ข้อพิจารณาในการคัดเลือกข้อคำถามที่มีอำนาจการจำแนก*

*บัณฑิตา อินสมบัติ***

บทคัดย่อ

อำนาจการจำแนกของข้อคำถาม (item discrimination power) เป็นสารสนเทศที่บ่งชี้ว่าข้อคำถามนั้นๆ สามารถจำแนกความแตกต่างของกลุ่มตัวอย่างตามคุณลักษณะที่ต้องการได้ดีเพียงใด หากข้อคำถามนั้นสามารถจำแนกความแตกต่างได้อย่างถูกต้องแสดงว่าข้อคำถามนั้นวัดได้ตรงตามคุณลักษณะที่ต้องการ ซึ่งเรียกว่าความเที่ยงตรงรายข้อ (item validity) โดยปกติแล้วผู้สร้างมักกำหนดเกณฑ์การคัดเลือกข้อคำถามที่มีค่าอำนาจการจำแนกตั้งแต่ .20 ขึ้นไปเป็นข้อคำถามที่สามารถนำไปใช้ได้โดยเข้าใจว่าเป็นเกณฑ์ที่ใช้ได้โดยทั่วไป หากพิจารณาค่าอำนาจการจำแนกที่มีนัยสำคัญ จะพบว่ามีความสัมพันธ์กับจำนวนกลุ่มตัวอย่างและค่าความยาก กล่าวคือ หากกลุ่มตัวอย่างมีจำนวนมากค่าอำนาจการจำแนกที่มีนัยสำคัญจะมีค่าน้อย แต่หากกลุ่มตัวอย่างมีจำนวนน้อยค่าอำนาจการจำแนกที่มีนัยสำคัญจะมีค่ามาก และอำนาจการจำแนกที่มีนัยสำคัญจะมีค่าสูงขึ้นหากข้อคำถามมีค่าความยากเข้าใกล้ .50 ดังนั้น ผู้สร้างจึงควรมีความรอบคอบในการพิจารณาคัดเลือกข้อคำถามที่มีค่าอำนาจการจำแนกที่เหมาะสมเพื่อให้สามารถจำแนกความแตกต่างของกลุ่มตัวอย่างตามคุณลักษณะที่ต้องการได้อย่างถูกต้อง

คำสำคัญ: ข้อพิจารณา/ การคัดเลือกข้อคำถาม/ อำนาจการจำแนก

*อาจารย์ ประจำคณะครุศาสตร์ มหาวิทยาลัยราชภัฏนครสวรรค์

**กศ.ด. (วิจัยและประเมินผลการศึกษา), มหาวิทยาลัยนเรศวร

Consideration tips in selecting test items with discrimination power*

Bantita Insombat, Ph.D.

Abstract

Item discrimination power is the information indicating how well a test item can discriminate between the sample according to the set characteristics. Ability to discriminate correctly shows that the test item is valid in measuring the characteristics as specified or has item validity. Generally test item makers would set the criteria in selecting test item with the discrimination power of .20 up with an understanding that it is accepted in general. However, having another look it can be found that the significant discrimination power is correlated to the number of samples and item difficulty. That is, the bigger the number of samples, the smaller the discrimination power, and vice versa. Also, the significant discrimination power will get higher if the test item difficulty value is nearer to .50. Therefore, test item makers ought to be careful in considering which item is suitable as far as discrimination power is concerned so as to be able to discriminate between the samples according to the required characteristics correctly.

Keywords: consideration tips, selecting test items, discrimination power

* Full-time lecturer for the Faculty of Education, Nakhon Swan Rajabhat University
Ph.D. (Educational Research and Evaluation), Naresuan University

บทนำ

การวิเคราะห์รายข้อ (item analysis) มีวัตถุประสงค์สำคัญเพื่อค้นหาสารสนเทศที่แสดงถึงคุณภาพของข้อคำถามแต่ละข้อ ซึ่งสารสนเทศที่กล่าวถึงนี้จะเกี่ยวข้องกับคุณภาพรายข้อของเครื่องมือที่สำคัญ 3 ประการ ได้แก่ สารสนเทศเกี่ยวกับตัวลวง (information about distractors) สารสนเทศเกี่ยวกับความยากของข้อคำถาม (information about item difficulty) และสารสนเทศเกี่ยวกับอำนาจการจำแนกของข้อคำถาม (information about item discrimination) สำหรับการพัฒนาเครื่องมือที่ใช้ในการวัดและประเมินผลการเรียนรู้ หรือเครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลการวิจัย หากสิ่งที่ต้องการวัดนั้นเป็นความสามารถทางสติปัญญาหรือคุณลักษณะทางจิตวิทยา (trait) ของกลุ่มตัวอย่าง อาทิ ผลสัมฤทธิ์ทางการเรียน แรงจูงใจใฝ่สัมฤทธิ์ ความสนใจ ความฉลาดทางอารมณ์ ความสุขในการเรียนรู้ ความซื่อสัตย์ ความคิดสร้างสรรค์ และความมีเหตุผล เป็นต้น ย่อมมีความจำเป็นต้องวิเคราะห์คุณภาพรายข้อเพื่อค้นหาค่าอำนาจการจำแนกของข้อคำถาม ซึ่งเป็นค่าที่บ่งชี้ว่าข้อคำถามแต่ละข้อนั้นสามารถจำแนกความแตกต่างของกลุ่มตัวอย่างตามคุณลักษณะที่ต้องการวัดได้มากน้อยเพียงไร การจำแนกความแตกต่างของกลุ่มตัวอย่างได้อย่างถูกต้องแน่นอนนั้น หมายความว่าข้อคำถามสามารถวัดได้ตรงตามความมุ่งหมายที่ต้องการซึ่งกล่าวอีกอย่างหนึ่งว่าข้อคำถามสามารถวัดในสิ่งเดียวกันหรือสามารถวัดได้ตรงตามคุณลักษณะที่ต้องการ เช่นนี้เรียกว่าข้อคำถามมีความเที่ยงตรงรายข้อ (item validity) และเมื่อข้อคำถามมีความเที่ยงตรงรายข้อสูงแล้วย่อมส่งผลให้เครื่องมือนั้นมีความเชื่อมั่น (reliability) สูงด้วย

ตามทฤษฎีการทดสอบแบบดั้งเดิม (classical theory) มีวิธีการอธิบายการวิเคราะห์อำนาจการจำแนกมากกว่า 20 วิธี อาทิ สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (pearson correlation coefficient : r_{XY}) สัมประสิทธิ์สหสัมพันธ์ไบซีเรียล (biserial correlation coefficient : r_{bis}) สัมประสิทธิ์สหสัมพันธ์พอยท์ไบซีเรียล (point-biserial correlation coefficient : r_{pbis}) สัมประสิทธิ์สหสัมพันธ์แบบฟี (phi coefficient : Φ) สัมประสิทธิ์สหสัมพันธ์เตตะคลอริค (tetrachoric correlation coefficient : r_{tet}) สัมประสิทธิ์สหสัมพันธ์แรงไบซีเรียล (rank biserial correlation coefficient : r_{rb}) สัมประสิทธิ์ของฟานาแกน (Fanagan's coefficient) และสัมประสิทธิ์ของเดวิส (Davis's coefficient) เป็นต้น อำนาจจำแนกที่วิเคราะห์จากค่าสัมประสิทธิ์สหสัมพันธ์จะเขียนแทนด้วยสัญลักษณ์ r สำหรับวิธีการวิเคราะห์อำนาจการจำแนกที่สะดวกและนิยมมากในปัจจุบันอีกวิธีการหนึ่งคือ การหารผลต่างระหว่างจำนวนคนในกลุ่มสูงและกลุ่มต่ำที่ตอบข้อสอบแต่ละข้อถูกต้องด้วยจำนวนคนในกลุ่มสูงหรือกลุ่มต่ำ ซึ่งอำนาจการจำแนกที่วิเคราะห์โดยวิธีนี้บ่อยครั้งจะเรียกว่าดัชนีอำนาจการจำแนกอย่างง่าย (simple discrimination index) เขียนแทนด้วยสัญลักษณ์ D เดิมเรียกดัชนี U-L (Upper-Lower Index: U-L Index) ผู้นำเสนอวิธีการนี้คือ Johnson A. Pemberton (1951: 499-504) โดยได้นำเสนอไว้ในบทความเรื่อง "Note on a suggested Index of Item Validity : The U-L Index" ใน Journal of Educational Psychology ซึ่ง Engelhart (1969: 72-74) ได้เปรียบเทียบดัชนีอำนาจการจำแนกจำนวน 9 ดัชนี กรณีที่มีค่าความยากต่างกันพบว่าดัชนี D มีประสิทธิภาพในการจำแนกข้อคำถามได้ดีกว่าวิธีอื่นๆ Engelhart จึงได้สนับสนุนให้มีการใช้ดัชนี D ใน

การพัฒนาข้อสอบ หลังจากนั้นในปี 1972 Robert L. Brennan ได้นำเสนอวิธีการวิเคราะห์อำนาจการจำแนกของข้อคำถามซึ่งพัฒนาจาก U-L Index ของ Pemberton เรียกว่าดัชนีการจำแนก B (discrimination index B) ซึ่งเป็นที่นิยมและมีการนำไปใช้อย่างกว้างขวางในการวิเคราะห์คุณภาพข้อสอบแบบอิงเกณฑ์ (criterion reference test)

ดังจะเห็นว่าการอธิบายอำนาจการจำแนกความแตกต่างกันตามคุณลักษณะที่ต้องการวัดของข้อคำถามจะมีหลายค่า แต่ทั้งนี้ค่าที่นิยมใช้กันมากในหมู่ผู้สร้างข้อสอบหรือเครื่องมือการเก็บรวบรวมข้อมูลการวิจัยทั้งครู และนิสิตนักศึกษา หรือคณาจารย์ในมหาวิทยาลัย ได้แก่ ค่า r ซึ่งอธิบายจากค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ของคะแนนรายข้อกับคะแนนรวมซึ่งโดยมากจะเรียกค่าสัมประสิทธิ์นี้ว่า **"ค่าอำนาจการจำแนก"** (discrimination power) และค่า D หรือดัชนี D ซึ่งอธิบายจากค่าความแตกต่างระหว่างสัดส่วนผู้ตอบถูกในกลุ่มที่ได้คะแนนสูงกับกลุ่มที่ได้คะแนนต่ำ (proportion of correct in upper-lower group) รวมทั้งดัชนี B กรณีการวิเคราะห์คุณภาพรายข้อแบบอิงเกณฑ์ซึ่งเรียกว่า **"ดัชนีการจำแนก"** (index of discrimination or discrimination index) เป็นที่ทราบกันดีว่าทั้งค่าอำนาจการจำแนกหรือดัชนีการจำแนกมีค่าอยู่ระหว่าง -1 ถึง +1 หากค่านี้เข้าใกล้ +1 จะเป็นค่าที่ดีสามารถจำแนกความแตกต่างของกลุ่มตัวอย่างได้ถูกต้องมาก หากค่านี้เข้าใกล้ 0 หรือมีค่าติดลบแสดงว่าจำแนกไม่ได้หรือไม่ดี และเป็นที่นิยมอย่างยิ่งของผู้สร้างและพัฒนาเครื่องมือการวัดที่มักกำหนดค่า r หรือค่า D ที่มีค่าตั้งแต่ .20 ขึ้นไป เป็นเกณฑ์การคัดเลือกข้อคำถามที่ดีสามารถนำไปใช้ได้ ซึ่งยังเป็นที่น่าสงสัยอยู่ว่าเกณฑ์การคัดเลือกข้อคำถามดังกล่าวเป็นเกณฑ์ทั่วไปที่สามารถใช้ได้กับทั้งค่า r และค่า D ในทุกกรณีเลยหรือไม่ และเกณฑ์ดังกล่าวเป็นเกณฑ์ที่เหมาะสมสำหรับการนำไปใช้ในการคัดเลือกข้อคำถามที่ดีแล้วหรือไม่ สำหรับการตอบข้อสงสัยดังกล่าวจึงได้นำเสนอข้อควรพิจารณาในการคัดเลือกข้อคำถามที่มีอำนาจการจำแนกดังต่อไปนี้

1. เกณฑ์การคัดเลือกข้อคำถามที่ดี (good item) มีคุณภาพด้านการจำแนกและมีการอ้างอิงในการนำไปใช้มากที่สุด เป็นเกณฑ์ที่นำเสนอโดย Ebel (1965: 364; 1979: 267) ซึ่งกำหนดเกณฑ์การคัดเลือกข้อคำถามที่มีค่าดัชนีการจำแนกในระดับต่าง ๆ ดังนี้

.40 ขึ้นไป	ข้อคำถามทำหน้าที่ในการจำแนกได้ดีมาก
.30 - .39	ข้อคำถามทำหน้าที่ในการจำแนกได้ดี แต่อาจต้องปรับปรุง
.20 - .29	ข้อคำถามทำหน้าที่ในการจำแนกได้เพียงเล็กน้อย ต้องพิจารณาปรับปรุง
ต่ำกว่า .19	ข้อคำถามทำหน้าที่ในการจำแนกไม่ดีต้องตัดทิ้ง หรือต้องสร้างใหม่

Ebel (1979: 273) ได้สรุปไว้ด้วยว่าข้อคำถามที่ดีควรมีค่าดัชนีการจำแนกตั้งแต่ .30 ขึ้นไป ซึ่ง Engelhart (1969: 73) ได้นำเสนอบทความที่สนับสนุนเกณฑ์การคัดเลือกข้อคำถามดังกล่าวโดยได้กล่าวว่าหากความแตกต่างของสัดส่วนของผู้ตอบถูกในกลุ่มสูงและกลุ่มต่ำมีค่าน้อยกว่า +.30 แสดงว่ามีอำนาจการจำแนกน้อยมากควรตัดข้อคำถามนั้นทิ้งหรือปรับปรุงข้อคำถามนั้น นอกจากนี้ยังมีผู้สนับสนุนการใช้เกณฑ์ดังกล่าวในการคัดเลือกข้อคำถาม อาทิ Aiken (1994: 66) , Wiersma and Jure (1990: 147), Kubiszyn and Borich (2000: 139) และChase (1974: 127) เป็นต้น สำหรับ Thorndike and Others (1991: 251) ได้

เสนอว่าข้อคำถามที่ดีมากควรมีค่าดัชนีอำนาจการจำแนกมากกว่า .50 หากข้อคำถามใดมีดัชนีอำนาจการจำแนกต่ำกว่า .20 ควรตัดทิ้งและให้ปรับปรับปรุงข้อคำถามที่มีค่าตั้งแต่ .20 ขึ้นไป และ Henressen (1971: 134 citing Alen and Yen.1979: 121) กล่าวว่าหากค่าสัมประสิทธิ์สหสัมพันธ์ไบซีเรียลระหว่างคะแนนรายข้อกับคะแนนรวมมีค่ามากกว่า .40 แสดงว่าข้อคำถามมีความเป็นเอกพันธ์สูง (the test is more homogeneous) คือวัดในสิ่งเดียวกัน และหากมีค่าสัมประสิทธิ์ต่ำกว่า .30 แสดงว่าข้อคำถามเป็นวิวิธพันธ์กันสูง (the test is more heterogeneous) กล่าวคือข้อคำถามไม่ได้วัดในสิ่งเดียวกัน

2. หากพิจารณาวิธีการทางสถิติที่ใช้ในการทดสอบ ซึ่งเรียกว่าการทดสอบนัยสำคัญ (testing of significant) จะเป็นวิธีการที่สามารถยืนยันได้ว่าคุณลักษณะทั้งสองสิ่งนั้นมีความสัมพันธ์กันจริงหรือไม่ เช่น ในกรณีการใช้ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคะแนนรายข้อกับคะแนนรวมเพื่ออธิบายค่าอำนาจการจำแนก เช่น สัมประสิทธิ์สหสัมพันธ์เพียร์สัน สัมประสิทธิ์สหสัมพันธ์พอยท์ไบซีเรียล หรือสัมประสิทธิ์สหสัมพันธ์ไบซีเรียลนั้น เป็นที่ทราบกันดีว่าใช้การทดสอบที (t-test) ในการทดสอบความมีนัยสำคัญ และหากเมื่อพิจารณาถึงค่าสัมประสิทธิ์สหสัมพันธ์ (r) ที่มีนัยสำคัญเป็นที่ทราบกันดีแล้วว่ามีความเกี่ยวข้องกับจำนวนกลุ่มตัวอย่าง ตัวอย่างดังตาราง 1

ตาราง 1 ค่าวิกฤติของสัมประสิทธิ์สหสัมพันธ์ (Furguson.1987: 523)

	ระดับนัยสำคัญ			***ระดับนัยสำคัญ***	
df	***.05***	***.01***	***df***	***.05***	***.01***
15	*.482*	*.606*	*50*	*.273*	*.354*
20	*.423*	*.537*	*60*	*.250*	*.325*
25	*.381*	*.487*	*70*	*.232*	*.303*
30	*.349*	*.449*	*80*	*.217*	*.283*
40	*.304*	*.393*	*90*	*.205*	*.267*
45	*.288*	*.372*	*100*	*.195*	*.245*

เมื่อพิจารณาค่า r ที่ระดับนัยสำคัญ .05 จะพบว่า เมื่อจำนวนกลุ่มตัวอย่างเท่ากับ 100 ค่า r จะมีนัยสำคัญเมื่อมีค่าเท่ากับ .195 เมื่อกลุ่มตัวอย่างเท่ากับ 90 ค่า r จะมีนัยสำคัญเมื่อมีค่าเท่ากับ .205 และเมื่อกลุ่มตัวอย่างเท่ากับ 40 ค่า r จะมีนัยสำคัญเมื่อมีค่าเท่ากับ .304 นั่นคือ หากกลุ่มตัวอย่างมีจำนวนน้อยลง ค่า r ที่มีนัยสำคัญจะมีค่าเพิ่มขึ้น เมื่อเป็นเช่นนี้แล้ว การนำค่าสัมประสิทธิ์สหสัมพันธ์มาใช้ในการอธิบายค่าอำนาจการจำแนกของข้อคำถามจึงมีความเกี่ยวข้องกับจำนวนกลุ่มตัวอย่าง และการกำหนดเกณฑ์การคัดเลือกข้อคำถามที่มีอำนาจการจำแนกที่ดีย่อมต้องเกี่ยวข้องกับจำนวนกลุ่มตัวอย่างที่นำข้อคำถามนั้นไปทดลองใช้ด้วย กล่าวคือ ในกรณีที่ผู้สร้างแบบทดสอบนำเครื่องมือไปทดลองใช้กับกลุ่มตัวอย่าง จำนวน 100 คน อาจพิจารณาค่า r ที่มีค่าตั้งแต่ .20 ขึ้นไป เป็นเกณฑ์การคัดเลือกข้อคำถามที่สามารถนำไปใช้ หากผู้สร้างแบบทดสอบนำเครื่องมือไปทดลองใช้กับกลุ่มตัวอย่าง จำนวน 60 คน ควรพิจารณาค่า r ที่มีค่าตั้งแต่ .25 หรือ .30 ขึ้นไป เป็นเกณฑ์การคัดเลือกข้อคำถาม ซึ่งโดยส่วนใหญ่แล้วในการทดลองใช้เพื่อ

วิเคราะห์คุณภาพรายข้อของแบบทดสอบหรือเครื่องมือการวิจัย ผู้สร้างมักกำหนดกลุ่มตัวอย่างจำนวน 30 คน ซึ่งเมื่อพิจารณาค่า r ที่มีนัยสำคัญแล้วพบว่ามีค่าเท่ากับ .349 ในกรณีเช่นนี้ควรกำหนดเกณฑ์การคัดเลือกข้อคำถามที่มีค่าอำนาจการจำแนก (r) ที่มีค่าตั้งแต่ .35 ขึ้นไป

3. กรณีการอธิบายอำนาจการจำแนกของข้อคำถามโดยใช้สัมประสิทธิ์สหสัมพันธ์แบบฟี (phi coefficient : r_Φ) ซึ่งเป็นความสัมพันธ์ของข้อมูลแบบ dichotomous เช่น การสอบผ่าน-ไม่ผ่าน หรือคะแนนกลุ่มสูง-กลุ่มต่ำ ผลการวัดอยู่ในรูปความถี่หรือจำนวนโดยแสดงในรูปตาราง 2 X 2 สำหรับการทดสอบนัยสำคัญของ r_Φ นั้นใช้การทดสอบไคสแควร์ (chi-square test) ดังนี้ (Anastasi.1990: 218)

$$r_{\Phi.05} = \frac{1.96}{\sqrt{N}}, \quad \text{N = จำนวนกลุ่มตัวอย่างทั้งหมด}$$

เช่น ถ้ามีกลุ่มตัวอย่างในกลุ่มได้คะแนนสูงจำนวน 50 คน และกลุ่มได้คะแนนต่ำจำนวน 50 คน ที่ระดับนัยสำคัญ .05 จะได้ $r_{\Phi.05} = \frac{1.96}{\sqrt{100}} = \frac{1.96}{10} = 0.196$

นั่นคือ หากค่า r_Φ ที่คำนวณได้มากกว่า 0.196 แสดงว่ามีนัยสำคัญ สำหรับตัวอย่าง ค่า r_Φ ที่มีนัยสำคัญที่ระดับ .05 ดังตาราง 2

ตาราง 2 ค่าวิกฤติของสัมประสิทธิ์สหสัมพันธ์แบบฟี ที่ระดับนัยสำคัญ .05

(อำนวย เลิศชยันตี. มปป.: 180-181)

N	$r_{\Phi'}$	*N*	$r_{\Phi'}$
20	*.44*	*55*	*.26*
25	*.39*	*60*	*.25*
30	*.36*	*70*	*.23*
35	*.33*	*80*	*.22*
40	*.31*	*90*	*.21*
45	*.29*	*100*	*.20*
50	*.28*		

จากตัวอย่างในตาราง 2 จะพบว่าเมื่อจำนวนกลุ่มตัวอย่างเท่ากับ 100 ค่า r_Φ ที่มีนัยสำคัญจะมีค่าเท่ากับ .20 หากจำนวนกลุ่มตัวอย่างเท่ากับ 40 ค่า r_Φ ที่มีนัยสำคัญจะมีค่าเท่ากับ .31 และเมื่อจำนวนกลุ่มตัวอย่างเท่ากับ 30 ค่า r_Φ ที่มีนัยสำคัญจะมีค่าเท่ากับ .36 นั่นคือ เมื่อจำนวนกลุ่มตัวอย่างลดลงค่า r_Φ ที่มีนัยสำคัญจะเพิ่มขึ้น โดยค่า r_Φ ที่มีนัยสำคัญจะสูงกว่าค่า r ในตาราง 1 เล็กน้อย เมื่อจำนวนตัวอย่างเท่ากัน เมื่อเป็นเช่นนี้ จึงพอเป็นเหตุผลสนับสนุนได้ว่าการกำหนดเกณฑ์การคัดเลือกข้อคำถามที่ดีนั้นมีความจำเป็นต้องพิจารณาจำนวนกลุ่มตัวอย่างและวิธีการที่นำมาใช้ในการวิเคราะห์คุณภาพเครื่องมือนั้นด้วย เช่น หากผู้สร้างนำเครื่องมือไปทดลองใช้กับลุ่มตัวอย่างจำนวน 30 คน และวิเคราะห์อำนาจการจำแนกโดยใช้ค่า r_Φ ควรกำหนดเกณฑ์การคัดเลือกข้อคำถามที่มีค่า r_Φ ตั้งแต่ .36 ขึ้นไป

4. กรณีการวิเคราะห์ความแตกต่างระหว่างสัดส่วนของผู้ตอบถูกในกลุ่มสูงและกลุ่มต่ำ ซึ่งเรียกว่าดัชนีการจำแนก หรือดัชนี D นั้น Aiken (1979: 821-824) ได้นำเสนอค่าวิกฤติ D ไว้สำหรับการคัดเลือกข้อคำถามโดยเขากล่าวว่า *บ่อยครั้งที่มีข้อเสนอว่าข้อคำถามที่ยอมรับได้จะมีค่าความยาก (p value) ระหว่าง .20 - .80 และมีค่า D เท่ากับ .20 หรือมากกว่า อย่างไรก็ตาม การกำหนดเกณฑ์เช่นนั้นอาจมีข้อจำกัดทั้งนี้ขึ้นอยู่กับจำนวนผู้ตอบในกลุ่มสูงและกลุ่มต่ำ และค่าความมีนัยสำคัญของ D ซึ่งจะแปรผันตามค่าความยาก..”* Aiken ได้นำเสนอค่าวิกฤติ D โดยกำหนดกลุ่มตัวอย่างในกลุ่มสูงจำนวน 10 ถึง 25 คน และกำหนดระดับนัยสำคัญที่.05 และ .10ค่าความยาก .10 ถึง .90 ในที่นี้จะนำเสนอเพียงตัวอย่างค่าวิกฤติ D ที่ระดับนัยสำคัญ .05 เท่านั้น สำหรับผู้สนใจสามารถศึกษาเพิ่มเติมจากเอกสารในรายการอ้างอิงท้ายบทความนี้

ตาราง 3 ค่าวิกฤติ D (Aiken.1979: 821)

Upper group (n_1)	*Item Difficulty Index (p)* ($\alpha = .05$)											
	.10	*.20*	*.25*	*.30*	*.40*	*.50*	*.60*	*.70*	*.75*	*.80*	*.85*	*.90*
10	-	*.35*	*.38*	*.40*	*.43*	*.44*	*.43*	*.40*	*.38*	*.35*	-	-
12	-	*.32*	*.35*	*.37*	*.39*	*.40*	*.39*	*.37*	*.35*	*.32*	*.29*	-
14	-	*.30*	*.32*	*.34*	*.36*	*.37*	*.36*	*.34*	*.32*	*.30*	*.26*	-
15	-	*.29*	*.31*	*.33*	*.35*	*.36*	*.35*	*.34*	*.33*	*.31*	*.29*	-
16	-	*.28*	*.30*	*.32*	*.34*	*.35*	*.34*	*.32*	*.30*	*.28*	*.25*	-
18	*.20*	*.26*	*.28*	*.30*	*.32*	*.33*	*.32*	*.30*	*.28*	*.26*	*.23*	*.20*
20	*.19*	*.25*	*.27*	*.28*	*.30*	*.31*	*.30*	*.28*	*.27*	*.25*	*.22*	*.19*
22	*.18*	*.24*	*.26*	*.27*	*.29*	*.30*	*.29*	*.27*	*.26*	*.24*	*.21*	*.18*
24	*.17*	*.23*	*.24*	*.26*	*.28*	*.28*	*.28*	*.26*	*.24*	*.23*	*.20*	*.17*
25	*.17*	*.22*	*.24*	*.25*	*.27*	*.28*	*.27*	*.25*	*.24*	*.22*	*.20*	*.17*
D_{max}	*.20*	*.40*	*.50*	*.60*	*.80*	*1.0*	*.80*	*.60*	*.50*	*.40*	*.30*	*.20*

จากตาราง 3 มีสิ่งที่ควรพิจารณา 2 ประการ ได้แก่ 1) ค่าดัชนี D ที่มีนัยสำคัญจะมีความสัมพันธ์กับค่าความยาก กล่าวคือ ค่าดัชนี D ที่มีนัยสำคัญจะมีค่าสูงขึ้นหากข้อคำถามมีค่าความยากเข้าใกล้ .50 และค่าดัชนี D จะมีค่าสูงที่สุดเมื่อค่าความยากเท่ากับ .50 ไม่ว่ากลุ่มตัวอย่างจะมีจำนวนเท่าใดก็ตาม และ 2) ค่าดัชนี D ที่มีนัยสำคัญจะมีความสัมพันธ์กับจำนวนกลุ่มตัวอย่าง กล่าวคือ เมื่อกลุ่มตัวอย่างเพิ่มมากขึ้น ค่าดัชนี D ที่มีนัยสำคัญจะลดลงไม่ว่าข้อคำถามจะมีค่าความยากเท่าใดก็ตาม อาทิเช่น ที่ค่าความยาก .50 หากกลุ่มตัวอย่างเท่ากับ 30 แบ่งเป็นกลุ่มสูง-ต่ำกลุ่มละ 15 ค่าดัชนี D ที่มีนัยสำคัญจะเท่ากับ .36 หรือหากกลุ่มตัวอย่างเท่ากับ 50 แบ่งเป็นกลุ่มสูง-ต่ำ กลุ่มละ 25 ค่าดัชนี D ที่มีนัยสำคัญจะเท่ากับ .28 นั่น

หมายความกรณีที่กลุ่มตัวอย่างมีจำนวน 30 หรือ 50 คน และข้อคำถามมีค่าความยากปานกลางผู้สร้างควรกำหนดเกณฑ์การคัดเลือกข้อคำถามที่มีค่าดัชนีอำนาจการจำแนกตั้งแต่ .36 หรือ .28 ขึ้นไป

อย่างไรก็ตาม ผู้สร้างข้อสอบพึงตระหนักเกี่ยวกับความเข้าใจที่ว่า เมื่อข้อคำถามมีค่าความยากเข้าใกล้ .50 หรือเท่ากับ .50 แล้ว ข้อคำถามนั้นจะมีอำนาจการจำแนกสูงนั้น อาจไม่เป็นจริงในทุกกรณี ดังที่ Murphy and Davidshofer (1994: 162) ได้กล่าวว่า *"... a p value near .50 does not guarantee that an item will be a good discriminator"*

5. กรณีการวิเคราะห์คุณภาพข้อสอบแบบอิงเกณฑ์ (criterion reference test) Brennan (1972 : unpaged; อ้างถึงใน โกวิท ประวาลพฤกษ์ และสมศักดิ์ สินธุระเวชญ์. 2527 : 286-287) ได้นำเสนอวิธีการวิเคราะห์ดัชนีการจำแนกของคำถามไว้โดยพัฒนาจาก U-L Index และเรียกว่าดัชนีการจำแนก B (discrimination index B) โดย Brennan เชื่อว่าผู้สร้างข้อสอบแบบอิงเกณฑ์นั้นคาดหวังว่านักเรียนส่วนใหญ่จะทำข้อสอบถูกมาก ดังนั้นการกระจายของคะแนนจึงเบ้ไปทางลบ จึงไม่มีความจำเป็นที่ต้องให้นักเรียนทั้งสองกลุ่มเท่ากัน เพราะควรให้อิสระผู้สร้างในการกำหนดจุดตัดของคะแนน (cut of score) โดยพิจารณาเนื้อหาของข้อสอบ จำนวนประชากรของข้อสอบ และความคาดหวังที่นักเรียนจะสามารถปฏิบัติได้ Brennan จึงได้นำเสนอการทดสอบความมีนัยสำคัญของดัชนี B (testing the significance of B) โดยนำเสนอค่าความมีนัยสำคัญของดัชนี B เมื่อ n_1 , n_2 มีจำนวนระหว่าง 10 ถึง 30 และทดสอบนัยสำคัญที่ระดับ .01 และ .05 โดยค่าความมีนัยสำคัญของดัชนี B จะมีความสัมพันธ์กับจำนวนกลุ่มตัวอย่าง กล่าวคือ เมื่อจำนวนกลุ่มตัวอย่างเพิ่มขึ้นค่าความมีนัยสำคัญจะลดลงและเมื่อพิจารณาตารางแสดงค่าความมีนัยสำคัญแล้วจะพบว่าค่าความมีนัยสำคัญต่ำสุดของดัชนี B คือ .250 ในกรณีที่ n_1, n_2 เท่ากับ 28 และ 30 เท่านั้น ดังนั้น หากผู้สร้างข้อสอบต้องการวิเคราะห์คุณภาพรายข้อด้านอำนาจการจำแนกตามแนวคิดของ Brennan ควรต้องพิจารณาคัดเลือกข้อสอบที่มีคุณภาพด้านการจำแนกตามค่าความมีนัยสำคัญของดัชนี B ซึ่งผันแปรตามจำนวนกลุ่มตัวอย่าง และหากผู้สร้างกำหนดเกณฑ์การคัดเลือกข้อคำถามที่มีค่าดัชนี B ตั้งแต่ .20 ขึ้นไปอาจได้ข้อคำถามที่ไม่สามารถจำแนกผู้เข้าสอบได้ หรือได้ข้อคำถามที่วัดได้ไม่ตรงตามความมุ่งหมายที่ต้องการ แสดงตัวอย่างค่าความมีนัยสำคัญของดัชนี B ดังตาราง 4

ตาราง 4 ค่าวิกฤติของดัชนี B (โกวิท ประวาลพฤกษ์ และสมศักดิ์ สินธุระเวชญ์.2527: 287-289)

n_1	n_2	.05	.01	n_1	n_2	.05	.01	n_1	n_2	.05	.01
10	**10**	.500	.700	**17**	**17**	.412	.471	**24**	**24**	.333	.417
10	**30**	.400	.500	**17**	**30**	.300	.396	**24**	**30**	.275	.358
11	**11**	.545	.636	**18**	**18**	.389	.500	**25**	**25**	.320	.400
11	**30**	.355	.455	**18**	**30**	.300	.389	**25**	**30**	.273	.353
12	**12**	.500	.583	**19**	**19**	.368	.474	**26**	**26**	.308	.385
12	**30**	.350	.450	**19**	**30**	.288	.371	**26**	**30**	.262	.346
13	**13**	.462	.538	**20**	**20**	.350	.450	**27**	**27**	.296	.370
13	**30**	.328	.426	**20**	**30**	.300	.383	**27**	**30**	.267	.341
14	**14**	.429	.571	**21**	**21**	.333	.429	**28**	**28**	.286	.393
14	**30**	.319	.419	**21**	**30**	.286	.367	**28**	**30**	**.250**	.345
15	**15**	.400	.533	**22**	**22**	.318	.409	**29**	**29**	.276	.379
15	**30**	.333	.433	**22**	**30**	.276	.364	**29**	**30**	.255	.325
16	**16**	.438	.500	**23**	**23**	.348	.435	**30**	**30**	.300	.367
16	**30**	.300	.392	**23**	**30**	.275	.355				

6. การวิเคราะห์คุณภาพรายข้อกรณีแบ่งกลุ่มสูงกลุ่มต่ำโดยส่วนใหญ่ใช้เทคนิค 27% แต่เดิมวิธีการที่เป็นที่นิยมอย่างยิ่งคือการใช้ตารางการวิเคราะห์ข้อสอบของ Chung Teh Fan ซึ่งเป็นตารางแสดงระดับความยากและอำนาจการจำแนกของข้อสอบจากผู้ตอบถูกในกลุ่มสูง 27% และกลุ่มต่ำ 27% โดยถือหลักประชากรมีการแจกแจงปกติ นั่นคือ การวิเคราะห์ข้อสอบโดยใช้ตารางสำเร็จรูปของ Chung Teh Fan นักเรียนในกลุ่มสูงและกลุ่มต่ำควรมีจำนวนกลุ่มละ 100 แสดงว่ามีกลุ่มผู้เข้าสอบทั้งสิ้น 370 คน แต่หากมีผู้เข้าสอบจำนวนน้อยกว่านี้ควรใช้เทคนิค 50% จะทำให้ค่าอำนาจการจำแนกมีความเชื่อมั่นสูง อย่างไรก็ตาม Chung Teh Fan ไม่ได้กำหนดจำนวนกลุ่มตัวอย่างที่แน่นอนในการวิเคราะห์คุณภาพข้อสอบโดยใช้ตารางสำเร็จรูป แต่ได้ให้ข้อเสนอแนะว่าประชากรต้องมีการแจกแจงปกติ สำหรับจำนวนกลุ่มตัวอย่างที่นำมาวิเคราะห์คุณภาพรายข้อนั้น Nunnally (1964: 194) ได้เสนอไว้ว่าในการวิเคราะห์คุณภาพรายข้อต้องเป็นไปตาม ***"The forty rule"*** นั่นคือ จำนวนกลุ่มตัวอย่างที่นำมาวิเคราะห์ไม่ควรต่ำกว่า 40 คน จึงจะทำให้ได้ผลการวิเคราะห์ที่มีความเชื่อมั่น ซึ่งเมื่อพิจารณาค่า r ที่มีนัยสำคัญเมื่อจำนวนกลุ่มตัวอย่างเท่ากับ 40 แล้วค่า r = .304, r_{Φ} =.36 , D=.29-.36 (เมื่อ p=.20-.80) Aiken (1994: 66) สนับสนุนว่าดัชนีอำนาจการจำแนกที่ยอมรับได้ (acceptable) ควรมีค่าเท่ากับ .30 หรือมากกว่า แต่บางกรณีดัชนีอำนาจการจำแนกที่น้อยกว่า .30 ก็สามารถยอมรับได้หากจำนวนกลุ่มสูงและกลุ่มต่ำที่นำมาเปรียบเทียบมีขนาดใหญ่พอ

เช่นเดียวกับ Ebel (1979: 263) ที่มีความเห็นว่า ไม่สามารถบอกได้ว่าค่าดัชนีอำนาจการจำแนกควรเป็นเท่าไหร่จึงจะเป็นค่าที่ดีที่สุด เพราะขึ้นอยู่กับจำนวนกลุ่มตัวอย่าง

เมื่อเป็นเช่นนี้แล้วผู้สร้างควรตระหนักและมีความรอบคอบในการกำหนดเกณฑ์การคัดเลือกข้อคำถามที่ดีโดยอาจต้องพิจารณาทั้งค่าความยากของคำถามแต่ละข้อและจำนวนกลุ่มตัวอย่างผู้เข้าสอบ เพื่อให้ได้ข้อสอบที่มีคุณภาพสามารถจำแนกความแตกต่างของผู้เข้าสอบตามคุณลักษณะที่ต้องการวัดได้อย่างถูกต้อง หากผู้สร้างละเลยความมีนัยสำคัญของอำนาจการจำแนกจะทำให้ข้อคำถามที่สร้างขึ้นไม่สามารถจำแนกความแตกต่างของผู้เข้าสอบได้จริง นั่นคือ ข้อคำถามไม่สามารถวัดได้ตรงตามคุณลักษณะที่ต้องการ เช่น หากกลุ่มตัวอย่างมีจำนวน 100 ค่าความมีนัยสำคัญจะเท่ากับ .195 กรณีนี้อาจพิจารณาคัดเลือกข้อคำถามที่มีค่าอำนาจการจำแนกตั้งแต่ .20 ขึ้นไป หากกลุ่มตัวอย่างจำนวน 40 ค่าความมีนัยสำคัญจะเท่ากับ .304 กรณีเช่นนี้ควรคัดเลือกข้อคำถามที่มีอำนาจการจำแนกตั้งแต่ .30 ขึ้นไป หรือหากกลุ่มตัวอย่างมีจำนวน 30 ดังที่ผู้สร้างส่วนใหญ่นิยม ค่าความมีนัยสำคัญจะเท่ากับ .349 กรณีนี้ควรคัดเลือกข้อคำถามที่มีอำนาจการจำแนกตั้งแต่ .35 ขึ้นไป และในบางกรณีผู้สร้างต้องคำนึงถึงค่าความยากของข้อคำถามด้วย ทั้งนี้เพราะอำนาจการจำแนกที่มีนัยสำคัญจะมีค่าสูงขึ้นหากข้อคำถามมีค่าความยากเข้าใกล้ .50

สรุป

อำนาจการจำแนกของข้อสอบหรือข้อคำถามเป็นสารสนเทศเกี่ยวกับคุณภาพของข้อคำถามที่สามารถบ่งชี้ความแตกต่างของกลุ่มตัวอย่างได้ตามคุณลักษณะที่ต้องการ ข้อคำถามที่มีความสามารถในการจำแนกได้ดี แสดงว่าข้อคำถามนั้นสามารถวัดได้ตรงตามคุณลักษณะที่ต้องการ ตามทฤษฎีการทดสอบแบบดั้งเดิมมีวิธีการอธิบายอำนาจการจำแนกของข้อคำถามหลายวิธี ทั้งการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์ประเภทต่าง ๆ และวิธีอย่างง่ายโดยการหาความแตกต่างระหว่างสัดส่วนผู้ตอบถูกในกลุ่มสูงและกลุ่มต่ำ หรือการใช้ตารางการวิเคราะห์สำเร็จรูปที่มีผู้นำเสนอไว้หลายตารางโดยเฉพาะตารางสำเร็จรูปที่เป็นที่นิยมของ Chung Teh Fan ทุกวิธี ล้วนมีวัตถุประสงค์เพื่อนำเสนอสารสนเทศที่ดีที่สุดเกี่ยวกับข้อคำถาม ขึ้นอยู่กับผู้สร้างว่าจะพิจาณาเลือกวิธีการใดที่เหมาะสมจะนำไปใช้อธิบายคุณภาพของข้อคำถามได้มากที่สุด สำหรับการพิจารณาคัดเลือกข้อคำถามนั้น การทดสอบนัยสำคัญจะเป็นวิธีการหนึ่งที่จะช่วยให้การกำหนดเกณฑ์การคัดเลือกข้อคำถามมีความเชื่อมั่นสูง ซึ่งหากพิจารณาค่าความมีนัยสำคัญของอำนาจการจำแนกประเภทต่าง ๆ ที่ได้นำเสนอไว้แล้ว จะพบว่ามีความเกี่ยวข้องกับจำนวนกลุ่มตัวอย่างทั้งสิ้น กล่าวคือ หากกลุ่มตัวอย่างมีจำนวนมากอำนาจการจำแนกที่มีนัยสำคัญจะมีค่าน้อย หากกลุ่มตัวอย่างมีจำนวนน้อยอำนาจการจำแนกที่มีนัยสำคัญจะมีค่ามากผู้สร้างเครื่องมือจึงควรมีความรอบคอบในการคัดเลือกข้อคำถามที่มีค่าอำนาจการจำแนกที่เหมาะสมเพื่อให้เครื่องมือมีคุณภาพสามารถจำแนกความแตกต่างของกลุ่มตัวอย่างตามคุณลักษณะที่ต้องการได้อย่างถูกต้อง

****

รายการอ้างอิง

โกวิท ประวาลพฤกษ์ และสมศักดิ์ สินธุระเวชญ์. (2527). **การประเมินในชั้นเรียน.** พิมพ์ครั้งที่ 2. กรุงเทพมหานคร: ไทยวัฒนาพานิช.

อำนวย เลิศชยันตี. (มปป). **การสร้างข้อสอบและการประเมินผลการศึกษา.** กรุงเทพมหานคร : อำนวยการพิมพ์.

Aiken R. Lewis. (1979). *"Relationships between the item difficulty and discrimination indexes"* in **Educational and Psychological Measurement.** Vol 39: 821-824.

Aiken R. Lewis. (1994). **Psychological testing and assessment**.8th ed. Massachusetts: A Division of Simon & Schuster. Inc.

Allen J. Mary and Yen M. Wendy. (1979). **Introduction to measurement theory**. California: Wadsworth, Inc.

Anastasi Anne. (1990). **Psychological testing**. 6thed. New Yok: Macmillan Publishing Company.

Chase I. Clinton. (1974). **Measurement for educational evaluation**. Massachusetts:Addison-Wesley Publishing Company, Inc.

Ebel L. Robert. (1965). **Mearsuring educational achievement.** Englewood Cliffs, N.J.: Prentice Hall.

Ebel L. Robert. (1972). **Essentials of Educational Measurement**. 3rded. New Jersey: Prentice-Hal Inc.

Engelhart D. Max. (1965). *"A comparison of several item discrimination indices"*. in **Journal of Educational Measurement.** Vol 2: 69-76.

Ferguson A. George. (1981). **Statistical analysis in psychology and education**. 5th ed. McGraw-Hill Book Co.

Findley G. Warren. (1956). *"A rationale for evaluation of item discrimination statistics"* in **Educational and Psychological Measurement**. Vol 16: 79-84.

Kubiszyn Tom and Borich Gary. (1984). **Educational testing and measurement**. 6th ed. Glenview Illinois; Scot Forsman and Company,Inc.

Murphy R. Kevin and Davidshofer O. Charles. (1994). **Psychological testing: principle and applications**. New Jersey: Prentice-Hall.Inc.

Nunnally C. Jum. (1972). **Educational and psychological measurement and evaluation**. New Yok : McGraw-Hill.

Pemberton A. Johnson. (1951). *"Note on a suggested index of item validity: The U-L Index"* in **Journal of Educational Psychology**. Vol 62: 499-504.

Thorndike M. Robert and others. (1991). **Measurement and evaluation in psychology and education.** 5th ed. New Yok: Macmilland**.**

Wiersma William and Jurs G. Stephen. (1990). **Educational measurement and testing**. 2nd ed. Massachusetts: Addison-Wesley Publishing Company, Inc.

****