that we may evaluate a test based on whether it is reflective of the teaching. For example, at present communicative trends is popular in language teaching. Tests that focus only on grammar, translation, and language manipulation without any element to do with communication may be said to be inappropriate tests. The tests may be accused of being not reflective of communicative English and out of teaching contexts.

ในมุมมองของการเรียนการสอน (Brown & Abeywickrama 2010) แบบสอบที่ดีอาจกล่าวโดยทั่วไปได้ว่า ควร

1) ช่วยระบุเนื้อหาที่ยากสำหรับชั้นเรียน หรือสำหรับผู้เรียนรายคน

2) ทำให้ครูสามารถเพิ่มประสิทธิผลในการสอน ด้วยการปรับการเรียนการสอนให้เหมาะสม

3) ให้โอกาสผู้เรียนได้แสดงความสามารถในการทำชิ้นงานให้สำเร็จลุล่วงได้

From the perspective of teaching (Brown & Abeywickrama 2010), a good test should

1) help to locate the precise areas of difficulty encountered by the class or by an individual student;

2) enable the teachers to increase their effectiveness by making adjustments in their teaching;

3) provide the students with an opportunity to show their ability to perform a certain task.

ทั้งนี้ ด้านที่ควรได้รับการทดสอบอาจแบ่งออกได้เป็น 3 กลุ่มใหญ่ๆ ได้แก่

1) ทักษะทั้ง 4 ด้านที่ใช้ในการสื่อสาร ได้แก่ การฟัง การพูด การอ่าน และการเขียน

2) ด้านทางภาษาที่เรียน เช่น ไวยากรณ์และการใช้ภาษา คำศัพท์ และด้านระบบเสียงในภาษา (phonology)

3) องค์ประกอบของภาษา เช่น คำนาม คำกริยา คำคุณศัพท์ เป็นต้น

The areas worth testing may be categorized into three groups, which are:

1) the four skills for communication, namely listening, speaking, reading, and writing;

2) the areas of the language learned, such as grammar and language usage, vocabulary, and phonology;

3) the language elements, such as nouns, verbs, and adjectives.

ในการวัดผล เป็นหน้าที่ของผู้สร้างข้อสอบ ที่จะประเมินว่าทักษะใดที่จะมีความสำคัญมากเพียงใดในแต่ละช่วงพัฒนาการของผู้เรียน แล้วจึงสร้างเครื่องมือที่แม่นยำและเหมาะสมที่จะวัดความสำเร็จของผู้เรียนที่ได้เรียนรู้และพัฒนาทักษะเหล่านั้นขึ้นมา อย่างเช่นกรณีที่ต้องการวัดด้านทางภาษาที่เรียนแยกออกมา ก็อาจจะมีข้อพิจารณาได้ดังนี้

- ด้านไวยากรณ์และการใช้ภาษา อาจวัดความสามารถผู้เรียนว่าจะสามารถแยกแยะโครงสร้างไวยากรณ์ที่เหมาะสม และการแต่งประโยคหรือโครงสร้างได้ตามที่ต้องการเพื่อการสื่อสารที่ถูกต้องหรือไม่
- ด้านคำศัพท์ อาจวัดความรู้ผู้เรียนว่ารู้ความหมายของคำ รูปแบบที่ใช้กันเป็นปกติของคำ (patterns) และคำปรากฏร่วม (collocations) หรือไม่ หรืออาจวัดความรู้คำศัพท์พร้อมใช้ (active vocabulary) และคำศัพท์รู้จัก (passive vocabulary) เป็นต้น
- ด้านระบบเสียงในภาษา อาจวัดทักษะย่อยได้ 3 ด้าน เช่น (ก) ความสามารถในการแยกแยะและออกเสียงคู่เทียบเสียงที่สำคัญ เช่น shore /ʃɔːr/และ chore /tʃɔːr/ เป็นต้น (ข) ความสามารถในการแยกแยะและใช้รูปแบบปกติของการเน้นเสียง (stress patterns) และ (ค) ความสามารถในการฟังและออกเสียงถ้อยคำสูงต่ำในประโยคได้ (rise and fall patterns)

In assessment, it is the teacher's or test writer's task to assess the relative importance of the skills at the various levels and to devise an accurate means of measuring the student's success in developing these skills. For example, in the case of the language areas learned, testing the following could be for consideration:

- grammar and language usage: the students' ability to recognize appropriate grammatical forms and to manipulate structures;
- vocabulary: the students' knowledge of the meaning of words, and the patterns and collocations in which they occur, or the knowledge of active and passive vocabulary;
- phonology: the students' ability to recognize and pronounce the significant sound contrasts, e.g., shore /ʃɔːr/ and chore /tʃɔːr/, ability to recognize and use the stress patterns, and ability to listen to and produce the rising and falling intonation patterns.

อย่างไรก็ดี ครูผู้สอนหรือผู้เขียนข้อสอบควรมุ่งเน้นไปที่ชนิดข้อสอบ (item type) ที่เกี่ยวข้องกับความสามารถในการใช้ภาษาเพื่อการสื่อสารในชีวิตจริงเป็นหลักด้วย โดยเฉพาะอย่างยิ่งการสื่อสารด้วยการฟังพูด โดยรูปแบบที่วัดความสามารถในทักษะทั้ง 4 ด้านอาจใช้ชนิดข้อสอบ เช่น

- ด้านการฟัง ควรเป็นรูปแบบเพื่อวัดความเข้าใจ เช่น ฟังข้อความ ฟังบทสนทนา ฟังการบรรยาย (lectures)
- ด้านการพูด มักมาในรูปแบบการสัมภาษณ์ การเล่นบทบาทสมมุติ การให้บรรยายรูปหรือการแก้ปัญหา ที่มักต้องทำเป็นงานคู่หรืองานกลุ่ม
- ด้านการอ่านเพื่อความเข้าใจ ข้อคำถามมักทดสอบความสามารถของผู้เรียนในการทำความเข้าใจใจความสำคัญ และสามารถดึงข้อมูลสำคัญออกมาจากเรื่องที่อ่านได้
- ด้านการเขียน มักมาในรูปแบบของการเขียนคำสั่งให้ปฏิบัติตาม การเขียนจดหมาย การเขียนรายงาน การเขียนข้อความ การเขียนบันทึก หรือการเขียนเล่าเหตุการณ์ที่เกิดขึ้นในอดีต

However, for communicative English class, the teachers or the test writers should focus on the item types that are related to language use for real-life communication, especially for oral communication. The forms of these item types may be as follows:

- listening: the item types should be for comprehension, e.g., listening to short utterances, dialogs, or talks and lectures;
- speaking: the item types should be in the form of an interview, a role-play, a picture description, or a problem-solving task that usually requires pair work or group work;
- reading: the item types should be questions for the students to get the gist of a text, and to extract key information out of a reading;
- writing: the item types should be in the form of instructions, letters, reports, messages, memos, and accounts of past events,

1.4 Exercise

1. Which does **NOT** involve evaluation?

A. assignment of code numbers to school subjects

B. qualitative description of student performance

C. ranking of students for grading

D. use of an achievement test for student progress

2. Which aspect is **NOT** in Messick's (1989) validity model?

A. communication

B. construct

C. utility

D. value implications

3. Which statement is true?

A. Only adequacy of evidence and theoretical rationales is enough for validity evaluation.

B. Score meaning underpins all validity investigations.

C. Suitability of test utility is the foundation of validity.

D. We can explore social consequences of a test without construct validity.

4. Which choice is directly related to English for communication?

A. adjectives

B. collocations

C. grammar

D. lectures

5. Which of the following test tasks may NOT be found focused on in a communicative English class?

A. choosing correct grammatical forms

B. listening to a dialog

C. performing a picture description

D. writing an email

Answer key: 1. A. 2. A. 3. B. 4. D. 5. A.

บทที่ 2 หลักการของการทดสอบและการวัดผลทางภาษา

Chapter 2 Basic Principles in Language Testing and Assessment

ในบทที่ 1 ได้กล่าวถึงแนวคิดของการทดสอบและการวัดผลทางภาษา เน้นไปที่ความแตกต่างระหว่างการทดสอบ การวัด และการประเมิน ทฤษฎีหลักด้านความตรง และความสำคัญของการทดสอบทางภาษา ในบทนี้จะได้กล่าวถึงหลักการของการทดสอบและการวัดผลทางภาษา โดยมีลำดับการนำเสนอได้แก่

2.1 ความสามารถใช้งานได้จริง

2.2 ความเที่ยง

2.3 ความตรง

2.4 ความสมจริง/การเป็นตามสภาพจริง

2.5 อิทธิพลย้อนกลับ

2.6 Exercise (แบบฝึกหัดท้ายบท)

In Chapter 1, principles in language testing and assessment have been dealt with. The emphasis is on the distinctions of key terms, a validity theoretical model, and the importance of language testing. In this chapter, basic concepts in language testing and assessment will be dealt with. The organization of the chapter is as follows:

2.1 Practicality

2.2 Reliability

2.3 Validity

2.4 Authenticity

2.5 Washback effect

2.6 Exercise

2.1 ความสามารถใช้งานได้จริง (Practicality)

ความสามารถใช้งานได้จริงคือ การที่แบบสอบหรือการวัดผลอื่นใดสามารถจัดทำ จัดการเก็บข้อมูล จัดสอบ ได้อย่างสำเร็จลุล่วง (Brown & Abeywickrama 2010) ความสามารถใช้งานได้จริงยังหมายรวมไปถึงการให้คะแนนคำตอบที่ได้มาจากเครื่องมือวัดด้วย เช่น แบบสอบอาจจะใช้เก็บข้อมูลมาได้เป็นอย่างดี แต่ข้อมูลจากผู้สอบมีจำนวนมาก ไม่สามารถตรวจให้คะแนนได้ทันตามกำหนดเวลา ก็อาจกล่าวได้ว่าการจัดสอบนั้นบกพร่องในด้านการใช้งานได้จริงอยู่นั่นเอง จึงอาจสรุปได้ว่า การใช้งานได้จริงหมายรวมไปถึง ค่าใช้จ่าย ระยะเวลาที่ใช้ในการสร้างแบบสอบขึ้นมาและนำมาใช้ ความยากง่ายในการตีความและการรายงานผล แบบสอบที่ใช้งานได้จริงจึงควร

1) อยู่ในข้อจำกัดเชิงงบประมาณ

2) สามารถทำสำเร็จลุล่วงได้โดยผู้สอบ ภายในกรอบระยะเวลาที่เหมาะสม

3) มีคำสั่ง/คำแนะนำชัดเจนในการนำมาใช้

4) ใช้ทรัพยากรบุคคลที่มีอยู่อย่างเหมาะสม

5) ไม่ใช้ทรัพยากรการจัดสอบเกินกว่าที่มี

6) พิจารณาทั้งเวลาและความทุ่มเทที่เกี่ยวข้อง ทั้งในการออกแบบและการให้คะแนน

Practicality is when a test or other forms of assessment can be made and administered successfully (Brown & Abeywickrama 2010). Practicality also entails when scores are obtained from the measures. For example, a test may be used well for collecting item responses, but it has a lot of parts that need to be scored manually. The scoring, therefore, cannot be completed within the time limit, and so the test may be considered impractical when it comes to scoring. It may thus be said that practicality normally covers the aspects of the budget, time, and efforts in making, administration, scoring, interpretation, and reporting of the results. A practical test thus:

1) does not exceed the budget limits;

2) can be completed by the test takers with the time limit;

3) has clear directions for the test organizer to administer;

4) uses human resources appropriately;

5) stays within the limit of material resources;

6) deals with both time and effort appropriately for design and scoring.

ตัวอย่างต่อไปนี้เป็นสถานการณ์ที่แสดงให้เห็นถึงความไม่สามารถใช้งานได้จริงในด้านต่างๆ เช่น แบบสอบวัดสมิทธิภาพฉบับหนึ่ง ผู้สอบต้องใช้เวลาถึง 5 ชั่วโมงในการทำให้แล้วเสร็จ ก็อาจจัดได้ว่าเป็นแบบสอบที่ไม่เหมาะใช้งานจริง เหตุเพราะใช้เวลามากเกินไปในการที่จะบรรลุวัตถุประสงค์ อีกหนึ่งตัวอย่างเช่น แบบสอบอีกฉบับหนึ่งต้องใช้ผู้คุมสอบตัวต่อตัวต่อผู้สอบแต่ละคน ก็จัดว่าไม่เหมาะสมที่จะนำมาใช้งานจริง หากมีผู้สอบหลายร้อยคน แต่มีผู้คุมสอบเพียงไม่กี่คน อีกหนึ่งตัวอย่างเช่น แบบสอบฉบับหนึ่งใช้เวลาไม่กี่นาทีผู้สอบก็ทำเสร็จเรียบร้อย แต่ผู้ตรวจข้อสอบกลับต้องใช้เวลาหลายชั่วโมงในการตรวจคำตอบของผู้สอบแต่ละคน ก็ไม่จัดว่าเหมาะกับบริบทห้องเรียนภาษาโดยทั่วไป อีกหนึ่งตัวอย่าง แบบสอบฉบับหนึ่งสามารถตรวจโดยใช้คอมพิวเตอร์เท่านั้น แต่หากการสอบเกิดขึ้นห่างออกไปจากคอมพิวเตอร์ที่ใกล้ที่สุดที่จะใช้ตรวจได้หลายพันกิโลเมตร แบบสอบเช่นนี้ก็อาจจัดได้ว่าไม่เหมาะแก่การใช้งานจริงเช่นกัน

The following examples are situations that exemplify impracticality in various aspects of testing and administration. The first example is when a proficiency test is administered and it has to take the test takers up to five hours to complete. The test may be considered impractical because the test takes too much time to achieve its goal of proficiency testing. Another example is when a test needs one proctor for each test taker. The test may be considered impractical if there are several hundreds of test takers, but there are only a handful of proctors. Another example is when a test takes the test takers minutes to complete, but it takes the teachers several hours to mark the answers of each test taker. The test, thus, is considered impractical for most classroom settings. The last example is when there is a test that has to be marked by a specific computer only. But the examination takes place hundreds of kilometers away from the nearest computer. Such a test is also considered impractical for real use.

2.2 ความเที่ยง (Reliability)

แบบสอบที่มีความเที่ยงคือ แบบสอบที่มีความคงเส้นคงวาและเชื่อถือได้ (Brown & Abeywickrama 2010) ถ้าเราให้ผู้สอบคนเดียวกันทำข้อสอบฉบับเดิมสองรอบ ผลสอบจากข้อสอบก็จะมีความใกล้เคียงกัน แบบสอบที่มีความเที่ยงจึงควร

1) มีความคงเส้นคงวาในเรื่องสภาพการณ์ ระหว่างการจัดสอบตั้งแต่สองครั้งหรือมากกว่า

2) มีแนวทางปฏิบัติชัดเจนในเรื่องการตรวจให้คะแนน หรือการประเมินผล

3) มีเกณฑ์การให้คะแนนที่ชัดเจน

4) มีเกณฑ์การให้คะแนนเหมาะสมสำหรับให้ผู้ตรวจให้คะแนนหรือผู้ประเมินใช้

5) มีตัวข้อสอบหรือชิ้นงานที่ไม่กำกวมสำหรับผู้สอบ

A reliable test is a test that has consistency and dependability in their administration and scoring (Brown & Abeywickrama 2010). If we let a test taker do a test twice, the results from the two administrations should be similar. A reliable test thus:

1) is consistent across two or more administrations;

2) has a clear guideline for how to score or evaluate;

3) has a clear rubric;

4) has a rubric that is suitable for a scorer to use;

5) has clear items or test tasks, which are unambiguous to the test takers.

ทั้งนี้ การพิจารณาความเที่ยงอาจแบ่งออกได้เป็น 4 ด้านที่มีผลต่อความเที่ยง ได้แก่ ด้านผู้สอบ ด้านผู้ตรวจให้คะแนน ด้านการจัดสอบ และด้านข้อสอบ ดังนี้

The factors that could affect reliability may be divided into four areas, namely the test takers, the scorers or raters, the test administration, and the test itself.

<u>ด้านผู้สอบ</u>

ปัจจัยด้านผู้สอบที่มีผลต่อความเที่ยงอย่างเช่น ความเจ็บป่วยที่เกิดขึ้นชั่วคราว ความเหนื่อยล้าของผู้สอบ ความประหม่าวิตกกังวล หรือปัจจัยทางด้านร่างกายด้านจิตใจอื่นๆ ที่มีผลทำให้คะแนนที่วัดได้สังเกตได้แปรปรวนและคลาดเคลื่อนห่างออกจากคะแนนที่แท้จริง (true score) คะแนนที่แท้จริงคือคะแนนที่สะท้อนความสามารถหรือสามัตถิยะที่กำลังวัดผ่านเครื่องมือวัดได้อย่างเที่ยงตรง ปัจจัยที่ส่งผลต่อความเที่ยงนี้รวมไปถึงความฉลาดในการทำข้อสอบ (test-wiseness) และกลวิธีการทำข้อสอบด้วย (test-taking strategies)

Test-taker Reliability

Test-taker factors that can have an impact on the test reliability include temporary illness, the test-takers' fatigue, nervousness, or other physical or mental issues. These issues could affect the measurement of the true score and cause the observed score to sway further off the true score. The true score is the score that is a function of the true ability or competence that is being assessed in a test. Test-taker factors also encompass test-wiseness and test-taking strategies that the test takers use while doing the test.

ด้านผู้ตรวจข้อสอบ

ปัจจัยด้านผู้ตรวจข้อสอบหรือผู้ประเมินผล ที่มีผลต่อความเที่ยง มักเกิดจากความผิดพลาดของมนุษย์ (human error) ความเป็นอัตนัย (subjectivity) หรืออคติ (bias) ที่เข้ามามีอิทธิพลในขั้นตอนของการตรวจให้คะแนน ปัจจัยด้านผู้ตรวจข้อสอบแบ่งออกได้เป็นสองมิติ ได้แก่ ความเที่ยงระหว่างผู้ประเมิน (inter-rater reliability) และความเที่ยงภายในของผู้ประเมิน (intra-rater reliability)

Rater Reliability

Rater factors that can have an impact on the test reliability include human error, subjectivity, and bias. These issues could affect the scoring. Rater factors can be divided into two dimensions: inter-rater reliability and intra-rater reliability.

ความเที่ยงระหว่างผู้ประเมิน เกิดขึ้นเมื่อผู้ประเมินตั้งแต่สองคนขึ้นไปตรวจให้คะแนนมีความคงเส้นคงวาสอดคล้องกัน ความไม่สอดคล้องกันมักเกิดขึ้นเมื่อมีการไม่ตรวจให้คะแนนตามเกณฑ์ การ

ขาดประสบการณ์ การขาดความตั้งใจ หรือแม้แต่อคติ ในการคำนวณความเที่ยงระหว่างผู้ประเมิน นิยมใช้สถิติแคปปา ซึ่งวิธีคำนวณสามารถดูรายละเอียดได้ที่วิดีโอลิงก์:

https://www.youtube.com/watch?v=DfNo32nL_fo

Inter-rater reliability is when two or more raters grade assignments consistently. Lack of inter-rater reliability normally occurs because of failure to adhere to the scoring criteria, inexperience, inattention, or even bias. In calculating inter-rater reliability, Cohen's kappa is often used. The following link shows how to calculate a kappa:

https://www.youtube.com/watch?v=DfNo32nL_fo

ความเที่ยงภายในผู้ประเมิน มักเป็นสถานการณ์ที่เกี่ยวข้องกับผู้สอนในชั้นเรียนทั่วไป ความย่อหย่อนเกี่ยวกับความเที่ยงภายในผู้ประเมินอาจเกิดได้จากเกณฑ์การให้คะแนนที่ไม่ชัดเจน ความเหนื่อยล้า อคติเกี่ยวกับผู้เรียนที่ดีและผู้เรียนที่ไม่ดี หรือแม้แต่ความสะเพร่าเลินเล่อ

Intra-rater reliability is deeply associated with classroom teachers. Violation of this type of reliability could occur because of unclear scoring criteria, fatigue, bias against or in favor of bad and good students, or even carelessness.

ด้านการจัดสอบ

ปัจจัยด้านการจัดสอบ ที่มีผลต่อความเที่ยง เช่น สถานที่สอบ คุณภาพของสำเนาข้อสอบ แสงสว่างในห้องสอบ อุณหภูมิ หรือแม้แต่สภาพของโต๊ะหรือเก้าอี้ที่ใช้ในการสอบ ตัวอย่างเช่น หากจัดสอบการฟังเพื่อความเข้าใจ แต่ห้องสอบอยู่ติดถนนใหญ่ที่มีเสียงดัง ก็อาจมีผลทำให้ผู้สอบทำข้อสอบการฟังคลาดเคลื่อนไปจากความสามารถที่แท้จริงได้

Test Administration Reliability

Test administration reliability is when the conditions allow a consistent and reliable collection of test and item responses. The conditions include the venues, photocopying quality of the examination papers, the lights and temperatures of the examination rooms, or even the desks and chairs. An example is when a listening comprehension test is administered in the examination room that is next to a crowded

street. Noises from the street could affect reliability of the collection of test and item responses that would otherwise reflect the students' true ability.

ด้านข้อสอบ

ปัจจัยด้านข้อสอบที่มีผลต่อความเที่ยง ตัวอย่างเช่น ลักษณะของตัวข้อสอบเองอาจก่อให้เกิดความคลาดเคลื่อนของการวัด (measurement error) ข้อสอบแบบเลือกตอบต้องออกแบบอย่างระมัดระวัง เช่น ข้อสอบต้องมีค่าความยากพอเหมาะและกระจายตัวอย่างเหมาะสม ตัวลวงต้องออกแบบไว้ดี เพื่อให้ข้อสอบมีค่าความคลาดเคลื่อนของการวัดต่ำ ส่วนในบริบทห้องเรียน ความไม่เที่ยงของแบบสอบอาจเกิดขึ้นได้จากหลายสาเหตุ ซึ่งรวมไปถึงอคติของผู้ประเมิน ดังที่ได้กล่าวไว้ข้างต้น นอกจากนี้ ข้อสอบอัตนัยที่มีคำถามปลายเปิด เช่น การเขียนเรียงความ ที่ต้องใช้ดุลพินิจของครูผู้สอนในการระบุว่าคำตอบใดถูก คำตอบใดผิด ข้อสอบอัตนัยเช่นนี้ ก็อาจมีความคลาดเคลื่อนของการวัด ในทางกลับกัน ข้อสอบปรนัยที่มีคำเฉลยคำตอบอยู่แล้วตายตัว ก็อาจเพิ่มความเที่ยงได้

Test Reliability

Test factors can also have an impact on reliability. The characteristics of the very test may cause measurement error, and so multiple-choice tests, as an example, have to be carefully designed. The difficulty level of the examination has to be appropriate and distributed evenly. The distractors have to be carefully chosen too, so that the measurement error would be minimal. In the classroom contexts, test unreliability may be caused by several factors, including rater bias as discussed earlier. Moreover, open-ended, constructed-response tests such as essay writing, which are called subjective tests and have to rely on the teachers' judgment, would experience measurement error. By contrast, selected-response tests, which are called objective tests, could increase test reliability.

นอกจากนี้ ความเที่ยงของข้อสอบยังถูกบั่นทอนด้วยข้อสอบที่เขียนขึ้นมาอย่างแย่ๆ เช่น ข้อสอบที่โจทย์มีความกำกวม หรือข้อสอบที่มีคำตอบที่ถูกต้องมากกว่าหนึ่งคำตอบ นอกจากนี้ ข้อสอบที่มีจำนวนข้อมากเกินไป ก็มักจะทำให้ผู้เรียนรู้สึกล้าเมื่อทำข้อสอบไปถึงข้อหลังๆ และอาจจะเร่งๆ ตอบข้อสอบไปอย่างผิดๆ ถูกๆ และประเด็นสุดท้ายก็คือ ข้อสอบที่มีการจับเวลาก็อาจไม่เหมาะกับผู้สอบที่ไม่ชอบหรือไม่

ถนัดกับการเร่งทำข้อสอบแข่งกับเวลา การทำข้อสอบของผู้เรียนกลุ่มนี้จึงอาจเพิ่มความคลาดเคลื่อนของการวัด

In addition, test reliability can be affected by poorly written test items, e.g., those with ambiguous task content, those with more than one correct answer. Moreover, an unusually long test may make the test takers tired and fatigued by the time they reach later items. They may guess blindly, accordingly. Another example is when the tests are timed, and some test takers do not perform well in such a situation. Their doing of the tests would just create more measurement error.

2.3 ความตรง (Validity)

ในบทที่ 1 (หน้า 5) ได้พูดถึงโมเดลด้านความตรงของ Messick (1989) ในฐานะโมเดลทางทฤษฎีที่ครอบคลุมทุกมิติของหลักการวัดและประเมินผลทางภาษา ในบทนี้จะกล่าวถึงแนวคิดเรื่องความตรงเมื่อนำมาสู่ภาคปฏิบัติ โดยเฉพาะในบริบทห้องเรียน

In Chapter 1 (page 5) a validity model by Messick (1989) is discussed as a theoretical model that covers all aspects of test-score interpretation and use in the principles of language testing and assessment. In this chapter, practical issues of validity, especially those related to classroom settings, will be dealt with.

ความตรง (validity) ในคำนิยามทั่วไป มักแปลกันว่า เมื่อแบบสอบวัดสิ่งที่ต้องการวัด ถือว่าแบบสอบมีความตรง อย่างไรก็ดี ในทางวิชาการ ความตรงหมายถึงขอบเขตที่การตีความผลของการวัดมีความเหมาะสม มีความหมายและมีประโยชน์สำหรับวัตถุประสงค์ในการวัดผล จึงอาจสรุปลักษณะความตรงได้ดังต่อไปนี้ (Brown & Abeywickrama 2010) ว่า แบบสอบที่มีความตรงควร

1) วัดสิ่งที่เสนอว่าวัด

2) ไม่วัดตัวแปรที่ไม่เกี่ยวข้อง

3) ยึดโยงอยู่บนหลักฐานเชิงประจักษ์ เช่น การแสดงออกของผู้เรียน ให้มากที่สุดเท่าที่จะทำได้

4) ใช้วัดการแสดงออกที่สุ่มตัวอย่างจากเครื่องมือเกณฑ์ (the test's criterion measure)

5) ให้ข้อมูลที่มีประโยชน์และมีความหมายเกี่ยวกับความสามารถของผู้สอบ

6) มีหลักการและเหตุผลทางทฤษฎี หรือข้อโต้แย้งทางทฤษฎีรองรับ

A general definition of *validity* usually deals with when a test can measure what it is purported to measure. But in academic terms, validity refers to the extent in which the inferences based on assessment results are appropriate, meaningful, and useful for the purposes of assessment. We may thus conclude the characteristics of a valid test (Brown & Abeywickrama 2010), in that it:

1) measures what it is supposed to measure

2) does not measure irrelevant variables

3) relies on empirical evidence, as much as possible, e.g., student's performance

4) involves performance that is sampled through the test's criterion measure

5) offers useful and meaningful information about the test taker's ability

6) is supported by an argument or a theoretical rationale

ความตรงของแบบสอบเป็นเรื่องเกี่ยวกับระดับ (degree) ว่ามากหรือน้อย มิใช่เรื่องว่ามีหรือไม่มี (not all or none) ตัวอย่างการพิจารณาความตรงของแบบสอบ อย่างเช่น หากเป็นแบบสอบที่บอกว่าวัดความสามารถในการเขียนภาษาอังกฤษ และครูผู้สอนก็ให้ผู้เรียนเขียนคำภาษาอังกฤษให้มากที่สุดเท่าที่จะเขียนได้ภายใน 15 นาที จากนั้นก็นับจำนวนคำว่าเขียนได้ทั้งหมดกี่คำเพื่อคิดเป็นคะแนน แบบสอบเช่นนี้ง่ายที่จะจัดสอบ จัดว่ามีความสามารถในการใช้งานได้จริงสูง (highly practical) และการให้คะแนนก็มีความเที่ยงสูงด้วย (highly reliable scoring) แต่อย่างไรก็ดี แบบสอบเช่นนี้จะไม่จัดเป็นข้อสอบที่วัดความสามารถในการเขียนภาษาอังกฤษที่มีความตรงสูง เหตุเพราะว่าไม่มีการวัดด้านการเขียนเพื่อสื่อสารให้รู้เรื่อง (lack of comprehensibility) ไม่ได้วัดความสามารถในการเรียงร้อยถ้อยคำเป็นข้อความให้เห็น (lack of rhetorical discourse elements) หรือแม้แต่การจัดเรียงความคิดก็ไม่ได้วัด (lack of organization of ideas)

Validity is a matter of degree, not a matter of all or none. An example for conceptualizing validity is a test for writing ability. The teacher asks the students to write as

many words in English as possible in 15 minutes. Then the number of words is counted for the scoring. This kind of test is easy to administer, deemed to be highly practical. Given that the scoring is very straightforward, the test is highly reliable, too. However, such a test like this will not be considered a valid test for writing ability. The reasons are that it has no assessment of writing for comprehensible communication, no assessment of rhetorical discouse elements, and no assessment of organization of ideas.

ในการหาหลักฐานมารองรับการตีความเกี่ยวกับความตรง เราอาจดูขอบเขตที่การแสดงออกในการสอบว่าสอดคล้องกับการแสดงออกในรายวิชาหรือบทเรียนที่กำลังทำการทดสอบอยู่ ในบางกรณีเราอาจพิจารณาดูว่า แบบสอบระบุได้ดีเพียงใดว่าผู้เรียนได้บรรลุถึงเป้าประสงค์ อีกรูปแบบของหลักฐานที่นิยมใช้กันทั่วไปคือค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างมาตรวัดที่ไม่เกี่ยวข้องกับแบบสอบที่เรากำลังทดสอบกับตัวแบบสอบ อีกหนึ่งรูปแบบของหลักฐานเกี่ยวกับความตรงได้แก่ผลกระทบของแบบสอบต่อไปนี้คือหลักฐานเกี่ยวกับความตรง 4 ประเภท

In searching for evidence for validity inquiries, we may investigate the extent to which the test performance agrees with the performance in the course or the unit that is being tested. In some other cases, we may investigate how well a test determines that the students have attained the course objectives. Another form of evidence that is widely used is statistical correlation with an independent measure. Another form of validity evidence is from test consequences. The following are four types of validity evidence.

<u>หลักฐานเกี่ยวกับเนื้อหา</u>

ถ้าเนื้อหาในแบบสอบดึงมาจากเนื้อหาที่เราจะทำการสรุปเกี่ยวกับผู้เรียน และถ้าแบบสอบก็ให้ผู้เรียนแสดงพฤติกรรมเดียวกันกับที่ผู้เรียนแสดงเกี่ยวกับเนื้อหานั้น จัดว่าแบบสอบนั้นมีความตรงเชิงเนื้อหา (content-related validity) หลักฐานเกี่ยวกับความตรงเชิงเนื้อหาโดยปกติสามารถสังเกตได้ไม่ยาก หากเราสามารถนิยามผลสัมฤทธิ์ที่เรามองหาได้ชัด ตัวอย่างเช่น หากเราต้องการวัดความสามารถผู้เรียนในการพูดภาษาอังกฤษในบริบทการสนทนาทั่วไป แต่เราให้ผู้เรียนตอบคำถามแบบเลือกตอบในข้อสอบข้อเขียน (paper-and-pencil questions) เกี่ยวกับไวยากรณ์ ข้อสอบเช่นนี้ก็อาจไม่มีความตรงเชิงเนื้อหามากนัก แต่หากเราจัดสอบให้ผู้เรียนได้พูดอะไรออกมาในสถานการณ์จริงๆ ข้อสอบเช่นนี้ก็จัดว่ามีความตรง

เชิงเนื้อหามากกว่า อีกกรณีหนึ่งก็เช่น หากเรามีวัตถุประสงค์การเรียนรู้อยู่ 10 ประการ แต่หากข้อสอบครอบคลุมแค่เพียงสองวัตถุประสงค์การเรียนรู้ ความตรงเชิงเนื้อหาก็อาจจัดว่าไม่ดี

Content-related evidence

If the content in an exam paper is selected from the subject matter that we wish to draw conclusions about, the test has content-related validity. If a test requires the students to perform the same behavior that they have done when studying a particular content, then the test can be claimed to have content-related validity. Normally, content-related evidence of validity is not difficult to take notice of, if we could clearly define the achievement that we are looking for. For example, we want to assess the students' ability to speak English in general contexts, but we make the students answer paper-and-pencil questions about English grammar. Such a test may not have much content validity. However, if we administer a test where they actually speak English, then the test would have higher content-related validity. Another situation is when we have 10 learning objectives, but the examination paper covers only two objectives. In such a case, the content validity suffers.

อีกมุมมองสำหรับการทำความเข้าใจความตรงเชิงเนื้อหาคือ การพิจารณาความแตกต่างระหว่างการทดสอบโดยตรงกับการทดสอบโดยอ้อม การทดสอบโดยตรงคือการให้ผู้สอบแสดงออก (perform) หรือทำชิ้นงานหนึ่งๆ ที่เป็นเป้าหมายจริงๆ ส่วนการทดสอบโดยอ้อม การทำชิ้นงานอาจจะมีความเกี่ยวข้องกับโดเมนเป้าหมายเพียงบางส่วน ตัวอย่างเช่น เราต้องการทดสอบการเน้นเสียงพยางค์ (syllable stress) และในการทดสอบ เราก็ให้ผู้สอบทำเครื่องหมายพยางค์เน้นเสียงในกระดาษที่มีคำภาษาอังกฤษอยู่ นับเป็นการทดสอบการออกเสียงพยางค์ทางอ้อม กรณีเช่นนี้ หากเราต้องการทดสอบโดยตรง เราก็จำต้องให้ผู้เรียนออกเสียงคำที่เป็นเป้าหมายออกมาจริงๆ ดังนั้น หัวใจสำคัญในการสร้างความตรงเชิงเนื้อหาสำหรับการวัดผลในชั้นเรียนก็คือการวัดการแสดงออกโดยตรงนั่นเอง

Another way of comprehending content validity is to consider a difference between direct testing and indirect testing. Direct testing is to make the test takers actually do the target task. Indirect testing is to involve the test takers in doing a task that is related only in some way to the target task. For example, we want to test the oral production of

syllable stress. In direct testing, we would ask the test takers to actually pronounce the words and take notice of the syllables that are stressed. In indirect testing, the test takers may mark stressed syllables of given words on a sheet of paper. Accordingly, the rule of thumb for content validity is the direct assessment of the performance intended for test-score interpretation and use.

หลักฐานเครื่องมือเกณฑ์

นอกเหนือจากหลักฐานเกี่ยวกับเนื้อหา อีกรูปแบบของหลักฐานที่ใช้สนับสนุนความตรง ได้แก่หลักฐานเกี่ยวกับเครื่องมือเกณฑ์ หลักฐานเครื่องมือเกณฑ์หมายถึงการที่คะแนนหรือผลลัพธ์ที่ได้จากการวัดในแบบสอบอันหนึ่ง มีความสอดคล้องกับคะแนนหรือผลลัพธ์ที่ได้จากอีกเครื่องมือหนึ่ง (criterion measure) ตัวอย่างเช่น ผู้เรียนทำแบบสอบที่ออกแบบมาเพื่อวัดการเรียนรู้ไวยากรณ์หัวข้อหนึ่งได้คะแนนสูง คะแนนจากแบบสอบนี้จะได้รับการสนับสนุนจากการสังเกตเห็นพฤติกรรมการใช้ไวยากรณ์หัวข้อนี้ในภายหลัง หรืออาจจะเป็นแบบสอบอื่นในหัวข้อไวยากรณ์หัวข้อเดียวกัน พฤติกรรมที่สังเกตเห็นและผลการแสดงออกในแบบสอบอื่นจัดเป็นเครื่องมือเกณฑ์

Criterion-related evidence

In addition to content-related validity evidence, another form of evidence that can be used in support of validity is criterion-related evidence. Criterion-related validity evidence is when the scores or results of one assessment agrees with the scores or results of another measure (criterion measure). For example, the students do a test on one grammar point. The scores of this test would gain criterion-related evidence from an observed behavior of the grammar point in real use, or the scores from another test on the same grammar point. The observed behavior and performance in another test are considered to be criterion measures.

หลักฐานเครื่องมือเกณฑ์มักแบ่งออกได้เป็นสองประเภท ประเภทแรกคือความตรงตามสภาพ (concurrent validity) และความตรงพยากรณ์ (predictive validity) ความตรงตามสภาพหมายถึงการที่ผลจากแบบสอบฉบับหนึ่งสอดคล้องกับผลจากอีกเครื่องมือหนึ่งในเวลาเดียวกัน ตัวอย่างเช่น คะแนนข้อสอบปลายภาควิชาภาษาอังกฤษมีความสอดคล้องกับสมิทธิภาพจริงของผู้เรียน เป็นต้น ส่วนความตรง

พยากรณ์คือการที่ผลจากแบบสอบฉบับหนึ่งสามารถคาดทำนายผลในอีกเครื่องมือหนึ่งในอนาคต เช่น ข้อสอบวัดระดับ (placement test) สามารถทำนายแนวโน้มที่ผู้เรียนจะประสบความสำเร็จในการเรียนรายวิชาที่เหมาะสมกับระดับของตน เป็นต้น

Criterion-related evidence can be categorized into two types: concurrent validity and predictive validity. Concurrent validity is when the results of a test are supported by another concurrent measure. For example, the results of a final English examination might accord with the students' English proficiency. Predictive validity is when the results of a test can predict the results of another measure that is in the future. For example, a placement test may be able to predict the likelihood of the students' success in the levels that they have been assigned to.

หลักฐานเชิงความหมายคะแนน

หลักฐานประเภทที่สามที่สามารถสนับสนุนความตรงได้แก่หลักฐานเชิงความหมายคะแนน (construct validity) ความหมายคะแนนคือทฤษฎี ข้อสมมุติฐาน หรือโมเดลใดๆ ที่พยายามอธิบายปรากฏการณ์ที่สามารถสังเกตเห็นได้ ความหมายคะแนนอาจจะถูกวัดโดยตรงหรืออาจจะไม่ถูกวัดโดยตรงก็ได้ การพิสูจน์ทราบบ่อยครั้งจึงต้องใช้ข้อมูลเชิงอนุมาน (inferential data) ตัวอย่างความหมายคะแนนเช่น สมิทธิภาพ (proficiency) สามัตถิยะในการสื่อสาร (communicative competence) ความคล่องแคล่ว (fluency) จัดเป็นความหมายคะแนนในทางภาษาศาสตร์ จะสังเกตได้ว่า แทบทุกประเด็นในการเรียนการสอนภาษาล้วนมีความหมายคะแนนเข้ามาเกี่ยวข้อง (ดูหัวข้อที่ 1.2 หน้า 5 ประกอบ) ตัวอย่างเช่นหากเราต้องการวัดผลความคล่องแคล่วด้านการพูดภาษาอังกฤษ แบบสอบควรจะต้องวัดหลากหลายองค์ประกอบของความคล่องแคล่ว เช่น ความเร็ว จังหวะจะโคน การหยุดพักเป็นช่วงๆ การพูดโดยไม่ลังเล และองค์ประกอบอื่นภายใต้ความหมายคะแนนว่าด้วยเรื่องความคล่องแคล่วในการพูด เป็นต้น แบบสอบต่างๆ จึงเป็นเหมือนคำนิยามความหมายเชิงปฏิบัติการ เพราะเป็นตัวต่อประกอบกันขึ้นมาเป็นความหมายคะแนน

Construct-related evidence

The third type of evidence that can support validity is construct-related evidence. A construct is any theory, hypothesis, or model which attempts to explain the phenomena that can be observed. The construct may or may not be directly observed, and so verifying it often involves inferential data. Examples of linguistic constructs include

proficiency, *communicative competence*, and *fluency*. It is worth highlighting that virtually every issue in language teaching and learning involves score meaning or a construct (see also section 1.2, page 5). For example, we wish to assess fluency in English speaking. The test should thus incorporate various components of fluency, e.g., speed, rhythm, juncture, lack of hesitation, other elements for oral fluency. Accordingly, the tests are like operational definitions because they are components for a construct.

ในบริบทห้องเรียนทั่วไป การวิจัยตรวจสอบความตรงต่อความหมายคะแนนอาจดูเป็นเรื่องยาก แต่กระนั้นก็ตาม การวิจัยแบบไม่เป็นทางการเป็นสิ่งที่ควรทำและสามารถกระทำได้สำหรับแบบสอบในบริบทห้องเรียน เช่น สมมุติว่าเราสร้างแบบทดสอบย่อย (quiz) สำหรับการเขียนคำศัพท์ แบบทดสอบครอบคลุมเนื้อหาของบทเรียนที่ผู้เรียนเพิ่งเรียนไปไม่นาน โดยให้ผู้เรียนเขียนคำนิยามของคำศัพท์ให้ถูกต้อง คำศัพท์ที่เราเลือกมาทำแบบทดสอบอาจจะสุ่มเลือกตัวแทนคำศัพท์ในบทเรียนได้อย่างเพียงพอ แต่หากวัตถุประสงค์การเรียนรู้ด้านคำศัพท์ของบทนั้นเป็นเรื่องการใช้คำศัพท์เพื่อการสื่อสาร แน่นอนว่าการเขียนคำนิยามคำศัพท์ก็ไม่เหมาะสมกับความหมายคะแนนที่ต้องมีการใช้ภาษาเพื่อการสื่อสาร

In classroom settings, construct validation research may seem difficult to do. However, an informal construct validation is worth doing and should be done in classroom settings. For example, we construct a quiz for written vocabulary. The content of the quiz is from the unit that the students have just studied. The students are required to write the definitions of the vocabulary in the quiz. The vocabulary that is chosen for the quiz may have been sampled adequately for the unit. But if the learning objective for lexical use of the unit is communicative vocabulary use, then the writing of vocabulary definition clearly does not fit the construct of communicative vocabulary use.

ในการทดสอบมาตรฐาน ความตรงต่อความหมายคะแนนเป็นประเด็นสำคัญสำหรับการศึกษาวิจัย ข้อจำกัดประการหนึ่งก็คือการที่ไม่สามารถทดสอบเนื้อหาทั้งหมดสำหรับสาขาวิชาใดสาขาวิชาหนึ่งหรือสำหรับทักษะใดทักษะหนึ่งได้ทั้งหมด ทั้งนี้เพราะการทดสอบมาตรฐานต้องยึดเหตุผลด้านความสามารถในการใช้งานได้จริงสาเหตุเพราะงบประมาณจำกัด และการเลือกโดเมนภาษามาใช้สอบก็เลือกได้เพียงจำนวนเล็กน้อยเท่านั้น ตัวอย่างเช่น ข้อสอบมาตรฐานขนาดใหญ่จำนวนมากในสมัยก่อนมักไม่

มีการสอบการแสดงออกทางการพูด (oral production) ทั้งๆ ที่การแสดงออกทางการพูดเป็นด้านที่สำคัญของความสามารถทางภาษา ข้อสอบเหล่านี้มักใช้งานวิจัยค่าสหสัมพันธ์ (correlation) เพื่อแสดงว่าการแสดงออกทางการพูดมีความสัมพันธ์เชิงบวกกับทักษะด้านอื่นๆ ที่วัดในข้อสอบ ในปัจจุบัน ด้วยความก้าวหน้าของการพัฒนาเกณฑ์การประเมินสำหรับการแสดงออกทางการพูด และซอฟต์แวร์ที่ตรวจการพูดได้โดยอัตโนมัติ ข้อสอบมาตรฐานที่วัดสมิทธิภาพทางภาษาทั่วไป จึงมีชิ้นงานการแสดงออกทางการพูดด้วย

In standardized testing, construct validity is an important issue for research. One of the limitations is that standardized tests cannot comprise all the content of a particular field of study. Nor can they assess all aspects of a skill. A reason is that standardized tests have to be practical due to financial limitations. Only a limited number of language domains could be sampled in the testing. For example, a number of standardized tests in the past had no examination part for oral production. This was despite the fact that oral production is an important part of language ability. These examinations were usually justified by correlation research, showing that oral production had positive relationships with other skills measured in the examinations. Nowadays, given the advancement in developing scoring rubrics for oral production, and automated software for speech recognition, standardized tests for general English proficiency have tasks for oral production.

<u>หลักฐานผลกระทบ</u>

นอกเหนือไปจากหลักฐานเกี่ยวกับเนื้อหา หลักฐานเครื่องมือเกณฑ์ และหลักฐานเชิงความหมายคะแนนแล้ว หลักฐานที่สามารถใช้สนับสนุนความตรงได้อีกประเภทหนึ่งคือหลักฐานผลกระทบ (consequential validity evidence) หลักฐานผลกระทบหมายถึง ผลกระทบของข้อสอบทุกรูปแบบ ซึ่งรวมไปถึงข้อพิจารณาเรื่องความแม่นยำในการวัดด้วยเกณฑ์ที่ต้องการ ผลกระทบของข้อสอบต่อการเตรียมตัวของผู้สอบ หรือแม้แต่ผลกระทบทางสังคม ทั้งที่เจตนาและที่ไม่ได้เจตนาของการตีความและการใช้คะแนน

Consequential evidence

Thus far, content-related evidence, criterion-related evidence and construct-related evidence have been discussed. Another type of evidence that can be used for supporting validation of test scores is consequential evidence. Consequential validity entails all types of consequence of a test, e.g., considerations about the accuracy of the intended criteria, test preparation of the test takers, and intended and unintended social consequences of test-score interpretation and use.

คำศัพท์ที่ใช้กันบ่อยในการพูดถึงหลักฐานผลกระทบคือคำว่า impact หรือก็คือผลกระทบ ผลกระทบอาจเกิดได้ทั้งก่อนและหลังการสอบ Bachman & Palmer (1996: 30) แบ่งผลกระทบออกเป็นระดับมหภาคและระดับจุลภาค ตัวอย่างผลกระทบระดับมหภาคเช่น การใช้ข้อสอบมาตรฐานแบบเลือกตอบเข้ามหาวิทยาลัย อาจทำให้ผู้เรียนละทิ้งโอกาสในการเรียนรู้และฝึกฝนทักษะในการส่งสาร (productive skills) ทั้งยังบิดเบือนภาพจำเกี่ยวกับการสอบภาษาต่างประเทศในหมู่ผู้นำผลคะแนนไปใช้ด้วย

A word frequently used in relation to consequential validitiy evidence is *impact*. An impact can occur either before a test administration or after a test administration, or both. Bachman & Palmer (1996: 30) categorize an impact into a macro level and a micro level. An example of a macro-level impact is when we use a multiple-choice standardized test for university entrance. A lot of students may not practice the productive skills for the English examination, and EFL testing would be plagued with a disillusion of non-productive language skills for such a high-stakes testing situation.

ในส่วนผลกระทบระดับจุลภาค คำที่ใช้กันบ่อยคืออิทธิพลย้อนกลับ (washback) ซึ่งจะกล่าวถึงในรายละเอียดในหัวข้อถัดไป (หัวข้อที่ 2.5 อิทธิพลย้อนกลับ)

Regarding the micro level of consequential evidence, the term that is used frequently is washback. Washback will be dealt with specifically in the next section (Section 2.5 Washback).

หลักฐานความตรงเชิงปรากฏ

ความตรงเชิงปรากฏ หมายถึงระดับที่แบบสอบหนึ่งๆ ดูใช้ได้และดูน่าจะวัดความรู้หรือความสามารถที่แบบสอบบอกว่าวัด โดยขึ้นอยู่กับดุลพินิจของตัวผู้สอบ เจ้าหน้าที่ที่ตัดสินใจใช้แบบสอบและผู้สังเกตการณ์อื่นๆ ที่อาจจะยังไม่เชี่ยวชาญมากนัก ในมุมมองของผู้เรียนความตรงเชิงปรากฏคือเมื่อผู้เรียนมองว่าการวัดผลมีความยุติธรรม เกี่ยวข้องและเป็นประโยชน์สำหรับการปรับปรุงพัฒนาการเรียนรู้

Face validity evidence

Face validity is when a test *appears* to work well and *looks* right for measuring the knowledge or abilities that it claims to measure, on the basis of the subjective judgment of the examinees, the test personnel who decide to use the test, and other observers who may not be psychometrically sophisticated. In the viewpoint of the students, face validity is when they consider the assessment to be fair, relevant, and potentially useful for the improvement of the learning.

แม้ว่าความตรงเชิงปรากฏจะมีความน่าสนใจ ความตรงชนิดนี้ในแวดวงวิชาการจัดว่ายังคงเป็นแนวคิดที่ไม่สามารถวัดเชิงประจักษ์ หรือให้เหตุผลเชิงทฤษฎีเกี่ยวกับความตรงได้ ความตรงเชิงปรากฏเป็นเรื่อง “แล้วแต่ใครจะมองว่าเหมาะสม” (eye of the beholder) ว่าผู้สอบหรือผู้นำข้อสอบมาใช้จะมองเครื่องมือที่นำมาใช้อย่างไร ด้วยเหตุผลนี้ ความตรงเชิงปรากฏจึงเป็นแค่เพียงปัจจัยฉายฉวย (superficial) ปัจจัยหนึ่งที่นักวิชาการมองว่าควรนำออกไปจากสารบบ

Face validity might be intuitively appealing, but the scholars still contend that it is a notion which cannot be empirically measured or theoretically justified for validity. It is a factor of the “eye of the beholder,” depending on how the test takers or the test giver would perceive an instrument. For this reason, face validity is a superficial factor that many believe should be banished from the testers’ lexicon.

อย่างไรก็ดี ลักษณะที่แบบสอบปรากฏกลับมีผลจริงๆ ตัวอย่างเช่น หากผู้เรียนรู้สึกว่าแบบสอบไม่ได้วัดสิ่งที่มันควรจะวัด ความรู้สึกเช่นนี้ก็อาจมีผลต่อการแสดงออกในข้อสอบ เกิดเป็นข้อบกพร่อง

ด้านความเที่ยงที่เกิดจากตัวผู้สอบดังที่ได้อภิปรายไว้แล้วในหัวข้อที่ 2.2 ดังนั้น การรับรู้ของผู้เรียนเกี่ยวกับความยุติธรรมของข้อสอบจึงมีความสำคัญต่อการวัดผลในชั้นเรียน เพราะสามารถส่งผลกระทบถึงการแสดงออกของผู้เรียนและความเที่ยงได้ ครูผู้สอนจึงสามารถเพิ่มการรับรู้เกี่ยวกับความยุติธรรมของแบบสอบได้โดยใช้

1) รูปแบบข้อสอบที่มีชิ้นงานที่คุ้นเคย

2) ชิ้นงานที่สามารถทำเสร็จในกรอบระยะเวลาที่กำหนด

3) ข้อสอบที่ชัดเจนและไม่ยุ่งยากซับซ้อน

4) คำสั่งข้อสอบที่ชัดเจน

5) ชิ้นงานที่เคยซ้อมทำมาก่อนในรายวิชาก่อนหน้า

6) ชิ้นงานที่เกี่ยวข้องกับรายวิชาปัจจุบัน (ความตรงเชิงเนื้อหา)

7) ระดับความยากที่พอเหมาะ

However, test appearance does have an effect on the test takers. For example, if they think that the test does not measure what it is supposed to test, then such a feeling may affect their performance in the test, resulting in a student-related unreliability that has been discussed in Section 2.2. Accordingly, student perception of a test's fairness can be important for classroom-based assessment, given that it can have an impact on the students' test performance and reliability. Teachers can thus increase the students' perception of fair tests by using:

1) a well-constructed test format with familiar tasks

2) tasks that can be finished with time limit

3) items that are clear and uncomplicated

4) directions that are clear

5) tasks that have been rehearsed in previous coursework

6) tasks that relate to their coursework (content validity)

7) an appropriate difficulty level.

ประเด็นเรื่องความตรงเชิงปรากฏ โดยเฉพาะเรื่องที่รูปแบบและชิ้นงานข้อสอบต้องเป็นที่คุ้นเคยสำหรับผู้สอบ ย้ำเตือนให้เราตระหนักถึงปัจจัยด้านจิตใจของผู้สอบที่มีผลต่อการแสดงออกในข้อสอบ ผู้สอบอาจจะเขวและมีความประหม่าเพิ่มขึ้น หากเรานำสิ่งที่ไม่คุ้นเคยเข้ามาในการสอบ อาจกล่าวได้ว่าแบบสอบในบริบทห้องเรียนไม่ใช่ที่ที่จะนำชิ้นงานการทดสอบใหม่เข้ามา เพราะจะทำให้เราไม่รู้ว่า ที่ผู้เรียนทำข้อสอบไม่ได้เกิดจากปัจจัยเรื่องชิ้นงานหรือเกิดจากการไม่บรรลุวัตถุประสงค์การเรียนรู้กันแน่

The issues of face validity, especially those of familiar tasks and formats, remind us that the psychological state of the test takers may have an effect on their task performance. The test takers may be distracted and have increased anxiety, if we introduce something new into the examination. It can be said that language classroom tests are not a place for trying something brand-new. The reason is that we would not know if the difficulty they experience is the result of the unfamiliar test tasks, or the non-achievement of the class objectives.

2.4 ความสมจริง/การเป็นตามสภาพจริง (Authenticity)

Bachman & Palmer (1996) นิยามการเป็นตามสภาพจริงไว้ว่า หมายถึงระดับของความสอดคล้องของลักษณะของชิ้นงานทางภาษากับลักษณะของชิ้นงานเป้าหมาย โดยมีประเด็นสำคัญอยู่ที่ชิ้นงาน หรือตัวอย่างทางภาษาต้องสะท้อนการใช้ภาษาในโลกความเป็นจริง ในด้านหนึ่ง การตัดสินว่าชิ้นงานมีความเสมือนจริงเป็นเรื่องอัตนัย แต่ในอีกด้านหนึ่ง การเป็นตามสภาพจริงนับเป็นหนึ่งในแนวคิดสำคัญที่แวดวงการทดสอบทางภาษาให้ความสนใจ สาเหตุหนึ่งอาจเป็นเพราะแบบสอบจำนวนมากชนิดไม่สามารถจำลองสถานการณ์การใช้ภาษาได้ดีมากพอ

Bachman & Palmer (1996) define authenticity as the degree of correspondence between the characteristics of a given language test task and the features of a target language task. The key point is that the test task or the sample of language must reflect the language use in the real world. On the one hand, judging which test tasks are authentic is subjective. Yet, on the other hand, authenticity is a key concept that receives attention widely in language testing. A reason for that could be that a lot of test types cannot adequately simulate real-world tasks.

ในการสอบ การเป็นตามสภาพจริงอาจอยู่ในรูปแบบต่อไปนี้ (Brown & Abeywickrama 2010)

1) มีภาษาที่เป็นธรรมชาติที่สุดเท่าที่จะเป็นไปได้

2) มีข้อสอบที่อิงกับบริบทมากกว่าจะแยกเดี่ยวไม่เกี่ยวข้องกับบริบทใดๆ

3) นำเสนอชิ้นงานที่เลียนแบบโลกความเป็นจริง

4) มีหัวข้อที่มีความหมาย เกี่ยวข้องกับผู้สอบ และน่าสนใจ

5) มีการจัดเรียงหัวข้อเรื่องให้กับข้อสอบ เช่นโดยใช้เรื่องราวหรือตอนของเรื่องต่างๆ

In testing, an authentic test (Brown & Abeywickrama 2010):

1) contains language that is as natural as possible

2) has contextualized items, rather than isolated ones

3) has tasks that replicate real-world tasks

4) has meaningful, relevant, and interesting topics

5) provides some thematic organization to items, e.g., through a story line or episode

ในอดีต ข้อสอบที่แต่งขึ้นมา ไม่น่าสนใจ และไม่เกี่ยวข้องกับตัวผู้สอบถูกนำมาใช้ โดยมองว่าเป็นส่วนที่เลี่ยงไม่ได้ของการสอบ ข้อสอบเหล่านี้มักจะไม่ได้ให้ผู้สอบแสดงทักษะการส่งสารออกมา ด้วยเหตุผลด้านข้อจำกัดของงบประมาณ ในช่วงหลังมานี้ ความเป็นตามสภาพจริงของชิ้นงานข้อสอบได้มีการปรับปรุงให้ดีขึ้น โดยเฉพาะอย่างยิ่งมีการให้ผู้สอบแสดงทักษะส่งสารได้แก่การพูด และการเขียนออกมาด้วยเรื่องที่อ่านก็ถูกเลือกมาจากแหล่งที่ใช้กันในชีวิตจริง แหล่งที่ผู้สอบมีโอกาสได้พบเจอ ส่วนด้านการฟังก็มีการใช้ภาษาที่เกิดขึ้นตามปกติ มีทั้งการลังเลที่จะพูด เสียงพื้นหลัง หรือการพูดแทรก เป็นต้น

In the past, the test items were contrived, uninteresting, and unconnected to the test takers. They were viewed as unavoidable in language testing. Productive skills were also missing for the reason of limited budgets. Recently, authenticity in test tasks has

improved, especially regarding productive skills of speaking and writing. Reading passages are from real-world sources which the test takers are likely to find. As for listening comprehension, natural language is featured, with hesitations, white noise, and interruptions.

2.5 อิทธิพลย้อนกลับ (Washback)

อิทธิพลย้อนกลับคือ ผลกระทบของการทดสอบที่มีต่อการเรียนและการสอน (Brown & Abeywickrama 2010) เป็นด้านย่อยด้านหนึ่งของความตรงผลกระทบ (consequential validity) โดยความแตกต่างสำคัญคือ อิทธิพลย้อนกลับมักจะใช้สื่อถึงบริบทห้องเรียนเป็นหลัก ส่วนความตรงผลกระทบมักกล่าวถึงบริบทที่ใหญ่กว่าห้องเรียนเป็นหลัก อิทธิพลย้อนกลับอาจเป็นได้ทั้งเชิงบวกคือส่งเสริมการเรียนรู้และเชิงลบคือบั่นทอนการเรียนรู้

Washback is an effect that assessment has on teaching and learning (Brown & Abeywickrama 2010). It is part of consequential validity. A key distinction is that the term *washback* is usually used for impact in classroom settings, whereas *consequential validity* usually refers to the impact beyond classroom settings. Washback can be both positive—promotion of learning—and negative—inhibition of learning.

แบบสอบหนึ่งๆ ที่มีอิทธิพลย้อนกลับเชิงบวก

1) มีอิทธิพลเชิงบวกในเรื่องว่าครูจะนำอะไรมาสอนและสอนอย่างไร

2) มีอิทธิพลเชิงบวกในเรื่องว่าผู้เรียนจะเรียนอะไรและเรียนอย่างไร

3) ให้โอกาสผู้เรียนได้เตรียมตัวอย่างเพียงพอ

4) ให้ข้อมูลสะท้อนกลับ (feedback) แก่ผู้เรียน ที่ช่วยปรับปรุงการพัฒนาภาษา

5) มีความเกี่ยวข้องกับการประเมินระหว่างเรียนมากกว่าการประเมินหลังเรียน

6) ช่วยจัดสภาพการณ์สำหรับการแสดงออกระดับสูงสุด (peak performance) ของผู้เรียน

A test that has beneficial washback

1) has a positive effect on what and how the teachers teach

2) has a positive effect on what and how the learners learn

3) gives the learners a chance to prepare adequately

4) offers the learners feedback which enhances their language development

5) is more formative in nature than summative

6) provides conditions for the learner's peak performance

ในบริบทการประเมินผลขนาดใหญ่ อิทธิพบย้อนกลับมักหมายถึงอิทธิพลย้อนกลับที่การสอบมีต่อการสอน โดยเฉพาะในแง่ที่ว่าผู้เรียนจะเตรียมตัวสำหรับการสอบอย่างไร คอร์สกวดวิชาและการสอนเพื่อเตรียมสอบโดยเฉพาะเป็นตัวอย่างของอิทธิพลย้อนกลับที่อาจจะมีทั้งผลดีและผลเสีย ตัวอย่างเช่น การใช้แบบสอบมาตรฐานทั่วโลกในปัจจุบันในฐานะผู้คุมประตูได้ทำให้ผู้เรียนมุ่งแต่จะทำคะแนนให้ได้ระดับที่ต้องการมากกว่ามุ่งพัฒนาภาษา กระนั้นก็อาจมีด้านบวกอยู่บ้าง เช่นว่า ผู้เรียนจำนวนมากในคอร์สกวดวิชามีสามัตถิยะเพิ่มขึ้นในชิ้นงานทางภาษาบางอย่าง (Chapelle, Enright, & Jamieson 2008)

In large-scale assessment, washback often refers to the effects that the tests have on instruction. This is especially regarding how the learners may prepare for the tests. Cram courses and teaching to the test are good examples of how examinations can have both positive and negative impact. For example, the current use of standardized tests worldwide functions as the gate keeper to universities. They make the students focus on scoring at a certain level, rather than actual language development. A positive side has been found, though. A lot of students in cram courses are found to have a higher competence in some language tasks (Chapelle, Enright, & Jamieson 2008).

ในบริบทการประเมินผลในห้องเรียน อิทธิพลย้อนกลับสามารถส่งผลเชิงบวกได้หลายประการ ตั้งแต่การเตรียมบทเรียนและการทบทวนสำหรับการสอบ จนถึงการเรียนรู้ที่เพิ่มพูนจากการได้รับข้อมูลสะท้อนกลับ (feedback) เกี่ยวกับผลงานของตนเอง โดยครูผู้สอนสามารถให้ข้อมูลที่จะส่งผลกลับไปยังผู้เรียนในรูปของผลการประเมินด้านที่ถนัดและไม่ถนัด อิทธิพลย้อนกลับยังหมายรวมไปถึงผลกระทบของการประเมินต่อการเรียนการสอนก่อนที่จะมีการประเมินด้วย เช่นการเตรียมตัวเพื่อรับการประเมิน ทั้งนี้การประเมินผลงานแบบไม่เป็นทางการโดยธรรมชาติแล้วมักมีอิทธิพลย้อนกลับ เพราะผู้สอนมักจะให้ข้อมูล

สะท้อนกลับแบบมีปฏิสัมพันธ์ด้วย ส่วนการทดสอบอย่างเป็นทางการก็สามารถมีอิทธิพลย้อนกลับเชิงบวกได้ แต่หากเป็นการประเมินที่มีแค่ตัวอักษรหรือตัวเลขเดี่ยวๆ ก็อาจไม่มีผลเชิงบวก

In classroom assessment, washback can have several positive effects. This ranges from preparation and review for a test, to the learning that is built on feedback on one's performance. Teachers can give information that is fed back in the form of diagnosis of strengths and weaknesses. Washback also covers the effects of an assessment on teaching and learning before the assessment, for example, on preparation for the assessment. By nature, informal performance assessment has higher likelihood to have washback effects because the teachers usually provide interactive feedback. Formal tests can also have positive washback, but if the assessments result in just a letter grade or a single numerical score, then they likely have little beneficial washback.

ความท้าทายสำคัญของครูผู้สอนคือ การสร้างการทดสอบในห้องเรียนที่ช่วยสนับสนุนอิทธิพลย้อนกลับเชิงบวก ครูสามารถชี้แนะและชมเชยเมื่อการตอบสนองแสดงให้เห็นถึงพัฒนาการในการเรียนรู้ภาษา ครูสามารถชี้แนะกลวิธีในการเรียนให้ประสบความสำเร็จในบทบาทแบบเดียวกับโค้ชผู้ฝึกสอน อิทธิพลย้อนกลับสามารถช่วยในเรื่องแรงจูงใจภายใน การพึ่งพาตนเองในการเรียนรู้ ความมั่นใจในตนเอง อัตลักษณ์ทางภาษา อันตรภาษา และการทุ่มเทแบบมีกลยุทธ์ เป็นต้น

An important challenge for the teachers is to foster beneficial washback through creating classroom tests. The teachers can praise the students for correct responses that show that they are developing language competence. The teachers can also suggest strategies for success in the role of a language coach. Washback can help with intrinsic motivation, learner autonomy, self-confidence, language identity, interlanguage, and strategic investment, among others.

ดังได้กล่าวไว้แล้วข้างต้น การประเมินที่มีแค่ตัวอักษรหรือตัวเลขเดี่ยวๆ ก็อาจไม่มีผลย้อนกลับเชิงบวก กลับกลายการเป็นการส่งเสริมให้ผู้เรียนนำเอาผลคะแนนตัวอักษรหรือตัวเลขคะแนนเดี่ยวๆ นี้มาเปรียบเทียบกัน เกิดเป็นการแข่งขันในการเรียนรู้ไป วิธีการอย่างเป็นรูปธรรมในการปรับปรุงอิทธิพลย้อนกลับ คือครูต้องเขียนแสดงความคิดเห็นให้มากและมีความเจาะจงมากพอสำหรับผลงานในการ

สอบ ชื่นชมจุดเด่นของผู้เรียนและวิพากษ์อย่างสร้างสรรค์ถึงจุดด้อย รวมถึงการชี้แนะวิถีทางที่ผู้เรียนอาจจะปรับปรุงบางองค์ประกอบของผลงานได้ กล่าวอีกนัยหนึ่ง ครูควรต้องทำให้ประสบการณ์การทดสอบของผู้เรียนช่วยส่งเสริมแรงจูงใจจากข้างใน เพื่อให้ผู้เรียนสัมผัสได้ถึงการประสบความสำเร็จและความท้าทาย

Mentioned earlier, the assessments that result only in a single letter grade or a single numerical score may not have many beneficial washback effects. Rather, they indirectly promote the students' competition. A tangible measure for improving washback is for the teachers to write a lot of comments, which must be specific to the students' performance. Teachers need to praise the students for the strengths in their performance, and constructively criticize their weaknesses. The teachers also need to give strategic hints on how to improve elements of their performance. That is to say, the teachers should make the test performance an intrinsically motivating experience, which would allow the students to sense the feelings of accomplishment and challenge.

2.6 Exercise

1. Which is in accordance with the principle of practicality?

A. The students can all finish the test in one hour, but it takes days for the raters to mark.

B. The students can complete an achievement test in two hours.

C. The test answers have to be marked with the computer that is 500 kilometers away.

D. The test needs one proctor for two test takers, and there are 200 test takers.

2. Which can decrease reliability?

A. Some raters do not follow the scoring rubrics.

B. The test is very long, taking 4 hours to complete.

C. The test takers are very anxious.

D. all of the above

3. Which of the following is **NOT** content-related validity evidence?

A. correlation coefficient with another test

B. the same test behavior that the students also do in classroom

C. when the test measures the same learning objectives as those in class

D. none of the above

4. Which type of validity evidence encompasses other types?

A. consequential evidence

B. construct-related evidence

C. content-related evidence

D. criterion-related evidence

5. Which may not create much washback?

A. assessment in which the teachers write a lot of comments for the students

B. assessment that provides conditions for the learners' peak performance

C. assessment that results in a single letter grade

D. assessment that the students have to prepare a lot for

Answer key: 1. B. 2. D. 3. A. 4. B. 5. C.

บทที่ 3 เป้าหมายในการวัดผลและวิธีวัดผล

Chapter 3 Assessment Purposes and Approaches

ในบทที่ 2 ได้กล่าวถึงหลักการพื้นฐานที่มักจะใช้กันในการทดสอบและการวัดผลทางภาษา โดยเน้นไปที่หลักการกว้างๆ ที่น่าจะเป็นประโยชน์สำหรับผู้อ่านทั่วไป ในบทนี้จะได้กล่าวถึงเป้าหมายในการวัดผลและวิธีในการวัดผล โดยแบ่งหัวข้อออกได้เป็นดังนี้

3.1 เป้าหมายในการวัดผล

3.2 วิธีวัดผล

3.3 Exercise (แบบฝึกหัดท้ายบท)

In Chapter 2, basic principles that are often used in language testing and assessment have been covered. The principles covered aims for accessibility and so should be useful for the general audience. In this chapter, purposes and approaches in language assessment will be dealt with, as follows:

3.1 Assessment purposes

3.2 Assessment approaches

3.3 Exercise

3.1 เป้าหมายในการวัดผล (Assessment Purposes)

<u>แบบสอบวัดผลสัมฤทธิ์ทางการเรียน</u>

เป้าหมายในการวัดผลที่พบได้บ่อยมากที่สุดในบริบทห้องเรียนคือ การวัดความสามารถของผู้เรียนในกรอบของบทเรียนในชั้นเรียน ของหน่วยการเรียนรู้หนึ่งๆ หรือแม้แต่กรอบของหลักสูตรทั้งหมด (Brown & Abeywickrama 2010) แบบสอบวัดผลสัมฤทธิ์ทางการเรียนมักจำกัดอยู่ในกรอบของเนื้อหาภายในหลักสูตร โดยมีกรอบระยะเวลาในการทำข้อสอบที่จำกัด และมักถูกนำมาใช้หลังจากที่รายวิชาได้สอนวัตถุประสงค์รายวิชานั้นๆ ไปแล้ว แบบสอบวัดผลสัมฤทธิ์ทางการเรียนอาจใช้เพื่อการวินิจฉัยว่า หัวข้อใดที่ผู้เรียนจะต้องไปศึกษาเรียนรู้ต่อได้ด้วย แต่วัตถุประสงค์หลักของแบบสอบวัดผล

สัมฤทธิ์ทางการเรียน คือการระบุลงไปให้ได้ว่า วัตถุประสงค์รายวิชาวัตถุประสงค์ใดบ้างที่ผู้เรียนได้เรียนบรรลุแล้ว ในตอนท้ายของการเรียนการสอน

Achievement tests

The most common purpose of assessment in the classroom settings is to use the tests to measure the learners' ability within a lesson, a unit, or even a curriculum (Brown & Abeywickrama 2010). Achievement tests are usually restricted within the content of a curriculum, with a time limit. They are also usually used after the learning objectives have been met in class. Still, the achievement tests may also be used for diagnostic purposes, in that the topics could be identified for the learners to work on. But the main purpose for achievement tests is to determine which objectives the learners have achieved at the end of instruction.

แบบสอบวัดผลสัมฤทธิ์ทางการเรียนมักมีลักษณะเป็นการประเมินผลสรุป (summative) เพราะวัดผล ณ ตอนสิ้นสุดบทเรียน หน่วยการเรียนรู้ หรือตอนปลายเทอม แบบสอบวัดผลสัมฤทธิ์ทางการเรียนสามารถเป็นการประเมินผลย่อยระหว่างเรียน (formative) ได้ หากให้ข้อมูลสะท้อนกลับ (feedback) เกี่ยวกับผลการเรียนของผู้เรียนในหน่วยย่อยของหน่วยการเรียนรู้หรือของรายวิชาหนึ่งๆ รายละเอียดเกี่ยวกับแบบสอบวัดผลสัมฤทธิ์ทางการเรียนควรได้รับการกำหนดโดย

- วัตถุประสงค์ของบทเรียน หน่วยการเรียนรู้ หรือรายวิชาที่กำลังประเมินผล
- ค่าน้ำหนักหรือความสำคัญของแต่ละวัตถุประสงค์
- ชิ้นงานที่ใช้ในบทเรียนของแต่ละคาบ
- กรอบเวลาของการสอบและการตรวจให้ผลคะแนน
- ความเป็นไปได้ที่จะให้ข้อมูลสะท้อนกลับระหว่างเรียน

Achievement tests are often summative. The reason is that they take place at the end of a lesson, a unit, or a course. Achievement tests can be formative if feedback about the learners' performance is given in subsets of a unit or a course. The specifications for achievement tests should be determined by:

- the objectives of the lesson, unit, or the course being assessed

- the weight or importance of each objective
- the tasks used in classroom lessons
- the time frame for the test and turnaround
- the potential for formative feedback

แบบสอบวินิจฉัยการเรียน

เป้าหมายของแบบสอบวินิจฉัยการเรียนคือ เพื่อวินิจฉัยแง่มุมหรือด้านของภาษาที่ผู้เรียนจำเป็นต้องพัฒนาหรือที่รายวิชาจะต้องนำมาสอน ตัวอย่างเช่น แบบสอบเกี่ยวกับการออกเสียงภาษาอังกฤษก็อาจวินิจฉัยลักษณะของระบบเสียงที่ยากสำหรับผู้เรียน และจะได้นำมาเป็นส่วนหนึ่งของหลักสูตรที่จะสอน โดยปกติ แบบสอบลักษณะนี้มักจะมีแบบรายการตรวจสอบสำหรับครูผู้สอน เพื่อใช้ระบุจุดที่ยากสำหรับผู้เรียน อีกหนึ่งตัวอย่างเช่น แบบสอบวินิจฉัยการเขียนจะเก็บตัวอย่างงานเขียนจากผู้เรียน เพื่อให้ครูสามารถระบุลักษณะทางการเขียนและทางภาษาศาสตร์ ที่รายวิชาจะต้องให้ความสนใจเป็นพิเศษ

Diagnostic tests

The purpose of diagnostic tests is to diagnose the aspects of language that the learners need to develop or that the course needs to teach. For example, a test in pronunciation may identify difficult phonological features for the students and inform the teachers to bring those as part of the curriculum for instruction. Normally, such tests have a checklist for the teachers to identify the difficult areas for the students. Another example is a writing diagnostic test, where a writing sample can be elicited from the learners. The aim is to identify rhetorical and linguistic features that the course needs to pay particular attention to.

โดยผิวเผิน อาจดูเหมาะสมที่จะรวมแบบสอบวินิจฉัยการเรียนเข้ากับแบบสอบวัดผลสัมฤทธิ์ทางการเรียน อย่างไรก็ดี แบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิเคราะห์ขอบเขตที่ผู้เรียนได้เรียนรู้ลักษณะทางภาษาที่ได้สอนไปแล้ว ส่วนแบบสอบวินิจฉัยการเรียนให้ข้อมูลว่าผู้เรียนควรจะต้องพัฒนาด้านใดเพิ่มเติมในอนาคต ดังนั้น แบบสอบวินิจฉัยการเรียนจะให้รายละเอียดได้มากกว่าในส่วนด้านย่อยๆ ของผู้เรียน เมื่อเทียบกับแบบสอบวัดผลสัมฤทธิ์ทางการเรียน ตัวอย่างเช่น ในช่วงหลักสูตรที่ต้องเน้นด้าน

ไวยากรณ์ แบบสอบวินิจฉัยการเรียนจะให้ข้อมูลเกี่ยวกับด้านต่างๆ ของไวยากรณ์ได้เป็นรายประเด็นได้ดีกว่า เช่นเดียวกัน ในรายวิชาเพื่อการพูดภาษาอังกฤษ การสอบวินิจฉัยอาจให้ผู้เรียนอ่านข้อความออกมาดังๆ หรือให้ลองพูดข้อความตามอิสระ การสอบวินิจฉัยลักษณะนี้จะให้มุมมองเกี่ยวกับความสามารถของผู้เรียนอย่างลึกซึ้ง ทั้งเกี่ยวกับการเน้นเสียงพยางค์ การลงจังหวะจะโคน หน่วยเสียง เป็นต้น

Superficially, it may seem appropriate to combine diagnostic tests with achievement tests. However, achievement tests analyze the extent that the learners have acquired language features. By contrast, diagnostic tests give information as to which aspects the learners will need to work on in the future. Therefore, when compared with achievement tests, diagnostic tests can give more details as to subcategorized aspects of the learner language. For example, in a grammatical, form-focused phase of a curriculum, diagnostic tests can give information about each learner's acquisition of verb tenses, modal auxiliaries, and the like. Likewise, in an oral production, the diagnostic tests may be to give the learner a passage to read aloud, or to ask the learner to say a free speech. Diagnostic tests like these can give the teachers a deep insight into the learners' ability to, say, produce stress, intonation, and segmental phonemes.

แบบสอบวัดระดับ

แบบสอบวัดผลสัมฤทธิ์ทางการเรียนและแบบสอบวัดสมิทธิภาพ (จะกล่าวถึงในหัวข้อย่อยถัดไป) บางแห่งสามารถทำหน้าที่เป็นข้อสอบวัดระดับได้ ข้อสอบวัดระดับมีวัตถุประสงค์เพื่อวัดว่าผู้เรียนมีความสามารถอยู่ในระดับใด หรือสอดคล้องกับสาขาใดในหลักสูตร โดยปกติข้อสอบวัดระดับมักจะรวมเอาตัวอย่างเนื้อหาที่จะเรียนในรายวิชาต่างๆ ของหลักสูตรเข้ามาด้วย ผลการทำข้อสอบของผู้เรียนจึงควรแสดงจุดที่ผู้เรียนพบว่า เนื้อหาไม่ยากหรือไม่ง่ายเกินไป แต่มีความท้าทายในระดับที่เหมาะสมด้วย

Placement tests

Some achievement tests and proficiency tests (explained later) can fuction as placement tests. The purpose of a placement test is to place a student into a particular section of a language curriculum or into a particular level. Usually a placement test contains a sampling of the material that will be covered in the curriculum. Accordingly, a student's

performance on the test would indicate the point at which the student will find the material appropriately challenging.

แบบสอบวัดระดับสามารถทำหน้าที่แบบเดียวกับแบบสอบวินิจฉัยการเรียนได้ เหตุผลสนับสนุนได้แก่ หากสถาบันการศึกษาจะต้องใช้ทั้งแรงงานและทรัพยากรจัดสอบวัดระดับความรู้ผู้เรียนให้เข้าไปตามระดับชั้นต่างๆ ตามหลักสูตรแล้ว ผลพลอยได้ที่พึงประสงค์ของการจัดสอบคือการแจกแจงด้านที่ถนัดและด้านที่ไม่ถนัดของนักศึกษาออกมา การแจกแจงข้อถูกข้อผิดเช่นนี้จะช่วยให้ครูผู้สอนทราบว่า หัวเรื่องใดบ้างที่ควรเน้นหรือไม่จำเป็นต้องเน้น เมื่อเริ่มการเรียนการสอน การใช้ผลจากการสอบวัดระดับในลักษณะนี้แสดงให้เห็นว่า ข้อสอบวัดระดับสามารถมีบทบาทช่วยประเมินระหว่างเรียนได้ (formative role)

Some argue that an effective placement test should be diagnostic too. The reason for this is that if an institution is going to put effort and resources for a placement test, then a beneficial side effect could be a breakdown of strengths and weaknesses that the students showed in the test. A tally of correct and incorrect responses would offer the teachers with useful information on what may or may not need to be emphasized when the term starts. As such, the placement test takes on a formative role.

แบบสอบวัดระดับสามารถมีได้หลากหลายรูปแบบ เช่น วัดผลการตอบสนองผ่านการเขียนและการพูด วัดผลความเข้าใจและการส่งสาร การตอบคำถามปลายเปิดและการเลือกตอบ การเติมคำลงในช่องว่าง และการทำข้อสอบแบบหลายตัวเลือก ทั้งนี้ล้วนขึ้นกับโครงสร้างหลักสูตรและความจำเป็น บางหลักสูตรอาจเลือกใช้แบบสอบวัดสมิทธิภาพมาตรฐานที่มีอยู่แล้ว ด้วยเหตุผลหลักคือความสามารถใช้งานได้จริง ทั้งเรื่องค่าใช้จ่าย ความเร็วในการตรวจให้คะแนน และการรายงานผลคะแนนได้อย่างมีประสิทธิภาพ ส่วนหลักสูตรอื่นอาจใช้การประเมินผลที่อิงตามรายวิชา ที่สามารถนำมาใช้เป็นเครื่องมือวัดระดับได้ด้วย แม้ว่าวัตถุประสงค์หลักจะยังคงเป็นการวัดว่าผู้เรียนควรจะอยู่ระดับใดที่เหมาะสม การนำมาใช้เป็นแบบสอบวินิจฉัยความสามารถของผู้เรียนด้วยทำให้ได้ประโยชน์ทั้งสองประการสำหรับครู

Placement tests can come in a variety of formats, e.g., responding through written and oral performance, assessing comprehension and production, responding to open-ended and limited responses, gap-filling and choice selection. This depends on the nature of a program and its needs. Some programs may use existing standardized proficiency

tests for practical reasons, e.g., budget, speed in scoring, and score-reporting efficiency. Other programs may prefer specific course-based assessments that can double as diagnostic tools. Although placing a student into a course or level is the ultimate goal of a placement test, diagnostic information on a student's performance is a useful secondary benefit for the teachers.

แบบสอบวัดสมิทธิภาพ

แบบสอบวัดสมิทธิภาพใช้วัดสามัตถิยะ (competence) ทั่วไปทางภาษา แบบสอบวัดสมิทธิภาพจึงมิได้จำกัดอยู่ที่วิชาใดวิชาหนึ่ง หลักสูตรใดหลักสูตรหนึ่ง หรือวัดทักษะใดทักษะหนึ่ง หากแต่วัดความสามารถทั่วไป โดยรูปแบบดั้งเดิมจะมีข้อสอบที่ประกอบด้วยข้อสอบแบบเลือกตอบมาตรฐาน ในด้านไวยากรณ์ คำศัพท์ การฟังเพื่อความเข้าใจ และการอ่านเพื่อความเข้าใจ ส่วนแบบสอบวัดสมิทธิภาพที่ผลิตขึ้นมาเชิงพาณิชย์ อย่างเช่น ข้อสอบโทเฟล (the Test of English as a Foreign Language – TOEFL test) ก็มีการเก็บตัวอย่างการเขียน และการพูดด้วย

Proficiency tests

Proficiency tests are used for assessing global competence in a language. They are thus not restricted to any particular course, curriculum, or any skill. They are rather used for testing overall language ability. Traditional proficiency tests are composed of standardized multiple-choice items on grammar, vocabulary, listening comprehension and reading comprehension. Many commercial proficiency tests, e.g., the Test of English as a Foreign Language – TOEFL test, also have writing and oral production, too.

แบบสอบวัดสมิทธิภาพโดยปกติจะมีลักษณะเป็นการสอบประเมินผลสรุป (summative) และอิงกลุ่มเป็นหลัก ข้อสอบให้ผลลัพธ์เป็นตัวเลขคะแนนเดี่ยวๆ หรืออาจมีคะแนนของตอนย่อยในข้อสอบด้วย แบบสอบวัดสมิทธิภาพลักษณะนี้หลายคนมองว่าเพียงพอสำหรับทำหน้าที่เป็นผู้รักษาประตู ในการให้หรือไม่ให้ใครคนใดคนหนึ่งผ่านขึ้นไปยังระดับชั้นถัดไป นอกจากนี้ เนื่องจากแบบสอบวัดสมิทธิภาพวัดผลเมื่อเทียบกันในกลุ่มผู้สอบ แบบสอบชนิดนี้จึงไม่อาจให้ข้อมูลสะท้อนกลับเชิงวินิจฉัยได้

Proficiency tests are normally summative and norm-referenced. They offer results of single numerical scores, which many view as sufficient for the gate-keeping role. Moreover, they measure performance against a norm, and so may not give diagnostic feedback.

ประเด็นสำคัญสำหรับการสอบวัดสมิทธิภาพคือ ความหมายคะแนนเกี่ยวกับการสอบได้ระบุไว้อย่างไร ชิ้นงานที่ผู้สอบทำจำเป็นต้องเป็นตัวอย่างของการใช้ภาษาในขอบเขตของบริบทที่ชัดเจน การสร้างชิ้นงานเช่นนี้และวิจัยความสมเหตุสมผลของชิ้นงานนับเป็นกระบวนการที่ต้องใช้เวลาและมีค่าใช้จ่ายมาก ครูสอนภาษาโดยทั่วไปไม่ควรสร้างแบบสอบวัดสมิทธิภาพทั่วไปด้วยตนเอง วิธีการที่เหมาะแก่การใช้งานได้จริงคือ การใช้แบบสอบวัดสมิทธิภาพที่มีบริการในเชิงพาณิชย์

A key point for proficiency tests is how the constructs of language ability are defined. The test tasks that the test takers will perform must be legitimate samples of English language use in a well-defined context. The creation of these tasks and their validation with empirical research is time-consuming and costly. In general, language teachers should not attempt to create the whole proficiency test on their own. Rather, a more practical approach is to choose a proficiency test that is commercially available.

<u>แบบสอบวัดความถนัด</u>

แบบสอบชนิดนี้ปัจจุบันไม่ได้เป็นที่นิยมเหมือนที่เคย แบบสอบวัดความถนัดถูกออกแบบมาให้วัดศักยภาพหรือความสามารถทั่วไปในการเรียนรู้ภาษาต่างประเทศ*ก่อน*ที่จะเรียนรายวิชาหนึ่งๆ และยังวัดความสำเร็จคาดทำนายในการเรียนรายวิชานั้นๆ แบบสอบวัดความถนัดทางภาษายังอ้างว่าได้ถูกออกแบบมาให้สามารถประยุกต์ใช้กับการเรียนภาษาใดๆ ในบริบทห้องเรียนได้ด้วย

<u>Aptitude tests</u>

This type of test is nowadays not as popular as it used to be. An aptitude test is designed to measure general ability and capacity in foreign language learning *prior to* the learning and seeks to predict the success in that learning too. Language aptitude tests were designed to be applicable to the classroom learning of any language.

มีแบบสอบวัดความถนัดมาตรฐานอยู่สองสำนักที่เคยใช้ในสหรัฐอเมริกา ได้แก่ แบบสอบวัดความถนัดทางภาษายุคใหม่ (Modern Language Aptitude Test – MLAT) และชุดข้อสอบวัดความถนัดทางภาษา Pimsleur (Pimsleur Language Aptitude Battery – PLAB) ทั้งสองแบบสอบเป็นข้อสอบภาษาอังกฤษและให้ผู้สอบทำชิ้นงานที่เกี่ยวข้องกับภาษา เช่น การนับตัวเลขเป็นภาษาอังกฤษ การแยกแยะเสียงในคำพูด การระบุหน้าที่ทางไวยากรณ์ และการจดจำคำที่เกี่ยวข้องกันเป็นคู่ๆ เป็นต้น ทั้ง MLAT และ PLAB มีค่าสหสัมพันธ์เชิงบวกกับผลการเรียนของผู้เรียนในรายวิชาภาษา แต่อย่างไรก็ดี ค่าสหสัมพันธ์เหล่านี้ขึ้นอยู่กับรายวิชาภาษาที่ก็วัดผลสำเร็จโดยการให้ออกเสียงเลียนแบบ การจดจำ และการแก้ปริศนา ซึ่งก็คล้ายๆ กับที่ปรากฏในทั้งสองแบบสอบ จนถึงปัจจุบัน ยังคงไม่มีงานวิจัยที่จะแสดงว่าชิ้นงานของทั้งสองแบบสอบจะสามารถทำนายและมีความเชื่อมโยงอย่างเด่นชัดกับการใช้ภาษาเพื่อการสื่อสารได้อย่างประสบความสำเร็จ โดยเฉพาะการเรียนรู้ภาษาที่มิได้มีการเตรียมการมาก่อน

There were two standardized aptitude tests that were used in the U.S. They are Modern Language Aptitude Test (MLAT) and Pimsleur Language Aptitude Battery (PLAB). Both of them are English language tests and require the students to do the tasks that are language-related, e.g., numbering, distinguishing speech sounds, identifying grammatical functions, memorizing word associates. The MLAT and PLAB exhibit some correlations with performance of students in language courses. However, those correlations depend on a language course in which the success is measured by similar processes of mimicry, memorization, and puzzle solving. There is still no empirical research to show that such tasks could predict communicative success in a language, especially untutored language acquisition.

ด้วยข้อจำกัดเรื่องที่ยังไม่มีงานวิจัยมารองรับดังกล่าวไว้ข้างต้น ปัจจุบันแบบสอบวัดความถนัดมาตรฐานจึงไม่ได้รับความนิยมอีกต่อไป ยกเว้นในกรณีเพื่อหาความพิการด้านการเรียนภาษาต่างประเทศ แบบสอบวัดความถนัดมักให้ข้อมูลแก่ผู้เรียนเกี่ยวกับสไตล์การเรียนที่ชอบ จุดแข็งและจุดอ่อน พร้อมด้วยคำแนะนำเกี่ยวกับกลวิธีในการใช้ประโยชน์จากจุดแข็ง และพัฒนาเติมเต็มจุดอ่อน ข้อสอบใดๆ ที่อ้างว่าจะสามารถคาดทำนายเกี่ยวกับความสำเร็จในการเรียนภาษาถือได้ว่ามีข้อบกพร่อง เพราะปัจจุบันเป็นที่ทราบกันอยู่แล้วว่า ด้วยการแสวงหาความรู้ด้วยตนเองอย่างเหมาะสม การเรียนอย่างมียุทธวิธีเชิงรุก และการสอนอย่างมีกลวิธีที่เหมาะสม ผู้เรียนแทบทุกคนสามารถประสบความสำเร็จได้ในที่สุด

การไปจำแนกแยกแยะผู้เรียนอย่างละเอียดก่อนที่แม้แต่กระทั่งจะได้พยายามเรียน เป็นการแขวนความสำเร็จหรือความล้มเหลวไว้ล่วงหน้าโดยไม่มีเหตุจำเป็น

Because of the limitation of lacking research support, standardized aptitude tests are rarely used nowadays. An exception is perhaps when identifying a disability in foreign language learning. Attempts to measure language aptitude often end up with providing information about the learners' preferred styles and their potential strengths and weaknesses, together with follow-up strategies for capitalizing on the strengths and overcoming weaknesses. We now know that with appropriate self-knowledge, active strategic involvement in learning and strategies-based instruction, nearly everyone can eventually have learning success. Therefore, any test that claims to predict success in language learning is flawed. Classifying the learners in detail *a priori*, even before they have attempted to learn a language, is to presume failure or success without appropriate cause.

3.2 วิธีวัดผล (Assessment Approaches)

วิธีวัดผลมีรูปแบบการแบ่งได้หลายวิธี วิธีแบ่งที่พบได้บ่อยมีสองรูปแบบคือ การแบ่งตามเป้าประสงค์ของการวัดผล แบ่งเป็นการประเมินผลเพื่อการเรียนรู้ การประเมินผลในขณะเรียนรู้ และการประเมินผลการเรียนรู้ (NSW Government 2019) รูปแบบการแบ่งนี้จะกล่าวถึงเป็นลำดับแรก อีกวิธีที่ใช้แบ่งคือ แบ่งตามประเภทคำตอบที่ผู้เรียนทำ ได้แก่ การประเมินผลแบบเลือกตอบ การประเมินผลแบบสร้างคำตอบ และการประเมินผลแบบคำตอบเฉพาะตัว (Brown & Hudson 1998) รูปแบบการแบ่งวิธีหลังนี้จะกล่าวถึงเป็นลำดับถัดไป

There can be several ways of classifying assessment approaches. Two ways can often be found in the literature of the field. The first one is to classify them according to the objectives of the assessment. They include assessment for learning, assessment as learning, and assessment of learning (NSW Government 2019). This classification will be dealt with in turn in the section that follows. The other way of classification is to classify the approaches in accordance with the responses that the learners or the test takers make. They are selected-response assessment, constructed-response assessment, and personal-

response assessment (Brown & Hudson 1998). The latter classification will be dealt with after the objective-based classification.

วิธีวัดผลตามวัตถุประสงค์

แรกเริ่มเดิมที จุดเน้นของการประเมินในชั้นเรียนอยู่ที่การประเมินผลการเรียนรู้ นั่นคือ การใช้ข้อมูลเพื่อตัดสินเกี่ยวกับผลการเรียนของผู้เรียน และรายงานผลการตัดสินนั้นต่อผู้เกี่ยวข้อง ในช่วงทศวรรษ 1990 งานวิจัยได้หันมาเน้นให้ความสำคัญของการประเมินผลเพื่อการเรียนรู้ (การประเมินระหว่างเรียน) ครูผู้สอนใช้การประเมินผลเพื่อการเรียนรู้ เมื่อมีการสอบวินิจฉัยการเรียน มีการประเมินระหว่างเรียน และมีการให้ข้อมูลสะท้อนกลับในขั้นตอนต่างๆ ระหว่างการเรียนการสอน ถึงแม้ว่าการประเมินผลเพื่อการเรียนรู้จะเป็นไปอย่างไม่เป็นทางการก็ตาม เมื่อมาถึงช่วงทศวรรษ 2000 เป็นต้นมา การประเมินผลเพื่อการเรียนรู้ได้แยกเป็นการประเมินผลเพื่อการเรียนรู้และการประเมินผลในขณะเรียนรู้ เพื่อเน้นบทบาทของผู้เรียนในกระบวนการประเมินผล ทั้งนี้ การประเมินผลในขณะเรียนรู้อย่างเป็นระบบได้กลายมาเป็นกิจกรรมการประเมินที่ได้รับการใช้เพื่อพัฒนาศักยภาพผู้เรียนในการประเมิน และปรับใช้กับการเรียนรู้ของตนเอง

Objective-based approaches to assessment

The focus of classroom assessment has traditionally been on assessment of learning. This means that assessment takes place after the classroom learning, for making judgments about students' performance and informing those involved of the judgments. In the 1990s, research turned its emphasis to assessment for learning (formative assessment). When teachers used diagnostic tests, and formative assessment, and gave feedback in stages of instruction, they used assessment for learning, even though the assessment for learning was often informal. Then in the 2000s onwards, assessment for learning was divided into assessment for learning and assessment as learning, in order to highlight the role of the students in the assessment processes. Systematic assessment as learning has become an assessment practice for developing the learners' potential in assessment and adapting for their own learning.

การประเมินผลเพื่อการเรียนรู้ การประเมินผลเพื่อการเรียนรู้เกี่ยวข้องกับการที่ครูผู้สอนใช้หลักฐานเกี่ยวกับความรู้ ความเข้าใจ และทักษะของผู้เรียน เพื่อให้ข้อมูลแก่การสอน บางครั้งจึงอาจเรียกรวมๆ ได้ว่าเป็นการประเมินระหว่างเรียน (formative assessment) การประเมินเพื่อการเรียนรู้สามารถเกิดขึ้นได้ตลอดกระบวนการเรียนการสอน เพราะทำให้ครูผู้สอนทราบว่าผู้เรียนรู้อะไรหรือสามารถทำอะไรได้บ้าง

Assessment for learning. Assessment for learning is concerned with the teachers using evidence of the learners' knowledge, understanding, and skills in order to inform their teaching. Because of this nature, assessment for learning is sometimes called formative assessment. Assessment for learning can happen throughout the process of instruction, because it enables the teachers to know what the learners know and can do.

การประเมินผลในขณะเรียนรู้ การประเมินผลในขณะเรียนรู้เกิดขึ้นเมื่อผู้เรียนเป็นผู้ประเมินตนเอง ผู้เรียนสามารถเฝ้าสังเกตการเรียนของตนเอง ตั้งคำถาม และใช้กลวิธีอย่างหลากหลายเพื่อที่จะตัดสินใจว่าอะไรที่รู้แล้ว อะไรที่สามารถทำได้ รวมไปถึงว่าจะใช้การประเมินผลเพื่อให้เกิดการเรียนรู้ใหม่ได้อย่างไร

Assessment as learning. This type of assessment is when the students become their own assessors. They can monitor their own learning, pose questions, and use a variety of strategies in order to decide what they already know and what they can do. They can also pose questions, and use strategies to decide how to use information from assessment for new learning.

การประเมินผลการเรียนรู้ การประเมินผลการเรียนรู้เกิดขึ้นเมื่อครูผู้สอนใช้หลักฐานของการเรียนรู้ของผู้เรียน เพื่อประเมินความสำเร็จหรือการบรรลุวัตถุประสงค์ เมื่อเทียบกับผลลัพธ์ในชั้นเรียนหรือมาตรฐานใดมาตรฐานหนึ่ง เนื่องจากการประเมินผลการเรียนรู้มักเกิดขึ้น ณ จุดเวลาที่ชัดเจน หรือในตอนท้ายของหน่วยการเรียนรู้ ตอนปลายเทอม หรือตอนสิ้นภาคการศึกษา รวมทั้งถูกใช้เพื่อจัดอันดับหรือประเมินเกรดผู้เรียน การประเมินผลการเรียนรู้จึงมักเรียกกันว่า เป็นการประเมินผลสรุป (summative assessment) ประสิทธิผลของการประเมินผลการเรียนรู้ในการจัดอันดับหรือประเมินเกรดผู้เรียนขึ้นอยู่กับความตรงและความเที่ยงของกิจกรรมการประเมินผล

Assessment of learning. Assessment of learning happens when the teachers use evidence of the learners' learning in order to assess success or achievement against outcomes or standards. Assessment of learning usually happens at a defined key point, often at the end of a unit, term or semester, and may be used for ranking or grading the students, so it is often called summative assessment. The effectiveness of this kind of assessment for grading and ranking depends on the validity and reliability of assessment activities.

วิธีวัดผลตามรูปแบบคำตอบ

การแบ่งวิธีวัดผลตามวัตถุประสงค์ได้กล่าวถึงไว้แล้วข้างต้น ในหัวข้อนี้จะกล่าวถึงวิธีวัดผลที่แบ่งตามรูปแบบคำตอบของผู้เรียน แบ่งเป็นการประเมินผลแบบเลือกตอบ การประเมินผลแบบสร้างคำตอบ และการประเมินผลแบบคำตอบเฉพาะตัว (Brown & Hudson 1998)

Response-based approaches to assessment

Objective-based approaches to assessment have been dealt with in the previous section. In this section, approaches based on the formats of student responses will be dealt with. They are selected-response assessment, constructed-response assessment, and personal-response assessment (Brown & Hudson 1998).

การประเมินผลแบบเลือกตอบ ในการประเมินผลแบบเลือกตอบ ผู้เรียนจะได้รับตัวอย่างภาษา โดยจะต้องเลือกคำตอบที่ถูกต้องจากตัวเลือกจำนวนหนึ่ง ตามปกติการประเมินผลแบบนี้ ผู้เรียนไม่ต้องส่งสารใดๆ ออกมา การประเมินผลแบบเลือกตอบจึงเหมาะกับการวัดทักษะรับสาร เช่น การฟังและการอ่าน การประเมินผลแบบเลือกตอบที่ใช้กันบ่อยๆ ได้แก่ ประเภทถูกผิด (true-false) ประเภทจับคู่ (matching) และประเภทเลือกตอบ (multiple-choice)

Selected-response assessment. In selected-response assessment, the learners are given language material. They are required to select one of the choices given. Normally, they do not create any language, and so this assessment type is suitable for

receptive skills such as listening and reading. The types of selected-response assessment that are used often include true-false, matching, and multiple-choice.

ข้อดีของการประเมินผลแบบเลือกตอบได้แก่ความรวดเร็วในการจัดสอบ การให้คะแนนก็สามารถทำได้อย่างรวดเร็ว สะดวก และมีความเป็นปรนัย อย่างไรก็ดี ข้อเสียสำคัญสองประการได้แก่ ความยากในการออกข้อสอบชนิดนี้ และการที่ไม่ได้ให้ผู้เรียนผลิตภาษาใดๆ ออกมา

Advantages of selected-response assessments are that they are quick to administer, and scoring them is also quick, easy and objective. However, disadvantages are difficulty to construct, and the fact that they do not require the test takers to produce any language.

การประเมินผลแบบสร้างคำตอบ ในการประเมินแบบสร้างคำตอบ ผู้เรียนจะต้องเขียน พูด หรือทำอะไรสักอย่างหนึ่งออกมา ดังนั้น การประเมินแบบนี้จึงเหมาะกับการวัดทักษะการส่งสาร ได้แก่ การพูดและการเขียน การประเมินแบบสร้างคำตอบยังมีประโยชน์ในการสังเกตปฏิสัมพันธ์ระหว่างทักษะการส่งสารและทักษะรับสาร เช่น ปฏิสัมพันธ์ระหว่างการฟังและการพูดในการสัมภาษณ์ปากเปล่า ปฏิสัมพันธ์ระหว่างการอ่านและการเขียน ในการประเมินผลการแสดงออกที่ผู้เรียนอ่านบทความวิชาการสองฉบับและเขียนเรียงความเปรียบเทียบความเหมือนและความต่าง เป็นต้น การประเมินผลแบบสร้างคำตอบที่ใช้กันบ่อยๆ ได้แก่ ประเภทเติมคำลงในช่องว่าง (fill-in) ประเภทตอบสั้น (short answer) และประเภทแสดงออก (performance)

Constructed-response assessment. In constructed-response assessment, the learners are required to write, say, or do something. Therefore, this type of assessment is suitable for measuring the productive skills of speaking and writing. The constructed-response assessment is also useful for observing the interactions between receptive and productive skills, e.g., the interaction between listening and speaking in an oral interview, the interaction of reading and writing in a performance assessment, in which the learners read two academic articles and write an essay comparing and contrasting the articles. The types of constructed-response assessment that are used often include fill-in, short answer, and performance.

ข้อดีของการประเมินผลแบบสร้างคำตอบก็คือ การประเมินแบบนี้ลดโอกาสเดาคำตอบลง แต่ก็เพิ่มปัญหาเรื่องความเป็นอัตนัยขึ้นมา โดยเฉพาะเมื่อดุลพินิจของผู้ตรวจต้องเข้ามาตัดสินใจว่าคำตอบที่ถูกต้องสำหรับช่องว่างช่องหนึ่งๆ คืออะไร หรือเมื่อผู้ตรวจประเมินตัวอย่างภาษาที่ผู้สอบได้แสดงออกมา

An advantage of constructed-response assessment is that it gets rid of some of the guessing factor. However, it creates problems of subjectivity, especially when raters score language samples or when human judgments are involved in deciding a correct answer.

การประเมินผลแบบคำตอบเฉพาะตัว ในการประเมินผลแบบคำตอบเฉพาะตัว ผู้เรียนจะต้องเขียน พูด หรือทำอะไรสักอย่างหนึ่งออกมาเหมือนในการประเมินผลแบบสร้างคำตอบ แต่ในการประเมินผลแบบคำตอบเฉพาะตัว คำตอบของผู้เรียนแต่ละคนสามารถแตกต่างกันได้มาก กล่าวคือ การประเมินผลแบบคำตอบเฉพาะตัวทำให้ผู้เรียนได้สื่อสารสิ่งที่ต้องการสื่อจริงๆ การประเมินผลแบบคำตอบเฉพาะตัวที่ใช้กันบ่อยๆ ได้แก่ การพูดคุยงาน (conferences) แฟ้มสะสมผลงาน (portfolios) และการประเมินตนเองและเพื่อนร่วมชั้น (self- and peer assessments) การประเมินผลแบบคำตอบเฉพาะตัวจะกล่าวถึงในรายละเอียดในบทที่ 7 ในหัวข้อการประเมินผลทางเลือก

Personal-response assessment. In personal-response assessment, the learners have to say, write, or perform something, just like in constructed-response assessment. A key difference, however, is that personal-response assessments allow the learners to communicate what they want to communicate, and so their responses can be quite different from one another. The types of personal-response assessment that are used often include conferences, portfolios, and self- and peer assessments. Personal-response assessment will be discussed more in detail in Chapter 7 as alternative assessment.

ข้อดีของการประเมินผลแบบคำตอบเฉพาะตัวคือ เป็นการวัดที่สะท้อนความเป็นปัจเจกบุคคลของผู้เรียนแต่ละคน สามารถผสมผสานกับหลักสูตรได้โดยตรง และสามารถใช้ประเมินกระบวนการเรียนรู้แบบต่อเนื่องตลอดการเรียนการสอน อย่างไรก็ดี การประเมินผลแบบคำตอบเฉพาะตัวก็มีข้อเสีย ได้แก่ ความยากที่จะจัดทำและบริหารจัดการ และการให้คะแนนก็มีความเป็นอัตนัย

Advantages of personal-response assessment are that it can provide individualized assessment, can be integrated directly into the curriculum, and can assess learning processes in an ongoing manner throughout the process of instruction. However, some disadvantages of personal-response assessment include relative difficulty to produce and organize, and subjective scoring.

3.3 Exercise

1. Which of the following is **UNLIKELY** to be practical?

A. A placement test also serves as a diagnostic test.

B. A proficiency test also functions as a diagnostic test.

C. An achievement test also serves as a diagnostic test.

D. An aptitude test also functions as a diagnostic test.

2. Which is usually designed to assess global competence?

A. achievement test

B. diagnostic test

C. placement test

D. proficiency test

3. Which is nowadays **NO LONGER** popular?

A. achievement test

B. aptitude test

C. diagnostic test

D. placement test

4. Which of the following is **NOT** an objective-based approach to assessment?

A. assessment as learning

B. assessment for learning

C. assessment of learning

D. selected-response assessment

5. Which test type is **NOT** a constructed-response assessment?

A. fill-in

B. matching

C. performance

D. short answer

Answer key: 1. B. 2. D. 3. B. 4. D. 5. B.

บทที่ 4 การออกแบบและการสร้างแบบทดสอบ

Chapter 4 Designing and writing tests

ในบทที่ 3 ได้นำเสนอเป้าหมายในการวัดผลและวิธีวัดผล ในบทนี้ จะกล่าวถึงการออกแบบและการสร้างแบบทดสอบ โดยมีลำดับการนำเสนอดังนี้

4.1 กำหนดเป้าหมายของแบบทดสอบ

4.2 กำหนดวัดถุประสงค์ในการวัดให้ชัดเจน

4.3 เขียนลักษณะเฉพาะของแบบทดสอบ

4.4 เขียนข้อสอบ

4.5 เขียนข้อสอบหลายตัวเลือก

4.6 จัดสอบ

4.7 ให้คะแนน ตัดเกรด และให้ข้อมูลสะท้อนกลับ

4.8 Exercise (แบบฝึกหัดท้ายบท)

In Chapter 3, assessment purposes and assessment approaches have been covered. In this chapter, designing and writing tests will be dealt with. The order of presentation is as follows:

4.1 Determining the purpose of a test

4.2 Drawing up clear objectives

4.3 Designing test specifications

4.4 Devising test items

4.5 Designing multiple-choice items

4.6 Administering the test

4.7 Scoring, grading, and giving feedback

4.8 Exercise

4.1 กำหนดเป้าหมายของแบบทดสอบ (Determining the purpose of a test)

ในการกำหนดเป้าหมายของแบบสอบ มีคำถามสำคัญดังต่อไปนี้ ทำไมเราจึงสร้างแบบทดสอบนี้ขึ้นมา หรือทำไมแบบทดสอบนี้จึงถูกสร้างขึ้นมาโดยผู้เขียนหนังสือ แบบทดสอบมีความสำคัญอย่างไรต่อรายวิชาหรือคอร์สของเรา เช่น เพื่อประเมินสมิทธิภาพโดยรวมหรือเพื่อจัดวางผู้เรียนเข้าไปในรายวิชาหนึ่งๆ แบบทดสอบนี้สำคัญอย่างไรเมื่อเทียบกับผลงานอื่นของผู้เรียน จะมีผลกระทบใดบ้างต่อครูผู้สอนและต่อผู้เรียนก่อนและหลังการประเมินผล (Brown & Abeywickrama 2010) การถามและตอบคำถามเหล่านี้จะช่วยให้เรากำหนดวัตถุประสงค์ในการวัดได้ง่ายขึ้น

In determining the purpose of a test, there are several critical questions that we have to ask ourselves. First, why do we create this test, or why has this test been written by the author of a book? How important is this test to my course, e.g., to assess overall proficiency, or to place a student in a course? What is the significance of this test when compared with other student performance? Will there be any impact of this evaluation on the teacher and the students before and after the assessment? (Brown & Abeywickrama 2010) Asking these questions helps us design the objectives of the assessment more easily.

การออกแบบรูปแบบการทดสอบใหม่ๆ มิได้เป็นเรื่องง่ายๆ หากแต่ต้องอาศัยความพยายามในการออกแบบ และใช้เวลามากในการทำให้ดีขึ้นจนใช้งานได้จริงผ่านการลองผิดลองถูก ส่วนการใช้เทคนิคการสอบแบบดั้งเดิมก็สามารถเข้ากันได้กับหลักสูตรภาษาที่เน้นการสื่อสารแบบมีปฏิสัมพันธ์ ดังนั้นวิธีการที่ดีที่สุดในฐานะครูมือใหม่คือ การทำงานภายในกรอบคำแนะนำในเรื่องเทคนิควิธีการทดสอบแบบดั้งเดิมที่ผู้เรียนรู้จักอย่างดีและได้รับการยอมรับ ต่อเมื่อมีประสบการณ์เพิ่มขึ้น ครูผู้สอนก็สามารถลองออกแบบรูปแบบการทดสอบที่ฉีกแนวมากขึ้นได้

New and innovative task formats are not easy to design. Rather, they require a lot of effort and are time-consuming before they are refined through trial and error. Traditional testing techniques can conform to an interactive, communicative language curriculum. Therefore, the best course of action for new teachers in designing task formats

is to work within the guidelines of known and accepted traditional testing techniques. With more experience, the teachers can attempt bolder test designs.

ในการออกแบบการวัดผลในชั้นเรียนใดๆ หรือแม้แต่ในการพิจารณาความเหมาะสมของแบบทดสอบที่มีอยู่ ขั้นตอนแรกและอาจจะเป็นขั้นตอนที่สำคัญที่สุดก็คือ การถอยมาหนึ่งก้าวและมองเป้าหมายรวมของกิจกรรมที่ผู้เรียนกำลังจะทำ เป้าหมายของการวัดอาจเรียกว่าเป็นประโยชน์ใช้สอยของแบบทดสอบ อาจพิจารณาได้จากรายการตรวจสอบ (checklist) ดังนี้

1) ฉันจำเป็นต้องจัดสอบ ณ ตอนนี้ในรายวิชาของฉันหรือไม่? ถ้าจำเป็น การทดสอบจะทำหน้าที่อะไรต่อผู้เรียน และ/หรือฉัน?

2) การทดสอบมีความสำคัญอย่างไรต่อรายวิชาของฉัน?

3) การทดสอบเป็นวิธีปกติในการจบบทเรียน หน่วยการเรียนรู้ หรือจบช่วงเวลาของรายวิชา?

4) การทดสอบสำคัญอย่างไร เมื่อเทียบกับผลงานด้านอื่นๆ ของผู้เรียน

5) ฉันอยากจะใช้ผลจากการสอบมาเพื่อพิจารณาว่า ผู้เรียนของฉันปฏิบัติตัวได้ตามกรอบมาตรฐานของหลักสูตรที่กำหนดไว้หรือไม่?

6) ฉันอยากจะให้ผู้เรียนเป็นผู้รับอิทธิพลย้อนกลับเชิงบวก (positive washback) จริงๆ หรือไม่?

7) ฉันจะใช้ผลจากการทดสอบเป็นหนึ่งในวิธีในการจัดสรรแนวการสอนของฉันในวันหรือสัปดาห์ที่กำลังจะมาถึง?

8) จะมีผลกระทบอะไรต่อสิ่งที่ฉันทำและที่ผู้เรียนทำ ก่อนและหลังการทดสอบ?

In designing any assessment, or even in determining the appropriateness of an existing test, the first and potentially most important step is to step back and consider the overall purpose of the activity that the students are about to do. The purpose of the assessment can also be called test usefulness. The following list is a checklist for determining the purpose of an assessment:

1) Do I need to administer a test at this point in my course? If so, what purpose will the test serve the students and/or me?

2) What is the significance of the test relative to my course?

3) Is the test an expected way to end a lesson, unit, or period of the course?

4) How important is the test, when compared with other student performance?

5) Do I want to use the results of the test to determine if my students have met predetermined curricular standards?

6) Do I really want my students to receive positive washback?

7) Will I use the results from the test as a means for allocating my teaching efforts in the days or weeks to come?

8) What will be the impact of the test on what I do and what the students do, before and after the test?

4.2 กำหนดวัตถุประสงค์ในการวัดให้ชัดเจน (Drawing up clear objectives)

นอกเหนือไปจากการทราบเป้าหมายของแบบทดสอบแล้ว เราจำต้องรู้ให้ชัดว่าเราต้องการวัดอะไร (Brown & Abeywickrama 2010) ในการกำหนดวัตถุประสงค์ของการวัด มีคำถามสำคัญดังต่อไปนี้ วัตถุประสงค์ของแบบสอบคืออะไร เรากำลังค้นหาสิ่งใดกันแน่ การตั้งวัตถุประสงค์ได้อย่างเหมาะสมเกี่ยวข้องกับหลายประเด็น ทั้งจากวัตถุประสงค์แบบพื้นฐานเกี่ยวกับรูปแบบภาษา (forms) และหน้าที่ที่ครอบคลุมอยู่ในหน่วยการเรียนรู้หนึ่งๆ ไปจนถึงวัตถุประสงค์ที่ซับซ้อนขึ้นมา เกี่ยวกับความหมายคะแนนที่นำเสนอในแบบสอบหนึ่งๆ ในการกำหนดวัตถุประสงค์ของการวัดจึงเป็นการตัดสินใจลงไปว่า เรากำลังวัดผลความสามารถทางภาษาด้านใด

In addition to knowing the purpose of the test constructed, we also need to know exactly what it is that we want to test (Brown & Abeywickrama 2010). In determining the objectives of a test, there are several questions that are to be asked. First, what are the objectives of the test, and what exactly are we trying to find out? Designing an appropriate and clear objective involves many issues, from simple objectives about forms and functions

that are covered in a course or in a unit, to more complex objectives about the test constucts. In designing test objectives, it is thus decisions about what language abilities we are going to assess.

ในการกำหนดวัตถุประสงค์ในการวัด ควรเริ่มจากการดูทุกอย่างที่ผู้เรียนน่าจะรู้หรือน่าจะสามารถทำได้ โดยดูจากเอกสารหรือหนังสือเรียนที่ผู้เรียนต้องใช้ กล่าวอย่างง่ายๆ ก็คือ ผู้สอนต้องดูวัตถุประสงค์ของหน่วยการเรียนรู้ที่กำลังจะทำการทดสอบนั่นเอง ทุกๆ หลักสูตรจะต้องมีวัตถุประสงค์ที่วัดได้และมีกรอบที่ชัดเจนเหมาะสมอยู่แล้ว วัตถุประสงค์ที่วัดได้และมีกรอบที่เหมาะสมได้แก่ วัตถุประสงค์ที่กล่าวถึงพฤติกรรมการแสดงออกของผู้เรียนที่ชัดแจ้ง ตัวอย่างเช่น วัตถุประสงค์หน่วยการเรียนรู้ที่กล่าวแค่ว่า “ผู้เรียนจะเรียนรู้กาลของคำกริยา” หรือ “มีจุดเน้นไวยากรณ์ในด้านกาลของคำกิริยา” นับว่าไม่สามารถนำมาทดสอบได้ วัตถุประสงค์ที่กล่าวถึงพฤติกรรมการแสดงออกที่ชัดแจ้งควรระบุถึงขั้นที่ว่า “ในบริบทแบบเดียวกับที่พบในการเรียน ผู้เรียนจะจำแนกรูปแบบปัจจุบันกาลของคำกริยา ในรูปการเขียนและการออกเสียงได้” เป็นต้น หลังจากดูขอบเขตวัตถุประสงค์เช่นนี้ ผู้สอนก็กำหนดวัตถุประสงค์ที่ชัดเจนขึ้นมา วัตถุประสงค์ที่ระบุชัดถึงพฤติกรรมการแสดงออกที่วัดได้ และด้านโดเมนภาษาที่เป็นเป้าหมาย

In designing assessment objectives, a starting point should be looking carefully at the material that the learners need to use, for what they should know or should be able to do. In other words, the teachers should have a look at the objectives of the unit they are going to test. Normally, every curriculum should have clear, assessable and appropriately framed objectives. Such an objective would be stated in relation to overt performance by students. For example, the objective of a learning unit that reads only, “the students will learn verb tenses” and “the grammatical focus is on verb tenses” would not be assessable. The objectives that engage overt performance should read, “In the contexts identical to the contexts studied in class, the students will recognize the written and oral forms of the present tense.” After checking out the class objectives, the teachers can design the assessment objectives that are assessable in terms of overt behavior and the target language domains.

4.3 เขียนลักษณะเฉพาะของแบบทดสอบ (Designing test specifications)

ในการเขียนลักษณะเฉพาะของแบบทดสอบ (test specifications – specs) เราต้องกำหนดลงไปว่า ลักษณะเฉพาะของแบบทดสอบจะสะท้อนทั้งเป้าหมายและและวัตถุประสงค์ในการวัดอย่างไร (Brown & Abeywickrama 2010) เพื่อที่จะออกแบบหรือประเมินแบบทดสอบหนึ่งๆ เราต้องทำให้แน่ใจว่า แบบทดสอบมีโครงสร้างที่สอดคล้องกับหน่วยการเรียนรู้หรือบทเรียนที่กำลังจะทำการทดสอบ วัตถุประสงค์ในชั้นเรียนควรจะต้องปรากฏชัดในแบบทดสอบ ผ่านชนิดของชิ้นงานที่เหมาะสม ให้ค่าน้ำหนักความสำคัญอย่างเหมาะสม มีการจัดเรียงอย่างเป็นเหตุเป็นผล รวมทั้งผ่านชิ้นงานที่หลากหลายด้วย

In drawing up test specifications, we need to determine how the test specifications will reflect both the purpose of the test and the objectives (Brown & Abeywickrama 2010). In order to design or evaluate a test, we need to be sure that the test has a logical structure that agrees with the unit or the lesson it is testing. The class objectives should be clearly present in the test, through suitable task types and weights, with a logical sequence and a variety of tasks.

ในบริบทห้องเรียน ลักษณะเฉพาะของแบบทดสอบสามารถทำหน้าที่เป็นเค้าโครงของการทดสอบ ว่าจะมีลักษณะ “หน้าตา” เป็นอย่างไร ลักษณะเฉพาะคือพิมพ์เขียวของแบบทดสอบ ที่มีองค์ประกอบดังต่อไปนี้

1) คำบรรยายเนื้อหา

2) ชนิดข้อสอบ (item types) หรือก็คือวิธีการ เช่น แบบเลือกตอบ แบบโคลซ (cloze) ฯลฯ

3) ชิ้นงาน (tasks) เช่น เรียงความ การอ่านข้อความสั้น ฯลฯ

4) ทักษะที่จะรวมในแบบทดสอบ

5) วิธีการให้คะแนนการทดสอบ

6) วิธีการที่จะรายงานผลการทดสอบแก่ผู้เรียน

In classroom context, test specifications (specs) can serve as an outline of the test – determining what it will look like. Test specs are blueprints of the test, with the following components:

1) a description of the content

2) item types or methods, e.g., multiple-choice, cloze etc.

3) tasks, e.g., written essay, reading a short passage

4) skills to be included in the test

5) how the test will be scored

6) how the score will be reported to the students

ในบริบทห้องเรียน ลักษณะเฉพาะของแบบทดสอบคือไกด์สำหรับการออกแบบแบบทดสอบ ที่จะช่วยจัดการเรื่องต่างๆ เช่นความตรง ได้อย่างมีประสิทธิภาพ ทั้งนี้ ในบริบทการสอบมาตรฐานขนาดใหญ่ ลักษณะเฉพาะของแบบทดสอบจะมีความเป็นทางการและรายละเอียดอีกมาก ทั้งยังมักจะเป็นเอกสารลับเฉพาะ ที่ช่วยให้องค์กรที่ออกข้อสอบสามารถคุมคุณภาพเรื่องความตรงของข้อสอบที่ออกมาต่อๆ กันได้ ส่วนในบริบทห้องเรียน ลักษณะเฉพาะของแบบทดสอบจะไม่ได้เป็นความลับ หากแต่ยังช่วยเตรียมผู้เรียนให้ได้ภาพที่ชัดเจนเรื่องชนิดข้อสอบและชิ้นงานที่ผู้เรียนจะต้องทำในการสอบ ยิ่งครูสามารถให้รายละเอียดเกี่ยวกับขั้นตอนการทดสอบและแบบทดสอบได้มากเท่าใด ก็ยิ่งเพิ่มโอกาสที่ผู้เรียนจะสามารถแสดงออกหรือทำผลงานได้ดีในการทดสอบมากขึ้นเท่านั้น

In classroom contexts, test specs are a guide for designing the test, helping to effectively fulfill principles like validity. In large-scale standardized testing contexts, test specs are often formal and full of details. They are usually confidential too, helping the organization or institution control test qualities such as test validity of the tests that are created subsequentially. In classroom settings, test specs are not confidential and can help prepare the learners for the test regarding the item types and the tasks that they are going to encounter in the test. The more detailed the teachers are in speficying the details of the assessment procedure, the better they are in providing the students with opportunities to perform well in the test.

4.4 เขียนข้อสอบ (Devising test items)

ในการเขียนข้อสอบ เราจะต้องกำหนดลงไปว่า จะเลือกชนิดข้อสอบ (item types) อย่างไร และจะเรียงลำดับข้อสอบแต่ละชนิดอย่างไร โดยชิ้นงานจำเป็นต้องใช้งานได้จริง (Brown & Abeywickrama 2010) และเพื่อให้ข้อสอบมีความตรงด้านเนื้อหา ชิ้นงานในการทดสอบก็ควรสะท้อนชิ้นงานในรายวิชา ในบทเรียน หรือในเนื้อหาท่อนที่ทำการวัด นอกจากนี้ ชิ้นงานควรจะมีความสมจริงด้วย และประเด็นสุดท้ายก็คือ ชิ้นงานควรจะสามารถใช้วัดผลได้อย่างเที่ยงตรงโดยครูผู้สอน หรือผู้ให้คะแนนด้วย

In devising test items, we will need to determine how the item types will be selected and how the separate items will be arranged. The tasks of the test have to be practical (Brown & Abeywickrama 2010). For content validity, the tasks have to be similar to the tasks of the course, the lesson, or the content segment that the learners use. The tasks should be authentic too. And the last key point is that the tasks should be able to assess reliably by the teachers or the scorers.

โดยทั่วไป เรามักเข้าใจกันว่า เมื่อกำหนดเป้าหมายของแบบทดสอบ กำหนดวัตถุประสงค์ในการวัด และเขียนลักษณะเฉพาะของแบบทดสอบแล้ว เราจะสามารถเขียนข้อสอบเสร็จได้ในคราวเดียว ในความเป็นจริง การเขียนข้อสอบมักมีลักษณะวน เพราะเมื่อเราพบเจอปัญหา เราต้องนำมาปรับแก้ในการเขียนข้อสอบใหม่ และนี่เป็นเหตุผลว่าเพราะเหตุใด เราจึงควรเผื่อเวลาในการเขียนข้อสอบให้มาก ต่อไปนี้คือตัวอย่างรูปแบบข้อสอบที่เราสามารถนำมาใช้ในการเขียนข้อสอบกลุ่มฟังพูด

In general, we often believe that we can finish test writing in one go once we have determined the test purpose, and the assessment objectives and have written the test specs. In reality, test design is often involved with several ‘loops’. The reason for this is that we find problems and shortcomings, and we have to correct them in newer test versions. For this very reason, we need to also spare ample time for test writing. The following are some examples of task formats that we could use in listening/speaking tests.

ตารางที่ 4.1 โหมดการดึงพฤติกรรมและการตอบสนองในการสร้างข้อสอบ (Elicatation and response modes in test construction)

โหมดการดึงพฤติกรรม (Elicitation mode):	**การพูด (ผู้เรียนฟัง) (Oral)**	**การเขียน (ผู้เรียนอ่าน) (Written)**
	- คำสั่งในการสอบ (administration directions) - ประโยค (sentence(s)), คำถาม (question) - คำ (word), คู่คำ (pair of words) - การพูดฝ่ายเดียว (monologue), การพูดแบบสุนทรพจน์ (speech) - บทสนทนาที่บันทึกไว้ (prerecorded conversation) - การสนทนาโต้ตอบแบบมีปฏิสัมพันธ์ (interactive dialog)	- คำสั่งในการสอบ (administration directions) - ประโยค (sentence(s)), คำถาม (question) - คำ (word), ชุดคำ (set of words) - ย่อหน้า (paragraph) - เรียงความ (essay), ข้อความที่ตัดมา (excerpt) - เรื่องสั้น (short story), หนังสือ (book)
โหมดการตอบสนอง (Response mode):	**การพูด (Oral)**	**การเขียน (Written)**
	- พูดตาม (repeat) - อ่านออกเสียง (read aloud) - ตอบใช่/ไม่ (yes/no) - ตอบด้วยคำตอบสั้น (short answer) - บรรยาย (describe) - เล่นบทบาทสมมุติ (role play) - พูดฝ่ายเดียว (monologue), พูดสุนทรพจน์ (speech) - การสนทนาโต้ตอบแบบมีปฏิสัมพันธ์ (interactive dialog)	- เลือกตอบในข้อสอบหลายตัวเลือก (mark a multiple-choice option) - เติมคำในช่องว่าง (fill in the blank) - สะกดคำ (spell a word) - ให้คำนิยาม (ในรูปแบบวลี) (define a term (with a phrase)) - ตอบด้วยคำตอบสั้น (2–3 ประโยค) (short answer (2–3 sentences)) - เขียนเรียงความ (essay)

ดังที่ได้แสดงในตารางที่ 4.1 การดึงพฤติกรรมออกมาด้วยคำพูดสามารถเข้ากันได้กับทั้งการตอบด้วยคำพูดและการตอบด้วยการเขียน และเช่นเดียวกัน การดึงพฤติกรรมออกมาด้วยการเขียนก็

สามารถเข้ากันได้กับทั้งการตอบด้วยคำพูดและการตอบด้วยการเขียน กระนั้น ก็ต้องมีการใช้ ‘สามัญสำนึก’ เล็กน้อยในการนำมาใช้ เช่น คงดูไม่เป็นการเหมาะสมนัก หากสิ่งที่ดึงพฤติกรรมเป็นคำคู่เทียบเสียง (“beat, bit”) โดยใช้การตอบสนองเป็นการตอบด้วยใช่หรือไม่ อีกตัวอย่างก็เช่น การใช้ตัวดึงพฤติกรรมเป็นการพูดฝ่ายเดียว โดยให้ตอบสนองด้วยการสะกดคำ เป็นต้น

As displayed in Table 4.1, eliciting behaviors orally could go with both oral responses and written responses. Likewise, written elicitations of behaviors could be used with both oral responses and written responses. However, common sense needs to be exercised when using the information in Table 4.1. For example, it does not seem to be suitable if the eliciting behavior is saying minimal pairs (“beat, bit”), and the response is to say yes or no. Another example for requiring a bit of common sense in using Table 4.1 is when the eliciting behavior is a monologue, and the response is to spell words.

4.5 เขียนข้อสอบหลายตัวเลือก (Designing multiple-choice items)

การออกแบบข้อสอบหลายตัวเลือกนั้นโดยผิวเผินอาจดูเป็นเรื่องง่าย แต่ความจริงแล้วเป็นเรื่องยากมากที่จะออกได้ถูกต้องเหมาะสม (Brown & Abeywickrama 2010) ข้อสอบหลายตัวเลือกมีจุดอ่อนหลายประการ เช่น

1) เทคนิคการเขียนข้อสอบหลายตัวเลือกวัดเฉพาะความรู้แบบรับสาร (recognition knowledge)

2) การเดาคำตอบอาจมีอิทธิพลมากต่อคะแนน

3) การใช้ข้อสอบหลายตัวเลือกทำให้มีข้อจำกัดอย่างมากเรื่องสิ่งที่จะวัดได้

4) เป็นการยากมากที่จะเขียนข้อสอบได้อย่างประสบความสำเร็จ

5) อิทธิพลย้อนกลับเชิงบวกอาจมีน้อย

6) อาจเอื้อต่อการโกงข้อสอบ

Writing multiple-choice items may superficially be easy. In reality, they are actually very difficult to write and design properly (Brown & Abeywickrama 2010). They also have several drawbacks, for example,

1) The multiple-choice item-writing technique tests only recognition knowledge.

2) Guessing may have a big impact on test scores.

3) What can be tested is restricted severely by the technique.

4) Writing successful items is very difficult.

5) Beneficial washback may be minimal.

6) This technique facilitates cheating.

ข้อดีสำคัญสองประการสำหรับการใช้รูปแบบข้อสอบแบบหลายตัวเลือกคือความสามารถใช้งานได้จริง และความเที่ยง การที่มีคำตอบที่ชัดเจนกำหนดไว้แล้วกับกระบวนการตรวจให้คะแนนที่สามารถทำได้อย่างรวดเร็ว รูปแบบข้อสอบหลายตัวเลือกเป็นทางเลือกที่น่าสนใจสำหรับครูผู้สอน คำถามสำคัญจึงคือ คุ้มค่ากับความพยายามหรือไม่ ถ้าวัตถุประสงค์คือเพื่อออกแบบข้อสอบมาตรฐานขนาดใหญ่เพื่อใช้ซ้ำได้ รูปแบบข้อสอบหลายตัวเลือกก็เป็นตัวเลือกที่น่าจะใช้งานได้เหมาะสมจริงๆ

There are two advantages for using multiple-choice formats: practicality and reliability. Multiple-choice formats have predetermined correct responses and time-saving scoring procedures, and so they are a tempting possibility for teachers. But is the preparation time worth the effort? If the teachers' objective is to design a large-scale standardized test for repeated administrations, then the multiple-choice format is indeed a viable option.

ในการออกแบบข้อสอบแบบหลายตัวเลือก เราควรทราบคำศัพท์สำคัญๆ เกี่ยวกับรูปแบบข้อสอบรูปแบบนี้ ได้แก่

1) ข้อสอบแบบหลายตัวเลือกเป็นแบบรับสารให้เลือกตอบ (receptive or selective response) นิยมเรียกว่าเป็นข้อสอบแบบจัดหาคำตอบมาให้ (supply type of response) ข้อสอบแบบรับสารให้เลือกตอบชนิดอื่นๆ รวมไปถึงแบบถูกผิด (true/false) และแบบจับคู่ (matching)

2) ข้อสอบแบบหลายตัวเลือกทุกข้อมีตัวคำถาม (stem) ซึ่งทำหน้าที่เป็นตัวดึงพฤติกรรม และมีหลายตัวเลือก (options or alternatives) ให้เลือก

3) หนึ่งในตัวเลือกเป็นคำตอบที่ถูกต้อง (key) ในขณะที่ตัวเลือกที่เหลือเป็นตัวลวง (distractors)

In designing a multiple-choice test, we should know some terminology used in association to this item format.

1) Multiple-choice items are receptive, or selective response, items in the sense that the test takers choose from a set of responses, rather than creating a response. Choosing from a set of responses is commonly called a supply type of response. Other receptive item types include true/false questions and matching lists.

2) Every multiple-choice item has a stem, which presents the stimulus, and several alternatives or options to choose from.

3) One of the options is the key or the correct response, and the others function as distractors.

ในสถานการณ์ที่การใช้ข้อสอบหลายตัวเลือกมีความเหมาะสม มีข้อควรคำนึงสี่ประการดังต่อไปนี้ในการออกแบบข้อสอบ

1) ออกแบบให้ข้อสอบหนึ่งข้อวัดเพียงหนึ่งวัตถุประสงค์

On occasions when multiple-choice items are appropriate, there are four guidelines for designing multiple-choice items:

1) Design each item to measure a single objective.

Excuse me, do you know ___?

A. where is the post office

B. where the post office is

C. where post office is

จากตัวอย่างข้างต้น ข้อสอบถูกออกแบบมาเพื่อทดสอบความรู้เกี่ยวกับลำดับคำของคำถามแบบอ้อม โดยตัวเลือก A ออกแบบมาเพื่อลวงผู้สอบที่ไม่รู้วิธีเรียงคำของคำถามแบบอ้อม จัดได้เป็นตัวลวงที่มีประสิทธิภาพ แต่ตัวเลือก C มีคำนำหน้านามแบบเฉพาะเจาะจง *the* หายไป จัดเป็นตัวช่วยแบบไม่ตั้งใจ (unintentional clue) เพราะอาจจะทำให้ผู้สอบสามารถตัดตัวเลือก C ได้ในทันที ซึ่งในการตัดตัวเลือก C นี้ ความรู้เกี่ยวกับลำดับคำของคำถามแบบอ้อมไม่ได้ถูกวัดเลย จากตัวอย่างนี้ จึงเป็นข้อสอบที่มิได้วัดเพียงหนึ่งวัตถุประสงค์ สมควรปรับแก้ตัวเลือก C ก่อนนำมาใช้

From the example above, the item is designed to test recognition of the correct word order in indirect questions. Option A is designed to lure the students who do not know how to frame indirect questions, and so could be considered an efficient distractor. By contrast, Option C does not have the definite article *the*. This presents an unintentional clue because it may help the students to eliminate Option C immediately. In this process of option elimination, no assessment is made of knowledge of word order in indirect questions. From this example, the item does not test only one objective, and Option C should be corrected before the item is actually used.

2) ระบุตัวคำถามและตัวเลือกให้ชัดเจนตรงไปตรงมาที่สุดเท่าที่จะเป็นไปได้

2) State both stem and options as directly and simply as possible.

My eyesight has really been deteriorating lately. I wonder if I need glasses. I think I'd better go to the ___ to have my eyes checked.

A. pediatrician

B. dermatologist

C. optometrist

บางครั้งเราทำให้ข้อสอบหลายตัวเลือกมีคำมากเกินจำเป็น กฎพื้นฐานในการเขียนคือ การเขียนเข้าประเด็นที่ต้องการ จากตัวอย่างข้างบน สองประโยคแรกถือว่าไม่จำเป็นหากว่าเราต้องการเพียงแค่ให้ผู้สอบระบุประเภทของบุคลากรทางการแพทย์ที่เกี่ยวกับปัญหาด้านสายตา นอกจากนี้ ในการเขียนบริบทที่เพิ่มขึ้น เราได้นำคำศัพท์ที่อาจทำให้ผู้สอบไขว้เขว คือคำว่า deteriorating เข้ามาด้วย

Sometimes we make multiple-choice items too wordy. A rule of thumb is to get directly to the point. From the example above, the first two sentences are unnecessary if we simply want the students to identify the type of medical professional who deals with eyesight problems. Furthermore, in the lengthened stem, we have introduced a potentially confounding word, *deteriorating*, which could distract the students unnecessarily.

อีกกฎเพื่อความกระชับชัดเจนคือการนำความซ้ำซ้อนออกไปจากตัวเลือก ในตัวอย่างข้างล่าง “which were” ถูกใช้ซ้ำในทั้งสามตัวเลือก วลีนี้จึงควรอยู่ในตัวคำถามเพื่อให้ข้อสอบออกมากระชับที่สุด

Another rule for succinctness is to remove redundancy from the options. In the example that follows, “which were” is repeated in all the three options. The phrase should thus be in the stem in order that the item would be succinct.

We went to visit the temples, ___ fascinating.

A. which were beautiful

B. which were especially

C. which were holy

3) ทำให้แน่ใจว่า คำตอบที่ต้องการเป็นคำตอบที่ถูกต้องตัวเลือกเดียวจริงๆ

3) Make sure that the intended answer is clearly the only correct one.

Test takers hear:	Where did George go after the party last night?
Test takers read:	A. Yes, he did.
	B. because he was tired
	C. to Elaine's place for another party
	D. He went home around eleven o'clock.

วัตถุประสงค์ของข้อสอบข้างบนคือวัดเรื่องคำถามที่ขึ้นต้นด้วย *wh-* (*wh-* questions) โดยตัวลวง A ออกแบบมาเพื่อให้แน่ใจว่าผู้สอบรู้ความแตกต่างระหว่างการตอบคำถามที่ขึ้นต้นด้วย *wh-* กับคำถามที่ตอบว่าใช่หรือไม่ (yes/no question) ส่วนตัวลวง B และคำตอบ C ทดสอบความเข้าใจความหมายของคำว่า *where* ในคำถาม อย่างไรก็ดี ตัวลวง D กลับเป็นคำตอบที่อาจจะถูกต้องได้ด้วย เพราะมีคำว่า *home* การกำจัดคำตอบที่ไม่ได้ตั้งใจที่อาจเป็นไปได้มักเป็นงานยากที่สุดในการออกแบบข้อสอบแบบหลายตัวเลือก

The objective of the multiple-choice item above is to test a *wh-* question. Distractor A is designed to ascertain that the students know the difference between an answer to a *wh-* question and an answer to a yes/no question. Distractor B and the key C test comprehension of the meaning of *where* as opposed to *why*. However, Distractor D may become a plausible answer because it has the mention of "home". Eliminating

unintended possible answers is often the most difficult problem of designing multiple-choice items.

4) (อาจจะไม่ใช้ก็ได้) ใช้ค่าดัชนีข้อสอบในการยอมรับ ตัดทิ้ง หรือแก้ไขข้อสอบ

4) (Optional) Use item indices to accept, discard, or revise items.

เมื่อใช้หลักเกณฑ์สามข้อแรก ในการตรวจคุณภาพข้อสอบแบบหลายตัวเลือก ข้อสอบที่ได้ก็น่าจะมีคุณภาพดีเพียงพอสำหรับการใช้งานในบริบทห้องเรียน แต่หากต้องการทำอะไรมากกว่านี้ ก็อาจใช้ค่าดัชนีข้อสอบเพื่อเกลาข้อสอบที่ออกแบบออกมา ค่าดัชนีข้อสอบที่นิยมใช้กันได้แก่ ค่าความยาก (item facility) ค่าอำนาจจำแนก (item discrimination) และค่าความมีประสิทธิภาพของตัวลวง (distractor efficiency) ในบทที่ 6 จะกล่าวถึงวิธีวิเคราะห์ค่าดัชนีข้อสอบเหล่านี้อย่างละเอียด ค่าดัชนีเหล่านี้ใช้ประกอบการพิจารณาในการยอมรับคือเก็บข้อสอบข้อใดไว้ใช้อีก ทิ้งข้อสอบข้อใดไม่นำมาใช้อีก และแก้ไขเพื่อปรับปรุงให้ข้อสอบดีขึ้น

When using the first three guidelines in checking the quality of multiple-choice items, the test items derived should be sufficiently good for classroom testing. But if you wish to take one step further, you may use item indices to refine the item design. Item indices that are widely used are item facility, item discrimination, and distractor efficiency. In Chapter 6, how to calculate these indices will be dealt with in detail. These indices are used for informing the decision-making processes for accepting, discarding, and revising items.

เมื่อจัดทำข้อสอบหลายตัวเลือกแล้วเสร็จ เราก็อาจปรับแก้ฉบับร่างของข้อสอบ ต่อไปนี้คือคำถามที่เราอาจใช้เพื่อการปรับแก้ฉบับร่างของข้อสอบ

1) คำสั่งของแต่ละส่วนของข้อสอบชัดเจนหรือไม่

2) มีข้อสอบตัวอย่างในแต่ละส่วนของข้อสอบหรือไม่ ถ้าไม่มีข้อสอบตัวอย่าง คำสั่งและรูปแบบข้อสอบเป็นที่คุ้นเคยสำหรับผู้เรียนมากอยู่แล้วใช่หรือไม่

3) ข้อสอบแต่ละข้อวัดวัตถุประสงค์ที่ระบุไว้ชัดหรือไม่

4) มีคำตอบที่ถูกต้องเพียงตัวเลือกเดียวสำหรับแต่ละคำถามใช่หรือไม่

5) ข้อสอบแต่ละข้อนำเสนอด้วยภาษาที่ชัดเจนและเรียบง่ายใช่หรือไม่

6) ข้อสอบหลายตัวเลือกแต่ละข้อมีตัวลวงที่เหมาะสมหรือไม่ ตัวลวงที่เหมาะสมคือผิดอย่างแน่นอน แต่ก็ “น่าสนใจ” มากพอที่จะไม่ง่ายจนเกินไป

7) ความยากของข้อสอบแต่ละข้อเหมาะสมกับผู้เรียนหรือไม่

8) ภาษาที่ใช้ในข้อสอบแต่ละข้อมีความสมจริงมากพอหรือไม่

9) มีความสมดุลระหว่างข้อง่ายและข้อยากหรือไม่

10) ข้อสอบโดยภาพรวมและแบบสอบทั้งฉบับสะท้อนวัตถุประสงค์การเรียนรู้ได้อย่างเพียงพอหรือไม่

When writing a multiple-choice test is completed, we may revise the draft. The following are some guidelines that we may use for revising the draft:

1) Are the directions to each examination section clear?

2) Is there an example item for each section? If not, are the directions and format very familiar to the students?

3) Does each item measure a specified objective?

4) Is there a single correct answer for each question?

5) Is each item stated in clear and simple language?

6) Does each multiple-choice item have appropriate distractors? Appropriate distractors are those that are clearly wrong but sufficiently alluring.

7) Is the difficulty of each item suitable for the students?

8) Is the language of each item sufficiently authentic?

9) Is there a balance between easy and difficult items?

10) Do the sum of the items and the whole test adequately reflect the learning objectives?

4.6 จัดสอบ (Administering the test)

ในการจัดสอบ มีคำถามสำคัญคือ มีรายละเอียดอะไรบ้างที่ควรจะต้องดูแล เพื่อช่วยให้ผู้เรียนทำข้อสอบได้ผลลัพธ์ที่สะท้อนความสามารถได้ดีที่สุด เมื่อข้อสอบถูกสร้างขึ้นมาพร้อมแล้ว ผู้เรียนควรจะรู้สึกพร้อมที่จะแสดงความสามารถ แบบทดสอบที่ดีอาจล้มเหลวที่จะบรรลุจุดมุ่งหมาย หากสภาพการณ์เงื่อนไขต่างๆ สำหรับการสอบมิได้มีการเตรียมไว้ดีพอ ตัวอย่างเช่น เราจะลดความวิตกกังวลในการสอบโดยไม่จำเป็นได้อย่างไร เราจะเพิ่มความมั่นใจในตนเองของผู้สอบได้อย่างไร และเราจะช่วยให้ผู้เรียนเห็นว่า การทดสอบเป็นโอกาสหนึ่งในการเรียนรู้ได้อย่างไร (Brown & Abeywickrama 2010)

In administering a test, a key question to ask is what details one should attend to in order to help the students achieve optimal performance. When the test has been written and is ready to administer, the students should feel well-prepapred for it. An otherwise good test may fail to achieve its goal if the conditions for taking the test are inadequately prepared. For example, how could we reduce unnecessary anxiety in the students? How could we raise their self-confidence for the test? And how could we make them see the test as an opportunity to learn? (Brown & Abeywickrama 2010)

ต่อไปนี้เป็นคำแนะนำบางส่วน โดยข้อ 1)–2) เป็นคำแนะนำก่อนวันสอบ ส่วนข้อ 3)–9) เป็นคำแนะนำทั่วไปในวันจัดสอบ รายการคำแนะนำเหล่านี้มิได้ครอบคลุมทุกสถานการณ์ หากแต่ทำหน้าที่เป็นรายการตั้งต้นในการจัดทำให้ครอบคลุมรายละเอียดของแต่ละบริบท

ข้อควรคำนึงก่อนการสอบ

1) ให้รายละเอียดก่อนการสอบที่เหมาะสมเกี่ยวกับ

ก) เงื่อนไขข้อกำหนดสำหรับการสอบ (กำหนดเวลา การห้ามใช้อุปกรณ์อิเล็กทรอนิกส์ การพัก ฯลฯ)

ข) สิ่งที่ผู้เรียนควรนำติดตัวเข้าสอบด้วย

ค) ชนิดข้อสอบที่จะอยู่ในการทดสอบ

ง) คำแนะนำเรื่องกลวิธีให้ทำข้อสอบได้ดีที่สุด

จ) เกณฑ์การประเมินผล (ตารางเกณฑ์การประเมิน ตัวอย่างการเปรียบเทียบสมรรถนะ)

2) ให้โอกาสผู้เรียนได้สอบถาม หากมีคำถามใดๆ และให้คำตอบให้ชัดเจน

<u>สิ่งที่พึงทำในวันสอบ</u>

3) มาถึงห้องสอบก่อนเวลา และตรวจตราดูความเรียบร้อยของห้องสอบ (แสงสว่าง อุณหภูมิ นาฬิกาที่ทุกคนเห็นได้ชัด การจัดโต๊ะเก้าอี้ ฯลฯ)

4) ถ้าไฟล์เสียง ไฟล์วีดิทัศน์ หรือเทคโนโลยีอื่นจำเป็นสำหรับการสอบ ให้ทดสอบล่วงหน้าก่อนการสอบ

5) เตรียมกระดาษ เครื่องเขียนสำรอง หรือสิ่งอื่นๆ ที่จำเป็นสำหรับการตอบข้อสอบให้พร้อม

6) แจกข้อสอบ

7) เริ่มการสอบตรงเวลา

8) นั่งอยู่เงียบๆ ที่โต๊ะครู พร้อมสำหรับคำถามจากผู้สอบขณะที่การสอบดำเนินไป

9) สำหรับการสอบที่มีกำหนดเวลา เตือนผู้สอบเมื่อเวลาใกล้หมด และสนับสนุนให้ผู้สอบทำข้อสอบ

The following are some guidelines for administering a test. Nos. 1–2 are guidelines intended before the testing day. Nos. 3–9 are those for the testing day. The list here is not exhaustive because it does not cover all possible testing situations. The list functions as a starting point for working on different contexts.

<u>Considerations before the testing day</u>

1) Provide appropriate information before the test about:

a) conditions for the test (time limits, no portable electronics, breaks etc.)

b) the materials that students should bring with them

c) the kinds of items (or item types) that will be on the test

d) suggestions of strategies for optimal performance

e) evaluation criteria (rubrics, benchmark samples)

2) Give the students an opportunity to ask any questions, and provide responses.

Test administration details

3) Arrive early and make sure that the classroom conditions (lighting, temperature, a clock that everyone can see, furniture arrangement etc.) are suitable for an examination.

4) If audio, video, or other technology is necessary for the test administration, try it out beforehand.

5) Prepare extra paper, writing instruments, or other materials for response.

6) Distribute the test.

7) Start on time.

8) Remain quietly seated at the teacher's desk, available for questions from the students.

9) In a timed test, warn students when time is running out, and encourage their completion.

4.7 ให้คะแนน ตัดเกรด และให้ข้อมูลสะท้อนกลับ (Scoring, grading, and giving feedback)

ในการให้คะแนน การตัดเกรด และการให้ข้อมูลสะท้อนกลับ มีคำถามสำคัญคือ การให้คะแนน การตัดเกรด และการให้ข้อมูลสะท้อนกลับเป็นไปในรูปแบบใด รูปแบบของข้อมูลสะท้อนกลับเกี่ยวกับการทดสอบมีความแตกต่างกันออกไป ขึ้นอยู่กับเป้าหมายในการวัดผล ในทุกการทดสอบ รูปแบบการรายงานผลการทดสอบก็มีความสำคัญ ในบางสภาพการณ์การตัดเกรดเป็นอักษรตัวเดียวหรือคะแนนแบบภาพรวมอาจมีความเหมาะสม ในสถานการณ์อื่นครูผู้สอนอาจต้องให้ข้อมูลสะท้อนกลับอย่างเป็นชิ้นเป็นอัน เพื่อให้เกิดอิทธิพลย้อนกลับเชิงบวก (Brown & Abeywickrama 2010)

In scoring, grading, and giving feedback, there is a key question to ask: What kind of scoring, grading, and feedback is expected? The form for test feedback can vary, depending on the purpose of assessment. In every test, how the test results are reported is also important. In some circumstances, a holistic score or a single letter grade may be appropriate. In other circumstances, the teacher may be required to write substantial feedback for positive washback effects on the students' learning (Brown & Abeywickrama 2010).

การให้คะแนน

ในขณะที่ออกแบบการทดสอบ ครูผู้สอนต้องคำนึงถึงว่า การสอบจะให้คะแนน และตัดเกรดอย่างไร แผนงานการให้คะแนนจะสะท้อนถึงค่าน้ำหนักที่เรามีให้ข้อสอบแต่ละส่วน ในฐานะครูผู้สอน หลังจากจัดสอบไปครั้งหนึ่ง เราอาจตัดสินใจปรับแผนการให้คะแนนสำหรับวิชานั้นๆ ในครั้งต่อไปที่เราสอน ณ ตอนนั้น เราจะมีข้อมูลที่มีค่าว่า แบบทดสอบยากหรือง่ายอย่างไร เวลาที่ใช้สอบเหมาะสมหรือไม่ ปฏิกิริยาจากผู้เรียนเป็นอย่างไร และผลงานในภาพรวมเป็นอย่างไร ตลอดจนข้อมูลที่ว่า แบบทดสอบได้วัดผู้เรียนได้อย่างถูกต้องหรือไม่ ครูควรจดบันทึกข้อมูลเหล่านี้ไว้ แม้ว่าจะไม่ใช่ข้อมูลเชิงประจักษ์ เพราะจะสามารถนำมาใช้ได้ในเทอมถัดไป

Scoring

In designing a test, the teachers have to consider how the test will be scored and graded. Their scoring plan will reflect the weight they give to each section of the test items. As a teacher, after administering a test once, we may decide to revise the scoring plan for the course the next time when we teach it. We would have valuable information then as to whether the course test is easy or difficult, the time limit is suitable or not, what the students' affective reaction to the test is, and the general performance. Valuable information includes whether the test has correctly assessed the students. The teachers should write down all this information, although it may not be empirical, because it will be useful for another term when we teach the course again.

การตัดเกรด

ในการตัดเกรด มีสองวิธีที่เราใช้กัน ได้แก่ การตัดเกรดแบบสัมบูรณ์ และการตัดเกรดแบบสัมพัทธ์ ตารางที่ 4.2 แสดงตัวอย่างการตัดเกรดแบบสัมบูรณ์ ที่ครูผู้สอนจะต้องระบุมาตรฐานผลงานหรือพฤติกรรมการแสดงออกไว้ล่วงหน้าอย่างละเอียด ตารางที่ 4.3 แสดงตัวอย่างการตัดเกรดแบบสัมพัทธ์ ที่ยอมให้มีการตีความเพิ่มเติมและปรับใช้ให้เข้ากับความยากง่ายของแบบสอบที่อาจมิได้คาดการณ์ไว้ เมื่อเทียบทั้งสองวิธี การตัดเกรดแบบสัมพัทธ์ใช้กันบ่อยกว่ามาก

Grading

In grading, there are two main approaches that we may use. They are absolute grading and relative grading. Table 4.2 shows an example of absolute grading, in which the teachers need to prespecify the standards of performance on a numerical point system in detail. Table 4.3 shows examples of relative grading, in which the teachers' interpretation and adjustment for unpredicted ease or difficulty of a test are allowed. When compared the two approaches, relative grading is used much more often.

ตารางที่ 4.2 สเกลการตัดเกรดแบบสัมบูรณ์ (Absolute grading scale)

เกรด (Grade):	การสอบกลางภาค (Midterm exam) (50 คะแนน (points))	การสอบปลายภาค (Final exam) (100 คะแนน (points))	งานอื่น (Other performance) (50 คะแนน (points))	คะแนนรวมทั้งหมด (Total no. of points) (200 คะแนน (points))
A	45–50	90–100	45–50	180–200
B	40–44	80–89	40–44	160–179
C	35–39	70–79	35–39	140–159
D	30–34	60–69	30–34	120–139
F	ต่ำกว่า 30 (below 30)	ต่ำกว่า 60 (below 60)	ต่ำกว่า 30 (below 30)	ต่ำกว่า 120 (below 120)

ตารางที่ 4.3 สเกลการตัดเกรดแบบสัมพัทธ์ (Relative grading scale)

	เปอร์เซ็นต์ผู้เรียน (Percentage of students)		
เกรด (Grade):	สถาบัน X (Institution X)	สถาบัน Y (Institution Y)	สถาบัน Z (Institution Z)
A	~15%	~30%	~60%
B	~30%	~40%	~30%
C	~40%	~20%	~10%
D	~10%	~9%	
F	~5%	~1%	

ในการใช้สเกลการตัดเกรดแบบสัมพัทธ์ อาจทำได้โดยตัดเกรดตามการกระจายตัวของคะแนนแบบโค้งปกติ (normal bell curve) วิธีนี้อาจกำหนดลงไปว่า เกรด A คือผู้เรียนที่ได้ 10% แรก เกรด B คือผู้เรียนที่ได้ 20% ถัดไป เกรด C คือผู้เรียนที่ได้ 40% ตรงกลาง เกรด D คือผู้เรียนที่ได้ 20% ถัดไป เกรด F คือผู้เรียนที่ได้ต่ำสุด 10% ในความเป็นจริง มักไม่ค่อยมีใครใช้การตีความเช่นนี้ เพราะมีความจำกัดตายตัวเกินไป และไม่ได้รวมการตีความผลการสอบวัดผลสัมฤทธิ์ทางการเรียนเข้ามาด้วยอย่างเหมาะสม

In using a relative grading scale, the teachers can simulate a normal “bell curve” distribution to assign grades. For example, A would be for the top 10 percent, B the next 20 percent, C the middle 40 percent, D the next 20 percent, and F the lowest 10 percent. In reality, however, almost no one adheres to an interpretation like this because it is too restrictive, and does not appropriately interpret achievement test results in classrooms.

นอกเหนือจากการใช้การกระจายตัวของคะแนนแบบโค้งปกติ อีกวิธีในการใช้การตัดเกรดแบบสัมพัทธ์คือ การเลือกใช้เปอร์เซ็นต์ไทล์ตามความคาดหวังของสถาบันดังได้แสดงตัวอย่างในตารางที่ 4.3 ในสถาบัน X ความคาดหวังคือการกระจายตัวของคะแนนที่เบ้ไปทางขวาเล็กน้อย (มีความถี่มากในช่วงระดับบน) เมื่อเทียบกับแบบโค้งปกติ ส่วนในสถาบัน Y แทบจะไม่มีใครตกในรายวิชา และผู้เรียนส่วนใหญ่ได้เกรด A และ B ในสถาบัน Y นี้ การกระจายตัวของคะแนนยิ่งเบ้ไปทางขวาอย่างเห็นได้ชัด และในสถาบัน

Z อาจเรียกได้ว่าเป็นตัวแทนของโปรแกรมบัณฑิตศึกษา เกรด C จัดว่าเป็นเกรดตก เกรด B เป็นเกรดพอใช้ และเกรด A เป็นเกรดเป้าหมายที่คาดหวังสำหรับผู้เรียนส่วนใหญ่

Apart from using a normal distribution for the scores, another way of using relative grading is to preselect percentiles in accordance with an institutional expectation, as exemplified in Table 4.3. In Institution X, a curve that is slightly skewed to the right (higher frequencies in the upper levels) is expected, compared with a normal bell curve. In Institution Y, virtually no one would fail a course, and most of the students would achieve As and Bs. In this Institution Y, the score distribution curve is even more skewed to the right. In Institution Z, which would be representative of a postgraduate program, grade C is considered a failing grade. Grade B is an acceptable grade, and Grade A an expected grade for most students.

ในการตัดเกรดแบบสัมพัทธ์ เราอาจคำนวณเกรดภายหลังจากการได้เห็นผลงานและพฤติกรรมการแสดงออกแล้ว การเลือกใช้วิธีนี้จะทำให้เราสามารถปรับการตัดเกรดให้เข้ากับระดับความยากของการทดสอบ รวมไปถึงปรับให้เข้ากับหลักการและแนวคิด เช่นในการให้เกรด A ทั้งยังสามารถปรับให้เหมาะสมกับความทุ่มเทพยายาม หรือการไม่ทุ่มเทให้การทดสอบของผู้เรียนทั้งกลุ่มเก่งและกลุ่มไม่เก่งได้

In relative grading, we may calculate grades *a posteriori* after we observe the performances. In doing so, we could adjust for the difficulty of the test, incorporate our own philosophical objection or agreement to awarding an A, for example, and can support our intuition that the students, strong and weak, took or really did not take seriously their mandate to prepare well for the test.

โดยสรุป วิธีที่เราให้เกรดเป็นผลลัพธ์ของปัจจัยต่อไปนี้

- ประเทศ กรอบวัฒนธรรม และบริบทของห้องเรียนภาษาอังกฤษ
- ความคาดหวังขององค์กร ซึ่งมักไม่ได้มีการเขียนไว้
- นิยามทั้งที่เห็นได้ชัด และที่แฝงอยู่ของเกรด ที่เราได้ตั้งไว้
- ความสัมพันธ์ที่เรามีต่อทั้งชั้นเรียน

- ความคาดหวังของผู้เรียนที่เกิดขึ้นจากการสอบและการสอบย่อยก่อนหน้า

In grading, how we assign letter grades to one particular test is a product of

- the country, culture, and context of the English classroom
- institutional expectations (most of them unwritten)
- explicit and implicit definitions of grades that we have explained
- the relationship we have established with the class
- student expectations that have been caused in previous tests and quizzes in the class

การให้ข้อมูลสะท้อนกลับ

ข้อมูลสะท้อนกลับที่เราต้องการได้แก่ ข้อมูลสะท้อนกลับที่ก่อให้เกิดอิทธิพลย้อนกลับเชิงบวก (beneficial washback) ต่อไปนี้เป็นตัวอย่างของข้อมูลสะท้อนกลับที่เกี่ยวข้องกับการทดสอบ

Giving feedback

The feedback that we want is the feedback that create beneficial washback. The following are some examples of the feedback that is related to tests.

การให้คะแนนหรือการตัดเกรดสำหรับการทดสอบ

1) เกรดเป็นตัวอักษร

2) คะแนนรวม

3) คะแนนส่วนย่อย เช่น ของแต่ละทักษะที่วัดหรือของข้อสอบแต่ละส่วน

Scoring or grading for a test

1) a letter grade

2) a total score

3) subscores, e.g., of separate skills or sections of a test

สำหรับการตอบสนองต่อข้อสอบการฟังและข้อสอบการอ่าน

1) การบ่งบอกว่าคำตอบผิดหรือถูก

2) ชุดคะแนนวินิจฉัย เช่น คะแนนเกี่ยวกับเรื่องไวยากรณ์บางเรื่อง

3) แบบรายการตรวจสอบ ที่มีด้านที่ต้องการการอธิบาย

For responses to listening and reading items

1) indication of correct/incorrect responses

2) diagnostic set of scores, e.g., scores on certain grammatical categories

3) checklist of areas needing work

สำหรับการทดสอบส่งสารด้วยการพูด

1) คะแนนของแต่ละองค์ประกอบที่กำลังให้คะแนน

2) แบบรายการตรวจสอบ ที่มีด้านที่ต้องการการอธิบาย

3) การให้ข้อมูลย้อนกลับด้วยวาจา หลังการทดสอบ

4) การนัดพบพูดคุยหลังการสัมภาษณ์ เพื่อคุยเกี่ยวกับผลการทดสอบ

For oral production tests

1) scores for each element being rated

2) checklist of areas that need work

3) oral feedback after performance

4) post-interview conference to go over the results

สำหรับการทดสอบด้วยการเขียนเรียงความ

1) คะแนนของแต่ละองค์ประกอบที่กำลังให้คะแนน

2) แบบรายการตรวจสอบ ที่มีด้านที่ต้องการการอธิบาย และเทคนิคที่แนะนำสำหรับการปรับปรุงการเขียน

3) ความคิดเห็น คำแนะนำ ที่เขียนไว้ที่ขอบกระดาษและตอนท้ายของเรียงความ

4) การนัดพบพูดคุยหลังการทดสอบ เพื่อพูดคุยเกี่ยวกับเรียงความ

For written essays

1) scores for each element being rated

2) checklist for areas that need work, and suggested techniques for improving writing

3) marginal and end-of-essay comments and suggestions

4) post-test conference to go over work

ข้อมูลสะท้อนกลับสำหรับการทดสอบ

1) เกี่ยวกับส่วนหนึ่งของข้อสอบหรือทั้งฉบับข้อสอบ การพูดคุยกันระหว่างเพื่อนร่วมชั้นเกี่ยวกับผลสอบ

2) การพูดคุยกับผู้เรียนรายคนเกี่ยวกับการทวนการทดสอบทั้งฉบับ

3) การพูดคุยกันในชั้นเรียนเกี่ยวกับผลของการทดสอบ

4) การประเมินตนเองในหลากหลายสถานการณ์

Additional feedback for a test

1) on all or selected parts of a test; peer conferences on results

2) individual conferences with each student to review a complete test

3) whole-class discussion of results of the test

4) self-assessment in various manifestations

4.8 Exercise

1. Which is the first step in designing and writing tests?

A. administering a test

B. designing clear objectives

C. determining the test purpose

D. writing test items

2. Which is the first step in designing the test objectives?

A. checking out the class objectives

B. drawing up test specs

C. scoring test items

D. writing the test objectives

3. Which is **NOT** included in the test spec?

A. content description

B. item types

C. scoring

D. All of the above are included.

4. Which is usually confidential in large-scale standardized testing contexts?

A. item types

B. test administration

C. test objective

D. test spec

5. In writing multiple-choice test items, which is **NOT** a good practice?

A. Design each item to measure multiple objectives.

B. Make sure the intended answer is the only correct one.

C. State the item stem simply

D. State the options as directly as possible

Answer key: 1. C 2. A 3. D 4. D 5. A

บทที่ 5 การวัดทักษะทางภาษาทั้งสี่

Chapter 5 Assessing four language skills

ในบทที่ 4 ได้นำเสนอแนวทางการออกแบบและสร้างแบบทดสอบ ในบทนี้จะกล่าวถึงการวัดทักษะทางภาษาทั้งสี่ โดยมีลำดับการนำเสนอดังนี้

5.1 การบูรณาการทักษะในการวัดและประเมินผลทางภาษา

5.2 ทักษะย่อยทักษะหลักในการฟัง

5.3 การออกแบบชิ้นงานการประเมินผลด้านการฟัง

5.4 ทักษะย่อยทักษะหลักในการพูด

5.5 การออกแบบชิ้นงานการประเมินผลด้านการพูด

5.6 ทักษะย่อยทักษะหลักในการอ่าน

5.7 การออกแบบชิ้นงานการประเมินผลด้านการอ่าน

5.8 ทักษะย่อยทักษะหลักในการเขียน

5.9 การออกแบบชิ้นงานการประเมินผลด้านการเขียน

5.10 Exercise (แบบฝึกหัดท้ายบท)

In Chapter 4, designing and writing tests has been discussed. In this chapter, assessing the four skills of language use will be presented. The order of presentation is as follows:

5.1 Integration of skills in language assessment

5.2 Micro- and macroskills of listening

5.3 Designing listening assessment tasks

5.4 Micro- and macroskills of speaking

5.5 Designing speaking assessment tasks

5.6 Micro- and macroskills of reading

5.7 Designing reading assessment tasks

5.8 Micro- and macroskills of writing

5.9 Designing writing assessment tasks

5.10 Exercise

5.1 การบูรณาการทักษะในการวัดและประเมินผลทางภาษา (Integration of Skills in Language Assessment)

ในโลกความเป็นจริง เราอาจใช้ทักษะทั้งสี่คือ ฟัง พูด อ่าน เขียน แยกกันได้ เช่น ฟังวิทยุ กล่าวสุนทรพจน์ อ่านหนังสือ หรือเขียนจดหมาย (Brown & Abeywickrama 2010) แต่กระนั้นสัดส่วนของเวลาส่วนใหญ่ เราใช้ทักษะแบบบูรณาการ เช่นการสนทนาก็ต้องมีทั้งการฟังและการพูด การใช้คอมพิวเตอร์พิมพ์จดหมายก็ต้องมีทั้งการอ่านและการเขียน ดังนั้น ในการวัดและประเมินผลทางภาษา การบูรณาการทักษะจึงเป็นเรื่องที่มีความสมจริง (authentic) การแบ่งทักษะทั้งสี่ในบทนี้จึงเป็นไปเพื่อความชัดเจนในการนำเสนอเป็นหลัก มิได้มุ่งบ่งชี้ว่า ทักษะเหล่านี้ควรวัดผลเป็นอิสระแยกจากกัน

In the real world, we may use four language skills separately. For example, we listen to the radio, deliver a speech, read a book, or write a letter (Brown & Abeywickrama 2010). Nonetheless, for most of our time, we integrate our language skills. For example, in conversations, we need to both listen and speak. When we type letters on the computer, we need to both read and write. Accordingly, in language assessment, skill integration is authentic. Skill division in this chapter, thus, is for clarity in content presentation, not meant for any argument that the skills should be separated in language assessment.

เช่นเดียวกัน การทดสอบไวยากรณ์หรือคำศัพท์ก็ไม่สามารถแยกเด็ดขาดออกจากการใช้ทักษะการฟัง ทักษะการพูด ทักษะการอ่าน และทักษะการเขียน เช่น แบบทดสอบไวยากรณ์ก็ต้องมีการฟัง การพูด การอ่าน และ/หรือการเขียน เพื่อทำความเข้าใจโจทย์และสามารถตอบสนองต่อคำถามได้ ในบทนี้จึงจะมิได้แยกการทดสอบไวยากรณ์หรือการทดสอบคำศัพท์ออกมาเป็นการเฉพาะ เพื่อเน้นย้ำถึงความเกี่ยวพันที่ไม่สามารถแยกได้ระหว่างรูป (ไวยากรณ์และคำศัพท์) และความหมาย (การวัดทักษะทั้งสี่)

Likewise, testing grammar and vocabulary cannot be separated completely from use of listening skills, speaking skills, reading skills, and writing skills. For example,

grammar tests need the four skills for question comprehension and responses to questions. In this chapter, testing grammar and vocabulary is not separated for the purpose, in order to emphasize the inextricable relationship of form (grammar and vocabulary) and meaning (assessing the four skills).

อีกประเด็นในเรื่องเกี่ยวกับทักษะทั้งสี่คือเรื่องการสังเกตการแสดงออกของทักษะทั้งสี่ การแสดงความสามารถใดๆ ออกมาต้องอาศัยสามัตถิยะ (competence) ที่อยู่ข้างใน เมื่อเราเสนอว่าเราจะวัดความสามารถของผู้เรียนในทักษะด้านหนึ่งหรือหลายด้านประกอบกัน เราวัดสามัตถิยะแต่เราสังเกตการแสดงความสามารถออกมา ซึ่งการแสดงออกบางครั้งก็มิได้สะท้อนสามัตถิยะที่แท้จริง อันเป็นผลจากการพักผ่อนไม่เพียงพอ ความเจ็บป่วย ความวิตกกังวลเรื่องการสอบ ฯลฯ ทำให้เป็นการวัดสามัตถิยะที่ไม่เที่ยงตรง

Another issue in relation to the four basic skills is the observation of the performance of the four skills. The learners perform any tasks using the underlying competence that is inside them. When we offer to assess the competence, we in fact observe the learners' performance. Performance, however, sometimes cannot reflect true competence due to insufficient sleep, illness, test anxiety etc., making the competence measurement unreliable.

จากประเด็นเรื่องการวัดสามัตถิยะที่อาจไม่เที่ยงตรงจากหลายๆ ปัจจัย เราจึงควรใช้การตรวจสอบสามเส้า (triangulation) ในการวัดและประเมินผล กล่าวคือ พิจารณาใช้ข้อมูลการแสดงออกอย่างน้อยสองครั้ง และ/หรือสองสถานการณ์ในการตัดสินผลการวัด โดยอาจมาในรูปการทดสอบหลายครั้ง การใช้ชิ้นงานหลายแบบเพื่อให้เหมาะกับสไตล์การเรียนรู้และตัวแปรการแสดงออก การใช้งานให้คะแนนทั้งจากในห้องเรียนและนอกห้องเรียน รวมไปถึงการใช้การประเมินผลทางเลือกประกอบควบคู่กันไปด้วย

In light of the possibility of unreliable assessment of competence, we should triangulate the measurements. This is usuallly done by considering at least two (or more) performances and/or contexts before the teachers draw a conclusion. The triangulation may mean a) several tests, b) a single test with multiple test tasks to account for learning styles and performance variables, c) in-class and extra-class graded work, and d) alternative forms of assessment.

ในการใช้ข้อมูลการแสดงออก เราต้องอิงกับข้อมูลที่วัดได้สังเกตได้เป็นหลัก การวัดได้สังเกตได้หมายถึง สามารถที่จะเห็นหรือได้ยินการแสดงออกของผู้เรียน เมื่อพิเคราะห์ในบริบทของทักษะทั้งสี่ เราไม่สามารถสังเกตเห็นทั้งกระบวนการและผลผลิตของทักษะรับสาร (receptive skills) ได้แก่ทักษะการฟังและทักษะการอ่าน ดังแสดงในตารางที่ 5.1 ตัวอย่างเช่น เราอาจจะกล่าวได้ว่า เรา*เห็น*ผู้เรียนคนหนึ่งกำลังฟังอยู่ เพราะเธอพยักหน้าและยิ้ม แล้วก็ถามคำถามที่เกี่ยวข้อง แต่ความเป็นจริงแล้ว เรากำลังสังเกตเห็น*ผลลัพธ์*ของการฟัง เราไม่สามารถสังเกตเห็นการฟังได้มากไปกว่าการที่เราเห็นลมพัดกระบวนการการแสดงออกซึ่งการฟังเป็นกระบวนการที่ไม่สามารถมองเห็นได้ของการซึมซับความหมายจากสัญญาณเสียงที่ถูกถ่ายทอดไปยังหูและสมอง ถึงแม้เราอาจจะแย้งว่า ผลผลิตของการฟังคือการตอบออกมาเป็นคำพูดหรือเป็นการเขียนกลับมา แต่ผลผลิตของการฟังและการอ่านก็ไม่ใช่ทั้งคือการตอบออกมาเป็นคำพูดหรือเป็นการเขียนกลับมา ผลผลิตของการฟังเป็นการเปลี่ยนแปลงที่เกิดขึ้นในสมอง หากไม่มีเทคโนโลยีที่สแกนการรับความหมายที่เกิดขึ้นในสมอง ก็เป็นไปไม่ได้ที่จะสังเกตผลผลิตในสมอง เราสังเกตได้เพียงผลลัพธ์ของการป้อนข้อมูลเข้าไปอย่างมีความหมายในรูปของการพูดหรือเขียนออกมา เช่นเดียวกันกับการเห็นผลลัพธ์ของลมพัดที่เกิดขึ้นกับใบไม้ที่แกว่งไปมา

In using performance, we need to rely on observable performance. *Observable* means being able to see or hear the performance of the learner. Considering the contexts of assessing the four skills of listening, speaking, reading, and writing, we cannot observe both the process and the product of receptive skills, namely the listening and reading. Table 5.1 show this. On the one hand, we may say that we see a learner listening, because we see her nodding her head and smiling, and then asking relevant questions. In reality, we are observing the results of the listening. We can no more observe listening (or reading) than we can see the wind blowing. The listening performance is an unobservable process of meaning take-in from the sound signals that are relayed to the ears and brain. We may argue that the product of listening is to give a verbal or written response. But, in fact, the product of the listening and reading is not the same as verbal or written response. The product of listening is the change occurring in the brain. Without a scanning technology that deals with meaningful intake in the brain, it is impossible to observe the change in the brain. We could only observe the result of meaningful input in the form of a verbal or written response, in just the same way that we see the result of wind blowing as the leaves waving back and forth.

ตารางที่ 5.1 การแสดงออกที่สังเกตได้ของทักษะทั้งสี่ (Observable performance of the four skills)

ครูสามารถสังเกต... ได้*โดยตรง* **(Can the teacher *directly* observe...)**		
	กระบวนการ? (the process?)	**ผลผลิต? (the product?)**
การฟัง (listening)	ไม่ได้ (No)	ไม่ได้ (No)
การพูด (speaking)	ได้ (Yes)	ไม่ได้ (No)*
การอ่าน (reading)	ไม่ได้ (No)	ไม่ได้ (No)
การเขียน (writing)	ได้ (Yes)	ได้ (Yes)

* ยกเว้นเมื่อมีการบันทึกเสียง (Except in the case of audio recording)

ดังนั้น การสังเกตและการวัดทักษะรับสารจึงต้องอาศัยการสังเกตจากการพูดหรือการเขียนออกมาของผู้สอบ ไม่ใช่สังเกตการฟังหรือการอ่านโดยตรง กล่าวโดยสรุป การวัดประเมินผลทักษะรับสารต้องอาศัยการอนุมาน จากการรับสารที่ไม่สามารถวัดได้สังเกตได้ สู่การสร้างข้อสรุปเกี่ยวกับสามัตถิยะความเข้าใจ

In light of the non-observable nature of the listening and reading, it may be said that observing and assessing the receptive skills has to rely on observing the test takers' speaking or writing, not observing the listening or reading directly. In sum, assessing the receptive skills has to rely on inference, from the unobservable reception of information to drawing a conclusion about comprehension competence.

5.2 ทักษะย่อยทักษะหลักในการฟัง (Micro- and Macroskills of Listening)

ในการแสดงออกในด้านการฟังเพื่อความเข้าใจ เราอาจแบ่งวัตถุประสงค์ในการฟังออกเป็นทักษะย่อยและทักษะใหญ่ รายชื่อทักษะเหล่านี้ (adapted from Richards 1983) ช่วยในการระบุวัตถุประสงค์ในการเรียนรู้ และช่วยให้คนที่ออกแบบการทดสอบระบุวัตถุประสงค์ในการวัดผลได้อย่างเฉพาะเจาะจงมากยิ่งขึ้น

In listening comprehension, we may divide the listening objectives into microskills and macroskills. The following list of microskills and macroskills (adapted from Richards 1983) helps identify objectives for learning, and helps test designers identify specific assessment objectives.

ทักษะย่อย

1) แยกแยะเสียงที่แตกต่างกันในภาษาอังกฤษ
2) จดจำข้อความในภาษาที่มีความยาวต่างกันไว้ในความจำระยะสั้น
3) แยกแยะรูปแบบการเน้นเสียงในภาษาอังกฤษ คำในตำแหน่งที่ถูกเน้นและที่ไม่ได้ถูกเน้น โครงสร้างของจังหวะ เส้นการเน้นเสียง และบทบาทของรูปแบบเหล่านี้ในการสื่อสารข้อมูล
4) แยกแยะการลดรูปของคำ
5) จำแนกขอบเขตของคำ แกนของคำ และตีความรูปแบบของการลำดับคำ และความสำคัญได้
6) ประมวลผลการพูดที่มีระดับความเร็วต่างกัน
7) ประมวลผลการพูดที่มีการหยุดพัก ข้อผิดพลาด การพูดแก้ และที่มีตัวแปรในการแสดงออกอื่นๆ
8) แยกแยะชนิดของคำทางไวยากรณ์ (คำนาม คำกริยา ฯลฯ) ระบบต่างๆ (เช่น กาล การผันคำกริยาตามประธาน การทำให้เป็นพหูพจน์) รูปแบบต่างๆ กฎเกณฑ์ และรูปการละต่างๆ
9) ตรวจแยกส่วนประกอบของประโยค และแยกแยะระหว่างส่วนประกอบหลักและส่วนประกอบย่อย
10) แยกแยะได้ว่า ความหมายหนึ่งๆ อาจแสดงออกได้ในหลายรูปทางไวยากรณ์
11) แยกแยะคำเชื่อมในวัจนกรรมภาษาพูด

ทักษะหลัก

12) แยกแยะหน้าที่ของถ้อยคำในการสื่อสาร ตามสถานการณ์ ผู้สนทนา และเป้าหมาย
13) อนุมานสถานการณ์ ผู้สนทนา และเป้าหมาย โดยใช้ความรู้ทั่วไปในโลกความเป็นจริง
14) จากเหตุการณ์และแนวคิดที่ให้มา ทำนายผลลัพธ์ อนุมานการเชื่อมโยงและความเกี่ยวข้องระหว่างเหตุการณ์ อนุมานสาเหตุและผลลัพธ์ และบอกความสัมพันธ์เช่น แนวคิดหลัก แนวคิดสนับสนุนรอง ข้อมูลใหม่ ข้อมูลที่ให้มา การสรุปรวม และการยกตัวอย่างได้
15) แยกแยะระหว่างความหมายตรงตัวและความหมายโดยนัยได้
16) ใช้สีหน้า การเคลื่อนไหว หรือภาษากาย หรืออวัจนภาษาอื่นๆ ในการถอดความหมายได้

17) พัฒนาและใช้กลยุทธ์ในการฟัง เช่น การระบุคำสำคัญ การเดาความหมายของคำจากบริบท การร้องขอความช่วยเหลือ และการส่งสารว่าเข้าใจหรือไม่เข้าใจ

Microskills

1) Discriminate among the distinctive sounds of English

2) Retain chunks of language of different lengths in short-term memory

3) Recognize English stress patterns, words in stressed and unstressed positions, rhythmic structure, intonation contours, and their role in signaling information

4) Recognize reduced forms of words

5) Distinguish word boundaries, recognize a core of words, and interpret word order patterns and their significance

6) Process speech at different rates of delivery

7) Process speech containing pauses, errors, corrections, and other performance variables

8) Recognize grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), patterns, rules, and elliptical forms

9) Detect sentence constituents and distinguish between major and minor constituents

10) Recognize that a particular meaning may be expressed in different grammatical forms

11) Recognize cohesive devices in spoken discourse

Macroskills

12) Recognize the communicative functions of utterances, according to situations, participants, and goals

13) Infer situations, participants, and goals using real-world knowledge

14) From events and ideas described, predict outcomes, infer links and connections between events, deduct causes and effects, and detect such relations as main idea, supporting idea, new information, given information, generalization, and exemplification

15) Distinguish between literal and implied meanings

16) Use facial, kinesic, body language and other nonverbal clues to decipher meanings

17) Develop and use a set of listening strategies, such as detecting key words, guessing the meaning of words from context, appealing for help, and signaling comprehension or a lack thereof

5.3 การออกแบบชิ้นงานการประเมินผลด้านการฟัง (Designing Listening Assessment Tasks)

หลังจากเรากำหนดวัตถุประสงค์ในการวัด ขั้นต่อไปก็คือการออกแบบชิ้นงาน ซึ่งก็รวมไปถึงการตัดสินใจว่า เราจะดึงพฤติกรรมการแสดงออกออกมาอย่างไร และเราคาดหวังว่าผู้เรียนจะตอบสนองอย่างไร (Brown & Abeywickrama 2010) ในหัวข้อนี้ เราจะดูตัวอย่างชิ้นงานด้านการฟัง ได้แก่การฟังแบบเข้มข้น (intensive listening) ไปจนถึงการฟังแบบกว้างขวาง (extensive listening)

After we have determined the assessment objectives, the next step is to design assessment tasks. This includes a decision as to how we are going to elicit the performance, and how we expect the learners to respond to the tasks (Brown & Abeywickrama 2010). In this section, we will have a look at some sample tasks for listening, ranging from intensive listening to extensive listening.

การฟังแบบเข้มข้น (Intensive listening)

Test-takers hear:	Is he living?
Test-takers read:	A. Is he leaving?
	B. Is he living?

ในตัวอย่างข้างบน คู่เทียบความแตกต่างเป็นเป้าหมายในการวัด โดยวัดเสียงสระแบบบูรณาการการฟังและการอ่าน จัดเป็นทักษะย่อยในการฟังเพื่อความเข้าใจ

In the above example, a phonemic pair of vowels is the target for assessment. It integrates listening and reading, and is a microskill for listening comprehension.

Test takers hear:	Hello, my name's Suphat. I come from Thailand.
Test takers read:	A. Suphat is comfortable in Thailand.
	B. Suphat wants to come to Thailand.
	C. Suphat is Thai.
	D. Suphat likes Thailand.

ในตัวอย่างข้างบน การฟังเพื่อความเข้าใจวัดผลโดยการให้ประโยค (สิ่งเร้า) และให้ผู้สอบเลือกประโยคพาราเฟรซ (paraphrase) ที่สอดคล้อง โดยเป็นการบูรณาการการอ่านและการฟัง

In the example above, listening comprehension is assessed by providing a stimulus sentence to the test takers and asking them to choose the correct paraphrase.

<u>การฟังเพื่อตอบสนอง (Responsive listening)</u>

Test-takers hear:	How much time did you take to do homework?
Test-takers read:	A. In about an hour.
	B. About an hour.
	C. About $10.
	D. Yes, I did.

วัตถุประสงค์ของข้อสอบข้างบนคือ เพื่อให้ผู้สอบแยกแยะคำถามแบบ *wh-* "how much" และการตอบที่เหมาะสม ตัวลวงถูกเลือกมาเพื่อเป็นตัวแทนของคำตอบของผู้เรียนที่มักผิดกันบ่อยๆ ในตัวลวง A เป็นการตอบสำหรับคำถาม "how much longer" ตัวลวง C เป็นการตอบคำถาม "how much" แบบเป็นตัวเงิน และตัวลวง D เป็นการตอบคำถามแบบ yes/no

The objective of the item above is for the test takers to recognize the *wh*-question “how much” and its appropriate response. The distractors are selected to represent common errors. In distractor A, the answer is for “how much longer”. In distractor C, the answer is for “how much” used with money. In distractor D, the answer is for a yes/no question.

Test takers hear:	How much time did you take to do homework?
Test takers write or speak:	_____.

ข้อสอบข้อก่อนหน้าสามารถนำมาปรับเป็นแบบคำถามปลายเปิดได้ (ตัวอย่างข้างบน) โดยจากเดิมที่วัดด้านการฟังและการอ่าน ก็จะกลายมาเป็นการวัดการฟังและการเขียนหรือพูดได้

The previous item could be adjusted into an open-ended question (exemplified above). The skills assessed would be from listening and reading to listening and writing or speaking.

<u>การฟังแบบเลือกหา (Selective listening)</u>

การแสดงออกซึ่งการฟังแบบที่สามนี้ คือการฟังแบบเลือกหาข้อมูลที่ต้องการจากข้อความที่ได้ฟัง เช่น ข้อสอบแบบโคลซเพื่อการฟัง (listening cloze) ในข้อสอบชนิดนี้ ผู้สอบได้รับข้อความที่บางส่วนถูกลบไป ผู้สอบต้องฟังเพื่อเติมคำลงในช่องว่าง ข้อควรระวังของข้อสอบชนิดนี้คือ ข้อสอบจะกลายเป็นข้อสอบที่วัดการอ่านเพื่อความเข้าใจแต่เพียงอย่างเดียว วิธีแก้ได้แก่ คำที่ลบต้องเป็นคำที่มีโหลดข้อมูลมาก (high information load) เพื่อให้คาดเดาได้ยาก ต้องอาศัยการฟังเพื่อเติมข้อความให้สมบูรณ์ ดังแสดงตัวอย่างด้านล่าง

The third type of listening performance is to listen and select the information desired from the listening texts. For example, in a listening cloze, the test takers receive texts parts of which are deleted. The test takers have to listen and fill out the blanks. A weakness of this test format is that the test would become a reading comprehension test if the blanks can be filled out without use of the information from the listening. A solution

is to delete only words and phrases with high information load. Filling out the blanks has to rely on the information in the listening texts, as shown in the example below.

Test takers hear:

Ladies and gentlemen, I now have some connecting gate information for those of you making connections to other flights out of San Francisco.

Test takers read the sentences and writing the missing words or phrases in the blanks.

Flight *seven-oh-six* to Portland will depart from gate *seventy-three* at *nine-thirty* P.M.

Flight *ten-forty-five* to Reno will depart at *nine-fifty* P.M. from gate *seventeen*.

การฟังแบบกว้างขวาง (Extensive listening)

ในโลกวิชาการ การบรรยายในชั้นเรียนโดยอาจารย์เป็นประสบการณ์ทั่วไปของผู้เรียนภาษาอังกฤษ รูปแบบหนึ่งของการสอบการฟังคือ การฟังบรรยายเป็นตัวดึงพฤติกรรมการตอบสนอง (สิ่งเร้า) โดยผู้สอบให้จดโน้ต โน้ตที่จดได้อาจถูกประเมินโดยระบบเช่น เต็ม 30 แต้มดังนี้

In the academic world, classroom lectures given by professors are a common experience of non-native English users. A form of listening examination is to use a lecture as a stimulus, and the students will take notes. These notes are then evaluated, for example, on a 30-point system, as shown in the following criteria:

0–15 คะแนน (points)

การนำเสนอที่มองเห็นได้: โน้ตของคุณชัดเจนและอ่านง่าย? มองหาข้อมูลและดึงมาใช้ได้ง่าย? ใช้พื้นที่บนหน้ากระดาษเพื่อจะนำเสนอความคิดแบบเห็นภาพ? คุณใช้ย่อหน้า หัวข้อเรื่อง หมายเลข หรือไม่?

Visual presentation: Are your notes clear and easy to read? Can you easily find and retrieve information from them? Do you use the space on the paper to visually represent ideas? Do you use indentation, headers, numbers, etc.?

0–10 คะแนน (points)

ความถูกต้อง: คุณนำเสนอแนวคิดหลักจากการบรรยายได้อย่างถูกต้อง? ได้โน้ตรายละเอียดสำคัญ ข้อมูลรอง และตัวอย่างหรือไม่? ได้ละรายละเอียดที่ไม่สำคัญออกไปหรือไม่?

Accuracy: Do you accurately indicate main ideas from lectures? Do you note important details and supporting information and examples? Do you leave out unimportant information?

0–5 คะแนน (points)

สัญลักษณ์และตัวย่อ: คุณได้ใช้สัญลักษณ์และตัวย่อต่างๆ ให้มากที่สุดเท่าที่จะเป็นไปได้ เพื่อประหยัดเวลาหรือไม่? คุณได้หลีกเลี่ยงการเขียนออกมาทั้งคำและหลีกเลี่ยงการเขียนออกมาคำต่อคำที่ผู้บรรยายพูดหรือไม่?

Symbols and abbreviations: Do you use symbols and abbreviations as much as possible to save time? Do you avoid writing out whole words, and do you avoid writing down every single word the lecturer says?

จากตัวอย่างเกณฑ์การใช้คะแนนข้างต้น กระบวนการตรวจให้คะแนนอาจจะกินเวลามาก ซึ่งเท่ากับการประเมินผลลักษณะนี้พร่องความสามารถในการใช้งานได้จริง (low practicality) และเพราะความเป็นอัตนัยในการให้คะแนน ความเที่ยงของการประเมินผลนี้จึงน้อยลงด้วย กระนั้น ข้อดีสำคัญคือชิ้นงานการวัดได้สะท้อนสิ่งที่เน้นในบริบทห้องเรียน งานจดโน้ตนี้จึงมีความตรงในด้านการวัดการฟังเพื่อความเข้าใจอย่างมาก ชิ้นงานตอบโจทย์เกณฑ์เรื่องการใช้ทรัพยากรด้านปริชาน (cognitive demand) การใช้ภาษาเพื่อการสื่อสาร และความสมจริงเป็นตามสภาพจริง (high authenticity)

From the scoring system above, the scoring process can be said to be time-consuming. This means that it loses some practicality. Because of subjectivity in the scoring system, the reliability decreases too. Nonetheless, the note-taking task offers an authentic task which reflects exactly what is emphasized in the classroom settings. The task thus has face validity in assessing global listening comprehension. The task also fulfills the criteria of cognitive demand, communicative language, and authenticity.

5.4 ทักษะย่อยทักษะหลักในการพูด (Micro- and Macroskills of Speaking)

ในการแสดงออกในด้านการพูดเพื่อการสื่อสาร เราอาจแบ่งวัตถุประสงค์ในการพูดออกเป็นทักษะย่อยและทักษะใหญ่ รายชื่อทักษะเหล่านี้ (Brown & Abeywickrama 2010: 186) ช่วยในการระบุวัตถุประสงค์ในการเรียนรู้ และช่วยให้ผู้ที่ออกแบบการทดสอบระบุวัตถุประสงค์ในการวัดผลได้อย่างเฉพาะเจาะจงมากยิ่งขึ้น

In oral production, we may divide the speaking objectives into microskills and macroskills. The following list of microskills and macroskills (Brown & Abeywickrama 2010: 186) helps identify objectives for learning, and helps test designers identify specific assessment objectives.

ทักษะย่อย

1) พูดหน่วยเสียงและหน่วยเสียงย่อยที่แตกต่างกันในภาษาอังกฤษได้
2) พูดข้อความในภาษาอังกฤษที่มีความยาวต่างกันได้
3) พูดรูปแบบการเน้นเสียงในภาษาอังกฤษ คำในตำแหน่งที่ต้องเน้นเสียงและที่ไม่ต้องเน้นเสียง โครงสร้างของจังหวะ และเส้นการเน้นเสียงได้
4) พูดคำและวลีในรูปแบบที่มีการลดรูปของคำได้
5) ใช้หน่วยคำจำนวนมากพอ เพื่อให้บรรลุเป้าหมายวัจนปฏิบัติ
6) พูดคล่อง เพื่อสื่อความในอัตราที่แตกต่างกันได้
7) สังเกตการส่งสารผ่านการพูดของตนเอง และใช้กลวิธีในการสื่อสารที่หลากหลาย เช่น การหยุด การใส่คำเติมเต็ม การพูดแก้เอง การขยักข้อมูล เพื่อเพิ่มความชัดเจนของสารที่สื่อ
8) ใช้คำที่มีหน้าที่ทางไวยากรณ์ต่างกัน (คำนาม คำกริยา ฯลฯ) ระบบ (เช่น กาล การผันกริยาตามประธาน การใช้คำพหูพจน์) การลำดับคำ รูปแบบ กฎ และรูปย่อได้

9) พูดได้แบบธรรมชาติ ด้วยวลีที่เหมาะสม ด้วยการหยุด การหายใจ และส่วนประกอบของประโยคที่เหมาะสม

10) พูดบ่งชี้ความหมายเฉพาะอย่าง ในกรอบไวยากรณ์ที่ต่างกัน

11) ใช้คำเชื่อมในการพูด

ทักษะหลัก

12) พูดสื่อสารได้อย่างเหมาะสม ต่อสถานการณ์ ผู้ร่วมสื่อสาร และเป้าหมาย

13) ใช้สไตล์ ทำเนียบภาษา ความหมายนัยแฝง ลักษณะความเกินจำเป็น ธรรมเนียมวัจนปฏิบัติ กฎเกณฑ์ในการสนทนา การครองและการส่งต่อการสนทนา การขัดจังหวะ และลักษณะทางภาษาศาสตร์สังคมอื่นๆ ในการสนทนาต่อหน้าได้อย่างเหมาะสม

14) ใช้ตัวเชื่อมและคำเชื่อมระหว่างเหตุการณ์ และสื่อสารใจความหลัก ใจความรองเหตุการณ์และความรู้สึก ข้อมูลใหม่และข้อมูลที่มีอยู่แล้ว การสรุปรวม และการยกตัวอย่างได้

15) ใช้การแสดงออกทางสีหน้า การเคลื่อนไหว ภาษากาย และอวัจนภาษาอื่นๆ ควบคู่กับวัจนภาษาได้

16) พัฒนาและใช้กลวิธีในการพูด เช่นการเน้นคำสำคัญ การแปลงวลี การบอกบริบทเพื่อการตีความความหมายของคำ การร้องขอความช่วยเหลือ และการประเมินว่าคู่สนทนากำลังเข้าใจคุณได้อย่างแม่นยำ

Microskills

1) Produce phonemes and allophones in the English language

2) Produce language chunks of different lengths

3) Produce English stress patterns, words in stressed and unstressed positions, rhythmic structure, and intonation contours

4) Produce reduced forms of words and phrases

5) Use an adequate number of lexical units (words) to accomplish pragmatic purposes

6) Produce fluent speech at different rates of delivery

7) Monitor one's own oral production and use various strategic devices—pauses, fillers, self-corrections, backtracking—to enhance the clarity of the message

8) Use grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), word order, patterns, rules, and elliptical forms

9) Produce speech in natural constituents: in appropriate phrases, pause groups, breath groups, and sentence constituents

10) Express a particular meaning in different grammatical forms

11) Use cohesive devices in spoken discourse

<u>Macroskills</u>

12) Accomplish communicative functions according to situations, participants, and goals

13) Use appropriate styles, registers, implicature, redundancies, pragmatic conventions, conversation rules, floor-keeping and -yielding, interrupting and other sociolinguistic features in face-to-face conversations

14) Convey links and connections between events and communicate such relations as focal and peripheral ideas, events and feelings, new information and given information, generalization and exemplification

15) Convey facial features, kinesics, body language, and other nonverbal cues along with verbal language

16) Develop and use a number of speaking strategies, such as emphasizing key words, rephrasing, providing a context for interpreting the meaning of words, appealing for help, and accurately assessing how well your interlocutor is understanding you

5.5 การออกแบบชิ้นงานการประเมินผลด้านการพูด (Designing Speaking Assessment Tasks)

หลังจากเรากำหนดวัตถุประสงค์ในการวัด ขั้นต่อไปก็คือการออกแบบชิ้นงาน ซึ่งก็รวมไปถึงการตัดสินใจว่า เราจะดึงพฤติกรรมการแสดงออกออกมาอย่างไร และเราคาดหวังว่าผู้เรียนจะตอบสนองอย่างไร (Brown & Abeywickrama 2010) ในหัวข้อนี้ เราจะดูตัวอย่างชิ้นงานด้านการพูด ได้แก่การพูดตาม (imitative speaking) ไปจนถึงการพูดแบบกว้างขวาง (extensive speaking)

After we have determined the assessment objectives, the next step is to design assessment tasks. This includes a decision as to how we are going to elicit the performance, and how we expect the learners to respond to the tasks. In this section, we will have a look at some sample tasks for speaking, ranging from imitative speaking to extensive speaking.

<u>การพูดตาม (Imitative speaking)</u>

ในอดีตเราเคยเชื่อกันว่า การพูดเลียนเสียงแบบไม่มีความหมายไม่มีประโยชน์ในการเรียนภาษา แต่เมื่อเราพบว่า หากเราเน้นเรื่องการพูดคล่องมากเกินไป บางครั้งก็นำไปสู่ความถูกต้องที่ลดน้อยลง เราจึงหันมาให้ความสนใจกับการออกเสียง โดยเฉพาะการเน้นเสียงระดับคำและระดับประโยค เพื่อให้ผู้เรียนพูดได้เข้าใจมากยิ่งขึ้น แต่ก็ต้องไม่ให้กินเวลาส่วนใหญ่ของการวัดทักษะการส่งสาร ต่อไปนี้เป็นตัวอย่างการวัดด้วยการให้พูดตาม

In the past, we used to believe that non-meaningful imitation of sounds was useless in communicative language teaching. But when we discovered that if we overemphasized fluency, then accuracy in speech suffered. So, we have paid attention to pronunciation, especially stress and intonation, which would help learners be more comprehensible. Nonetheless, such imitative tasks should not be the majority of oral production assessment. The following is an example of imitative speaking assessment.

Test takers hear:	Repeat after me:	
	eat *pause* it *pause*	
	chat *pause* rat *pause*	etc.
	I bought a boat yesterday.	
	The glow of the candle is growing.	etc.
	When did they go on vacation?	
	Do you like coffee?	etc.
Test takers repeat the stimulus.		

<u>การพูดแบบเข้มข้น (Intensive speaking)</u>

ในการออกแบบชิ้นงานเพื่อใช้ในการวัดผล สำหรับการพูดในระดับเข้มข้นนั้น ผู้สอบจะพูดข้อความสั้นๆ ไม่เกินหนึ่งประโยค เพื่อแสดงความสามารถทางภาษาในระดับที่กำหนด ชิ้นงานจำนวนมากเป็นแบบควบคุมคำตอบ กล่าวคือ โจทย์ทำให้ผู้สอบสามารถตอบได้ไม่กี่แบบเท่านั้น ต่อไปนี้เป็นตัวอย่างการวัดการพูดแบบเข้มข้น

In designing assessment tasks for intensive speaking, the test takers would say a short response, which normally is not longer than one sentence. This is in order to demonstrate linguistic ability at a specified level of language. Many tasks are controlled, in that they allow a limited band of possibilities for their answers. The following are some examples for assessing intensive speaking.

Test takers see:

Interviewer: What did you do last weekend?

Test taker: _____

Interviewer: What will you do after you graduate from this program?

Test taker: _____

Test taker: _____?

Interviewer: I was in Japan for two weeks.

Test taker: _____?

Interviewer: It's ten-thirty.

Test takers respond with appropriate lines.

ข้อดีประการหนึ่งของการใช้เทคนิคการประเมินผลแบบนี้คือ การสามารถควบคุมสิ่งที่ผู้สอบจะพูดออกมาได้ในระดับหนึ่ง นอกจากนี้ การที่โจทย์มาในรูปของข้อความบนกระดาษแทนที่จะเป็นข้อความที่อ่านให้ผู้สอบฟังยังช่วยตัดปัญหาเรื่องการฟังโจทย์ไม่เข้าใจ ส่วนข้อเสียประการแรกคือ การที่ต้อง

พึ่งการอ่านออก และความสามารถในการถ่ายโอนจากการเขียนไปสู่การพูดในการตอบคำถาม ข้อเสียประการต่อมาคือความไม่สมจริงของชิ้นงาน ซึ่งหากทำเป็นการเล่นบทบาทสมมุติก็อาจสมจริงกว่านี้มาก

An advantage of this technique is its moderate control of the test taker's output. Moreover, the written form of the stimulus gives the test taker more time and removes potential ambiguity that may be created by aural misunderstanding. As for disadvantages, it has to rely on literacy and the test taker's ability to transfer easily from written to spoken English. Another disadvantage is the inauthenticity of the task, which would otherwise be accomplished better with a role-play.

การพูดตอบสนอง (Responsive speaking)

การวัดและประเมินผลชิ้นงานที่ต้องมีการพูดแบบตอบสนองเกี่ยวข้องกับการมีปฏิสัมพันธ์สั้นๆ กับคู่สนทนา แตกต่างจากชิ้นงานที่วัดการพูดแบบเข้มข้น ในแง่ที่ให้อิสระความคิดสร้างสรรค์แก่ผู้สอบมากกว่า และแตกต่างจากชิ้นงานแบบมีปฏิสัมพันธ์ตรงที่ข้อความที่พูดจะมีความยาวจำกัดกว่า ต่อไปนี้เป็นตัวอย่างการวัดการพูดแบบตอบสนอง

Assessing responsive tasks is involved with brief interactions with a conversation partner or a rater. They are different from intensive speaking assessment tasks, in that they allow the test takers more creativity. They also differ from interactive assessment tasks, in the way that they involve limited length of utterances. The following are some examples of assessing responsive speaking.

Test takers hear:

1. What do you think about the weather today?
2. What do you like about the English language?
3. Why did you choose your academic major?
4. What kind of strategies have you used to help you learn English?
5. a. Have you ever been to the United States before?
   b. What other countries have you visited?
   c. Why did you go there? What did you like best about it?
   d. If you could go back, what would you like to do or see?
   e. What country would you like to visit next, and why?

Test takers respond with a few sentences at most.

คำถามในระดับการวัดแบบพูดตอบสนองมักเป็นคำถามแบบต้องการข้อมูล กล่าวคือ ผู้สอบจะมีโอกาสในการให้คำตอบที่มีความหมายเป็นการตอบสนองต่อคำถาม ตัวอย่างเช่น “รัฐบาลควรทำอย่างไรบ้างเพื่อแก้ปัญหาการตัดไม้ทำลายป่า” ในชั้นเรียนเพื่อการสื่อสาร คำถามลักษณะนี้มักเป็นที่นิยมมากกว่าคำถามที่ผู้ถามทราบคำตอบอยู่แล้ว ตัวอย่างคำถามที่ผู้ถามทราบคำตอบอยู่แล้ว เช่น “สิ่งนี้เรียกว่าอะไรในภาษาอังกฤษ” ในการออกแบบคำถาม จึงเป็นเรื่องสำคัญที่เราต้องทราบว่าเราถามคำถามไปทำไม เราแค่จะหยั่งเชิงดูสามัตถิยะทางวัจนกรรมทั่วไปว่าอยู่ในระดับใด หรือเรากำลังรวมถามทั้งสามัตถิยะทางวัจนกรรมและสามัตถิยะทางไวยากรณ์ไว้ในคำถามเดียวกัน ดังเช่นในตัวอย่างการวัดข้างบน คำถามที่ 5 แตกเป็นคำถามแยกย่อยตามแต่สถานการณ์ ขึ้นอยู่กับคำตอบของคำถามในข้อที่ 4 จากผู้สอบ คำถามที่ 4 จึงถือว่ามีเป้าหมายเป็นคำถามนำของคำถามที่ 5 ด้วย

Questions for responsive speaking tend to seek information. The test takers get a chance to give meaningful responses to questions, e.g., “What are the steps the government should take to stop deforestation?” In communicative language class, such questions are more inclined to be used than known-information questions, e.g., “What is this called in English?” When designing questions, it is thus important that we know why we ask particular questions. Do we want to gain a general sense of discourse competence of the test takers? Or do we want to combine both discourse competence and grammatical competence in the same question? Like question 5 in the above example, situationally

linked questions are dependent on the answer to question 4. Question 4 thus has a goal of introducing question 5.

<u>การพูดแบบมีปฏิสัมพันธ์ (Interactive speaking)</u>

อีกประเภทของการวัดการส่งสารผ่านการพูดเกี่ยวข้องกับการพูดแบบมีปฏิสัมพันธ์ โดยเป็นวัจนกรรมโต้ตอบขนาดยาวระหว่างผู้สอบและคู่สนทนา ต่อไปนี้เป็นตัวอย่างการวัดการพูดแบบมีปฏิสัมพันธ์

Another category of oral production assessment involves interactive speaking. It entails relatively long stretches of interactive discourse. The following is an example of an interactive speaking task.

การสัมภาษณ์คือการที่ผู้จัดสอบและผู้สอบสนทนาแลกเปลี่ยนกันต่อหน้า และดำเนินการสนทนากันต่อไปผ่านคำถามและคำสั่ง จากนั้น การสนทนาซึ่งอาจบันทึกเทปไว้เพื่อฟังอีกรอบหนึ่ง ก็ถูกให้คะแนนตามรายการเกณฑ์การประเมิน เช่น ด้านความถูกต้องในการออกเสียงหรือในด้านไวยากรณ์ การใช้คำศัพท์ ความคล่องแคล่วในการพูด ความเหมาะสมทางด้านภาษาและสังคมและด้านวัจนปฏิบัติ การทำชิ้นงานได้ลุล่วง และแม้แต่ความเข้าใจ

An interview is when a test administrator and a test taker have a direct face-to-face exchange, and proceed with questions and directives. After that, the conversation, which may be tape-recorded for relistening, is scored on a list of rubrics, e.g., accuracy in pronunciation or grammar, vocabulary usage, fluency, sociolinguistic and pragmatic appropriateness, task completion, and even comprehension.

Michael Canale (1984) เสนอกรอบการทดสอบด้านการพูด โดยเสนอว่า ผู้สอบจะแสดงออกได้ดีที่สุดถ้าการสัมภาษณ์มี 4 ขั้นตอนดังนี้

Michael Canale (1984) proposed a framework for testing oral proficiency. He suggested that the test takers could perform their best if the interview has the following four steps:

1) *อุ่นเครื่อง* ผู้สัมภาษณ์เกริ่นนำและช่วยผู้สอบให้รู้สึกคลายกังวล และแจ้งรูปแบบการสัมภาษณ์ให้ทราบ ไม่มีการเก็บคะแนนในขั้นตอนนี้ ซึ่งกินเวลาประมาณหนึ่งนาที

1) *Warm-up*. For about one minute, the interviewer introduces the test taker to the interview format, and helps reduce the test anxiety. No scoring takes place in this phase.

2) *ทดสอบระดับ* ผู้สัมภาษณ์ถามคำถามให้ผู้สอบตอบด้วยรูปประโยคและหน้าที่ที่คาดเดาได้ ถ้าจากข้อมูลพบว่า ผู้สอบอยู่ในระดับใดระดับหนึ่ง คำถามของผู้สัมภาษณ์ก็จะพยายามยืนยันว่าผู้สอบได้อยู่ระดับนั้นๆ จริง คำถามมักออกแบบมาให้วัดไวยากรณ์ การใช้คำศัพท์ และ/หรือปัจจัยทางภาษาและสังคม เช่น การใช้คำสุภาพ ภาษาทางการ/ไม่เป็นทางการ เป็นต้น ด้านที่เป็นเป้าหมายเหล่านี้จะถูกให้คะแนนในขั้นตอนนี้

2) *Level check*. The interviewer asks the test taker with predicted forms and functions. If the information such as grades, or other data shows that the test taker is at one particular level, the questions will seek to confirm that the test taker is really at the specified level. Questions are usually designed so as to assess grammar, vocabulary usage, and/or sociolinguistic factors such as language for politeness, and formal/informal language. These target criteria are scored in this phase.

3) *เจาะลึก* คำถามเจาะลึกท้าทายผู้สอบให้ไปถึงระดับที่สูงที่สุดของความสามารถผู้สอบ ตัวคำถามอาจซับซ้อนในแง่ตัวภาษา และ/หรืออาจซับซ้อนในแง่ปริชานหรือภาษาที่ต้องใช้ตอบ ในระดับสมิทธิภาพต่ำ คำถามเจาะลึกอาจเพียงแค่ว่าผู้สอบต้องใช้คำศัพท์หรือไวยากรณ์ที่สูงขึ้นมากว่าที่คาดไว้ ในระดับสมิทธิภาพสูง คำถามเจาะลึกอาจจะให้ผู้สอบแสดงความคิดเห็นหรือประเมินค่า พูดคุยเกี่ยวกับด้านที่ถนัด เล่าเรื่อง หรือตอบคำถามที่มีความซับซ้อน การตอบสนองต่อคำถามเจาะลึกอาจให้คะแนนได้ หรืออาจจะไม่ต้องสนใจหากว่าผู้สอบไม่สามารถรับมือกับความซับซ้อนที่ถามได้

3) *Probe*. Probe questions challenge the test takers to go to the heights of their ability. The questions may be complex in terms of language, and/or complex in terms of cognitive or linguistic demand. At the low levels of proficiency, probe questions might simply require a higher range of grammar or vocabulary. At the high levels of proficiency, probe questions will ask the test takers to give an opinion or judgment, to talk about areas of interest, to give a narrative, or to answer complex questions. Responses to probe questions may be scored, or ignored if the test takers cannot handle the questions.

4) *เบาเครื่อง* ขั้นตอนสุดท้ายของการสัมภาษณ์เป็นช่วงเวลาสั้นๆ ที่ผู้สัมภาษณ์ถามคำถามง่ายๆ เพื่อให้ผู้สอบผ่อนคลาย และให้ข้อมูลเกี่ยวกับการประกาศคะแนน ขั้นตอนนี้ไม่ให้คะแนน

4) *Wind-down*. This is the final phase of an interview, in which the interviewer asks easy questions to relax the test taker, and give information as to where and when the test taker could get the results of the interview. This phase is not scored.

ต่อไปนี้เป็นตัวอย่างคำถามสำหรับ 4 ขั้นตอนของการสัมภาษณ์ (Brown & Abeywickrama 2010)

The following are sample questions for the four stages of an oral interview (Brown & Abeywickrama 2010).

1. Warm-up:

How are you?

What's your name?

What city/town are you from?

Let me tell you about this interview.

2. Level check:

How long have you been in this country/city?

Tell me about your family.

What are your hobbies or interests?

Why do you like your hobby/interest?

3. Probe:

What are your goals for learning English in this program?

What is your opinion of [a recent headline news event]?

If you could redo your education all over again, what would you do differently?

What career advice would you give to your younger friends?

4. Wind-down:

Did you feel okay about this interview?

What are your plans for [the weekend, the rest of today, the future]?

You'll get your results from this interview [tomorrow, next week].

Do you have any questions you want to ask me?

<u>การพูดแบบกว้างขวาง (Extensive speaking)</u>

การพูดแบบกว้างขวางเกี่ยวข้องกับวัจนกรรมที่ซับซ้อนและมีความยาว มักมีลักษณะเป็นแบบการพูดคนเดียว โดยมีการโต้ตอบกันไม่มากนัก ต่อไปนี้เป็นตัวอย่างการวัดการพูดแบบกว้างขวาง

Extensive speaking tasks entail complex and long stretches of discourse. They are often monologues, with little verbal interaction. The following is an example of an extensive speaking task.

ในแวดวงวิชาการและการประกอบอาชีพ เป็นปกติที่ต้องมีการนำเสนอรายงาน สำหรับการพูดนำเสนอ แบบรายการตรวจสอบเป็นเครื่องมือปกติในการวัดและให้คะแนน ซึ่งมีสององค์ประกอบหลัก ได้แก่ เนื้อหา และการสื่อสาร ต่อไปนี้เป็นตัวอย่างแบบรายการตรวจสอบสำหรับการนำเสนอรายงาน

In academic and professional contexts, it is normal there to be oral presentations. A checklist is a usual tool for assessing and scoring, which has two main components: content and delivery. The following is an example of a checklist for oral presentation.

Evaluation of oral presentation

Assign a number to each box according to your assessment of the various aspects of the speaker's presentation.

3	Excellent	2	Good
1	Fair	0	Poor

Content:

- ☐ The purpose or objective of the presentation was accomplished.
- ☐ The introduction was lively and got my attention.
- ☐ The main idea or point was clearly stated toward the beginning.
- ☐ The supporting points were clearly expressed and supported well by facts.
- ☐ The conclusion restated the main idea or purpose.

Delivery:

- ☐ The speaker used gestures and body language well.
- ☐ The speaker maintained eye contact with the audience.
- ☐ The speaker used notes (and did not read a script verbatim).
- ☐ The speaker's language was natural and fluent.
- ☐ The speaker's volume of speech was appropriate.
- ☐ The speaker's rate of speech was appropriate.
- ☐ The speaker's pronunciation was clear and comprehensible.
- ☐ The speaker's grammar was correct and didn't prevent understanding.
- ☐ The speaker used visual aids, handouts, etc. effectively.
- ☐ The speaker showed enthusiasm and interest.
- ☐ (If appropriate) The speaker responded to audience questions well.

5.6 ทักษะย่อยทักษะหลักในการอ่าน (Micro- and Macroskills of Reading)

ในการอ่าน การพิจารณาทักษะที่เกี่ยวข้องเป็นสิ่งสำคัญสำหรับการประเมินความสามารถในการอ่าน ต่อไปนี้เป็นทักษะย่อยและทักษะหลัก (Brown & Abeywickrama 2010: 227) ที่อาจใช้เป็นวัตถุประสงค์ในการประเมินการอ่านเพื่อความเข้าใจ

In reading, considering related skills is important for assessing reading ability. The following are microskills and macroskills (Brown & Abeywickrama 2010: 227) that could be used for objectives of assessing reading comprehension:

ทักษะย่อย

1) แยกแยะหน่วยอักขระ (ตัวอักษร ชุดตัวอักษรที่ประกอบมาเป็นหน่วยเสียง) ที่แตกต่างกันได้ และแยกแยะระบบการสะกดคำในภาษาอังกฤษได้
2) ทรงจำข้อความที่มีความยาวต่างกันในความทรงจำระยะสั้นได้
3) ประมวลผลข้อความที่เขียนมาได้อย่างมีประสิทธิภาพ
4) แยกแยะคำที่เป็นแกนคำศัพท์ และตีความลำดับคำและความสำคัญของลำดับคำได้
5) แยกแยะชนิดของคำทางไวยากรณ์ (คำนาม คำกริยา ฯลฯ) ระบบต่างๆ (เช่น กาล การผันคำกริยาตามประธาน การทำเป็นพหูพจน์) รูปแบบต่างๆ กฎ และรูปละ
6) แยกแยะได้ว่า ความหมายหนึ่งอาจมีหลายรูปทางไวยากรณ์ที่แสดงออกได้
7) แยกแยะคำเชื่อมในวัจนกรรมที่เขียนขึ้นมา และแยกแยะบทบาทของคำเชื่อมในความสัมพันธ์ระหว่างอนุประโยคต่างๆ

ทักษะหลัก

8) แยกแยะขนบนิยมในงานเขียน และความสำคัญของขนบนิยมนั้นในการตีความ
9) แยกแยะหน้าที่ทางการสื่อสารของข้อความที่เขียนขึ้น ขึ้นอยู่กับรูปแบบและเป้าหมาย
10) อนุมานบริบทที่ไม่ชัดเจน โดยการกระตุ้นใช้เค้าร่างทางความคิด (schemata – ใช้ความรู้รอบตัว)

11) อนุมานความเชื่อมโยง จากเหตุการณ์ที่บรรยาย อนุมานสาเหตุและผลลัพธ์ และมองหาความสัมพันธ์ เช่น ใจความหลัก ใจความเสริม ข้อมูลใหม่ ข้อมูลเก่า การสรุปรวม และการยกตัวอย่าง

12) แยกแยะความหมายตามตัวอักษร และความหมายแฝงที่อนุมานได้

13) มองหาสิ่งที่อ้างถึงซึ่งมีความจำเพาะทางวัฒนธรรม และนำมาตีความในบริบทเค้าร่างทางความคิดเชิงวัฒนธรรม (cultural schemata) ได้

14) พัฒนาและใช้กลวิธีในการอ่าน เช่น การกวาดตาเพื่อมองหาข้อมูล การอ่านอย่างคร่าวๆ การมองหาตัวเชื่อมในระดับวัจนกรรม การเดาความหมายของคำจากบริบท การกระตุ้นเค้าร่างทางความคิดสำหรับการตีความข้อความ (schemata for interpretation) ได้

Microskills

1) Discriminate among distinctive letters and letter combinations, and orthographic patterns

2) Retain chunks of the English language of differing lengths in short-term memory

3) Process pieces of writing at an efficient rate of speed

4) Recognize a core of words, and interpret patterns of word order and their significance

5) Recognize grammatical word classes (nouns, adjectives, etc.), systems (e.g., tense, agreement, pluralization), patterns, rules, and elliptical forms

6) Recognize that different grammatical forms can express a particular meaning

7) Recognize cohesive devices used in written discourse, and their role in indicating relationship between and among clauses

Macroskills

8) Recognize the rhetorical traditions of written texts and their significance for interpretation

9) Recognize the communicative functions of discourse in writing, in accordance with form and purpose

10) Infer implicit context by activating schemata (i.e., using background knowledge)

11) Infer links and connections between events from given ideas, events, etc., infer causes and effects, and detect relations such as main idea, supporting idea, given information, new information, generalization and exemplification

12) Differentiate between literal and implied meanings

13) Identify culturally specific references and interpret them in a context of appropriate cultural schemata

14) Develop and use reading strategies, e.g., skimming and scanning, detecting discourse markers, making a guess about meaning of unknown words from context, and activating schemata for interpretation

5.7 การออกแบบชิ้นงานการประเมินผลด้านการอ่าน (Designing Reading Assessment Tasks)

ในการวัดความสามารถในการอ่าน เรามักบูรณาการทักษะการอ่านเข้ากับการพูด และ/หรือการเขียน แต่ดังที่จะพบในส่วนนี้ รูปแบบข้อสอบหลายรูปแบบไม่จำเป็นต้องผสมผสานเข้ากับการฟัง การพูดหรือการเขียน (Brown & Abeywickrama 2010) ทั้งนี้เพราะเราสามารถใช้รูปแบบข้อสอบแบบเลือกตอบ เพื่อใช้ในการวัดความสามารถทางการอ่าน ในหัวข้อนี้ เราจะดูตัวอย่างชิ้นงานด้านการอ่าน ได้แก่ การอ่านแบบรับรู้ (perceptive reading) ไปจนถึงการอ่านแบบกว้างขวาง (extensive reading)

In assessing reading ability, we often integrate reading skills with speaking and/or writing. But as it will be shown in this part, several item formats for reading do not have to integrate with listening, speaking, or even writing. This is so because we can use selected-response item formats to assess reading ability. In this part, we will see some examples of item formats for assessing reading, ranging from perceptive reading to extensive reading.

การอ่านแบบรับรู้ (Perceptive reading)

ในขั้นต้นและพื้นฐานของการอ่านภาษาที่สองคือ การแยกแยะสัญลักษณ์ที่เป็นตัวอักษรได้ แยกแยะตัวพิมพ์ใหญ่ตัวพิมพ์เล็กได้ รวมไปถึงเครื่องหมายวรรคตอน คำต่างๆ ความสอดคล้องกันระหว่างอักขระและเสียง เป็นต้น ชิ้นงานในระดับนี้จึงมักเรียกว่า ชิ้นงานรู้หนังสือ ว่าผู้เรียนกำลังอยู่ในขั้นต้นของการอ่านออกเขียนได้ ต่อไปนี้เป็นตัวอย่างของชิ้นงานในระดับนี้

At the fundamental level of reading a foreign language are recognition of the alphabet, differentiation of capital and lowercase letters, punctuation, and phoneme-grapheme correspondences. The tasks at this level are often called literacy tasks, suggesting that the learners are in the early stage of becoming literate.

*Test takers read:**		Circle S for same and D for different	
1. bet	bed	S	D
2. let	let	S	D
3. beat	bet	S	D
4. too	to	S	D

**In the case of learners at a very low level, the teacher or test administrator reads directions.*

<u>การอ่านแบบเลือกอ่าน (Selective reading)</u>

การอ่านแบบเลือกอ่านมุ่งไปที่ด้านรูปแบบของภาษา (formal aspects) เช่น คำศัพท์ ไวยากรณ์ และด้านวัจนกรรมบางส่วน การวัดผลประเภทนี้รวมถึงด้านที่หลายคนมักเรียกว่าการทดสอบคำศัพท์และไวยากรณ์ ทั้งที่ในความเป็นจริง คำศัพท์และไวยากรณ์เป็นเพียงแค่รูปแบบที่เราใช้เพื่อบรรลุการแสดงออกซึ่งทักษะทั้งสี่คือ ฟัง พูด อ่าน เขียน ต่อไปนี้เป็นตัวอย่างชิ้นงานด้านคำศัพท์และไวยากรณ์ของความสามารถในการอ่าน

Selective reading involves formal aspects of language, e.g., vocabulary, grammar and some discourse features. Assessing selective reading covers what many call testing vocabulary and grammar, despite the fact that vocabulary and grammar are simply the forms which we employ to perform all the four skills of listening, speaking, reading and writing. The following are some example tasks of vocabulary and grammar for reading ability.

1. John: Do you like wine?
   Mary: No, I can't _____ it!
   A. feel
   B. hate
   C. prefer
   D. stand

2. John: Do you have a jacket like this?
   Jack: Yes, mine is _____ yours.
   A. as same as
   B. so same as
   C. the same as
   D. the same like

การอ่านแบบมีปฏิสัมพันธ์ (Interactive reading)

ชิ้นงานการอ่านแบบมีปฏิสัมพันธ์เหมือนชิ้นงานการอ่านแบบเลือกอ่าน ตรงที่มีวัตถุประสงค์ที่เน้นทั้งรูปแบบและความหมาย กระนั้นชิ้นงานการอ่านแบบมีปฏิสัมพันธ์จะเน้นความหมายมากกว่า และมีการประมวลผลแบบบนลงล่าง (top-down processing) มากกว่าเล็กน้อย รวมทั้งใช้ข้อความที่ยาวกว่า แผนภูมิ หรือรูปประกอบอื่นก็จะมีความซับซ้อนมากกว่าด้วย ต่อไปนี้เป็นตัวอย่างของชิ้นงานการอ่านแบบมีปฏิสัมพันธ์

Interactive reading tasks are like selective reading tasks, because both have objectives that are form-focused and meaning-focused. However, interactive reading tasks focus more on meaning, and are more inclined to top-down processing and longer texts. Graphs, other graphics, and charts might be more complex in terms of their format than those used in selective reading tasks. The following is an example of an interactive reading task.

The recognition that one's feelings (1) _____ happiness (2) _____ unhappiness can coexist much like love and hate (3) _____ a close relationship may offer valuable clues (4) _____ how to lead a happier life. It suggests, (5) _____ example, that changing (6) _____ avoiding things that make you miserable may well make you less miserable (7) _____ probably no happier.

ในตัวอย่างข้างบนคือข้อสอบแบบโคลซ วิธีการลบคำเป็นแบบลบตามลักษณะเฉพาะของข้อสอบ (rational deletion – test spec-determined) โดยในตัวอย่างนี้ คำที่ลบถูกออกแบบให้เป็นคำบุพบทและคำสันธาน วิธีการลบแบบอื่นก็เช่น การลบทุกเจ็ดคำ (โดยปกติ 7±2)

In the above example is a cloze test. The method of deletion is rational deletion, which is to delete words as predetermined in the test spec. In this example, it is

prepositions and conjunctions that are deleted. Other methods for deletion include every 7th word being deleted (normally 7±2 words).

การอ่านแบบกว้างขวาง (Extensive reading)

ชิ้นงานการอ่านแบบกว้างขวางเกี่ยวข้องกับการอ่านข้อความที่มีขนาดยาว เช่น บทความในวารสารวิชาการ บทความขนาดยาว รายงานเชิงเทคนิค เรื่องสั้น และหนังสือ เป็นต้น การอ่านข้อความเหล่านี้มักมีจุดเน้นอยู่ที่ความหมายมากกว่าตัวรูปแบบ และมักเกี่ยวข้องกับการประมวลผลแบบจากบนลงล่าง ต่อไปนี้เป็นตัวอย่างชิ้นงานการอ่านแบบกว้างขวาง

Extensive reading tasks involve reading a long text, such as journal articles, longer essays, technical reports, short stories, and books. Reading such texts usually focuses on meaning, rather than form, and often involves top-down processing. The following is an example of an extensive reading task.

- What is the main idea of this text?
- What is the author's purpose in writing the text?
- What kind of writing is this (newspaper article, manual, novel, etc.)?
- What type of writing is this (expository, technical, narrative, etc.)?
- How easy or difficult do you think this text will be?
- What do you think you will learn from the text?
- How useful will the text be for your (profession, academic needs, interests)?

ในตัวอย่างข้างบน ชิ้นงานเป็นการอ่านแบบข้าม การวัดก็ตรงไปตรงมาคือ ผู้สอบอ่านข้อความหนึ่งแบบข้าม หลังจากนั้นก็ตอบคำถามในชิ้นงาน โดยอาจตอบเป็นแบบปากเปล่าหรือเป็นแบบข้อเขียนขึ้นอยู่กับบริบท การอ่านลักษณะนี้มักมีประโยชน์ตรงที่อาจใช้นำการอภิปรายในชั้นเรียน หรือการอ่านอย่างละเอียดที่จะตามมา ดังนั้น อิทธิพลย้อนกลับสู่การเรียนจึงเป็นบวก (positive washback)

นอกจากนี้ เพราะชิ้นงานสะท้อนเป้าหมายที่ต้องการให้ผู้เรียนมีความเข้าใจเรื่องที่อ่าน ชิ้นงานจึงสมจริงเป็นตามสภาพจริง (authentic) ด้วย

In the above example, the task is to skim a reading. The assessment is straightforward: The test takers skim the text and answer the questions orally or in writing, depending on the context. Skimming like this is often useful in that it can lead a class discussion or a careful reading that may follow. So, washback effects towards the learning is positive. Moreover, because the skimming task reflects the goal in which the test takers understand the reading, the task is authentic too.

5.8 ทักษะย่อยทักษะหลักในการเขียน (Micro- and Macroskills of Writing)

ในการเขียน การพิจารณาทักษะที่เกี่ยวข้องเป็นสิ่งสำคัญสำหรับการประเมินความสามารถในการเขียน ต่อไปนี้เป็นทักษะย่อยและทักษะหลัก (Brown & Abeywickrama 2010: 262) ที่อาจใช้เป็นวัตถุประสงค์ในการประเมินการเขียน

In writing, considering related skills is important for assessing writing ability. The following are microskills and macroskills (Brown & Abeywickrama 2010: 262) that could be used for objectives of assessing writing ability.

ทักษะย่อย

1) เขียนอักขระและรูปแบบการสะกดคำในภาษาอังกฤษ

2) เขียนด้วยอัตราความเร็วที่มีประสิทธิภาพ

3) เขียนคำศัพท์หลักและใช้รูปแบบลำดับคำได้เหมาะสม

4) ใช้ไวยากรณ์ (เช่น กาล การผันกริยาตามประธาน การทำเป็นรูปพหูพจน์) รูปแบบ และกฎต่างๆ ได้อย่างเหมาะสม

5) ใช้รูปแบบไวยากรณ์ที่ต่างกัน เพื่อสื่อความหมายหนึ่งๆ

6) ใช้คำเชื่อมในวัจนกรรมการเขียน

ทักษะหลัก

7) ใช้รูปแบบการเรียงร้อยถ้อยคำ และขนบนิยมในงานเขียน

8) เขียนสื่อสารได้อย่างเหมาะสมตามเป้าหมาย

9) ใช้ตัวเชื่อมและคำเชื่อมระหว่างเหตุการณ์ และสื่อสารใจความหลัก ใจความรอง ข้อมูลใหม่และข้อมูลที่มีอยู่แล้ว การสรุปรวม และการยกตัวอย่างได้

10) แยกแยะระหว่างความหมายตามตัวอักษรและความหมายโดยนัยในขณะเขียนได้

11) เขียนสื่อสิ่งที่มีความจำเพาะทางวัฒนธรรมในงานเขียนได้

12) พัฒนาและใช้กลวิธีในการเขียน เช่น การประเมินการตีความของผู้อ่านได้อย่างถูกต้อง การใช้เครื่องมือก่อนการเขียน การเขียนอย่างคล่องแคล่วในฉบับร่างแรก การใช้การพาราเฟรซและคำความหมายเหมือน การขอข้อมูลสะท้อนกลับจากเพื่อนร่วมชั้นและผู้สอน และการใช้ข้อมูลสะท้อนกลับเพื่อการเกลาและการปรับแก้

Microskills

1) Produce English graphemes and orthographic patterns

2) Produce writing at an efficient rate of speed

3) Produce an acceptable core of words and use appropriate patterns of word order.

4) Use appropriate grammatical systems (e.g., tense, agreement, pluralization), patterns, and rules

5) Use different grammatical forms to convey a particular meaning

6) Employ cohesive devices in written discourse

Macroskills

7) Use rhetorical forms and conventions of written discourse

8) Use the communicative functions of written texts appropriately, according to form and purpose

9) Accomplish links and connections between events, and use relations such as main idea, supporting idea, new information, given information, generalization, and exemplification

10) Differentiate between literal and implied meanings in writing

11) Use culturally specific references in written texts

12) Develop and use writing strategies, such as assessing the reader's interpretation accurately, employing prewriting devices, writing in the first drafts with fluency, using synonyms and paraphrases, asking for peer and instructor feedback, and using feedback for revising and editing

5.9 การออกแบบชิ้นงานการประเมินผลด้านการเขียน (Designing Writing Assessment Tasks)

ในการวัดความสามารถส่งสารด้วยการเขียน เราต้องทราบประเภทของภาษา (genre) เพื่อที่จะได้กำหนดบริบทและเป้าหมายของการวัดได้อย่างชัดเจน รวมทั้งเราต้องทราบประเภทของการเขียน เพื่อที่ลำดับขั้นของการพัฒนาการเขียนจะได้ถูกระบุได้อย่างถูกต้อง (Brown & Abeywickrama 2010) ต่อไปนี้เป็นตัวอย่างของชิ้นงานการเขียนจำแนกตามประเภทของงานเขียน

In assessing writing production, we need to know the genre so that we could determine the context and purpose of the assessment clearly. Moreover, we need to know types of writing, so that stages of writing development can be determined correctly (Brown & Abeywickrama 2010). The following are some examples of writing assessment tasks according to the types of writing.

<u>การเขียนตาม (Imitative writing)</u>

ปัจจุบันภาษาอังกฤษถูกสอนให้กับผู้เรียนตั้งแต่อายุยังน้อย จึงเป็นเรื่องน่าสนใจที่จะต้องวัดว่า ผู้เรียนทราบวิธีการเขียนตัวอักษรโรมันหรือไม่ ต่อไปนี้เป็นตัวอย่างการวัดการเขียนตาม

At present, English is taught to young learners. It is thus interesting to assesss if they know how to write Roman letters. The following is an example of imitative writing assessment.

Test takers hear:

Write the missing word in each blank. Below the story is a list of words to choose from.

Have you ever visited San Francisco? It is a very nice city. It is cool in the summer and warm in the winter. I like the cable cars and bridges.

Test takers see:

Have _____ ever visited San Francisco? It _____ a very nice city. It is _____ in _____ summer and _____ in the winter. I _____ the cable cars _____ bridges.

is	you	cool	city
like	and	warm	the

ในตัวอย่างชิ้นงานข้างบน การบอกจด (dictation) นำมาผสมผสานกับข้อความที่มีการลบคำค่อนข้างถี่ การทดสอบให้รายชื่อคำที่หายไปที่ผู้สอบจะต้องเลือกมาเติม เป้าหมายของการทดสอบแบบนี้ไม่ได้อยู่ที่การทดสอบการสะกดคำ หากแต่อยู่ที่การฝึกเขียน

In the example above, dictation is combined with a written script which has a frequent deletion ratio. The testing gives a list of missing words that the test taker has to choose from. The purpose of this test is not to assess spelling, but to give a writing practice.

การเขียนแบบเข้มข้น (มีการควบคุม) Intensive (controlled) writing

การเขียนขั้นนี้ในคู่มือฝึกหัดครูทั่วไปเรียกว่า การเขียนแบบมีการควบคุม ซึ่งจริงๆ แล้วก็คือการเขียนแบบเน้นรูปแบบ การเขียนเชิงไวยากรณ์ หรือการเขียนแบบมีการนำ (guided writing) ผู้เรียนเขียนข้อความภาษาขึ้นมา เพื่อแสดงสามัตถิยะในเชิงไวยากรณ์ คำศัพท์ หรือการแต่งประโยค แต่มิได้จำเป็นต้องเขียนเพื่อสื่อความหมายในการสื่อสารอย่างแท้จริง

Writing in this level is called by teacher training manuals as controlled writing. It is in fact form-focused writing, grammar writing, or guided writing. The test takers produce writing in order to show their competence in terms of grammar, vocabulary, or sentence formation, rather than actually conveying meaning for a communicative purpose.

Test takers read:

Put the words below into a possible order to make a grammatical sentence:

1. cold / winter / is / weather / the / in / the
2. studying / what / you / are

Test takers write:

1. The weather is cold in the winter. (or) In the winter the weather is cold.
2. What are you studying?

ในตัวอย่างชิ้นงานข้างบน ผู้สอบต้องเรียงคำให้เป็นประโยคที่ถูกต้อง แม้ว่าชิ้นงานเช่นนี้จะไม่มีความสมจริง (authentic) มากนัก แต่ก็ถือได้ว่าวัดความสามารถในการเขียนและอาจกล่าวได้ว่าวัดเรื่องกฎไวยากรณ์ในการเรียงคำ

In the above example, the test takers have to rearrange the words into a correct sentence. Even though the task is not really authentic, it taps into writing ability and the grammatical rule for word ordering.

การเขียนเพื่อตอบสนอง (Responsive writing)

ในการเขียนเพื่อตอบสนอง ผู้สอบจะมีอิสระมากขึ้นเมื่อเทียบกับในการเขียนแบบเข้มข้น กล่าวคือ ผู้สอบสามารถเขียนแบบสร้างสรรค์ได้หลายแบบในการตอบสนองต่อโจทย์ ผู้สอบสามารถใช้ทางเลือกในเรื่องคำศัพท์ ไวยากรณ์ และวัจนกรรม โดยมีเงื่อนไขหรือข้อจำกัดบ้าง ซึ่งเกณฑ์การประเมินก็จะเริ่มมีเรื่องวัจนกรรมและขนบนิยมในการเขียนของโครงสร้างของย่อหน้าเข้ามา หรือขนบนิยมของการเชื่อมย่อหน้า เป็นต้น ประเภทของงานเขียนที่เกี่ยวข้องก็เช่น รายงานขนาดสั้น การเขียนตอบสนองต่อเรื่องหรือบทความที่อ่าน เป็นต้น

In responsive writing, the tasks give the test takers more possibilities of responding to the tasks, when compared with intensive writing. The test takers can be creative in using vocabulary, grammar, and discourse, with some constraints and conditions. As for criteria, discourse and rhetorical conventions begin to apply to the paragraph structure and to connecting paragraphs. The genres of texts typically involved include short reports, and responses to the reading of a story or an article.

การพาราเฟรซ (paraphrasing) จัดเป็นการเขียนเพื่อตอบสนองรูปแบบหนึ่ง การพาราเฟรซเป็นหนึ่งในเรื่องที่ยากสำหรับผู้เรียน การสอนลำดับแรกต้องเน้นย้ำความสำคัญของการพาราเฟรซ ซึ่งก็คือการกล่าวบางสิ่งด้วยคำพูดของตนเอง เพื่อหลีกเลี่ยงการลอกเลียนผลงาน และเพื่อเพิ่มความหลากหลายในการแสดงออกผ่านงานเขียน ในการให้คะแนนการพาราเฟรซต้องเน้นว่าการคงความหมายที่เหมือนหรือคล้ายกันเป็นประเด็นหลัก ส่วนเรื่องวัจนกรรม ไวยากรณ์ และคำศัพท์เป็นเรื่องรอง

Paraphrasing is a format of responsive writing. It is one of the difficulties to teach to the learners. The first step to teach paraphrasing is to highlight its importance:

saying something in their own words to avoid plagiarizing and to add variation in expression. In scoring paraphrasing, the test takers should be aware that retaining similarity in terms of meaning is primary, and discourse, grammar, and vocabulary are secondary.

การเขียนแบบกว้างขวาง (Extensive writing)

การเขียนแบบกว้างขวาง หรือการเขียนอิสระใช้หลักการและแนวทางทั้งหมดของการเขียนเพื่อตอบสนอง โดยเขียนเป็นข้อความที่ยาวขึ้นมาก เช่น เรียงความ งานเขียนประจำภาค รายงานประจำวิชา และวิทยานิพนธ์ ในการเขียนแบบกว้างขวาง ผู้สอบอาจได้รับอิสระมากยิ่งขึ้นไปอีก เช่น เลือกหัวเรื่องเอง กำหนดความยาวเอง ออกแบบสไตล์การเขียนเอง หรือแม้แต่ขนบนิยมในการจัดรูปแบบเอง ในขั้นนี้ กฎเรื่องการเขียนให้มีประสิทธิผลสามารถถูกนำมาใช้ได้ทั้งหมด และผู้เขียนที่ใช้ภาษาอังกฤษเป็นภาษาที่สองก็ถูกคาดหวังว่าจะทำตามมาตรฐานเช่นเดียวกับผู้เขียนที่ใช้ภาษาอังกฤษเป็นภาษาแม่ด้วย

Extensive writing or free writing puts all the guidelines and principles of responsive writing into practice. The texts written are much longer, e.g., essays, project reports, term papers, and dissertations and theses. The test takers may obtain even more freedom in choosing the topics, determining the length, designing the writing style, and even designing the formatting conventions. In this stage of writing, all the rules for effective writing can be put into practice. Second language users are also expected to follow the same standards as those who are native speakers of English.

5.10 Exercise

1. Which of the following is true?

A. Four language skills tend to be used separately.

B. Triangulation should not be used in competence assessment because of its complexity.

C. We assess competence by observing performance.

D. We can assess grammar and vocabulary separately from the four language skills.

2. Which of the following is **NOT** a listening skill?

A. discriminate among distinctive sounds of English

B. infer situations and goals using real-world knowledge

C. process speech at different rates of delivery

D. produce English phonemes and allophones

3. Which steps of testing oral proficiency is compulsory in scoring?

A. level check

B. probe

C. warm-up

D. wind-down

4. Which of the following is **NOT** a level of task formats for assessing reading?

A. extensive reading

B. imitative reading

C. perceptive reading

D. selective reading

5. Which task has high authenticity?

A. an imitative speaking task of minimal pairs

B. asking about the main idea in an extensive reading task

C. rational cloze testing on function words

D. unscrambling words into sentences

Answer key: 1. C. 2. D. 3. A. 4. B. 5. B.

บทที่ 6 การวิเคราะห์แบบทดสอบ และสถิติข้อสอบ

Chapter 6 Item Analysis and Test Statistics

ในบทที่ 5 ได้กล่าวถึงการวัดทักษะทางภาษาทั้งสี่ทักษะคือ ฟัง พูด อ่าน และเขียน ในบทนี้จะกล่าวถึงการวิเคราะห์ข้อสอบและแบบทดสอบ และสถิติข้อสอบที่ใช้กันบ่อยๆ โดยลำดับในการนำเสนอมีดังนี้

6.1 เป้าหมายของการวิเคราะห์ข้อสอบ

6.2 การวิเคราะห์ข้อสอบสำหรับการทดสอบแบบอิงกลุ่ม

6.3 การวิเคราะห์ข้อสอบสำหรับการทดสอบแบบอิงเกณฑ์

6.4 สัมประสิทธิ์สหสัมพันธ์พอยต์ไบซีเรียล

6.5 การวิเคราะห์ประสิทธิภาพของตัวลวง

6.6 การวิเคราะห์ค่าความเที่ยง

In Chapter 5, assessing the four language skills is dealt with. In this chapter, item analysis and test statistics that are used often will be dealt with. The order of presentation is as follows:

6.1 Purpose of item analysis

6.2 Item analysis for norm-referenced tests

6.3 Item analysis for criterion-referenced tests

6.4 Point-biserial correlation coefficients

6.5 Distractor efficiency analysis

6.6 Reliability analysis

6.1 เป้าหมายของการวิเคราะห์ข้อสอบ (Purpose of Item Analysis)

การวิเคราะห์ข้อสอบเป็นขั้นตอนหนึ่งในกระบวนการพัฒนาแบบสอบ กระบวนการพัฒนาแบบสอบมีขั้นตอนหลักๆ (Brown 2016: 63) ดังนี้

(1) รวบรวมข้อสอบและชนิดข้อสอบปริมาณค่อนข้างมาก ที่จะนำไปอยู่ในแบบสอบ

(2) วิเคราะห์ข้อสอบอย่างระมัดระวัง เพื่อให้แน่ใจว่า ข้อสอบเขียนมาดีและชัดเจน

(3) ทดสอบข้อสอบโดยใช้ผู้เรียนกลุ่มที่มีลักษณะเหมือนกลุ่มเป้าหมายที่จะต้องทำข้อสอบในสภาพการณ์ที่แย่กว่านี้ การทดสอบข้อสอบนี้คือการใช้ข้อสอบเป็นครั้งแรก

(4) วิเคราะห์ผลของการทดสอบข้อสอบ โดยใช้เทคนิคการวิเคราะห์ข้อสอบ เทคนิคสำหรับการทดสอบแบบอิงกลุ่มจะกล่าวถึงในหัวข้อที่ 6.2 ส่วนเทคนิคสำหรับการทดสอบแบบอิงเกณฑ์จะกล่าวถึงในหัวข้อที่ 6.3

(5) เลือกข้อสอบที่มีประสิทธิผลมากที่สุดและทิ้งข้อสอบที่ไม่มีประสิทธิผล เพื่อให้ได้เวอร์ชันแบบสอบที่สั้นกระชับกว่าและมีประสิทธิผลกว่า

We analyze items as a step in the test development process. A test development process (Brown 2016: 63) is as follows:

(1) Assemble a relatively large number of items and item types that you wish to include in your test.

(2) Analyze the items carefully, to ensure the items are clear and well-written.

(3) Pilot the items using the group of students that are comparable to the target students. In less desirable circumstances, the pilot testing is the first operational test administration.

(4) Analyze the results of the pilot test administration. Use techniques for item analysis. The techniques for norm-referenced tests will be discussed in Section 6.2. The techniques for criterion-referenced tests will be discussed in Section 6.3.

(5) Choose the items that are the most effective and discard the items that are ineffective. This will result in a shorter and more effective test version.

6.2 การวิเคราะห์ข้อสอบสำหรับการทดสอบแบบอิงกลุ่ม (Item Analysis for Norm-referenced Tests)

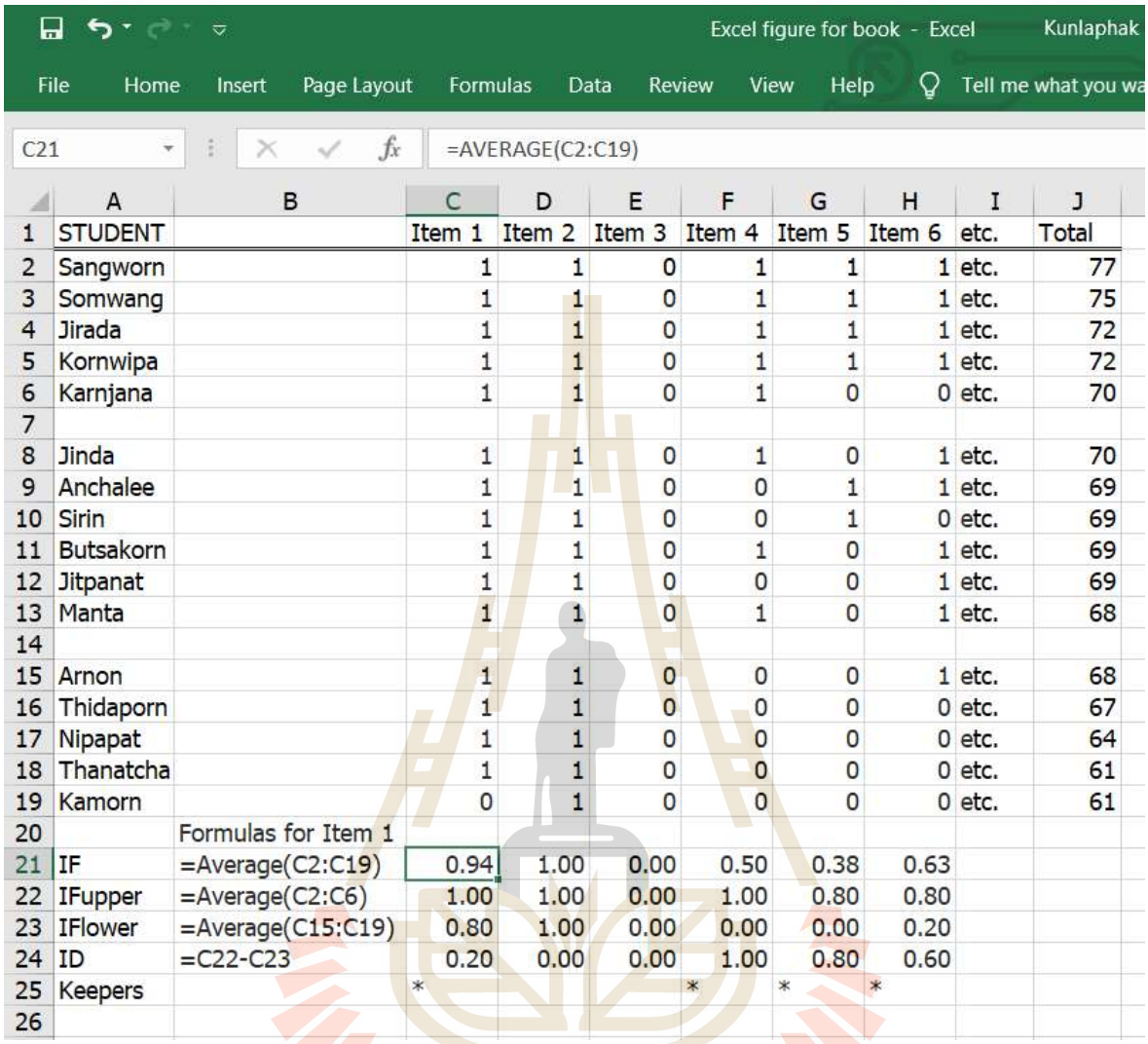
เป้าหมายของการทดสอบแบบอิงกลุ่มโดยทั่วไปคือเพื่อกระจายผู้เรียนออกบนเส้นต่อเนื่อง (continuum) ของความสามารถทางภาษา โดยปกติเพื่อทำการตัดสินใจเกี่ยวกับความถนัด สมิทธิภาพ หรือวัดระดับก่อนเข้าชั้นเรียน สถิติข้อสอบที่ใช้บ่อยได้แก่ ค่าความยาก และค่าอำนาจจำแนก (Brown 2016)

The general purpose of norm-referenced tests is to spread the learners out along a continuum of language ability, normally for making aptitude, proficiency, or placement decisions. Item statistics that are used often are item facility and item discrimination (Brown 2016).

ค่าความยาก (Item facility – IF)

ค่าความยากคือสัดส่วนจำนวนผู้เรียนที่ตอบข้อสอบข้อหนึ่งๆ ได้อย่างถูกต้อง เช่น ถ้าในข้อสอบข้อหนึ่ง มีผู้เรียน 40 คนจาก 50 คนตอบข้อสอบข้อนี้ถูกต้อง สัดส่วนจำนวนผู้เรียนที่ตอบข้อสอบข้อนี้ถูกต้องเท่ากับ 40/50 = 0.8 ค่าความยากที่ 0.8 จึงหมายความว่า 80% ของผู้เรียนตอบข้อสอบข้อนั้นถูกต้อง ซึ่งก็หมายความด้วยว่าข้อสอบข้อนี้ง่าย ในรูปที่ 6.1 แสดงวิธีการคำนวณค่าความยากวิธีหนึ่ง โดยใช้โปรแกรมสเปรดชีท Excel® สำหรับข้อสอบข้อที่ 1 ในตัวอย่างชุดข้อมูลที่ซึ่ง 1 มีค่าเท่ากับคำตอบที่ถูกต้อง และ 0 เท่ากับคำตอบที่ผิด ทั้งนี้ ในเซลที่ C21 คือค่าความยากข้อสอบข้อแรกในชุดข้อมูล คือ 0.94 โดยสูตรคำนวณแสดงอยู่ที่เซลที่ B21 ซึ่งก็คือ =AVERAGE(C2:C19) เป็นชุดคำสั่งให้โปรแกรมหาค่าเฉลี่ยของคะแนนตั้งแต่เซลที่ C2 ถึง C19 ค่าความยากที่เหมาะสมอยู่ระหว่าง 0.3 ถึง 0.7

Item facility (IF) is the proportion of students who can answer a particular test item correctly. For example, in one particular item, 40 students out of 50 students can answer this item correctly. The proportion will be 40/50 = 0.8. The IF of 0.8 thus means that 80% of the students can answer that particular item correctly, and so the item is easy. In Figure 6.1, how to calculate IF by using the spreadsheet program Excel® is shown for item 1, in cell C21. The answer 1 is a correct answer, and the answer 0 is an incorrect answer. The formula, =AVERAGE(C2:C19) is shown in cell B21, in which the average is sought from cell C2 to cell C19 for item 1. Appropriate IFs range from 0.3 to 0.7.



	A	B	C	D	E	F	G	H	I	J
1	STUDENT		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	etc.	Total
2	Sangworn		1	1	0	1	1	1	etc.	77
3	Somwang		1	1	0	1	1	1	etc.	75
4	Jirada		1	1	0	1	1	1	etc.	72
5	Kornwipa		1	1	0	1	1	1	etc.	72
6	Karnjana		1	1	0	1	0	0	etc.	70
7										
8	Jinda		1	1	0	1	0	1	etc.	70
9	Anchalee		1	1	0	0	1	1	etc.	69
10	Sirin		1	1	0	0	1	0	etc.	69
11	Butsakorn		1	1	0	1	0	1	etc.	69
12	Jitpanat		1	1	0	0	0	1	etc.	69
13	Manta		1	1	0	1	0	1	etc.	68
14										
15	Arnon		1	1	0	0	0	1	etc.	68
16	Thidaporn		1	1	0	0	0	0	etc.	67
17	Nipapat		1	1	0	0	0	0	etc.	64
18	Thanatcha		1	1	0	0	0	0	etc.	61
19	Kamorn		0	1	0	0	0	0	etc.	61
20		Formulas for Item 1								
21	IF	=Average(C2:C19)	0.94	1.00	0.00	0.50	0.38	0.63		
22	IFupper	=Average(C2:C6)	1.00	1.00	0.00	1.00	0.80	0.80		
23	IFlower	=Average(C15:C19)	0.80	1.00	0.00	0.00	0.00	0.20		
24	ID	=C22-C23	0.20	0.00	0.00	1.00	0.80	0.60		
25	Keepers		*			*	*	*		
26										

รูปที่ 6.1 วิเคราะห์ข้อสอบแบบอิงกลุ่ม (Norm-referenced item analysis)

<u>ค่าอำนาจจำแนก (Item discrimination – ID)</u>

ค่าอำนาจจำแนกคือการที่ข้อสอบแต่ละข้อสามารถแยกแยะผู้สอบที่มีความสามารถสูงออกจากผู้สอบที่มีความสามารถต่ำได้ ในรูปที่ 6.1 การคำนวณเริ่มจากการใช้คะแนนรวม (คอลัมน์ J) แยกแยะว่าผู้สอบคนใดอยู่กลุ่มสูงและผู้สอบคนใดอยู่กลุ่มต่ำ โดยทั้งสองกลุ่มควรมีจำนวนเท่าๆ กัน และมีจำนวนกลุ่มละประมาณหนึ่งในสามของจำนวนผู้สอบทั้งหมด จากนั้นก็คำนวณค่าความยากของกลุ่มสูง โดยสำหรับข้อ 1 ใช้สูตรคำนวณ =AVERAGE(C2:C6) ดังแสดงในแถวที่ 22 จากนั้น คำนวณค่าความยากของกลุ่มต่ำสำหรับข้อ 1 โดยใช้สูตรคำนวณ =AVERAGE(C15:C19) ดังแสดงในแถวที่ 23 เมื่อได้ค่าความยากของทั้งกลุ่มสูงและกลุ่มต่ำ ก็นำค่ากลุ่มสูงลบด้วยค่ากลุ่มต่ำ IFupper – IFlower (=C22-C23) ดังแสดงใน

แถวที่ 24 ได้ผลลัพธ์เป็นค่าอำนาจจำแนกของข้อ 1 เท่ากับ 0.2 โดยทั่วไป เมื่อเราได้ข้อสอบที่มีค่าความยากเหมาะสมแล้ว เราจะเลือกเก็บข้อสอบที่มีค่าอำนาจจำแนกสูงสุดก่อน ดังแสดงในแถวที่ 25

Item discrimination (ID) is when each item can discriminate those that have high proficiency from those that have low proficiency. In Figure 6.1, the calculation begins with sorting out the test takers using their total scores (column J). They are then arranged into a high proficiency group, middle proficiency group, and low proficiency group, each of which has around one third of all the test takers. Then the IF of the high group is calculated for item 1 using the following formula: =AVERAGE(C2:C6), as displayed in row 22. The IF of the low group is calculated for item 1 using the following formula: =AVERAGE(C15:C19), as displayed in row 23. With the IFupper and IFlower ready, the ID of item can be obtained by substracting IFupper with IFlower, =C22-C23, resulting in 0.2 in row 24. Generally, once we get items with appropriate difficulty, we select those with the highest ID first from them. This is shown in row 25.

6.3 การวิเคราะห์ข้อสอบสำหรับการทดสอบแบบอิงเกณฑ์ (Item Analysis for Criterion-referenced Tests)

เป้าหมายหลักของการทดสอบแบบอิงเกณฑ์คือเพื่อวัดปริมาณหรือเปอร์เซ็นต์ของเนื้อหาในรายวิชาหรือหลักสูตรหนึ่งๆ ที่ผู้เรียนรู้หรือสามารถทำได้ โดยปกติเพื่อทำการตัดสินใจในการทดสอบวินิจฉัยการเรียน ในการทดสอบความก้าวหน้า หรือในการวัดผลสัมฤทธิ์ทางการเรียน สถิติข้อสอบที่ใช้บ่อยได้แก่ ดัชนีความแตกต่าง และดัชนีบี (Brown 2016)

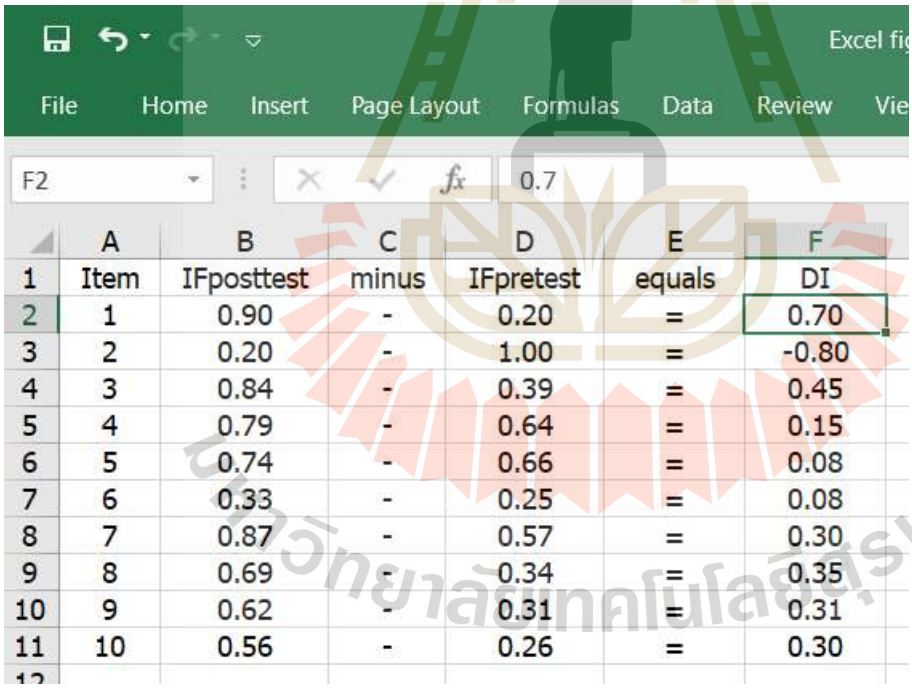
The main purpose of criterion-referenced tests is to assess the amount or percent of the material in a particular course or program that students know. This is usually for decision making in diagnostic tests, in progress tests, or in achievement tests. Statistics that are used often include the difference index and the *B*-index (Brown 2016).

<u>ดัชนีความแตกต่าง (Difference index – DI)</u>

ดัชนีความแตกต่างคำนวณโดยค่าความยากของข้อสอบข้อหนึ่งๆ ในการทดสอบหลังเรียนลบด้วยค่าความยากของข้อสอบข้อนั้นในการทดสอบก่อนเรียน กล่าวอีกนัยหนึ่ง ดัชนีความแตกต่างแสดงส่วนเพิ่มหรือความแตกต่างในการแสดงออกของข้อสอบแต่ละข้อระหว่างการทดสอบก่อนเรียนและการ

ทดสอบหลังเรียน ตัวอย่างวิธีคำนวณเช่น ในการทดสอบก่อนเรียน มีผู้เรียน 10 คนจาก 50 คนตอบข้อสอบข้อที่ 1 ได้ถูกต้อง ค่าความยากของการทดสอบก่อนเรียนในข้อสอบข้อนี้ (IFpretest) จึงเท่ากับ 10/50 = 0.20 ถ้าในข้อสอบข้อเดียวกัน มีผู้เรียน 45 คนจาก 50 คน ตอบได้ถูกต้อง ค่าความยากของการทดสอบหลังเรียน (IFposttest) จึงเท่ากับ 45/50 = 0.90 ค่าดัชนีความแตกต่างของข้อสอบข้อนี้ จึงเท่ากับ 0.70 (DI = IFposttest – IFpretest = 0.90 – 0.20 = 0.70) รูปที่ 6.2 แสดงการคำนวณค่าดัชนีความแตกต่างของข้อสอบข้อนี้

The difference index (DI) is calculated by subtracting an IFposttest with an IFpretest of the same item. That is, the DI shows the gain or difference of each test item between a pretest and a posttest. For example, in item 1 of Figure 6.2, 10 out of 50 test takers answer item 1 correctly in the pretest (IFpretest = 10/50 = 0.2). In the posttest, 45 out of the same 50 test takers answer item 1 correctly (IFposttest = 45/50 = 0.9). The DI of item 1 equals IFposttest – IFpretest = 0.90 – 0.20 = 0.70.



	A	B	C	D	E	F
1	Item	IFposttest	minus	IFpretest	equals	DI
2	1	0.90	-	0.20	=	0.70
3	2	0.20	-	1.00	=	-0.80
4	3	0.84	-	0.39	=	0.45
5	4	0.79	-	0.64	=	0.15
6	5	0.74	-	0.66	=	0.08
7	6	0.33	-	0.25	=	0.08
8	7	0.87	-	0.57	=	0.30
9	8	0.69	-	0.34	=	0.35
10	9	0.62	-	0.31	=	0.31
11	10	0.56	-	0.26	=	0.30

รูปที่ 6.2 วิเคราะห์ข้อสอบแบบอิงเกณฑ์ (Criterion-referenced item analysis)

วิธีพิจารณาเหมือนค่าอำนาจจำแนก กล่าวคือ ค่าดัชนีความแตกต่างยิ่งสูงก็ยิ่งดี ในรูปที่ 6.2 ข้อที่ 1, 3 และ 7-10 เกี่ยวข้องกับหลักสูตรดีกว่าข้อที่ 2 และ 4-6 เพราะมีค่าดัชนีความแตกต่างสูงกว่า

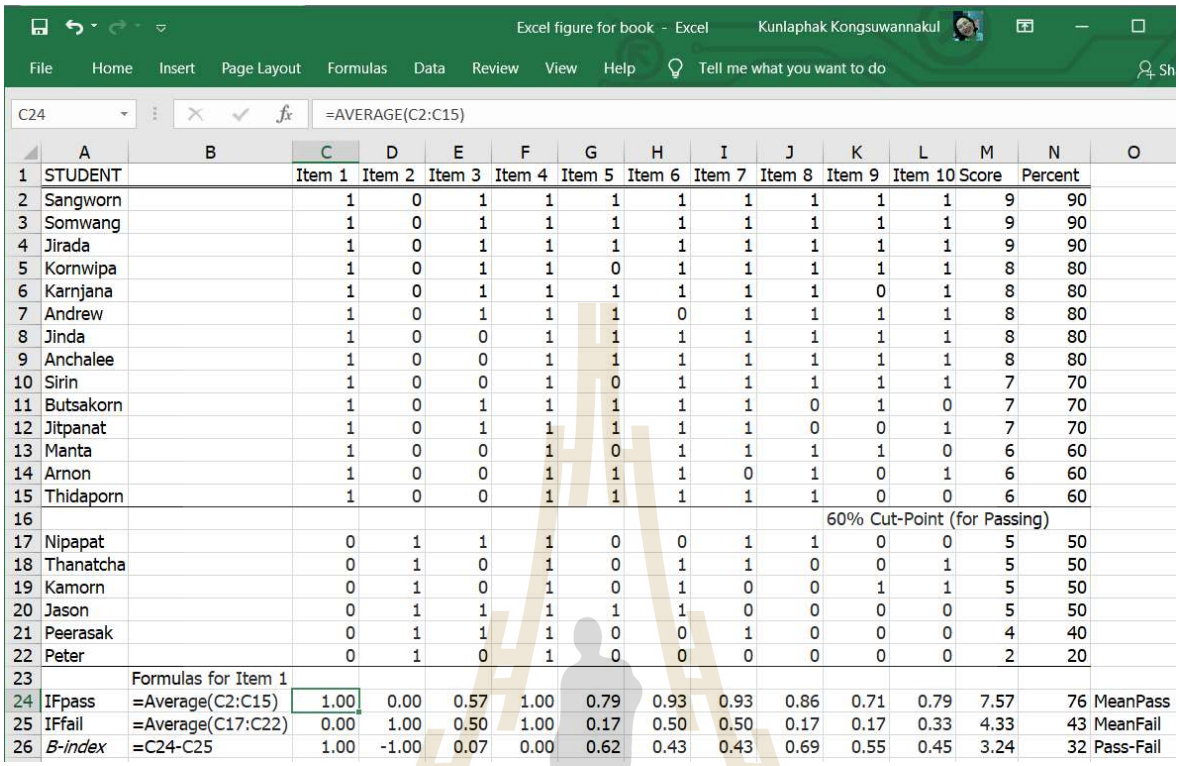
ข้อที่ 4-6 มีความไม่เหมาะสมเพราะมีส่วนเพิ่มเพียงเล็กน้อย หรือก็คือค่าดัชนีต่ำมาก ส่วนข้อ 2 ซึ่งมีค่าดัชนีความแตกต่างติดลบ แสดงให้เห็นว่า ระหว่างคอร์สเรียน 80% ของผู้เรียนที่ตอนเริ่มเรียนมีความรู้ในข้อนี้ได้เพิกถอนสิ่งที่รู้ ณ ตอนปลายคอร์ส

Like item discrimination, the higher the value of a DI, the better. In Figure 6.2, items 1, 3 and 7-10 are better related to the curriculum than items 2, and 4-6. This is so because they have higher DI values. Items 4-6 are not suitable because they have low DI values and thus only small gains. Item 2, with a negative value, indicates that 80% of the students who started out knowing the content of this item unlearned the content by the end of the course.

ดัชนีบี (*B*-index)

ดัชนีบีคำนวณโดยค่าความยากในข้อสอบข้อหนึ่งๆ ของผู้สอบที่สอบผ่าน ลบด้วยค่าความยากของข้อสอบข้อเดียวกันของผู้สอบที่สอบตก ดังนั้น ดัชนีบีแสดงว่าข้อสอบแต่ละข้อสนับสนุนการสอบตกหรือสอบผ่านมากน้อยเพียงใด ในรูปที่ 6.3 ผู้สอบที่สอบผ่านทั้ง 14 คนทำข้อสอบข้อที่ 1 ถูกต้อง ค่าความยากสำหรับผู้ที่สอบผ่าน (IFpass) จึงเท่ากับ 14/14 = 1.00 ส่วนคนที่สอบตกทุกคนทำข้อสอบข้อนี้ผิด ค่าความยากสำหรับผู้ที่สอบตก (IFfail) จึงเท่ากับ 0/6 = 0 จากข้อมูลนี้ ดัชนีบีสำหรับข้อที่ 1 จึงเท่ากับ 1 (*B*-index = IFpass – IFfail = 1.00 – 0.00 = 1.00)

The *B*-index is calculated by substracting the item facility of those who pass the test with the item facility of those who fail the test. So, the *B*-index shows how well each item contributes to the pass/fail decisions. In Figure 6.3, those 14 test takers who pass the test do item 1 correctly, and their item facility (IFpass) is thus 14/14 = 1.00. Those 6 test takers who fail the test do item 1 incorrectly, and their item facility (IFfail) is thus 0/6 = 0. With IFpass = 1.00 and IFfail = 0 in hand, the *B*-index for item 1 is thus 1 (*B*-index = IFpass – IFfail = 1.00 – 0.00 = 1.00).



Excel figure for book - Excel Kunlaphak Kongsuwannakul

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

C24 =AVERAGE(C2:C15)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	STUDENT		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Score	Percent	
2	Sangworn		1	0	1	1	1	1	1	1	1	1	9	90	
3	Somwang		1	0	1	1	1	1	1	1	1	1	9	90	
4	Jirada		1	0	1	1	1	1	1	1	1	1	9	90	
5	Kornwipa		1	0	1	1	0	1	1	1	1	1	8	80	
6	Karnjana		1	0	1	1	1	1	1	1	0	1	8	80	
7	Andrew		1	0	1	1	1	0	1	1	1	1	8	80	
8	Jinda		1	0	0	1	1	1	1	1	1	1	8	80	
9	Anchalee		1	0	0	1	1	1	1	1	1	1	8	80	
10	Sirin		1	0	0	1	0	1	1	1	1	1	7	70	
11	Butsakorn		1	0	1	1	1	1	1	0	1	0	7	70	
12	Jitpanat		1	0	1	1	1	1	1	0	0	1	7	70	
13	Manta		1	0	0	1	0	1	1	1	1	0	6	60	
14	Arnon		1	0	0	1	1	1	0	1	0	1	6	60	
15	Thidaporn		1	0	0	1	1	1	1	1	0	0	6	60	
16											60% Cut-Point (for Passing)				
17	Nipapat		0	1	1	1	0	0	1	1	0	0	5	50	
18	Thanatcha		0	1	0	1	0	1	1	0	0	1	5	50	
19	Kamorn		0	1	0	1	0	1	0	0	1	1	5	50	
20	Jason		0	1	1	1	1	1	0	0	0	0	5	50	
21	Peerasak		0	1	1	1	0	0	1	0	0	0	4	40	
22	Peter		0	1	0	1	0	0	0	0	0	0	2	20	
23		Formulas for Item 1													
24	IFpass	=Average(C2:C15)	1.00	0.00	0.57	1.00	0.79	0.93	0.93	0.86	0.71	0.79	7.57	76	MeanPass
25	IFfail	=Average(C17:C22)	0.00	1.00	0.50	1.00	0.17	0.50	0.50	0.17	0.17	0.33	4.33	43	MeanFail
26	*B-index*	=C24-C25	1.00	-1.00	0.07	0.00	0.62	0.43	0.43	0.69	0.55	0.45	3.24	32	Pass-Fail

รูปที่ 6.3 คำนวณดัชนีบีในสเปรดชีท (Calculating the *B*-index in a spreadsheet)

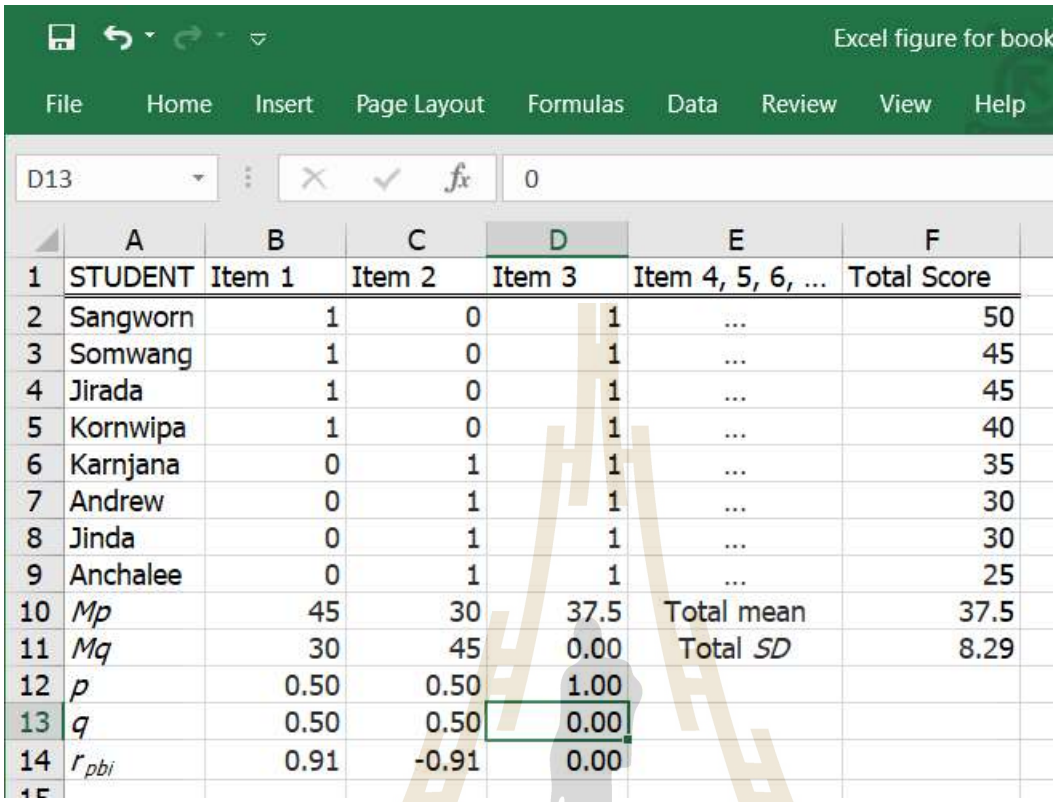
ดัชนีความแตกต่างและดัชนีบีใช้เพื่อวิเคราะห์ข้อสอบในการทดสอบแบบอิงเกณฑ์ เป้าหมายคือเพื่อการปรับแก้ข้อสอบได้อย่างตรงจุด ในทั้งสองกรณี ข้อที่ได้ค่าดัชนีสูงสุดควรเก็บไว้ โดยค่าดัชนีความแตกต่างจะบอกเราว่าข้อใดไม่เข้ากับวัตถุประสงค์ของหลักสูตรเรา ส่วนค่าดัชนีบีจะบอกเราว่าข้อสอบแต่ละข้อสนับสนุนการตัดสินใจว่าใครสอบตกสอบผ่านมากน้อยเพียงใด

The difference index and the *B*-index are used for analyzing items in criterion-referenced tests. The purpose is to revise the items efficiently because areas needing improvement are identified. In both cases, the items that gain high indices should be kept. The difference index will tell us how well each of the items fits the objectives of the curriculum. The *B*-index will tell us how well each of the items is contributing to the pass/fail decisions.

6.4 สัมประสิทธิ์สหสัมพันธ์พอยต์ไบซีเรียล (Point-biserial correlation coefficients)

สัมประสิทธิ์สหสัมพันธ์พอยต์ไบซีเรียลเป็นสถิติที่ใช้ประมาณค่าความสัมพันธ์ระหว่างตัวแปรในมาตรวัดแบบนามบัญญัติทวิวิภาค (dichotomous nominal scale) ที่เกิดเองตามธรรมชาติ กับตัวแปรในมาตรวัดแบบช่วงชั้น (interval scale) (Brown 2016: 75ff.) ในรูปที่ 6.4 คำตอบถูกคือ 1 คะแนน และคำตอบผิดคือ 0 คะแนน อยู่ในรูปแบบนามบัญญัติทวิวิภาค (natural dichotomous nominal scale) ส่วนคะแนนในคอลัมน์ขวามือสุดเป็นตัวแปรในมาตรวัดแบบช่วงชั้น สัมประสิทธิ์สหสัมพันธ์พอยต์ไบซีเรียลเป็นอีกวิธีหนึ่งที่สามารถใช้ในการคำนวณค่าอำนาจจำแนกของข้อสอบได้

The point-biserial correlation coefficient (r_{pbi}) is a statistic that is used for estimating the degree of relationship between a naturally occurring dichotomous nominal scale and an interval scale (Brown 2016: 75ff.). In Figure 6.4, a correct answer is scored as 1, and an incorrect answer as 0, both of which are naturally in a dichotomous nominal scale. In the rightmost column, the scores are in an interval scale. The point-biserial correlation coefficient is another statistic that can be used for calculating item discrimination.



	A	B	C	D	E	F
1	STUDENT	Item 1	Item 2	Item 3	Item 4, 5, 6, ...	Total Score
2	Sangworn	1	0	1	...	50
3	Somwang	1	0	1	...	45
4	Jirada	1	0	1	...	45
5	Kornwipa	1	0	1	...	40
6	Karnjana	0	1	1	...	35
7	Andrew	0	1	1	...	30
8	Jinda	0	1	1	...	30
9	Anchalee	0	1	1	...	25
10	*Mp*	45	30	37.5	Total mean	37.5
11	*Mq*	30	45	0.00	Total *SD*	8.29
12	*p*	0.50	0.50	1.00		
13	*q*	0.50	0.50	0.00		
14	r_{pbi}	0.91	-0.91	0.00		

รูปที่ 6.4 คำนวณ r_{pbi} (Calculating the r_{pbi})

ในการคำนวณ r_{pbi} สำหรับแต่ละข้อ ให้ใช้สูตรดังนี้

In calculating r_{pbi} for each item, use the formula that follows:

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq}$$

โดยที่

r_{pbi} = ค่าสัมประสิทธิ์สหสัมพันธ์พอยต์ไบซีเรียล

M_p = ค่าเฉลี่ยของทั้งข้อสอบของผู้เรียนที่ตอบข้อสอบข้อนั้นๆ ถูกต้อง

M_q = ค่าเฉลี่ยของทั้งข้อสอบของผู้เรียนที่ตอบข้อสอบข้อนั้นๆ ผิด

S_t = ค่าส่วนเบี่ยงเบนมาตรฐานของทั้งข้อสอบ

p = สัดส่วนผู้เรียนที่ตอบถูก (ที่ใส่รหัสไว้เป็น 1)

q = สัดส่วนผู้เรียนที่ตอบผิด (ที่ใส่รหัสไว้เป็น 0)

Where:

r_{pbi} = point-biserial correlation coefficient

M_p = whole-test mean for students who answer the item correctly

M_q = whole-test mean for students who answer the item incorrectly

S_t = standard deviation for the whole test

p = proportion of students that answer correctly (those coded as 1s)

q = proportion of students that answer incorrectly (those coded as 0s)

ตัวอย่างในการคำนวณ เช่น ในข้อ 1 ในรูปที่ 6.4 ค่าเฉลี่ยของคะแนนทั้งแบบสอบของผู้เรียนที่ตอบถูกคือ 45 คะแนน ส่วนค่าเฉลี่ยของคะแนนทั้งแบบสอบของผู้ที่ตอบผิดคือ 30 คะแนน ค่าส่วนเบี่ยงเบนมาตรฐานของทั้งแบบสอบคือ 8.29 สัดส่วนของผู้ตอบถูกคือ 0.5 และสัดส่วนของผู้ตอบผิดคือ 0.5

For example, in item 1 in Figure 6.4, the whole-test mean of the students that get this item right is 45. The whole-test mean of the students that get this item wrong is 30. The standard deviation of the whole test is 8.29. The proportion of those answering correctly is 0.5, and the proportion of those answering incorrectly is 0.5.

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq} = \frac{45 - 30}{8.29}\sqrt{(0.5)(0.5)} = \frac{15}{8.29}\sqrt{0.25} = 1.81(0.5) = 0.91$$

ฉะนั้น ค่าสหสัมพันธ์ระหว่างข้อที่ 1 กับคะแนนรวมทั้งหมดสูงมากที่ 0.91 และข้อสอบข้อนี้กระจายผู้เรียนออกในลักษณะเดียวกันกับคะแนนรวม ในลักษณะนี้ ค่าสัมประสิทธิ์สหสัมพันธ์พอยต์ไบซีเรียลบ่งชี้ว่า ข้อสอบข้อที่ 1 จำแนกแยกแยะผู้เรียนในกลุ่มนี้ได้ดี

Therefore, the correlation between item 1 and the total score is very high, at 0.91. This item appears to spread the test takers out in the same way that the total scores can. In this way, the point-biserial correlation coefficient indicates that item 1 can discriminate this group of students well.

อีกหนึ่งตัวอย่างในการคำนวณได้แก่ข้อ 3 ในรูปที่ 6.4 ค่าเฉลี่ยคะแนนของทั้งแบบสอบของผู้ที่ตอบถูกคือ 37.5 ส่วนค่าเฉลี่ยของคะแนนของผู้ที่ตอบผิดคือ 0 ส่วนเบี่ยงเบนมาตรฐานของทั้งแบบสอบยังคงเป็น 8.29 สัดส่วนนักศึกษาที่ตอบถูกคือ 1.00 และสัดส่วนนักศึกษาที่ตอบผิดคือ 0

Another example for calculation is item 3 in Figure 6.4. The whole-test mean for those answering correctly is 37.5, and the whole-test mean for those answering incorrectly is 0. The standard deviation of the whole test is still 8.29. The proportion of those answering correctly is 1.00, and the proportion of those answering incorrectly is 0.

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq} = \frac{37.5 - 0}{8.29}\sqrt{(1.00)(0)} = \frac{37.5}{8.29}\sqrt{0.00} = 4.52(0.00) = \mathbf{0.00}$$

ดังนั้น ค่าสหสัมพันธ์ระหว่างข้อที่ 3 และคะแนนรวมคือ 0 และข้อสอบข้อนี้มิได้กระจายผู้เรียนออกไปในลักษณะเดียวกันกับคะแนนรวม กล่าวอีกนัยหนึ่ง ข้อที่ 3 ไม่ได้จำแนกจำแนกผู้เรียนเลย ในกรณีนี้เพราะไม่มีความแตกต่างในคำตอบระหว่างกันเลย

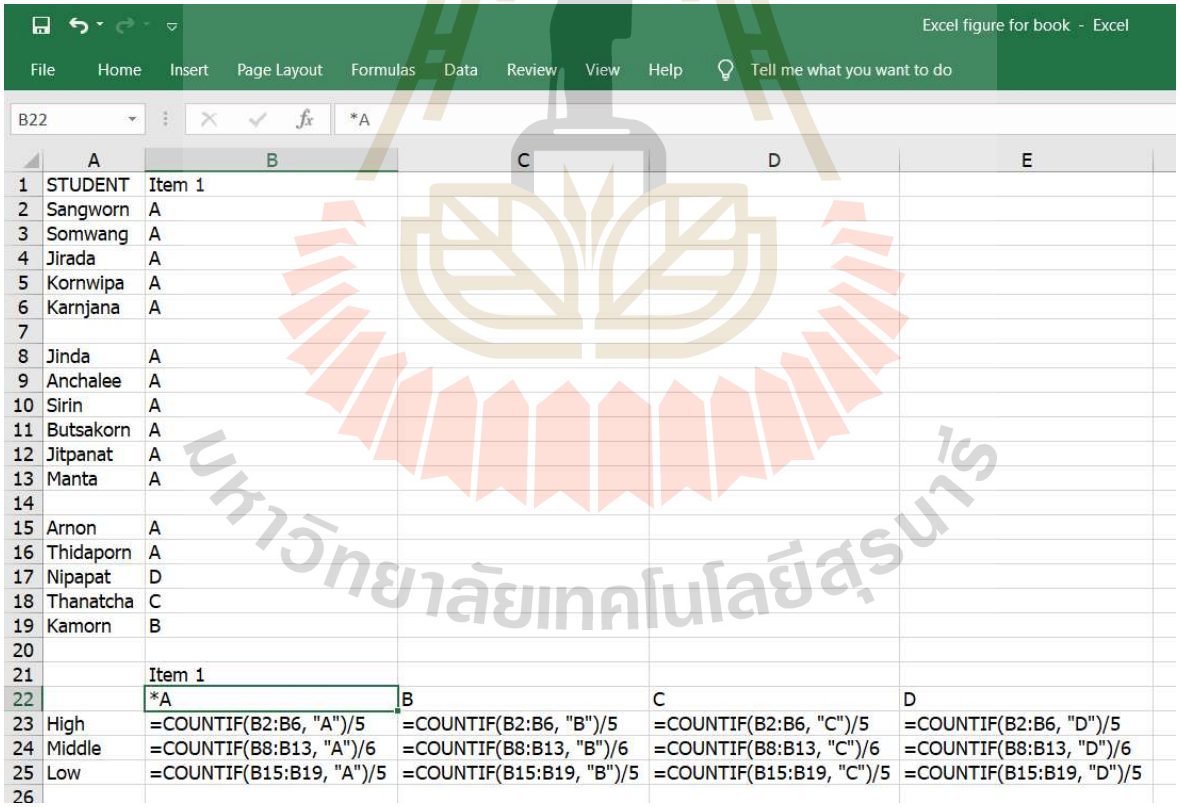
Therefore, the correlation between item 3 and the total scores is 0. The item does not spread the test takers out in the same way as the total scores. In other words, item 3 does not discriminate the test takers at all because, in this case, there is not any variation in terms of the answers.

6.5 การวิเคราะห์ประสิทธิภาพของตัวลวง (Distractor Efficiency Analysis)

ในการวิเคราะห์ประสิทธิภาพของตัวลวง (Brown 2016) เราใช้สูตรที่ปรากฏในสามแถวสุดท้ายของรูปที่ 6.5 ตัวอย่างเช่น ผู้สอบแบ่งออกเป็น 3 กลุ่มจากคะแนนรวมทั้งหมด (มิได้แสดงในรูปที่ 6.5) ได้แก่ กลุ่มสูง (Sangworn–Karnjana) กลุ่มกลาง (Jinda–Manta) และกลุ่มต่ำ (Arnon–Kamorn) ในข้อที่ 1 คำตอบที่ถูกต้องคือตัวเลือก A ผู้สอบกลุ่มต่ำเลือกตัวเลือก B หนึ่งคน จากสูตรคำนวณ =COUNTIF(B15:B19, “B”)/5 คิดเป็นร้อยละ 20 ของผู้สอบกลุ่มต่ำ ค่าประสิทธิภาพของตัวลวง B จึง

เท่ากับ 0.2 (1/5) เช่นเดียวกัน ตัวเลือก C และตัวเลือก D มีผู้สอบกลุ่มต่ำตัวเลือกละหนึ่งคน ค่าประสิทธิภาพตัวลวงคิดเป็น 0.2 ทั้งสองตัวเลือก

In analyzing distractor efficiency, we use the formulas shown in the last three rows in Figure 6.5. For example, the test takers can be grouped into three groups based on their total scores (not shown in the figure), namely the high group (Sangworn–Karnjana), the middle group (Jinda–Manta), and the low group (Arnon–Kamorn). In item 1, the key is option A, and one test taker in the low group chooses option B. From the formula =COUNTIF(B15:B19, “B”)/5, the distractor efficiency for this distractor is thus 0.2 (1/5). Similarly, option C and option D are chosen each by one test taker in the low group. The distractor efficiency of these options is thus 0.2.



	A	B	C	D	E
1	STUDENT	Item 1			
2	Sangworn	A			
3	Somwang	A			
4	Jirada	A			
5	Kornwipa	A			
6	Karnjana	A			
7					
8	Jinda	A			
9	Anchalee	A			
10	Sirin	A			
11	Butsakorn	A			
12	Jitpanat	A			
13	Manta	A			
14					
15	Arnon	A			
16	Thidaporn	A			
17	Nipapat	D			
18	Thanatcha	C			
19	Kamorn	B			
20					
21		Item 1			
22		*A	B	C	D
23	High	=COUNTIF(B2:B6, "A")/5	=COUNTIF(B2:B6, "B")/5	=COUNTIF(B2:B6, "C")/5	=COUNTIF(B2:B6, "D")/5
24	Middle	=COUNTIF(B8:B13, "A")/6	=COUNTIF(B8:B13, "B")/6	=COUNTIF(B8:B13, "C")/6	=COUNTIF(B8:B13, "D")/6
25	Low	=COUNTIF(B15:B19, "A")/5	=COUNTIF(B15:B19, "B")/5	=COUNTIF(B15:B19, "C")/5	=COUNTIF(B15:B19, "D")/5
26					

รูปที่ 6.5 คำนวณประสิทธิภาพของตัวลวง (Calculating distractor efficiency)

6.6 การวิเคราะห์ค่าความเที่ยง (Reliability Analysis)

ในการประมาณค่าความเที่ยง มีสามวิธีที่สามารถใช้ได้แก่ การทดสอบด้วยแบบสอบฉบับเดียวสองรอบแล้วหาค่าสหสัมพันธ์ระหว่างคะแนนในทั้งสองรอบ การใช้แบบสอบที่เทียบเท่า (equivalent or parallel forms) ทดสอบแล้วหาค่าสหสัมพันธ์ระหว่างคะแนนในทั้งสองแบบสอบ และวิธีที่สามได้แก่ การหาค่าความคงเส้นคงวาภายในแบบสอบ ในเชิงปฏิบัติ ข้อสอบส่วนใหญ่รายงานค่าความคงเส้นคงวาภายในเป็นสัมประสิทธิ์ความเที่ยง (Fulcher & Davidson 2007: 106ff.) ค่าเหล่านี้จริงๆ แล้วเป็นการวัดค่าสหสัมพันธ์เฉลี่ยระหว่างข้อสอบ หรือการวัดว่า ข้อสอบมีสหสัมพันธ์กันมากน้อยเพียงใด นี่คือคำนิยามหนึ่งของความคงเส้นคงวา อย่างไรก็ดี สัมประสิทธิ์ความเที่ยงภายในได้รับผลกระทบจากอีกหลายปัจจัย เช่น

1) จำนวนข้อในข้อสอบ การเพิ่มจำนวนข้อจะเพิ่มค่าความเที่ยงได้

2) ความหลากหลายในเรื่องค่าความยาก ข้อสอบควรยากเท่าๆ กัน เพื่อเพิ่มความเที่ยง ข้อสอบที่มีความยากหลากหลาย ค่าความเที่ยงจะลดลง

3) การกระจายของคะแนน ถ้ากลุ่มตัวอย่างความสามารถเท่าๆ กัน ไม่มีการกระจายตัวของคะแนนในกลุ่ม ค่าความเที่ยงจะลดลง

4) ระดับของความยากข้อสอบ ข้อสอบที่มีค่าความยาก 0.5 จะเพิ่มค่าความเที่ยงได้มากที่สุด

In estimating reliability, there are three different strategies that can be used for the purpose. First, correlation can be sought from one test being administered twice. Second, correlation can be sought from one test and its equivalent or parallel form. Finally, internal consistency can be sought from one test being administered only once. Most tests, in practice, report measures of internal consistency as reliability coefficients (Fulcher & Davidson 2007: 106ff.). The measures, in fact, are simply measures of average inter-item correlation or those of how well items correlate with one another. However, internal reliability coefficients can be affected by other factors, namely:

1) The number of items on the test. Increasing the number of items usually increases reliability.

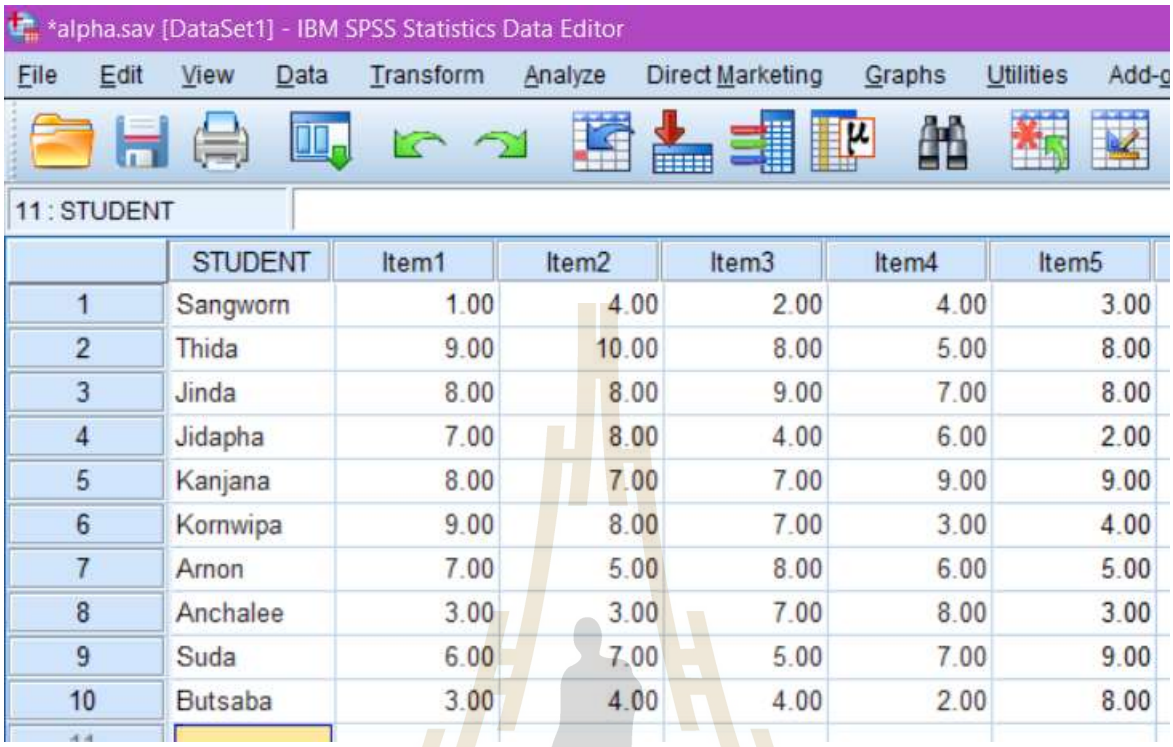
2) Variation in terms of item difficulty. To increase reliability, items should be equally difficult. If items are of a range of difficulty, reliability will decrease.

3) The dispersion of scores. If the sample of the test is homogeneous, reliability will decrease. If the sample is of a range of abilities, reliability will increase.

4) The level of item difficulty. The items whose facility values are 0.5 can increase item variance and thus test reliability.

ต่อไปนี้เป็นวิธีคำนวณค่าความคงเส้นคงวาภายในโดยใช้การคำนวณความเที่ยงครอนบาคแอลฟาในโปรแกรม SPSS ขั้นตอนแรกคือการจัดเตรียมข้อมูลคะแนนในโปรแกรม SPSS ในตัวอย่างรูปที่ 6.6 มีผู้สอบทั้งสิ้น 10 คน แต่ละคนทำข้อสอบคนละ 5 ข้อ แต่ละข้อคะแนนเต็ม 10 คะแนน

The following is a way to calculate internal consistency using the Cronbach alpha reliability estimation in SPSS. The first step is to prepare data in the SPSS program. In the example, Figure 6.6, there are 10 test takers, each of whom do five items. Each of the items is 10 points maximum.

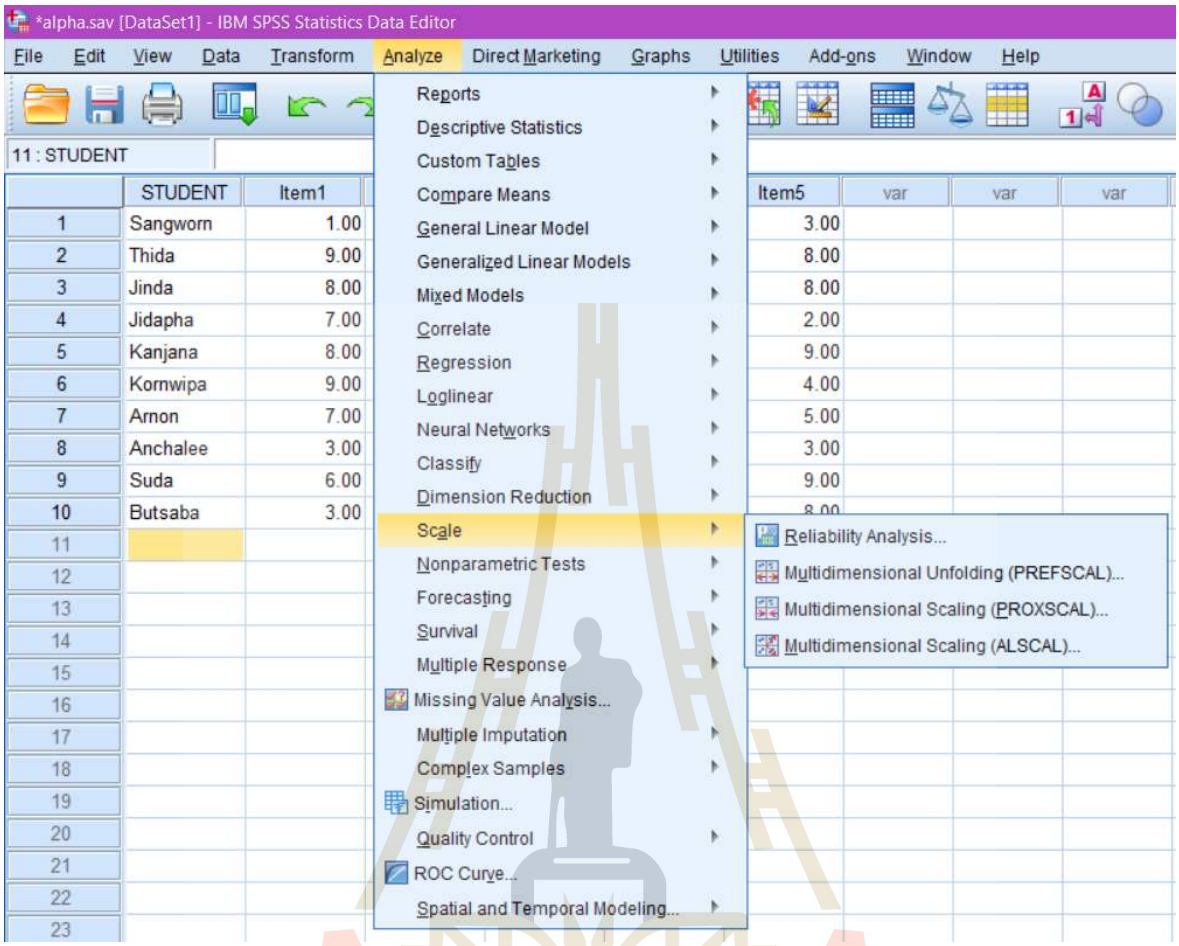


	STUDENT	Item1	Item2	Item3	Item4	Item5
1	Sangworn	1.00	4.00	2.00	4.00	3.00
2	Thida	9.00	10.00	8.00	5.00	8.00
3	Jinda	8.00	8.00	9.00	7.00	8.00
4	Jidapha	7.00	8.00	4.00	6.00	2.00
5	Kanjana	8.00	7.00	7.00	9.00	9.00
6	Kornwipa	9.00	8.00	7.00	3.00	4.00
7	Arnon	7.00	5.00	8.00	6.00	5.00
8	Anchalee	3.00	3.00	7.00	8.00	3.00
9	Suda	6.00	7.00	5.00	7.00	9.00
10	Butsaba	3.00	4.00	4.00	2.00	8.00

รูปที่ 6.6 เตรียมข้อมูลคะแนนใน SPSS (Preparing data in SPSS)

ในขั้นตอนที่ 2 (รูปที่ 6.7) ให้เลือกแท็บวิเคราะห์ (Analyze) เลือกฟังก์ชันสเกล (‘Scale’ function) และเลือก Reliability Analysis

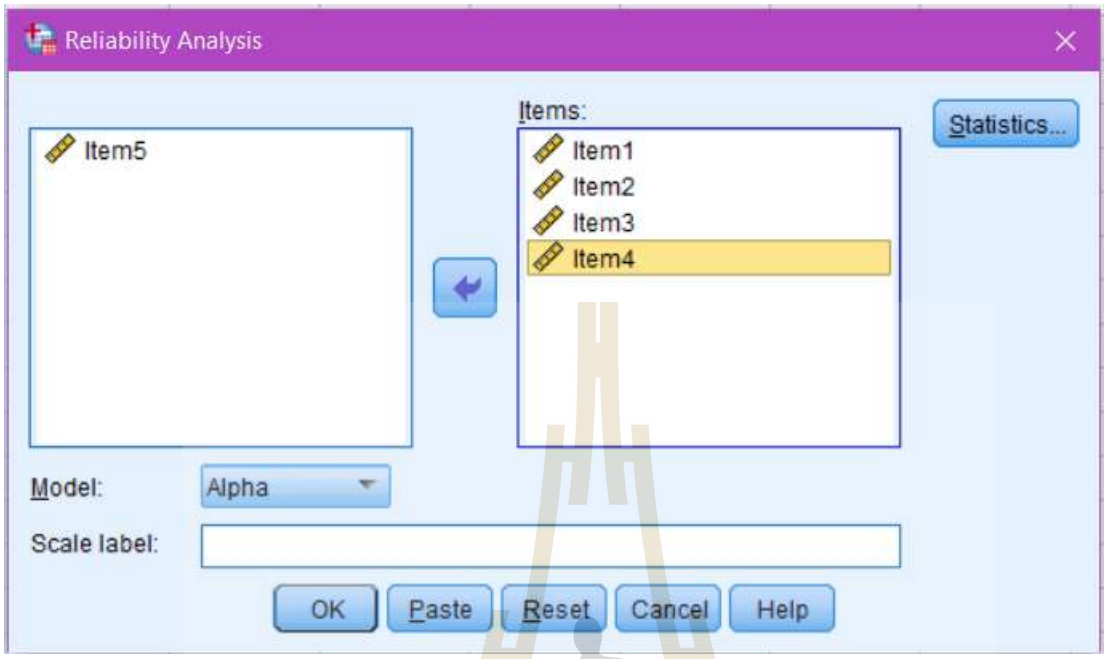
In the second step (Figure 6.7), choose the tab ‘Analyze’, then the function ‘Scale’, and ‘Reliability Analysis’.



รูปที่ 6.7 เลือกฟังก์ชันการคำนวณค่าความเที่ยง (Choosing the reliability function)

ในขั้นตอนที่ 3 (รูปที่ 6.8) ย้ายข้อสอบทั้ง 5 ข้อไปที่ช่องตัวแปรที่จะวิเคราะห์ทางด้านขวามือ เลือกโมเดล Alpha และกด OK

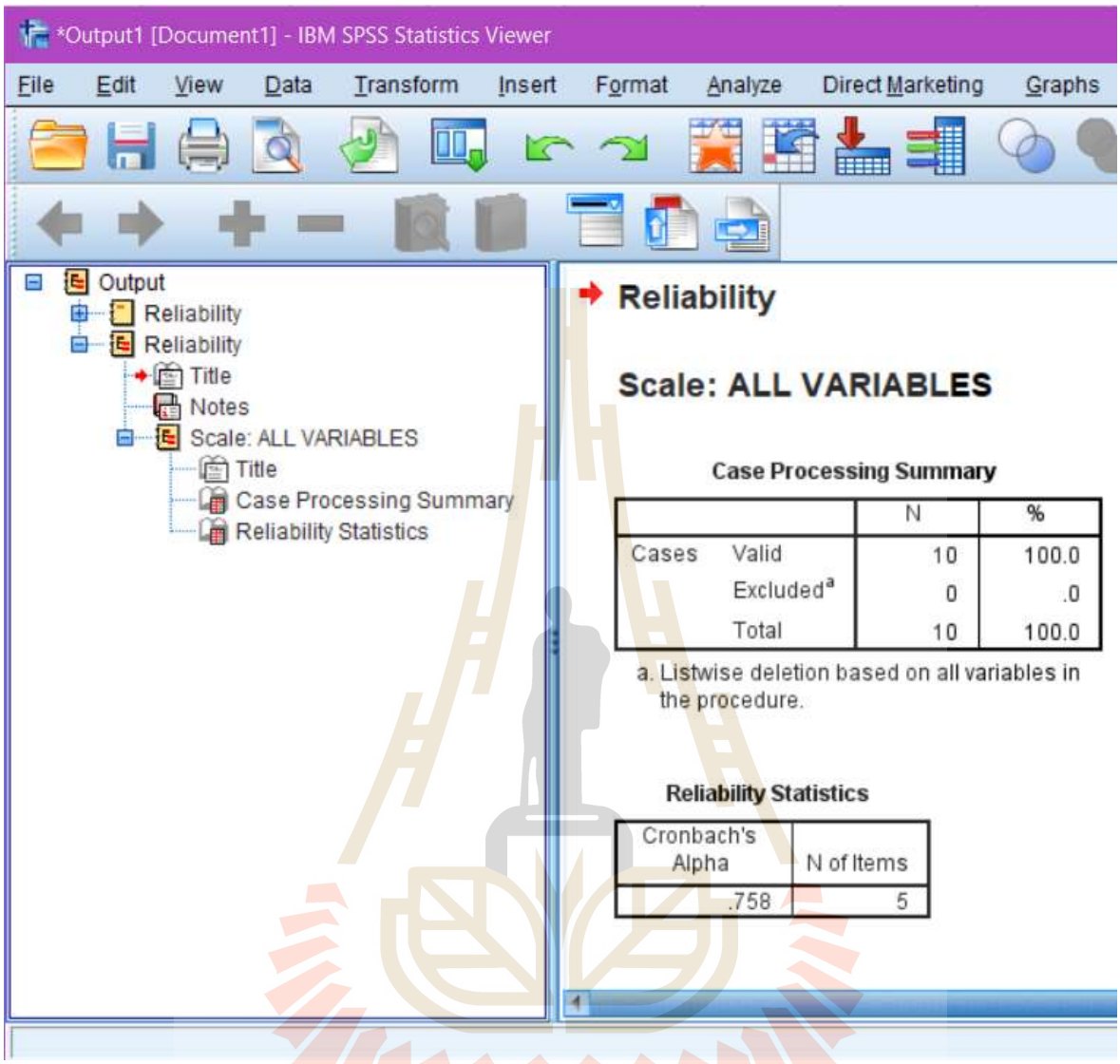
In the third step (Figure 6.8), move all the five items into the right box for variable selection. Choose the Alpha model and click ‘OK’.



รูปที่ 6.8 เลือกข้อมูลคำนวณด้วยวิธีสัมประสิทธิ์แอลฟา (Choosing the data for alpha calculation)

ในขั้นตอนสุดท้าย (รูปที่ 6.9) ค่าสัมประสิทธิ์ความเที่ยงแอลฟาจะถูกแสดงในรายงานผลลัพธ์ ในตัวอย่างค่าสัมประสิทธิ์คือ 0.758

In the last step (Figure 6.9), the alpha reliability coefficient is reported in the output report. In the example, the reliability coefficient is 0.758.



รูปที่ 6.9 รายงานค่าความเที่ยง (Reporting the reliability estimate)

6.7 Exercise

1. Which of the following is the first step in test and item analysis?

A. analyzing the items carefully

B. assembling items

C. choosing the items that are the most effective

D. piloting the items

2. In a test administration, there are 100 students taking the test. In an item, 60 students answer it correctly. What is the item facility value of this item?

A. 0.1 B. 0.4 C. 0.6 D. 1.0

3. Which of the following is the calculation for item discrimination?

A. IFpass – IFfail B. IFposttest – IFpretest

C. IFupper – IFlower D. proportion correct

4. Which of the following is used for the purpose of item discrimination?

A. *B*-index

B. difference index

C. item facility

D. point-biserial correlation coefficient

5. Which of the following will increase test reliability?

A. increasing the variation in item difficulty

B. reducing the number of items

C. seeking a sample of homogeneous abilities

D. using only the items with item facility values near 0.5

Answer key: 1. B 2. C 3. C 4. D 5. D

บทที่ 7 การประเมินผลทางเลือก

Chapter 7 Alternative Assessments

ในบทที่ 6 ได้กล่าวถึงการวิเคราะห์แบบสอบ และสถิติที่ใช้ในการวิเคราะห์ข้อสอบทางภาษา ในบทนี้ จะกล่าวถึงรูปแบบการวัดผลที่นอกเหนือไปจากการใช้แบบสอบ ลำดับในการนำเสนอมีดังนี้

7.1 ความสำคัญของการประเมินผลทางเลือก

7.2 การประเมินการแสดงออก

7.3 แฟ้มสะสมผลงาน

7.4 สมุดบันทึก

7.5 การสังเกต

7.6 การประเมินตนเองและการประเมินเพื่อนร่วมชั้น

7.7 เกณฑ์การประเมิน

7.8 Exercise (แบบฝึกหัดท้ายบท)

In Chapter 6, test and item analyses and statistics used for language testing have been dealt with. In this chapter, other techniques apart from tests will be dealt with. The order of presentation is as follows.

7.1 Importance of alternative assessment

7.2 Performance assessment

7.3 Portfolios

7.4 Journals

7.5 Observations

7.6 Self- and peer assessment

7.7 Rubrics

7.8 Exercise

7.1 ความสำคัญของการประเมินผลทางเลือก (Importance of Alternative Assessment)

การทดสอบโดยใช้แบบสอบ (tests) เป็นรูปแบบอย่างเป็นทางการของการเก็บตัวอย่างการแสดงความรู้ความสามารถของผู้สอบในโดเมนใดโดเมนหนึ่ง (Brown & Abeywickrama 2010) เช่น แบบสอบรายวิชาภาษาอังกฤษเพื่อวัตถุประสงค์เชิงวิชาการ เป็นต้น การใช้แบบสอบมักกระทำในสถานการณ์ที่มีการจำกัดเวลา เช่น นักศึกษาต้องทำข้อสอบรายวิชาหนึ่งๆ ให้แล้วเสร็จภายในสองชั่วโมง ส่วนคำว่าการประเมินผล (assessment) ให้ความหมายที่กว้างกว่ามาก เช่น ในขณะที่ครูท่านหนึ่งกำลังสอน ก็มักมีการประเมินผลในขณะเดียวกัน (real-time assessment) ควบคู่ไปด้วย เช่น การประเมินสีหน้าหรือท่าทางของผู้เรียน เพื่อประเมินว่าผู้เรียนมีความเข้าใจสิ่งที่ครูกำลังพูดอธิบายอยู่มากน้อยเพียงใด จำเป็นต้องอธิบายหัวข้อที่กำลังพูดอยู่เพิ่มเติมหรือไม่ เป็นต้น การประเมินผลจึงเป็นคำที่กว้างขวางครอบคลุมหลากหลายสถานการณ์กว่ามาก ซึ่งรวมไปถึงการสังเกตอย่างไม่เป็นทางการ หรือการทดสอบโดยใช้แบบสอบ ดังที่ได้กล่าวไว้แล้วข้างต้น

Tests are a formal way of collecting samples of performance from the test takers. They are usually based on a specific content domain, e.g., a test on English for academic purposes (Brown & Abeywickrama 2010). Tests are often administered in a situation where the time for taking the test is restricted. For example, the students must finish the test on English for academic purposes within two hours. In contrast, the term *assessment* refers much more broadly to any form of collecting samples of performance. For example, while teaching, teachers usually and, perhaps, unknowingly take notice of the facial expressions and bodily gestures of the students. They may be assessing the extent that the students understand or do not understand what they are explaining. They may even take action by explaining more or giving some more examples, in order to ensure the students' understanding of the subject matter being covered. Accordingly, *assessment* is a term that incorporates a variety of situations, including informal observations and tests.

ในช่วงทศวรรษ 1990 วัฒนธรรมคิดต่างได้ปฏิเสธแนวคิดที่ว่า ความสามารถทุกๆ ด้านของมนุษย์ทุกคนสามารถวัดได้ด้วยการใช้แบบสอบ จึงเกิดเป็นแนวคิดใหม่ที่เรียกรวมๆ ว่า การประเมินผลทางเลือก (alternative assessment) แนวคิดนี้ได้รวมเครื่องมือวัดอย่างหลากหลาย เพื่อที่จะสอบทวน

ผลลัพธ์ที่ได้จากการทดสอบผู้เรียน เช่น แฟ้มสะสมผลงาน สมุดบันทึก การสังเกต การประเมินตนเอง การประเมินเพื่อนร่วมชั้น เป็นต้น ทั้งนี้ Brown & Hudson (1998: 657) สนับสนุนให้เรียกการใช้เครื่องมือวัดเหล่านี้ว่า ทางเลือกในการประเมินผล (alternatives in assessment) เพราะการเรียกว่าการประเมินผลทางเลือก (alternative assessments) ทำให้เกิดความรู้สึกไปว่า การนำทางเลือกเหล่านี้มาใช้เป็นเรื่องใหม่ที่ไม่เคยมีมาก่อน (ลักษณะเดียวกันกับดนตรีทางเลือก alternative music) แปลกแยก และไม่จำต้องใช้กระบวนการที่เคร่งครัดแบบเดียวกันกับการสร้างแบบสอบในการทดสอบแบบขนบนิยมทั่วไป Brown & Hudson ระบุว่า ทางเลือกเหล่านี้มีมานานแล้ว และครูสอนภาษาก็ใช้กันมานานแล้วเช่นกัน การหันมานำทางเลือกเหล่านี้มาใช้อย่างเป็นกิจจะลักษณะเป็นเพียงแค่พัฒนาการใหม่ในวัฒนธรรมการประเมินผลแต่เพียงเท่านั้น อย่างไรก็ดี เพื่อให้เข้ากันได้กับการใช้โดยทั่วไป ในเอกสารประมวลฯ นี้จะยังคงใช้คำว่าการประเมินผลทางเลือกโดยอนุโลม

In the 1990s, there was a social movement against the traditional belief that all aspects of human competence could be measured through tests. The movement gave rise to the notion *alternative assessment*, which encompasses a variety of measures for triangulating tests of students' performance, e.g., portfolios, journals, observations, self-assessment, peer assessment. It is worth stating that Brown & Hudson (1998: 657) encourage scholars to call these measures *alternatives in assessment*, because the term *alternative assessment* implicitly conveys false impressions that (a) it is completely new (in much the same way as *alternative music* conveys newness, (b) it is isolated, and (c) it does not have to go through a rigorous procedure like that of traditional tests. Brown & Hudson state that alternatives in assessment have been used for a long time and such a movement in relation to alternatives in assessment is just a new development in the field. Despite this suggestion, the term *alternative assessment* will be used in this course document for practicality.

Brown & Hudson (1998: 654–655) ได้สรุปด้านบวกของการประเมินผลทางเลือกไว้ดังต่อไปนี้

1) ให้ผู้เรียนได้แสดงออก สร้างสรรค์ หรือผลิตสิ่งใดสิ่งหนึ่งออกมา
2) ใช้บริบทหรือสถานการณ์ในโลกความเป็นจริง
3) ไม่สร้างภาระจนเกินไป ในแง่ที่ว่า เป็นการประเมินที่เสริมเพิ่มเติมจากกิจกรรมในห้องเรียนที่ต้องทำอยู่แล้ว
4) ทำให้ผู้เรียนมีโอกาสได้รับการประเมินจากสิ่งที่ทำเป็นปกติในชั้นเรียน

5) ใช้ชิ้นงานที่สื่อถึงกิจกรรมในการเรียนการสอนอย่างมีความหมาย

6) มุ่งเน้นทั้งกระบวนการและผลลัพธ์

7) ใช้กระบวนการคิดขั้นสูงและทักษะการแก้ปัญหา

8) ให้ข้อมูลทั้งด้านเด่นด้านด้อยของผู้เรียนแต่ละคน

9) หากทำอย่างเหมาะสม จะสะท้อนความต่างทางวัฒนธรรมของผู้เรียนแต่ละคนได้

10) เป็นการให้คะแนนโดยใช้วิจารณญาณของมนุษย์ แทนที่จะเป็นเครื่องจักรให้คะแนน

11) สนับสนุนมาตรฐานและเกณฑ์การให้คะแนนที่โปร่งใส

12) ทำให้ครูผู้สอนได้แสดงบทบาทการสอนและการประเมินผลที่แตกต่างออกไป

Brown & Hudson (1998: 654–655) summarize the advantages of alternative assessment as follows:

1) require students to perform, create, produce, or do something;

2) use real-world contexts or simulations;

3) are nonintrusive in that they extend the day-to-day classroom activities;

4) allow students to be assessed on what they normally do in class every day;

5) use tasks that represent meaningful instructional activities;

6) focus on processes as well as products;

7) tap into higher-level thinking and problem-solving skills;

8) provide information about both the strengths and weaknesses of students;

9) are multiculturally sensitive when properly administered;

10) ensure that people, not machines, do the scoring, using human judgment;

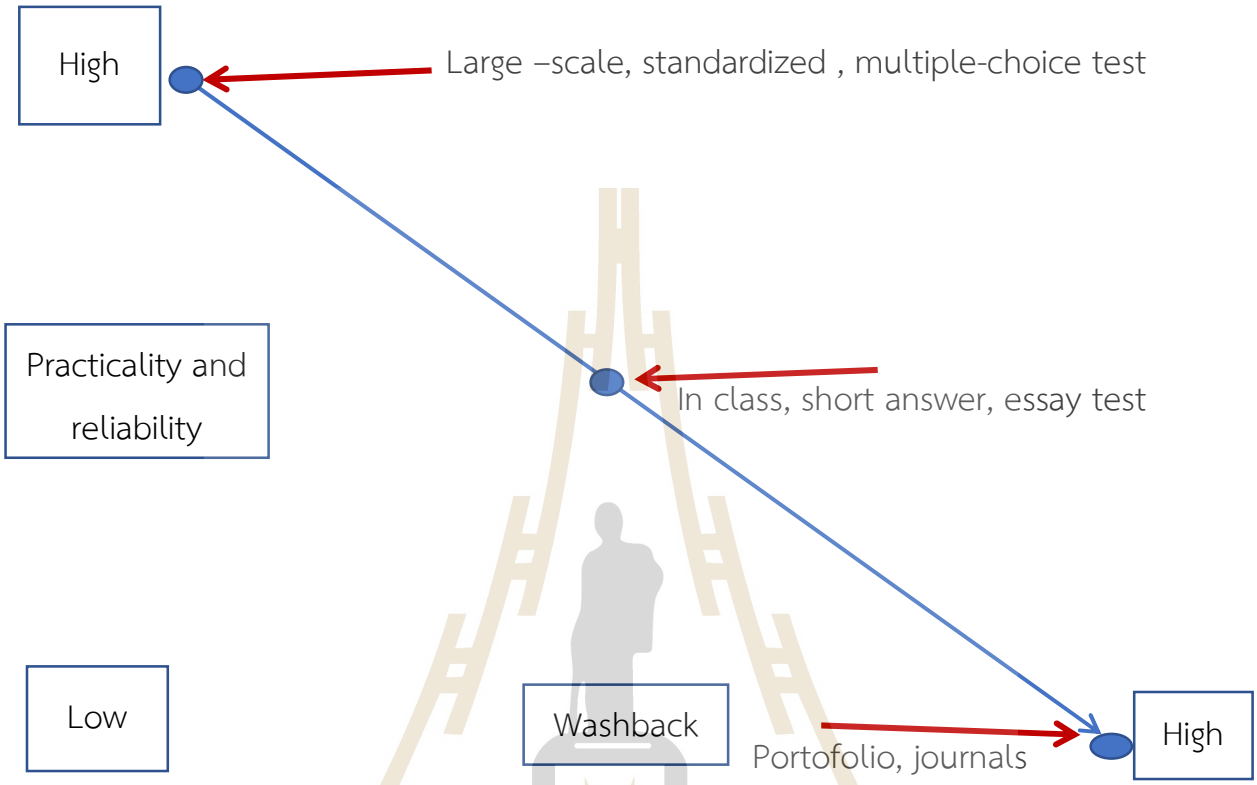
11) encourage open disclosure of standards and rating criteria; and

12) call upon teachers to perform new instructional and assessment roles.

โดยคำนิยามแล้ว การทดสอบมาตรฐานเน้นเชิงการนำมาใช้งานได้จริง กล่าวคือ ผู้เขียนข้อสอบและผู้สอบจะสามารถประหยัดเวลาและทรัพยากรได้มาก เช่น หากสร้างแบบสอบขึ้นมาหนึ่งชุด อาจใช้สอบคนหนึ่งพันคนพร้อมกันได้ เป็นต้น นอกจากนี้ การทดสอบมาตรฐานยังมุ่งเน้นให้แบบสอบมีค่าความเที่ยงสูงๆ และแม่นยำในการให้คะแนนด้วย เช่น หากแบบสอบมีคำเฉลยคำตอบอยู่หนึ่งชุด ไม่ว่าจะให้ใครตรวจก็ตาม ก็จะได้คะแนนของผู้สอบแต่ละคนออกมาเท่ากันทุกครั้ง เป็นต้น ในทางกลับกัน การ

ประเมินผลทางเลือก เช่น การใช้แฟ้มสะสมผลงาน หรือการสังเกตพฤติกรรมที่แสดงออกมา ต้องใช้เวลาและความพยายามอย่างมาก ทั้งในส่วนของครูผู้สอนและผู้เรียน แต่กระนั้นเทคนิคการประเมินผลทางเลือกก็มีข้อได้เปรียบกว่าตรงที่ว่า เป็นเครื่องมือที่เหมาะสำหรับการเก็บคะแนนเก็บ และเพราะการที่การประเมินผลทางเลือกใช้ข้อมูลจากกิจกรรมการเรียนการสอนจริง การประเมินผลทางเลือกจึงมักมีความตรงเชิงปรากฏ (face validity) สูงกว่าการทดสอบมาตรฐาน

By definition, standardized tests are practice-oriented. The test designers and test takers can save a lot of time and resources. For example, when one test has been constructed, it may be used for testing thousands of test takers at the same time. Moreover, standardized tests are usually designed so as to achieve a high reliability and precision in scoring and grading. For example, one standardized test has an answer key, and whoever marks it may get the same result. In contrast, alternative assessments such as portfolios and in-class observation require a lot of time and efforts from the instructors and the students alike. Still, alternative assessments have an edge over standardized testing, in that they are very suitable for formative assessment. Because alternative assessments use input from real-world, classroom contexts, they usually have higher face validity than standardized testing too.



รูปที่ 7.1 ความสัมพันธ์ระหว่างการนำมาใช้งานได้จริงและอิทธิพลย้อนกลับ (Relationship between practicality and washback)

ในรูปที่ 7.1 แสดงความสัมพันธ์ระหว่างการทดสอบมาตรฐาน การตอบคำตอบในบริบทห้องเรียน และการประเมินผลทางเลือก กับการนำมาใช้งานได้จริงและอิทธิพลย้อนกลับ โดยด้านบนสุดแบบสอบเช่นข้อสอบมาตรฐานแบบเลือกตอบจัดว่ามีความสามารถในการนำมาใช้งานได้จริงสูง (high practicality) และความเที่ยงสูงด้วย แต่แบบสอบมาตรฐานมักจะมีอิทธิพลย้อนกลับหรือการยึดโยงสู่รูปแบบการจัดการเรียนการสอนไม่มากนัก ในทางกลับกัน ด้านล่างสุดของรูปที่ 7.1 แฟ้มสะสมผลงานและสมุดบันทึก ซึ่งเป็นรูปแบบการประเมินผลทางเลือก มีอิทธิพลย้อนกลับหรือการยึดโยงสู่รูปแบบการจัดการเรียนการสอนสูง เพราะมักเป็นรูปแบบที่เกี่ยวเนื่องกับกิจกรรมในชั้นเรียนอยู่แล้ว แต่อาจจะมีความสามารถในการนำมาใช้งานได้จริงและความเที่ยงต่ำกว่าแบบสอบมาตรฐาน โดยเฉพาะในกรณีเช่นชั้นเรียนเป็นชั้นเรียนขนาดใหญ่ เป็นต้น

In Figure 7.1 the relationship between standardized testing, in-class assessment, and alternative assessment, and practicality and washback is illustrated. The uppermost node represents standardized testing, which is usually highly practical and reliable but does not affect many classroom activities. By contrast, alternative assessments such as portfolios and journals usually have a high washback effect because they are connected to classroom activities. Still, alternative assessments may have lower practicality and reliability than standardized tests, especially in the contexts of large classroom.

ตัวอย่างเทคนิคที่ใช้ในการประเมินผลทางเลือก ได้แก่ การประเมินการแสดงออก (performance-based assessment หรือ performance assessment) แฟ้มสะสมผลงาน (portfolios) แบบประเมินตนเอง (self-assessment) แบบประเมินเพื่อนร่วมชั้น (peer-assessment) สมุดบันทึกสะท้อนความคิดเห็น (reflective journals) การถกอภิปรายออนไลน์ (online discussions) แผนผังมโนทัศน์ (concept maps) โครงงานเดี่ยวหรือโครงงานกลุ่ม (individual/group projects) รายงานและการสัมมนาประจำภาค (term papers and seminars) การนำเสนอปากเปล่า (oral presentations) การโต้วาที (debates) และการสัมภาษณ์ (interviews) ทั้งนี้จะได้กล่าวถึงบางตัวอย่างข้างต้นโดยสังเขปในลำดับถัดไป

Examples of techniques for alternative assessment include performance-based assessment (aka. performance assessment), portfolios, self-assessment, peer-assessment, reflective journals, online discussions, concept maps, individual/group projects, term papers and seminars, oral presentations, debates, and interviews. In this course document, some of these examples will be dealt with in the sections that follow.

7.2 การประเมินการแสดงออก (Performance Assessment)

การประเมินการแสดงออก หมายถึงการแสดงทักษะออกมาให้วัดได้สังเกตได้ ในชิ้นงานที่มีความตรงเชิงเนื้อหา (content validity) เช่น หากเนื้อหาที่ต้องการวัดเป็นเรื่องการบอกทิศทาง ผู้เรียนที่บอกทิศทางได้ก็ถือว่าแสดงความสามารถออกมาได้ตรงกับเนื้อหาที่ต้องการวัด เป็นต้น การแสดงทักษะหรือความสามารถออกมาเช่นนี้มักมีความสมจริง (authentic) เพราะสถานการณ์การใช้ภาษาเช่นการบอกทิศทางเป็นสถานการณ์การใช้ภาษาที่มีใช้อยู่จริงในบริบทนอกห้องเรียน การแสดงทักษะหรือความสามารถยังเป็นการบูรณาการทักษะทางภาษาย่อยๆ หลายๆ ด้านเข้ามา เพื่อให้การแสดงความสามารถหรือทักษะออกมาสำเร็จลุล่วงด้วย

Performance assessment is involved with performing actual behaviors, such that they become observable and measurable. The tasks usually require content-valid performance. For example, if the content domain is about telling the directions, then we would expect the students to tell the directions to somewhere. Doing so would also imply authenticity of the test task because telling the directions is a language-use situation existing in the real world. Performance assessment is also concerned with an integration of language skills, as the students need to get the task done with more than one aspect of language use.

การประเมินผลการแสดงออกประกอบด้วยลักษณะสำคัญหลายประการ เช่น ผู้เรียนเป็นฝ่ายสร้างสรรค์การแสดงออกมา รูปแบบชิ้นงานเป็นแบบเดียวกันกับคำถามปลายเปิด คือมีรูปแบบที่ให้อิสระในการตอบโจทย์ได้มากกว่าหนึ่งรูปแบบตายตัว ชิ้นงานที่มอบหมายให้นักเรียนทำจึงมีความหมายต้องใช้ความพยายาม ทั้งยังสะท้อนสถานการณ์การใช้ภาษาในชีวิตจริงได้ดี ดังที่ได้กล่าวไว้แล้วข้างต้น ทำให้สามารถประเมินได้ทั้งกระบวนการทำชิ้นงานให้ลุล่วงและผลลัพธ์สุดท้ายได้ทั้งสองประการ ตลอดจนเป็นการให้ความสำคัญแก่การที่ผู้เรียนได้เรียนรู้เชิงลึก มากกว่าที่จะเป็นเชิงขยายแต่เพียงอย่างเดียว เพราะสถานการณ์ที่ผู้เรียนแสดงออกมาจะถูกวิเคราะห์อย่างลึกซึ้งไม่ฉาบฉวยมากกว่า

Performance assessment has several characteristics. First, it is the students who construct the responses. This makes the task meaningful and engaging, given that it has to reflect the real world. Secondly, because it is of the constructed-response format, the students have more freedom in engaging in the assessment task. Moreover, performance assessment allows the students to learn with depth, rather than with breadth alone. The reason for this is the deep engagement that is activated. The teacher can thus assess both their process and the product of the performance.

ในการประเมินการแสดงออก ควรต้องมีระเบียบแบบแผนในการประเมินเช่นเดียวกับการทดสอบมาตรฐานแบบปกติ ผู้สอนควร

1) ระบุเป้าหมายโดยรวมของการให้ผู้เรียนแสดงออก

2) ระบุเป้าประสงค์ รวมไปถึงเกณฑ์ในการแสดงออกอย่างละเอียด

3) เตรียมผู้เรียนสำหรับการแสดงออกอย่างเป็นขั้นเป็นตอน

4) ใช้แบบประเมินผล แบบรายการตรวจสอบ (checklist) แบบวัดประมาณค่า (rating form) อย่างคงเส้นคงวาระหว่างกลุ่มหรือระหว่างผู้เรียน

5) ให้มองการแสดงออกเป็นโอกาสที่จะให้ผลย้อนกลับ (feedback) รวมทั้งต้องให้ผลย้อนกลับอย่างเป็นระบบคงเส้นคงวาระหว่างกลุ่ม หรือระหว่างผู้เรียน

6) หากเป็นไปได้ ควรใช้แบบประเมินตนเองและแบบประเมินเพื่อนร่วมชั้นเรียนควบคู่ไปด้วยอย่างเหมาะสม

In performance assessment, procedures should be implemented rigorously, just as in standardized tests. The teachers should

1) clearly state the goal expected of the performance assigned;

2) clarify the objectives and criteria for the performance;

3) prepare the students step-by-step;

4) use an evaluation form, checklist, or rating form consistently;

5) treat the students' performance as a chance to give feedback, and provide the feedback consistently;

6) if possible, use self-assessment and peer-assessment too.

7.3 แฟ้มสะสมผลงาน (Portfolios)

แฟ้มสะสมผลงานเป็นการรวบรวมงานของผู้เรียนอย่างเป็นระบบ มักใช้แสดงถึงความพยายาม ความก้าวหน้า และการประสบผลสำเร็จในด้านใดด้านหนึ่ง (Genesee & Upshur, 1996) แฟ้มสะสมผลงานมักจะประกอบไปด้วยเอกสาร อย่างเช่น

1) เรียงความ หรือเรื่องที่แต่ง เรียบเรียงขึ้นมา ทั้งฉบับร่างและฉบับสมบูรณ์

2) รายงาน หรือโครงร่างของโครงงานต่างๆ

3) แถบบันทึกเสียง หรือไฟล์บันทึกภาพเคลื่อนไหวของการนำเสนองาน การสาธิต ฯลฯ

4) สมุดบันทึก สมุดอนุทิน (diaries) และงานเขียนแสดงความคิดเห็นส่วนตัว

5) แบบสอบ คะแนนแบบสอบ และแบบฝึกหัดที่เขียนเป็นการบ้าน

6) แบบประเมินตนเอง และแบบประเมินเพื่อนร่วมชั้น ที่มีทั้งความคิดเห็น การประเมินค่า และแบบรายการตรวจสอบ

A portfolio is a collection of students' work that demonstrates their efforts, progress, and achievements in a given area (Genesee and Upshur, 1996). Portfolios usually include materials such as:

1) draft and final forms of essays and compositions;

2) reports, and project outlines;

3) audio and/or video recordings of presentations, demonstrations, etc.;

4) journals, diaries, and other personal reflections;

5) tests, test scores, and written homework exercises;

6) self- and peer-assessments, including comments, evaluations, and checklists.

ทั้งนี้ Gottlieb (1995) ได้นำเสนอโมเดลในการพิจารณาลักษณะของแฟ้มสะสมผลงาน โดยใช้ตัวย่อ CRADLE ดังนี้

- เก็บรวบรวม (Collecting) เป็นการแสดงชีวิตและอัตลักษณ์ของผู้เรียน
- สะท้อนความคิดเห็น (Reflecting) เป็นการถ่ายทอดความคิดเห็นเกี่ยวกับประสบการณ์และกิจกรรมต่างๆ
- ประเมินผล (Assessing) เป็นการประเมินคุณค่าและพัฒนาการผ่านช่วงเวลาหนึ่งๆ
- จัดเก็บ (Documenting) เป็นการแสดงความสำเร็จในการทำชิ้นงานต่างๆ
- เชื่อมโยง (Linking) เป็นการเชื่อมโยงผู้เรียนกับครู กับผู้ปกครอง กับชุมชน และกับเพื่อนร่วมชั้นเรียน
- ประเมินค่า (Evaluating) เป็นสร้างสรรค์ผลลัพธ์อย่างมีความรับผิดชอบ

Gottlieb (1995) suggested a scheme for considering the nature and purpose of portfolios, using the acronym CRADLE to describe six attributes of a portfolio:

- Collecting: an expression of students' lives and identities
- Reflecting: thinking about experiences and activities
- Assessing: evaluating quality and development over time
- Documenting: demonstrating student achievement
- Linking: connecting the students and the teacher, their parents, their community, and their peer
- Evaluating: generating responsible outcomes.

ในการประเมินแฟ้มสะสมผลงาน ครูผู้สอนควร

1) ระบุเป้าประสงค์อย่างชัดเจน

2) ชี้แนะแนวทางว่า ควรนำเอกสารใดบ้างรวมมาไว้ในแฟ้มสะสมผลงาน

3) สื่อสารเกณฑ์การวัดและประเมินผลไปยังผู้เรียน

4) ระบุกรอบเวลาที่สอดคล้องกับหลักสูตร ให้ผู้เรียนมีเวลาสร้างสรรค์แฟ้มสะสมผลงาน

5) กำหนดช่วงเวลาเป็นระยะๆ สำหรับรอบการพิจารณาและพูดคุย

6) กำหนดจุดจัดเก็บแฟ้มสะสมผลงานให้เข้าถึงได้ง่าย

7) ให้ข้อมูลย้อนกลับเชิงบวกเมื่อประเมินผลขั้นสุดท้าย

In assessing the students' portfolios, the teachers should

1) State objectives clearly;

2) Give guidelines on what materials to include;

3) Communicate assessment criteria to the students;

4) Designate time within the curriculum or program for portfolio development;

5) Establish periodic schedules for review and conferencing;

6) Designate an accessible place to keep portfolios;

7) Provide positive feedback when giving final assessments.

ในการประเมินแฟ้มสะสมผลงาน ครูผู้สอนไม่ควรให้ผลการประเมินย้อนกลับเป็นแค่อักษร สัญลักษณ์หรือตัวเลขประเมินผล เช่น A, B, C หรือ 1, 2, 3 ผู้สอนควรให้ผลประเมินเป็นเชิงคุณภาพ เกี่ยวกับภาพรวมสุดท้ายของแฟ้มสะสมผลงาน อาจตั้งคำถามเพื่อให้ผู้เรียนประเมินตนเอง ตลอดจนประเมินออกมาเป็นคำบรรยายข้อเด่นข้อด้อยของแฟ้มสะสมผลงาน

In assessing portfolios, the teachers should not give just a number or letter grade for the hard work the students have done. Rather, the teachers should give a qualitative evaluation such as an appraisal of the work, questions for self-assessment, or a narrative evaluation of strengths and weaknesses.

7.4 สมุดบันทึก (Journals)

สมุดบันทึกเป็นอีกหนึ่งรูปแบบของการประเมินผลทางเลือก สมุดบันทึกทำหน้าที่บันทึกความรู้สึกนึกคิด ปฏิกิริยาตอบสนอง การประเมินผล ความคิดสร้างสรรค์ หรือความก้าวหน้า โดยมักเขียนแบบไม่ต้องมีโครงสร้าง แบบฟอร์ม หรือกังวลถึงความถูกต้องมากนัก ในส่วนของการเรียนการสอน สมุดบันทึกมีหน้าที่สำคัญคือ การฝึกฝนกลวิธีการเขียน โดยใช้การเขียนเป็นกระบวนการนึกคิด เป็นการแสดง

ตัวตนความเป็นปัจเจกบุคคลออกมา และรวมไปถึงเป็นการสื่อสารกับครูผู้สอน (Brown & Abeywickrama 2010)

A journal is a log of one's thought, feelings, reactions, assessments, ideas, or progress, toward goals, usually written with little attention to structure, form, or correctness. Journals serve important pedagogical purposes: practice in the mechanics of writing, using writing as a thinking process, individualization, and communications with the teacher (Brown & Abeywickrama 2010).

การสะท้อนความคิดเห็นออกมามีความสำคัญต่อการเรียนรู้ การสะท้อนความคิดเห็นแบ่งได้เป็นสองประเภทหลักๆ คือ การสะท้อนความคิดเห็นขณะที่กำลังทำกิจกรรมอย่างใดอย่างหนึ่งอยู่ (reflection-in-action) กับการสะท้อนความคิดเห็นหลังจากที่ได้ทำกิจกรรมนั้นๆ แล้วเสร็จ (reflection-on-action) การสะท้อนความคิดเห็นเป็นกระบวนการสร้างความหมายอย่างเป็นระบบ จำต้องมีทัศนคติเชิงบวกในการเห็นคุณค่าของการเติบโตในส่วนบุคคลและด้านสติปัญญา ของทั้งตนเองและของผู้อื่น ดังนั้น การเขียนสมุดบันทึกความคิดเห็นจึงมิใช่เป็นเพียงแค่การเขียนสมุดบันทึกประจำวันหรือสมุดอนุทิน (writing a diary) แต่เพียงเท่านั้น หากแต่การเขียนสมุดบันทึกความคิดเห็นยังครอบคลุมไปถึงการคิดเชิงวิพากษ์ ที่ต้องใช้การตั้งคำถามอย่างเช่น อะไร? อย่างไร? และทำไม? ด้วย การเขียนสมุดบันทึกความคิดเห็นจึงเป็นโอกาสในการปูบริบทสำหรับผู้เขียน วางกรอบทฤษฎี และทำให้สะท้อนตัวตนได้ด้วย

Reflection is important for learning. Reflection can be classified into two types. The first type is reflection-in-action, which refers to when one gives reflection while doing something or doing some action. The second type of reflection is reflection-on-action, which refers to when one gives reflection after they have done something or have finished doing the action. Reflection is a systematic meaning-making process that requires attitude to value the personal and intellectual growth of oneself and others. Writing a reflective journal, therefore, is not just writing a diary. It covers a critical approach using questions such as what, how and why too. It gives an opportunity to contextualize, theorize, and personalize.

ขั้นตอนในการนำสมุดบันทึกมาใช้เป็นการประเมิน ได้แก่

1) แนะนำความคิดรวบยอดเรื่องการเขียนสมุดบันทึกอย่างสมเหตุสมผลแก่ผู้เรียน

2) แจ้งให้ทราบถึงวัตถุประสงค์ในการนำสมุดบันทึกมาใช้ในการประเมิน เช่น เพื่อเก็บบันทึกการเรียนรู้ภาษา เพื่อประเมินการใช้ไวยากรณ์ เพื่อบันทึกการตอบสนองต่อเรื่องที่อ่าน เพื่อบันทึกการเรียนรู้การใช้กลวิธีในการเรียนรู้ภาษา เพื่อสะท้อนข้อคิดเห็นเกี่ยวกับการประเมินตนเอง ฯลฯ

3) ชี้แนะแนวทางว่า มีหัวข้อใดบ้างที่ควรจะนำมาเขียนในสมุดบันทึก

4) ระบุเกณฑ์ที่ใช้ในการประเมินผลสมุดบันทึกอย่างระมัดระวัง เกณฑ์อย่างเช่นความพยายามที่เห็นได้จากความละเอียดถี่ถ้วน หรือขั้นตอนการทำความเข้าใจเนื้อหารายวิชาก็สำคัญเช่นกัน

5) ให้ข้อมูลย้อนกลับอย่างเหมาะสม เช่น ข้อมูลย้อนกลับที่ให้กำลังใจ ที่สอนชี้แนะ หรือข้อมูลย้อนกลับที่ระบุข้อเท็จจริง เป็นต้น

6) กำหนดกรอบเวลาสำหรับการเขียนสมุดบันทึก ตารางนัดหมายสำหรับรอบการพิจารณาติดตามความก้าวหน้า

7) ให้ความเห็นเพื่อส่งเสริมพัฒนาการสำหรับการเก็บคะแนนต่อไป

Regarding steps for using journals, the teacher should

1) introduce the students to the concept of journal writing sensibly

2) state the objective(s) of keeping a journal, e.g., language-learning logs, grammar journals, responses to readings assigned, strategy-based learning logs, self-assessment reflections etc.

3) give guidelines on what kinds of topics to include

4) carefully specify the criteria for grading journals. Effort as exhibited in the thoroughness of students' content or the extent to which content reflects the processing of the course content will be important.

5) provide positive feedback in your responses: cheerleading feedback, instructional feedback, or reality-check feedback

6) determine appropriate time frames and schedules for periodic review

7) provide formative comments for creating positive washback

7.5 การสังเกต (Observations)

การสังเกตเป็นวิธีการเก็บบันทึกพฤติกรรมของผู้เรียนทั้งด้านวัจนภาษาและอวัจนภาษา ต้องมีการเตรียมการอย่างเป็นระบบ หนึ่งในวัตถุประสงค์ของการสังเกตก็คือเพื่อประเมินผู้เรียนแบบไม่ให้

รู้ตัวว่ามีการสังเกตพฤติกรรมอยู่ การทำเช่นนี้ช่วยลดอาการประหม่าของผู้เรียนได้ และยังเป็นการให้ผู้เรียนได้แสดงออกทางด้านภาษาอย่างเป็นธรรมชาติอีกด้วย

Observation is a planned procedure in systematically recording the students' verbal and non-verbal behavior. One of the objectives is to assess students' performance without their awareness that they are being observed, which might otherwise make them anxious. Observation, therefore, maximizes the students' naturalness in linguistic performance.

การสังเกตสามารถมุ่งเป้าได้หลายด้าน เช่น

1) ด้านทักษะในการพูดระดับประโยค
2) ด้านการออกเสียงในประเด็นต้องการวัดโดยเฉพาะ หรือด้านการออกเสียงสูงต่ำในประโยค (intonation)
3) ด้านไวยากรณ์ เช่น กาลของคำกริยา การสร้างรูปประโยคคำถาม ฯลฯ
4) ด้านทักษะวัจนกรรม (discourse-level skills) เช่น กฎเกณฑ์ในการสนทนา การแบ่งผลัดกันพูด ตลอดจนทักษะระดับวัจนกรรมอื่นๆ
5) ปฏิสัมพันธ์ระหว่างเพื่อนร่วมชั้น เช่น การร่วมไม้ร่วมมือกัน ความถี่ในการเป็นฝ่ายพูด
6) ความถี่ในการตอบสนองที่ผู้เรียนเป็นฝ่ายริเริ่ม ในชั้นเรียน หรือในงานกลุ่ม

There can be several foci in making observation, for example,

1) oral production skills at the sentence level;
2) pronunciation of the sounds that are the target of assessment, or intonation;
3) grammar points, e.g., verb tenses, question formation;
4) discourse-level skills such as conversation rules, turn-taking, and other macroskills;
5) interaction with classmates such as cooperation, frequency of oral production;
6) frequency of responses that are student-initiated (whole-class situation, group work).

ขั้นตอนในการนำการสังเกตมาใช้เป็นการประเมิน ได้แก่

1) กำหนดวัตถุประสงค์ในการสังเกตให้ชัด

2) กำหนดจำนวนผู้เรียนที่จะสังเกตในแต่ละครั้ง

3) คิดหากลวิธีที่จะสังเกตโดยไม่ให้ผู้เรียนรู้

4) ออกแบบระบบบันทึกการแสดงออกที่สังเกตได้

5) ไม่กำหนดจำนวนด้านที่ต้องการสังเกตในแต่ละครั้งให้มากเกินไป

6) วางแผนว่าต้องการสังเกตกี่ครั้ง

7) กำหนดให้ชัดว่าจะใช้ผลลัพธ์ที่ได้จากการสังเกตอย่างไร

Regarding steps for making observations, the teacher should

1) specify the objectives of the observations;

2) decide on the number of students that will be observed at a time;

3) set up a plan to make unnoticed observations;

4) design a system for recording performances that will be observed;

5) do not overestimate the number of aspects or elements that can be observed at a time;

6) plan the number of observations that you will make;

7) determine how exactly you will use the results of observations.

ทั้งนี้ ในการสังเกตก็อาจใช้เครื่องมือทางเลือกในการบันทึกผลลัพธ์ที่ได้จากการสังเกต เช่น รายการตรวจสอบ (checklist) และมาตรวัดประเมินค่า (rating scales) เป็นต้น

In making observations, a checklist or a rating scale may be used for recording the observation results.

7.6 การประเมินตนเองและการประเมินเพื่อนร่วมชั้น (Self- and Peer Assessment)

การประเมินตนเองมีรากฐานสำคัญมาจากหลักการของการเรียนรู้ภาษาที่ 2 คือ การให้ผู้เรียนเรียนรู้แบบพึ่งพาตนเอง (learner automony) การเรียนรู้แบบพึ่งพาตนเองได้ หมายถึง ความสามารถของผู้เรียนในการกำหนดเป้าหมายปลายทางที่ต้องการ ทั้งภายในโครงสร้างหลักสูตรของวิชาที่เรียน และในระดับที่เกินกว่าโครงสร้างหลักสูตรของวิชาที่เรียน นอกเหนือจากการกำหนดเป้าหมายแล้ว การเรียนรู้แบบพึ่งพาตนเองยังหมายรวมไปถึงการประพฤติปฏิบัติตนโดยไม่ต้องมีใครมาบังคับ ตลอดจนเฝ้าประเมินสถานการณ์เกี่ยวกับการปฏิบัติตนเหล่านั้นเองได้ การพัฒนาแรงจูงใจใฝ่สัมฤทธิ์ในตนเองจึงเป็นปัจจัยสำคัญยิ่งในการเรียนรู้ทักษะใดๆ (Brown & Abeywickrama 2010)

Self-assessment stems from a principle of second language acquisition. The principle is that the learners should have autonomy in learning. The learners can set the goal for their learning that they desire, both within and beyond the scope of their curriculum. Besides goal setting, learner autonomy entails the situations in which the learners choose what to do and how to do that properly, including self-motivating for monitoring their own progress while learning (Brown & Abeywickrama 2010).

การประเมินเพื่อนร่วมชั้นก็มีรากฐานที่มาคล้ายๆ กับการประเมินตนเอง หลักการสำคัญ คือ การเรียนรู้แบบร่วมมือ ผู้เรียนจำนวนมากเรียนตั้งแต่ชั้นอนุบาลจนจบถึงระดับปริญญาแต่อาจไม่เห็นคุณค่าของการประสานความร่วมมือกันในการเรียนรู้ การประเมินเพื่อนร่วมชั้นจึงเป็นเพียงชิ้นงานรูปแบบหนึ่งในการจัดการเรียนรู้ที่ผู้เรียนเป็นศูนย์กลางและการเรียนรู้แบบร่วมมือ

Peer assessment has a similar background to self-assessment. A key principle is collaborative learning. It would be no surprise that a lot of students learn in the kindergarten level up to a graduate degree, and never appreciate the value of collaborative learning. On the whole, peer assessment is one of the many tasks and tools that learner-centered education has.

ชนิดของการประเมินตนเองและการประเมินเพื่อนร่วมชั้น ตัวอย่างเช่น

1) การประเมินการแสดงออกอย่างใดอย่างหนึ่ง

2) การประเมินผลสามัตถิยะทั่วไป (general competence) ทางอ้อม

3) การประเมินผลระดับปริชาน (metacognitive assessment) เพื่อการวางเป้าหมาย

4) การประเมินทางสังคมและจิตพิสัย (socioaffective assessment)

5) ข้อสอบที่ออกโดยผู้เรียน

Types of self- and peer assessments include

1) Assessment of a specific performance

2) Indirect assessment of general competence

3) Metacognitive assessment for setting goals

4) Socioaffective assessment

5) Test generated by the students

ขั้นตอนในการให้ผู้เรียนประเมินผลตนเอง และผู้เรียนประเมินเพื่อนร่วมชั้น มีแนวทางดังนี้

1) บอกวัตถุประสงค์ในการประเมินผลให้ผู้เรียนทราบ
2) ระบุชิ้นงานที่จะให้ทำให้ชัดเจน
3) ส่งเสริมการประเมินผลความสามารถหรือการแสดงออก แบบไม่เอนเอียง มีอคติหรือเลือกข้าง
4) ใช้งานติดตามประเมินผลทำให้เกิดอิทธิผลย้อนกลับเชิงบวก (beneficial washback through follow-up tasks)

Some guidelines for self and peer assessment are as follows:

1) Tell students the purpose of the assessment.
2) Define the task(s) clearly.
3) Encourage impartial evaluation of performance or ability
4) Ensure beneficial washback through follow-up tasks.

ทั้งนี้รูปแบบชิ้นงานที่สามารถใช้กับการประเมินตนเองและการประเมินเพื่อนร่วมชั้น ตัวอย่างเช่น

<u>ทักษะการฟัง</u>

ดูโทรทัศน์หรือฟังวิทยุออกอากาศ และเช็คความเข้าใจกับเพื่อน
ดูคลิปเป็นภาษาอังกฤษ และพูดคุยเกี่ยวกับความเข้าใจเนื้อเรื่องกับเพื่อน
ฟังการบรรยายวิชาการ เช็คความเข้าใจด้วยแบบทดสอบย่อย (quiz) เกี่ยวกับเนื้อหาในการบรรยาย

<u>ทักษะการพูด</u>

ใช้แบบรายการตรวจสอบ (checklist) และแบบสอบถาม
ประเมินการนำเสนอปากเปล่า

<u>ทักษะการอ่าน</u>

อ่านเรื่องและทำคำถามเช็คความเข้าใจ
ทำแบบทดสอบย่อยเกี่ยวกับคำศัพท์

<u>ทักษะการเขียน</u>

แก้งานที่เขียนไว้เองหรือให้เพื่อนร่วมชั้นช่วยตรวจทาน
แก้งานก่อนส่งฉบับสมบูรณ์

There are several types of task that can be used with self- and peer-assessment, for example,

Listening tasks

Listening to TV or radio broadcasts and checking comprehension with a partner
Watching English clips and checking comprehension with a partner
Listening to an academic lecture and checking yourself on a content quiz

Speaking tasks

Using peer checklists and questionnaires
Rating someone's oral presentation

Reading tasks

Reading passages with self-check comprehension questions
Taking vocabulary quizzes

Writing tasks

Revising written work on your own or with a peer
Proofreading

7.7 เกณฑ์การประเมิน (Rubrics)

เกณฑ์การประเมินมิใช่รูปแบบหนึ่งของการประเมินผลทางเลือก หากแต่การใช้เกณฑ์การประเมินเป็นเครื่องมือชิ้นหนึ่งในการประเมิน จะทำให้ผู้ประเมินเช่นครูผู้สอนสามารถประเมินผลได้อย่างมีประสิทธิภาพและมีความรับผิดชอบ เกณฑ์การประเมินมักประกอบไปด้วยมาตรวัดประมาณค่า (rating scale) หรือแนวทางการให้คะแนน (scoring guide) เกณฑ์การประเมินนิยมใช้กันอยู่สองรูปแบบ ได้แก่ เกณฑ์การประเมินแบบองค์รวม และเกณฑ์การประเมินแบบแยกส่วน (Brown & Abeywickrama 2010: 128f.) ตารางที่ 7.1 แสดงตัวอย่างเกณฑ์การประเมินแบบองค์รวม ตารางที่ 7.2 แสดงลักษณะเกณฑ์การประเมินแบบแยกส่วน

Rubrics are not a form of alternative assessment. But rubrics are a tool in assessment, which allows teachers to assess the students' performance effectively and with responsibility. A rubric is usually composed of a rating scale and a scoring guide. There are two forms of rubrics that are used in general: a holistic rubric and an analytic rubric (Brown & Abeywickrama 2010: 128f.). Table 7.1 shows an example of a holistic rubric, and Table 7.2 shows the format of an analytic rubric.

ตารางที่ 7.1 ตัวอย่างเกณฑ์การประเมินแบบองค์รวม (example of a holistic rubric)

Score	Description
5	Demonstrate complete understanding of the problem. All requirements of task are included in response
4	Demonstrates considerable understanding of the problem. All requirements of task are included
3	Demonstrates partial understanding of the problem. Most requirements of task are included
2	Demonstrates little understanding of the problem. Many requirements of task are included
1	Demonstrates no understanding of the problem
0	No response/ task not attempted

ตารางที่ 7.2 ลักษณะเกณฑ์การประเมินแบบแยกส่วน (example of an analytic rubric)

Criteria	Beginning 1	Developing 2	Accomplished 3	Exemplary 4	Score
Criteria 1					
Criteria 2					
Criteria 3					
Criteria 4					

ขั้นตอนในการนำเกณฑ์การประเมินมาใช้ ได้แก่

1) พิจารณาเนื้อหาตารางแจกแจงข้อสอบ (table of specification) ในพิมพ์เขียวข้อสอบอีกครั้งหนึ่ง โดยเฉพาะอย่างยิ่งวัตถุประสงค์รายวิชา

2) พิจารณารูปแบบและกรอบการให้คะแนนอย่างถี่ถ้วน

3) แจกแจงวัตถุประสงค์ที่ครอบคลุมโดยข้อสอบ เพื่อนำมาใช้เป็นเกณฑ์การให้คะแนน

4) ระบุลักษณะที่สังเกตได้สำหรับแต่ละวัตถุประสงค์ หรือสำหรับแต่ละเกณฑ์การให้คะแนน

5) สำหรับเกณฑ์การประเมินแบบองค์รวม ให้เขียนคำบรรยายเพื่อเป็นคำชี้แนะการให้คะแนน สำหรับแต่ละระดับของกรอบการให้คะแนน โดยคำบรรยายควรมีลักษณะเฉพาะเจาะจงสำหรับแต่ละระดับ

6) สำหรับเกณฑ์การประเมินแบบแยกส่วน ให้เขียนคำบรรยายเพื่อเป็นแนวทางการให้คะแนน สำหรับแต่ละเกณฑ์แต่ละประเด็นในกรอบการให้คะแนน โดยคำบรรยายควรมีลักษณะเฉพาะเจาะจงสำหรับแต่ละเกณฑ์และในแต่ละระดับ

7) ให้เพื่อนร่วมงานดูเกณฑ์การประเมิน (peer review) ที่จัดทำขึ้นมา และลองใช้กับชิ้นงานของผู้เรียนจำนวนเล็กน้อย เพื่อหาข้อบกพร่อง

8) แจ้งให้ผู้เรียนทราบถึงเกณฑ์การประเมิน ก่อนหน้าการเก็บคะแนนที่จะใช้ในการประเมินผล

The steps for bringing a scoring rubric into use are as follows:

1) consider the table of specifications in your blueprint (re-examine the learning objectives);

2) consider the grading pattern and schemes carefully;

3) list the objectives covered in the item/test, as criteria;

4) identify the observable attributes for each of the objectives/criteria;

5) for holistic rubrics, write narrative descriptions for each level of the grading scheme incorporating specific attributes;

6) for analytic rubrics, write narrative descriptions for each individual criterion for the grading scheme incorporating specific attributes;

7) conduct peer-review of the rubrics and apply to some representative student works to identify problems, if any;

8) notify the students of the rubrics prior to the assessment task.

7.8 Exercise

1. Which of the following is **NOT** an alternative assessment?

A. journal

B. performance-based assessment

C. portfolio

D. standardized test

2. Which statement is true?

A. Alternative assessment has never been used before the 1990s.

B. Most teachers assess only in the examinations.

C. The term *assessment* is larger than *testing*.

D. The term *assessment* usually means standardized testing.

3. Which statement is **FALSE** about alternative assessment?

A. Alternative assessment focuses on both processes and products.

B. Alternative assessment has high practicality and reliability.

C. Alternative assessment taps into problem-solving skills.

D. Alternative assessment uses real-world contexts.

4. What can help alternative assessment to have increased objectivity in scoring?

A. observation

B. peer assessment

C. portfolio

D. rubric

5. Which is **NOT** a task for self-assessment?

A. proofreading

B. reading passages and using self-check comprehension questions

C. setting up a plan to make unnoticed observations

D. watching clips and checking comprehension

Answer key: 1. D 2. C 3. B 4. D 5. C

บทที่ 8 แนวโน้มและประเด็นในการวัดและประเมินผลภาษาอังกฤษ

Chapter 8 Current Trends and Issues in Language Assessment

ในบทที่ 7 ได้กล่าวถึงการประเมินผลทางเลือก ซึ่งสามารถนำมาใช้ร่วมกับรูปแบบการทดสอบแบบดั้งเดิม เพื่อให้เกิดการประเมินผลแบบสามเส้า (triangulation) ในบทนี้ จะได้กล่าวถึงแนวโน้มและประเด็นในการวัดและประเมินผลในภาษาอังกฤษ ลำดับในการนำเสนอมีดังนี้

8.1 แนวโน้มในการประเมินผลทางภาษา

8.2 ประเด็นในการประเมินผลทางภาษา

8.3 Exercise (แบบฝึกหัดท้ายบท)

In Chapter 7, alternative assessments have been dealt with. They could be used with traditional assessment tasks, thereby creating triangulation for a more valid assessment. In this chapter, trends and issues for language assessment will be tackled. The order of presentation is as follows.

8.1 Trends in language assessment

8.2 Issues in language assessment

8.3 Exercise

8.1 แนวโน้มในการประเมินผลทางภาษา (Trends in Language Assessment)

ในเชิงประวัติศาสตร์ แนวโน้มและแนวปฏิบัติในการทดสอบทางภาษาได้แปรเปลี่ยนไปตามวิธีการสอน ตัวอย่างเช่น ในช่วงทศวรรษ 1940 และ 1950 เป็นยุคของพฤติกรรมนิยมและการศึกษาเปรียบเทียบความแตกต่าง การทดสอบทางภาษาก็เน้นไปที่หน่วยทางภาษา เช่น การเปรียบเทียบหน่วยเสียง หน่วยไวยากรณ์ และหน่วยคำศัพท์ระหว่างสองภาษา เป็นต้น พอช่วงทศวรรษ 1970 และ 1980 ทฤษฎีทางภาษาเพื่อการสื่อสารได้นำมาซึ่งมุมมองเชิงบูรณาการมากขึ้น และสถานการณ์การสื่อสารก็ยอมรับว่ามีองค์ประกอบมากกว่าแค่นำเอาองค์ประกอบทางภาษามาประกอบกัน ทุกวันนี้ ผู้ออกแบบข้อสอบก็ยังคงต้องพยายามคิดหาเครื่องมือที่มีความตรง (validity) และมีความสมจริงเป็นตามสภาพจริง

(authentic) มากขึ้น ที่จะสะท้อนการมีปฏิสัมพันธ์กันในโลกความเป็นจริง (Brown & Abeywickrama 2010: 12ff.)

In historical terms, trends and practices in language testing change in accordance with teaching methodology. For example, behaviorism and contrastive analysis were popular in the 1940s and 1950s, and so language tests emphasized specific language elements, e.g., contrasts of phonological, grammatical, and lexical units, between two languages. Then in the 1970s and 1980s, theories of language for communication brought a more integrative perspective to language testing, and communicative events are recognized as containing more than just the sum of their language elements. Nowadays, test designers still have to look for a more valid and authentic instruments that would reflect real-world interaction (Brown & Abeywickrama 2010: 12ff.).

ในช่วงกลางของศตวรรษที่ 20 การสอนและการทดสอบภาษาได้รับอิทธิพลอย่างมากจากจิตวิทยาพฤติกรรมนิยมและภาษาศาสตร์เชิงโครงสร้าง เน้นไปที่มุมมองไวยากรณ์ระดับประโยค นิยามคำศัพท์ และการแปลจากภาษาหนึ่งไปเป็นอีกภาษาหนึ่ง และความสนใจส่วนน้อยอยู่ที่การสื่อสารในโลกความเป็นจริง แบบสอบมักเป็นแบบเลือกตอบสำหรับไวยากรณ์และคำศัพท์ รวมทั้งมีแบบฝึกหัดแปลตั้งแต่ระดับคำเรื่อยไปจนถึงเป็นย่อหน้าสั้นๆ

In the mid-twentieth century, language teaching and language testing were greatly influenced by behavioral psychology and structural linguistics. They focused much on sentence-level grammar, vocabulary definition, and translation, and little on real-world communication. Tests, typically in multiple-choice format, touched upon grammar and vocabulary, and had translation exercises which ranged from the word level to the level of short paragraphs.

แบบสอบข้างต้นนิยมเรียกกันว่าเป็นข้อสอบแบบเรื่องเดี่ยว (discrete point) ปัจจุบันก็ยังคงนิยมใช้กันในการสอบเข้ามหาวิทยาลัยทั่วโลก การทดสอบแบบนี้ตั้งอยู่บนสมมุติฐานที่ว่า ภาษาสามารถแยกออกเป็นองค์ประกอบย่อยๆ และองค์ประกอบย่อยๆ นี้สามารถนำมาทดสอบได้อย่างดี การทดสอบเช่นนี้ใช้วิธีการวัดทางจิตวิทยาและเชิงโครงสร้างของการวัดผลเป็นฐาน ซึ่งผู้ออกแบบข้อสอบใช้เครื่องมือทางการวัดผลเช่น ความตรง ความเที่ยง และความเป็นปรนัย ในการวิเคราะห์ข้อสอบได้

The above tests are usually called discrete-point testing. Nowadays they are still used widely in standardized entrance examinations. The examinations are based on the assumptions that language is componential, and those components can be tested well. Such tests are based on psychometric-structuralist approach, in which the test designers use tools to deal with validity, reliability, and objectivity in measurement.

วิธีการวัดแบบเรื่องเดี่ยวพบว่าขาดความสมจริง (inauthentic) เพราะขาดบริบทการสื่อสารในการทดสอบ ในขณะที่การสอนภาษาเริ่มหันมาเน้นการสื่อสาร ความสมจริง และบริบทมากขึ้น การทดสอบจึงกลายมาเป็นแบบบูรณาการ (integrative testing) กันมากขึ้น ตัวอย่างของการทดสอบแบบบูรณาการได้แก่ ข้อสอบแบบโคลซ (cloze tests) และข้อสอบแบบบอกจด (dictations)

The discrete-point approach proved to be inauthentic because it lacked communicative contexts in testing. As language pedagogy turned to a more communicative, authentic approach with contexts, language testing became integrative testing. Some examples of integrative testing include cloze tests and dictations.

ข้อสอบแบบโคลซคือข้อความที่ทุกเจ็ดคำ (โดยปกติ ±2 คำ) ถูกลบออกไป ผู้สอบต้องคิดคำเพื่อเติมลงในช่องว่างเหล่านั้น ข้อสอบแบบโคลซเชื่อว่าวัดสมิทธิภาพโดยรวมได้ดี

A cloze test is a passage in which every seventh word (usually ±2 words) is deleted. The test takers need to supply words that would fit into those blanks. Cloze tests were claimed to be good measures of overall proficiency.

ข้อสอบแบบบอกจด (dictation) คือข้อสอบที่ผู้สอบฟังข้อความสั้นๆ และเขียนตามที่ได้ยิน ข้อสอบแบบบอกจดเป็นข้อสอบแบบบูรณาการเพราะว่าวัดด้านไวยากรณ์และด้านวัจนกรรมที่จำเป็นสำหรับการแสดงออกด้านอื่นๆ ในการรู้ภาษา การทำข้อสอบแบบบอกจดได้ดีต้องอาศัยการฟังอย่างตั้งใจ การเขียนสิ่งที่ได้ยินออกมา ความจำระยะสั้นที่มีประสิทธิภาพ และความสามารถในการคาดการณ์เพื่อช่วยความจำระยะสั้น

Dictation is a test in which the test takers listen to a short passage and write what they hear. Dictation was argued to be an integrative test because it used grammatical and discourse competence, which were essential in expressing other aspects of language.

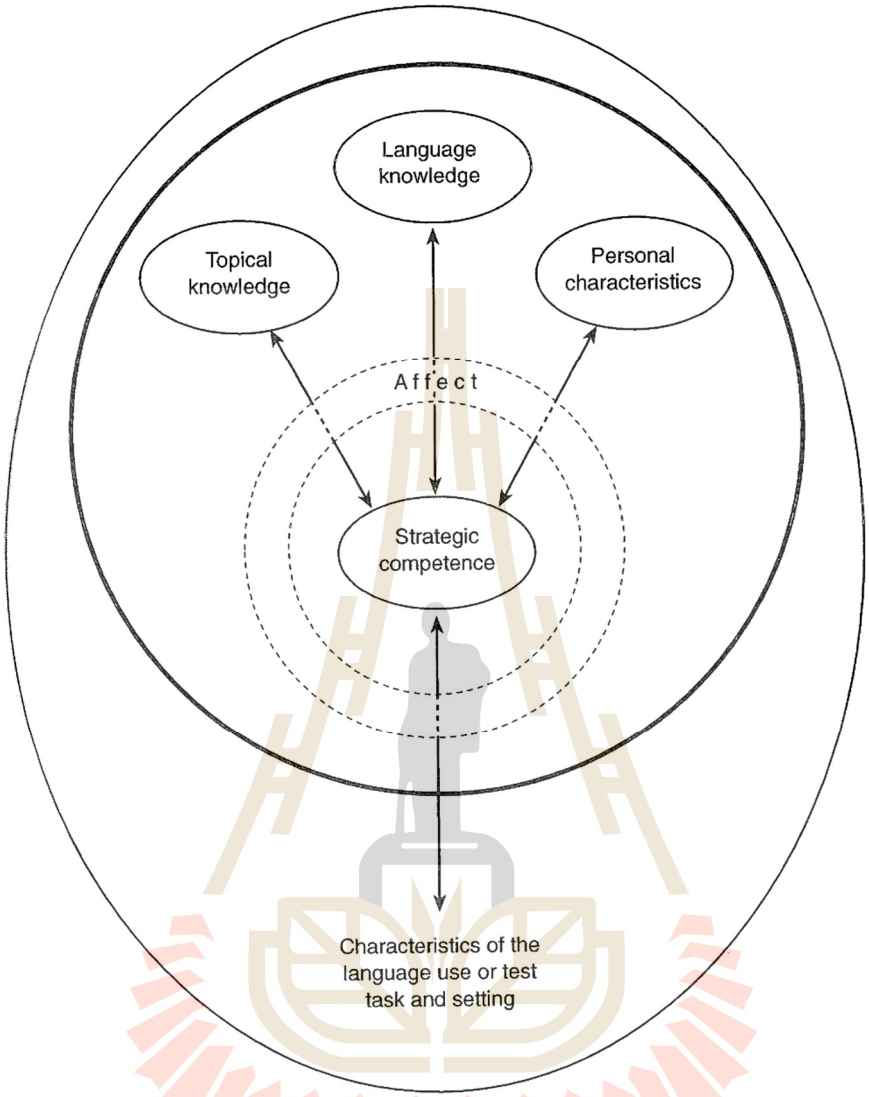
Doing dictation well needed careful listening, reproduction of what was heard in writing, efficient short-term memory, and expectancy ability to aid short-term memory.

นักวิชาการที่สนับสนุนวิธีการทดสอบแบบบูรณาการ ในไม่ช้าก็หันมาสนับสนุนข้อสมมุติฐานว่าด้วยความสามารถหนึ่งเดียว (unitary trait hypothesis) ซึ่งเกี่ยวกับการมองว่าสมิทธิภาพภาษาแบ่งแยกไม่ได้ และมีตัวแปรสมิทธิภาพทางภาษาทั่วไปที่ครอบคลุมองค์ประกอบย่อย แต่ในที่สุดหลังการโต้แย้งและหลักฐานจากการวิจัย ข้อสมมุติฐานว่าด้วยความสามารถหนึ่งเดียวก็ถูกละทิ้งไป

Proponents of integrative testing soon turned to support the unitary trait hypothesis, in which language proficiency was indivisible, and there was a general factor of language proficiency in which all the discrete points did not add up to that whole. But after debates and evidence from research, the unitary trait hypothesis was abandoned.

ในช่วงกลางทศวรรษที่ 1980 การทดสอบทางภาษาได้เริ่มหันมาสนใจการออกแบบชิ้นงานเพื่อการสื่อสาร Bachman & Palmer (1996) ได้กล่าวถึงความจำเป็นที่ต้องมีความสอดคล้องกันระหว่างการแสดงออกในการทดสอบทางภาษาและการใช้ภาษาในสถานการณ์ที่มิใช่การสอบ Bachman & Palmer ยังเน้นย้ำถึงความสำคัญของสามัตถิยะกลยุทธ์ในการสื่อสาร (strategic competence) เช่น ความสามารถชดเชยการสื่อความที่ไม่สมบูรณ์ รูปที่ 8.1 แสดงความสัมพันธ์นี้

In the mid-1980s, the field of language testing began to pay attention to designing communicative assessment tasks. Bachman & Palmer (1996) mentioned the necessity of correspondence between language in test performance and non-test language use. They also highlighted the importance of strategic competence in, for example, compensating for breakdowns in communication. Figure 8.1 illustrates this point.



รูปที่ 8.1 สามัตถิยะผู้สอบกับการใช้หรือการทดสอบทางภาษา (Bachman & Palmer 1996: 63)
(Test users' language competence and language use or test task)

จากการที่จำเป็นต้องมีความสอดคล้องกันระหว่างการแสดงออกในการทดสอบทางภาษาและการใช้ภาษาในสถานการณ์ที่มิใช่การสอบ การทดสอบภาษาเพื่อการสื่อสารจึงเป็นความท้าทายสำหรับผู้ออกแบบข้อสอบ ผู้เขียนข้อสอบเริ่มระบุลักษณะชิ้นงานในโลกความเป็นจริงที่ผู้เรียนภาษาต้องแสดงออกและต้องนำมาตรวจพิสูจน์ (validate) ด้วยสิ่งที่ผู้ใช้ภาษาทำจริงๆ กับตัวภาษา ในการวัดสมิทธิภาพทางภาษา ผู้เขียนข้อสอบจึงต้องระบุให้ได้ว่า ผู้เรียนใช้ภาษาเช่นนั้นเช่นนี้ที่ใด เมื่อใด อย่างไร กับใคร ทำไม ใน

หัวข้อใด และมีผลเช่นไร การวัดผลก็เริ่มหันมาใส่ใจกับความสมจริงของชิ้นงาน และความเป็นตามสภาพจริงของข้อความที่ใช้ในการทดสอบ

From the necessity of a correspondence between language test performance and non-test language use, communicative language testing became a challenge to test designers. The test designers began to identify the real-world tasks that the test takers had to perform and validate them with what the test takers actually do with the language. In language proficiency testing, the test designers had to specify where, when, how, why, in what topic, and with what effect the learners use the language. The assessment turned to involve authenticity of the test tasks and authenticity of the texts used for the assessment.

ปัจจุบัน คอร์สสอนภาษาและหลักสูตรทั่วโลกกำลังจัดทำวัตถุประสงค์ที่มีผู้เรียนเป็นศูนย์กลาง แทนที่จะใช้การทดสอบที่ใช้ปากกากับกระดาษ ก็ใช้การวัดผลด้วยการแสดงออก (performance-based assessment) การวัดผลทางภาษาด้วยการแสดงออกมักจะเกี่ยวข้องกับการส่งสารด้วยคำพูด การส่งสารด้วยการเขียน การตอบสนองต่อคำถามปลายเปิด การแสดงออกแบบบูรณาการหลายทักษะ การแสดงออกแบบกลุ่ม รวมไปถึงชิ้นงานที่มีปฏิสัมพันธ์อื่นๆ โดยเฉพาะอย่างยิ่ง การสัมภาษณ์ปากเปล่า การวัดผลแบบนี้กินเวลามากและมีค่าใช้จ่ายสูง แต่ก็ทำให้วัดได้โดยตรงและแม่นยำมากขึ้น เพราะผู้เรียนถูกประเมินในขณะที่แสดงออกในชิ้นงานจริงๆ หรือในชิ้นงานที่จำลองสถานการณ์จริง ในแง่การวัดผล การวัดแบบนี้ถือว่ามีความตรงเชิงเนื้อหาสูงกว่าเพราะผู้สอบถูกวัดในขณะที่กำลังแสดงชิ้นงานที่เป็นเป้าหมาย อีกชื่อเรียกของการวัดผลด้วยการแสดงออกจึงได้แก่ การวัดผลที่อิงชิ้นงาน (task-based assessment)

At present, language courses and programs around the world are handling this new and student-centered approach to assessment. Instead of paper-and-pencil tests, performance-based assessment deals with oral and written production, open-ended responses, integrated performance, group performance, and other interactive tasks, especially oral interviews. Performance-based assessment, thus, is time-consuming and expensive. But, in return, the assessment is more direct as well as more accurate, because the learners' performance is assessed while they are actually doing the tasks, or simulated tasks reflecting the real world. In assessment, performance-based assessment is considered to have higher content validity because the test takers are assessed while doing the target tasks. An alternative term for performance-based assessment is thus task-based assessment.

8.2 ประเด็นในการประเมินผลทางภาษา (Issues in Language Assessment)

พหุปัญญา (multiple intelligences)

เชาวน์ปัญญาครั้งหนึ่งเคยเชื่อกันว่าคือความสามารถแก้ปัญหาทางภาษาและตรรกะ-คณิตศาสตร์ ไอคิว (IQ – intelligence quotient) ได้มีอิทธิพลอย่างกว้างขวางในการทดสอบมาเกือบศตวรรษ เชาวน์ปัญญาวัดโดยแบบสอบที่จับเวลาและเป็นแบบข้อเดี่ยวประกอบด้วยข้อย่อยๆ เป็นลำดับขั้น สาขาวิชาอื่นจึงน่าจะทำแบบสอบเพื่อวัดแบบเดียวกันได้ โดยใช้ข้อสอบอิงกลุ่มมาตรฐานที่จับเวลาและเป็นแบบเลือกตอบ ซึ่งหลายข้อก็อาจไม่มีความสมจริง (Brown & Abeywickrama 2010: 17ff.)

Intelligence was once believed to be linguistic and logical-mathematical problem-solving ability. IQ (intelligence quotient) has influenced the field of measurement and testing for almost a century. An IQ test is a timed, discrete-point test, containing a hierarchy of items, and so tests of other fields of study could be designed similarly. For example, we have witnessed a world of standardized, norm-referenced tests, which are timed in a multiple-choice format, but many items of which are inauthentic (Brown & Abeywickrama 2010: 17ff.).

อย่างไรก็ดี งานวิจัยเกี่ยวกับเชาวน์ปัญญาในช่วงปลายศตวรรษที่ 20 ได้เปลี่ยนมุมมองแห่งโลกการวัดทางจิตวิทยา หนึ่งในงานวิจัยดังกล่าวได้แก่ เรื่องพหุปัญญา (multiple intelligences) ที่ได้แตกเชาวน์ปัญญาออกเป็นแปดด้าน อีกงานหนึ่งได้แก่ เชาวน์ปัญญาทางด้านอารมณ์ (emotional quotient – EQ) งานเหล่านี้มีอิทธิพลต่อด้านการวัดและประเมินผลทางภาษาทางอ้อม เช่น กิจกรรมในชั้นเรียนเพื่อการสื่อสารในหนังสือตำราและหลักสูตรต่างๆ ได้เพิ่มความหลากหลายด้านความสามารถในการเรียนและสไตล์ความถนัดในการเรียน ส่วนด้านการเรียนการสอนได้ลดทอนการพึ่งพาแต่ข้อสอบแบบเลือกตอบและจับเวลาลงในการวัดความสามารถทางภาษา ครูผู้สอนก็ได้รับการสนับสนุนให้ลดความสนใจแต่เรื่องความเป็นปรนัยในการวัดลง นอกจากนี้ ครูและผู้จัดสอบยังได้รับการสนับสนุนให้วัดทักษะภาษาในภาพรวม กระบวนการเรียนรู้ และความสามารถในการต่อรองความหมาย ความท้าทายในปัจจุบันจึงได้แก่การออกแบบการประเมินที่จะวัดด้านระหว่างบุคคล การสื่อสาร ความคิดสร้างสรรค์ และการมีปฏิสัมพันธ์ ซึ่งก็จะต้องมีความเชื่อมั่นในความเป็นอัตนัยและสัญชาตญาณของเราเพิ่มขึ้น

Nonetheless, research on intelligence in the late-twentieth century has changed perspectives of psychometrics. One of the studies is on multiple intelligences, which divide intelligence into eight different aspects. Another is on emotional quotient (EQ).

These studies have an indirect influence on language assessment. For example, communicative classroom activities in books and programs have increased in terms of diversity in learning abilities and learning styles. In language pedagogy, reliance on timed and multiple-choice format has decreased for language assessment. Moreover, teachers and test administrators have been given support for assessing whole language skills, learning processes, and the students' ability to negotiate meaning. Current challenges thus lie in assessing interpersonal, communicative, creative, and interactive skills, which would rely more on our intuition and subjectivity.

<u>การประเมินผลแบบดั้งเดิมและการประเมินผลทางเลือก (Traditional and Alternative Assessment)</u>

ในการประเมินผลการแสดงออกในชั้นเรียน มีแนวโน้มที่ว่า การออกแบบการทดสอบแบบดั้งเดิมจะถูกเสริมด้วยทางเลือกที่มีความสมจริงมากกว่าในแง่ของการดึงพฤติกรรมการสื่อสารอย่างมีความหมาย ตารางที่ 8.1 แสดงลักษณะของการประเมินทั้งสองแบบอย่างคร่าวๆ

In performance-based classroom assessment, there is a trend for traditional assessment to be supplemented by alternative assessment, which is more authentic in its elicitation of meaningful communication. Table 8.1 shows differences between the two approaches.

ตารางที่ 8.1 การประเมินผลแบบดั้งเดิมและการประเมินผลทางเลือก (Traditional and Alternative Assessment)

การประเมินผลแบบดั้งเดิม (Traditional Assessment)	**การประเมินผลทางเลือก (Alternative Assessment)**
การทดสอบมาตรฐาน (standardized exams)	วัดผลระยะยาวอย่างต่อเนื่อง (continuous long-term assessment)
แบบเลือกตอบและจับเวลา (timed, multiple choice format)	แบบคำถามปลายเปิดและไม่จับเวลา (untimed, open-ended responses)
ข้อสอบแยกบริบท (decontextualized test items)	ชิ้นงานการสื่อสารอิงบริบท (contextualized communicative tasks)

การประเมินผลแบบดั้งเดิม (Traditional Assessment)	การประเมินผลทางเลือก (Alternative Assessment)
คะแนนเป็นข้อมูลสะท้อนกลับ (scores as feedback)	ให้ข้อมูลสะท้อนกลับเป็นรายคน (individualized feedback)
คะแนนอิงกลุ่ม (norm-referenced scores)	คะแนนอิงเกณฑ์ (criterion-referenced scores)
มุ่งไปที่คำตอบที่ชัดเจนตายตัว (focus on discrete answers)	คำตอบปลายเปิด สร้างสรรค์ (open-ended, creative answers)
สรุปรวม (summative)	ระหว่างเรียน (formative)
มุ่งไปที่ผลงาน (oriented to product)	มุ่งไปที่กระบวนการ (oriented to process)
การแสดงออกที่ไม่มีปฏิสัมพันธ์ (noninteractive performance)	การแสดงออกที่มีปฏิสัมพันธ์ (interactive performance)
ส่งเสริมแรงจูงใจภายนอก (fosters extrinsic motivation)	ส่งเสริมแรงจูงใจภายใน (fosters intrinsic motivation)

แต่กระนั้น ก็มีข้อควรระวังสองประการในการใช้ตารางที่ 8.1 ประการแรก แนวคิดรวบยอดที่ปรากฏในตารางเป็นการสรุปรวมอย่างคร่าวๆ ในความเป็นจริง เป็นเรื่องยากที่จะแยกแนวคิดออกจากกันได้อย่างเด็ดขาด การประเมินผลหลายแบบก็รวมลักษณะจากทั้งของการประเมินผลแบบดั้งเดิมและแบบทางเลือก และข้อควรระวังประการที่สองคือ ตารางที่ 8.1 มีความลำเอียงไปทางการประเมินผลทางเลือก จึงไม่ถูกต้องนักหากจะมองว่าทุกอย่างทางด้านซ้ายมือมีตำหนิ และทุกอย่างทางด้านขวามือสมบูรณ์แบบ ดังนั้น รูปแบบใดในการประเมินผลแบบดั้งเดิมที่ทำหน้าที่ได้ดีอยู่แล้ว เราก็สามารถนำมาใช้ได้ ส่วนทางเลือกในการประเมินผลทางเลือกใดที่สามารถนำมาใช้อย่างสร้างสรรค์ เราก็สามารถนำมาปรับใช้ได้ในชั้นเรียน

Nonetheless, there are some caveats in using Table 8.1. First, the concepts in the table are just generalizations. In reality, it is difficult to differentiate the two columns completely. Many forms of assessment may combine features from both traditional assessment and alternative assessment. Second, Table 8.1 contains bias toward alternative assessment. It thus is not right to see that everything on the left-hand side is flawed, and everything on the right-hand side perfect. Therefore, assessment traditions that already work

well should be valued. And alternatives in assessment which we can use constructively in our classroom contexts could be adapted.

การทดสอบผ่านคอมพิวเตอร์ (Computer-based testing)

ในช่วงหลายปีมานี้ ได้มีการใช้เทคโนโลยีคอมพิวเตอร์และการประยุกต์ใช้ในการเรียนการสอนภาษามากขึ้น ผู้เรียนภาษาแทบทุกคนทั่วโลกกลายมาเป็นผู้ใช้คอมพิวเตอร์ จึงไม่น่าแปลกใจที่รายวิชาจำนวนมากใช้คอมพิวเตอร์ช่วยเรียนภาษา (computer-assisted language learning – CALL) ในด้านการประเมินผลการเรียนภาษาก็เช่นกัน บ้างเป็นแบบทดสอบขนาดเล็กตามเว็บไซต์ต่างๆ บ้างเป็นแบบทดสอบมาตรฐานขนาดใหญ่ที่มีผู้สอบหลายหมื่นคน ผู้เรียนได้รับตัวกระตุ้น (prompts) ในรูปคำพูดหรือข้อความที่เขียนขึ้นจากอัลกอริทึมที่โปรแกรมไว้ล่วงหน้า และต้องพิมพ์ตอบ ข้อสอบผ่านคอมพิวเตอร์ส่วนใหญ่ต้องการคำตอบที่ตายตัว แต่ข้อสอบเช่นโทเฟล (Test of English as a Foreign Language – TOEFL) ปัจจุบันมีส่วนที่เขียนเรียงความ และส่วนที่ต้องพูด ซึ่งทั้งสองส่วนให้คะแนนโดยผู้ตรวจให้คะแนน

In recent years, computer technology grows and is applied more in language teaching and learning. Nearly all language learners have become computer users. Thus, it is no surprise that a large number of subjects or programs use computer-assisted language learning (CALL). Similarly, language assessment uses CALL in test making and administration. Some websites host small-scale computer-based tests. Other testing programs provide standardized, large-scale tests which thousands of test takers are involved in. The test takers receive prompts in the form of written or spoken stimuli predetermined by algorithms, and have to type their responses. Most of these computer-based tests require fixed and closed-ended responses. But examinations like the Test of English as a Foreign Language (TOEFL) presently have essay and spoken parts, which are both scored by human raters.

พัฒนาการใหม่ๆ ในการประเมินผลด้วยคอมพิวเตอร์รวมไปถึงการใช้ภาษาศาสตร์คลังข้อมูลเพิ่มความสมจริงในการสอบ การออกแบบชิ้นงานที่ซับซ้อนขึ้นในการทดสอบผ่านคอมพิวเตอร์ การใช้ซอฟต์แวร์ที่แยกแยะการพูดและการเขียน เพื่อตรวจให้คะแนนการพูดและการเขียนเพื่อส่งสาร เป็นต้น เกิดเป็นประเด็นคำถามสำหรับการวิจัยใหม่ๆ ว่า การทดสอบทางภาษาโดยใช้คอมพิวเตอร์เป็นสื่อกลางเปลี่ยนธรรมชาติของสิ่งที่ถูกวัดหรือไม่และอย่างไร

New developments in computer-based assessment involve contributions of corpus linguistics, which increases authenticity of the texts used for testing, designs of more

complex test tasks that are delivered by a computer, use of software that recognizes speech and writing to score oral and written production. These developments pose issues in terms of the construct being measured, for example, whether and how the computer-based test delivery changes the nature of the test construct.

การทดสอบผ่านคอมพิวเตอร์อีกชนิดที่เริ่มเป็นที่นิยมกันมากขึ้นได้แก่ ข้อสอบแบบปรับเหมาะ (computer-adaptive test – CAT) ในการทดสอบแบบปรับเหมาะ ผู้สอบแต่ละคนจะได้ชุดคำถามที่ตรงกับลักษณะเฉพาะของแบบทดสอบ (test spec) และเหมาะสมกับระดับการแสดงออกของผู้สอบแต่ละคน ข้อสอบแบบปรับเหมาะเริ่มต้นด้วยคำถามที่มีความยากปานกลาง ขณะที่ผู้สอบตอบแต่ละคำถาม คอมพิวเตอร์จะตรวจให้คะแนนคำถามนั้นและใช้ข้อมูลนั้น รวมไปถึงข้อมูลจากการตอบสนองต่อข้อสอบก่อนหน้านั้น เพื่อกำหนดว่าควรจะส่งคำถามใดมาเป็นคำถามถัดไป ตราบเท่าที่ผู้สอบยังตอบได้ถูกต้อง คอมพิวเตอร์ก็จะเลือกคำถามที่ยากเท่ากันหรือยากยิ่งขึ้นมาให้ผู้สอบ แต่หากผู้สอบตอบผิด คอมพิวเตอร์ก็จะส่งคำถามที่ยากเท่ากันหรือยากน้อยกว่ามาให้ผู้สอบ ในการสอบแบบปรับเหมาะ ผู้สอบจะเห็นคำถามเพียงครั้งละหนึ่งคำถาม ดังนั้น ผู้สอบจะทำข้อสอบข้ามข้อไม่ได้ รวมถึงไม่สามารถย้อนกลับไปทำข้อที่ทำไปแล้วได้

Another type of computer-based assessment is a computer-adaptive test (CAT). In a CAT, which is gaining momentum, each test taker will receive a set of questions which match the test specification and are suitable for the performance level. The CAT begins with questions of moderate difficulty. While the test taker is answering each question, the computer will score the question and use that information, together with responses to other previous items, to determine which question should be the next. As long as the test taker gets it right, the computer will send items of equal or greater difficulty. But if the test taker gets it wrong, the computer will send items of equal or lesser difficulty. In a CAT, the test taker will see only one question at a time. Therefore, he/she cannot skip questions or return to the items that have been answered.

ในบทที่ 7 กล่าวถึงการประเมินผลทางเลือกว่า มีความพยายามใช้การประเมินผลทางเลือกเสริมเข้ามากับการประเมินผลแบบดั้งเดิม มีคนกล่าวไว้ว่า การทดสอบผ่านคอมพิวเตอร์ หากพัฒนาจนถึงขีดสุด จะทำลายความพยายามในการใช้การประเมินผลทางเลือก รวมไปถึงการประเมินที่ปรับให้เข้ากับชั้นเรียน อย่างไรก็ดี เทคโนโลยีคอมพิวเตอร์ไม่จำเป็นต้องทำให้เกิดผลกระทบเช่นนั้น ครูผู้สอน

และผู้เขียนข้อสอบปัจจุบันถือว่ามีเครื่องมือหลากหลายมากยิ่งขึ้นที่จะทำให้การทดสอบผ่านคอมพิวเตอร์ไม่เป็นแบบตายตัว ด้วยการใช้นวัตกรรมและเทคโนโลยีอย่างสร้างสรรค์ ครูและผู้สอบสามารถเพิ่มความสมจริง เพิ่มการแลกเปลี่ยนปฏิสัมพันธ์กัน ตลอดจนเพิ่มการพึ่งพาตนเองของผู้เรียนได้

In Chapter 7, it is said that alternative assessments are supplementing traditional assessment. Computer-based testing is said to eliminate such efforts as well as teacher-tailored classroom assessment if it is developed to the ultimate level. However, this does not have to be the case. Teachers and test makers are in a position to use a variety of tools that will make computer-based testing less formulaic. By using technology and innovations constructively, they can increase authenticity, enhance interactive exchange, and encourage student autonomy.

8.3 Exercise

1. Which is the least related to the developments in the 1940s and 1950s?

A. behavioral psychology

B. contrastive analysis

C. pragmatic expectancy

D. structuralist linguistics

2. Which approach in test writing is the most authentic?

A. behavioralist

B. communicative

C. integrative

D. structuralist

3. According to Bachman & Palmer (1996), with what does language in test performance need to have correspondence?

A. language knowledge

B. non-test language use

C. personal characteristics

D. strategic competence

4. Which is **NOT** a characteristic of alternative assessment?

A. criterion-referenced scores

B. formative

C. oriented to process

D. scores as feedback

5. What is the most important issue in computer-based testing?

A. cost of test administration

B. nature of the test construct

C. software availability

D. test length

Answer key: 1. C. 2. B. 3. B. 4. D. 5. B.

บรรณานุกรม (Bibliography)

Bachman, L.F., 2000. Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1–42.

Bachman, L. F. 1997. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. 1996. *Language testing in practice: design and developing useful language tests*. Oxford: Oxford University Press.

Brown, H. D., & Abeywickrama, P., 2004. *Language assessment: Principles and classroom practices*. London: Pearson Education.

Brown, J.D., 2016. *Statistics corner*. Tokyo, Japan: JALT Testing and Evaluation Special Interest Group.

Brown, J. D. , 2000. What is construct validity? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*(2), 8–12.

Brown, J. D., & Hudson, T., 1998. The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653-675. doi:10.2307/3587999

Canale, M. 1984. Considerations in the testing of reading and listening proficiency. *Foreign Language Annals, 17,* 349–357.

Chapelle, C., Enright, M., & Jamieson, J. 2008. *Building a validity argument for the test of English as a foreign language*. New York: Routledge.

EFL Learning, 2015, Dec 25. *Beyond test: Alternatives in assessment*. Available: https://www.slideshare.net/eflearners/language-assessment-beyond-testalternatives-assessment-by-efl-learners

Fulcher, G. 2014. Philosophy and language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1431–51). West Sussex, UK: John Wiley & Sons. doi:10.1002/9781118411360.wbcla032

Fulcher, G. and Davidson, F., 2007. *Language testing and assessment*. Oxon, UK: Routledge.

Genesee, F., & Upshur, J. A. 1996. *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Gottlieb, M. 1995. Nurturing student learning through portfolios. *TESOL Journal*, *5*(1), 12-14.

Heaton, J. B. 1988. *Writing English language tests*. London: Longman.

Kane, M.T., 2012. All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives, 10*(1–2), 66–70.

Kane, M.T., 2006. Validation. In: Brennan, R.L. (ed.), *Educational measurement*. 4th ed. Connecticut, US: American Council on Education; Praeger, 17–64.

McNamara, T., 2006. Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly: An International Journal, 3*(1), 31–51.

Mertler, C. A. 2001. Designing scoring rubrics for your classroom. *Practical Assessment Research & Evaluation, 7*(25). Available from http://pareonline.net/getvnasp?v=7&n=25

Minh, P. P. K. 2015, Mar 31. *Fundamental concepts and principles in language testing*. Available: https://www.slideshare.net/khanhminhlala/fundamental-concepts-and-principles-in-language-testing?from_action=save

Moss, P.A., 2007. Reconstructing validity. *Educational Researcher, 36*(8), 470–6.

NSW Government. 2019, Nov 6. Approaches to assessment. *Aspects of Assessment*. Available: https://education.nsw.gov.au/teaching-and-learning/professional-learning/teacher-quality-and-accreditation/strong-start-great-teachers/refining-practice/aspects-of-assessment/approaches-to-assessment.

Olteanu, C. 2017. Reflection-for-action and the choice or design of examples in the teaching of mathematics. *Mathematics Education Research Journal*, *29*, 349–367. https://doi.org/10.1007/s13394-017-0211-9

Richards, J.C., 1983. Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*, 219–239.

Rigney, S.L., Wiley, D.E. and Kopriva, R.J., 2008. The past as preparation: Measurement, public policy and implications for access. In: Kopriva, R.J. (ed.), *Improving testing for English language learners*. New York: Routledge, 37–63.