ผลของการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออกต่อคะแนนการสอบ และความเชื่อมั่นของแบบทดสอบ: คณะพยาบาลศาสตร์ มหาวิทยาลัยสวนดุสิต

THE EFFECTS ON TEST SCORES AND THE RELIABILITY OF THE TEST AFTER EXCLUDING THE ITEMS WITH NEGATIVE DISCRIMINATION POWER: FACULTY OF NURSING, SUAN DUSIT UNIVERSITY

เบ็ญจา เตากล่ำ ปร.ด. (Benja Taoklam, Ph.D)[1]
ประกาย จิโรจน์กุล ปร.ด. (Pragai Jirojanakul, Ph.D)[2]
สวงค์ บุญปลูก (Sawong Boonprook)[3]

บทคัดย่อ

การวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อ ศึกษาผลของการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออกต่อคะแนนการสอบ การเปลี่ยนแปลงของลำดับที่ของคะแนนสอบ และความเชื่อมั่นของแบบทดสอบ กลุ่มตัวอย่างเป็นแบบทดสอบแบบเลือกตอบ 4 ตัวเลือก ที่ใช้ในการสอบกลางภาค และปลายภาคของนักศึกษา คณะพยาบาลศาสตร์ มหาวิทยาลัยสวนดุสิต ภาคเรียนที่ 1 ปีการศึกษา 2555 จาก 11 รายวิชา จำนวน 26 ฉบับ เครื่องมือที่ใช้ในงานวิจัยเป็นแบบบันทึกข้อมูล มีลักษณะเป็นตารางสำหรับบันทึกคะแนนของนักศึกษารายบุคคล และโปรแกรมสำเร็จรูปสำหรับการวิเคราะห์ข้อสอบรายข้อ สถิติที่ใช้ประกอบด้วยสถิติพื้นฐาน ได้แก่ การแจกแจงความถี่ ร้อยละ ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน สถิติที่ใช้ในการทดสอบสมมุติฐาน ได้แก่ Paired t-test (one-tailed), Sign Ranked test และ Pearson's Product Moment Correlation ผลการวิเคราะห์ข้อมูลจากข้อสอบทั้งหมด 1,970 ข้อ จากจำนวนนักศึกษาที่เข้าสอบทั้งหมด 332 คนพบว่า

1) การเปรียบเทียบคะแนนที่นักศึกษาได้รับ ก่อน และหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก พบว่า คะแนนของนักศึกษาจากแบบทดสอบ จำนวน 21 ฉบับ (ร้อยละ 80.77) สูงขึ้นอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

2) การเปลี่ยนแปลงของลำดับที่ของคะแนนสอบ ก่อน และหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก พบว่า ลำดับที่ของคะแนนสอบ จากจำนวนแบบทดสอบ 21 ฉบับ (ร้อยละ 80.77) มีการเปลี่ยนแปลงอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

3) ผลการเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก พบว่า ค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับ หลังการตัดข้อสอบที่มีค่าอำนาจจำแนก

[1] *คณบดีคณะพยาบาลศาสตร์ มหาวิทยาลัยสวนดุสิต E-mail: benja_tao@hotmail.com*
[2] *รองคณบดีฝ่ายวิชาการ คณะพยาบาลศาสตร์ มหาวิทยาลัยสวนดุสิต E-mail: pragaij@gmail.com*
[3] *ผู้ช่วยอธิการบดี ฝ่ายวิชาการ มหาวิทยาลัยสวนดุสิต E-mail: sawong_boo@dusit.ac.th*

ติดลบออก มีค่าเพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติ ที่ระดับ 0.05 และ ผลการทดสอบความสัมพันธ์ระหว่างร้อยละของจำนวนข้อสอบที่ค่าอำนาจจำแนกติดลบ กับค่าความเชื่อมั่นที่เพิ่มขึ้น มีความสัมพันธ์กันในเชิงบวกอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

คำสำคัญ: ค่าอำนาจจำแนก คะแนนสอบ ค่าความเชื่อมั่นของแบบทดสอบ

Abstract

The objectives of this study were to examine the effects of excluding test items with negative discrimination power on the test scores and ranks students obtained, and the reliability of the test. The samples were test items, with multiple choices from 11 subjects, used in the mid-term and final examinations of the first semester in academic year 2012. The test items were analyzed for the index of difficulty, index of discrimination and reliability by a computer program. The objectives of the study were later tested by Paired t-test (one-tailed), Sign Ranked test and Pearson's Product Moment Correlation. The total of 1,969 items from 26 sets of test were analyzed for its quality.

It was found that 1) After excluding the items with negative discrimination power, students had obtained significant higher scores at α=0.05 in 21 sets of test (80.77%) 2) In the same way, the Signed Rank Test showed that student's ranks were significantly changed in 21 sets of test (80.77%). 3) The one-tailed Paired t-test revealed that the reliability of the test was significantly increased at α=0.05 after excluding the items with negative discrimination power.

Keywords: quality of test items, negative discrimination power, test score, reliability of the test

ความเป็นมาและความสำคัญ

การวัดและประเมินผลเป็นองค์ประกอบที่สำคัญของการจัดการศึกษาที่มีคุณภาพ และจุดมุ่งหมายของการวัดผลทางการศึกษา เพื่อให้ได้มาซึ่งข้อมูลที่เที่ยงตรง และเชื่อถือได้ ที่แสดงถึงความสามารถที่แท้จริง (True ability) ของผู้เรียน (อุทุมพร จามรมาน, 2535) เครื่องมือสำคัญที่ใช้ ในการประเมินว่าผู้เรียนมีความรู้ ตามที่หลักสูตรคาดหวังหรือไม่ คือ ข้อสอบ ดังนั้น ข้อสอบที่มีคุณภาพ จึงเป็นปัจจัยสำคัญของการจัดการเรียนการสอน คุณสมบัติที่สำคัญของข้อสอบที่ดีควรมีค่าความยากง่ายที่เหมาะสมและมีค่าอำนาจจำแนกสูง ค่าเฉลี่ยความยากง่ายของข้อสอบทั้งฉบับควรมีค่าประมาณ 0.50 แต่อย่างไรก็ตามในการคัดเลือกข้อสอบจะต้องคำนึงถึงความสมดุลระหว่างข้อสอบที่มีสถิติเหมาะสม(อิงกลุ่ม) กับข้อสอบที่วัดครอบคลุมจุดประสงค์และเนื้อหาที่ต้องการด้วย (อิงเกณฑ์) บางครั้งอาจมีความจำเป็นที่จะต้องใช้ข้อสอบที่มีอำนาจการจำแนกไม่สูงนัก เพื่อให้มีข้อสอบที่วัดครอบคลุมเนื้อหา (อิงเกณฑ์) ที่ต้องการ (ปวรส บุตะเขียว, 2555) ในการวิเคราะห์ข้อสอบ ถ้าพบว่า ข้อสอบไม่มีคุณภาพควรกลับมาพิจารณาว่า เกิดจากสาเหตุใด ข้อคำถามมีความเป็นปรนัยหรือไม่ รวมทั้งต้องพิจารณาที่ตัวเลือก

ด้วยว่า เป็นไปตามหลักการเขียนข้อสอบที่ดีหรือไม่ และดำเนินการปรับปรุงข้อสอบ รวมทั้งนำไปทดลองใช้ (Tryout) อีกครั้ง เพื่อวิเคราะห์ประสิทธิภาพของข้อสอบต่อไป ในส่วนของตัวลวง ประสิทธิภาพของตัวลวง เมื่อสร้างข้อสอบอิงเกณฑ์แบบหลายตัวเลือก (Multiple choices) ผู้ที่เลือกตัวลวง ถือว่า ตอบผิด ซึ่งแสดงให้เห็นว่า ผู้นั้นยังไม่สัมฤทธิ์ผลตามเป้าหมายของการวัด ในการวิเคราะห์ประสิทธิภาพตัวลวงทำโดยการตรวจสอบความถี่ของผู้ที่เลือกตัวลวงนั้น ๆ ถ้าตัวลวงใดมีผู้เลือกในสัดส่วนที่สูง ถือว่าเป็นตัวลวงที่ใช้ได้ แต่ถ้าตัวลวงใดมีผู้เลือกน้อย แสดงว่าตัวลวงนั้นขาดประสิทธิภาพ สมควรที่จะต้องมีการปรับปรุงแก้ไขให้ดีขึ้น (ศิริชัย กาญจนวาสี, 2552) และตัวลวงที่จัดว่าเป็นตัวลวงที่ดีนั้น ผู้ที่เรียนอ่อนควรจะเลือกตอบมากกว่าผู้ที่มีผลการเรียนดี

การจัดการเรียนการสอนในหลักสูตรพยาบาลศาสตรบัณฑิต ของมหาวิทยาลัยสวนดุสิต ใช้ระเบียบการวัดและประเมินผลของมหาวิทยาลัย ควบคู่กับแนวทางการวัดและประเมินผลที่คณะได้กำหนดขึ้น คือ ใช้การประเมินผลแบบอิงเกณฑ์ในรายวิชาภาคปฏิบัติ ส่วนรายวิชาภาคทฤษฎีใช้การประเมินผลแบบอิงเกณฑ์ควบคู่กับอิงกลุ่ม กล่าวคือ นักศึกษาต้องมีคะแนนดิบถึงเกณฑ์ 60% จึงจะได้รับเกรด C นอกจากนั้น ยังได้มีการพัฒนามาตรการในการช่วยเหลือผู้เรียนที่เรียนอ่อน เพื่อพัฒนานักศึกษาให้มีความรู้ความสามารถตามเกณฑ์ที่หลักสูตรกำหนด และได้พัฒนาเต็มตามศักยภาพของตน

คณะพยาบาลศาสตร์ มีระบบในการควบคุมคุณภาพของข้อสอบ โดยการกำหนดให้มีการจัดทำตารางวิเคราะห์ข้อสอบ (Test blueprint) พร้อมกับรายละเอียดของรายวิชา (มคอ.3) ที่ผ่านความเห็นชอบของทีมผู้สอน และคณะกรรมการบริหารหลักสูตรทุกวิชาที่ดำเนินการสอนโดยอาจารย์ของคณะพยาบาลศาสตร์ และมีการวิพากษ์ข้อสอบโดยคณะกรรมการ ที่ประกอบไปด้วย ทีมผู้สอน ประธานสาขาวิชา และผู้แทนจากคณะกรรมการบริหารหลักสูตร นอกจากนั้น หลังการสอบจะต้องมีการส่งข้อสอบไปวิเคราะห์ด้วยโปรแกรมสำเร็จรูปที่ สำนักส่งเสริมวิชาการและงานทะเบียนของมหาวิทยาลัย และนำผลมาประกอบการพิจารณาในการตัดเกรดของทีมผู้สอน การทวนสอบมาตรฐานผลสัมฤทธิ์ของนักศึกษา และการพิจารณาเกรดโดยคณะกรรมการบริหารหลักสูตรก่อนส่งให้คณบดี คณะพยาบาลศาสตร์ให้ความเห็นชอบและส่งเกรดให้สำนักส่งเสริมวิชาการและงานทะเบียนของมหาวิทยาลัย เพื่อประกาศผลการเรียนให้นักศึกษาทราบผ่านระบบบริหารการศึกษาของมหาวิทยาลัย

จากการดำเนินงานตามระบบดังกล่าว พบว่า ผลการวิเคราะห์ข้อสอบ ยังพบข้อสอบที่มีค่าอำนาจจำแนกติดลบ อยู่ระหว่าง ร้อยละ 1–13 ซึ่งแสดงว่าคุณภาพของข้อสอบข้อนั้น ๆ ไม่สามารถจำแนกผู้เรียนเก่งและอ่อนได้ ผู้เรียนอ่อนส่วนมากตอบข้อนั้นถูก แต่ผู้เรียนเก่งส่วนมากไม่ได้คะแนนจากข้อนั้น การนำคะแนนจากข้อสอบที่มีค่าอำนาจจำแนกติดลบไปประเมินผลการศึกษาของนักศึกษาย่อมส่งผลกระทบต่อนักศึกษาโดยตรง ผู้วิจัยจึงพิจารณาเห็นความสำคัญของการศึกษาวิจัยว่า ข้อสอบที่มีค่าอำนาจจำแนกติดลบจะส่งผลต่อคะแนนการสอบ และลำดับที่ของคะแนนที่นักศึกษาได้รับอย่างไร และมีผลต่อความเชื่อมั่นของแบบทดสอบอย่างไร โดยการศึกษาวิจัยนี้ประกอบไปด้วยโครงการวิจัยย่อย 11 โครงการ ได้เลือกกลุ่มตัวอย่างจากแบบทดสอบที่มาจากหลากหลายวิชาในทุกชั้นปีในภาคการศึกษาเดียวกัน จำนวน 11 รายวิชา และดำเนินการภายหลังจากที่ได้ดำเนินการจัดการเรียนการสอน และวัดประเมินผลตามปกติเสร็จเรียบร้อยแล้ว เพื่อไม่ให้เกิดผลกระทบต่อนักศึกษา ผลที่ได้จากการศึกษาครั้งนี้จะเป็นองค์ความรู้ที่เป็นหลักฐาน

เชิงประจักษ์ ที่นำมาสู่ข้อสรุป หรือข้อเสนอแนะเชิงนโยบายในด้านการวัดและประเมินผลการจัดการเรียนการสอนในแต่ละรายวิชาของหลักสูตรพยาบาลศาสตรบัณฑิตต่อไป

วัตถุประสงค์ของการวิจัย

เพื่อศึกษาผลของการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก ต่อคะแนนการสอบ และความเชื่อมั่นของแบบทดสอบทั้งฉบับ โดยมีวัตถุประสงค์เฉพาะเพื่อศึกษา

1. คุณภาพของข้อสอบที่ใช้สอบกับนักศึกษาภาคการศึกษาที่ 1 ปีการศึกษา 2555 ของคณะพยาบาลศาสตร์ จำนวน 11 รายวิชา

2. การเปลี่ยนแปลงของคะแนนสอบรายวิชาและอันดับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

3. การเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

ขอบเขตการวิจัย

ประชากร คือ ข้อสอบ และคะแนนการสอบรายวิชาต่าง ๆของนักศึกษา หลักสูตรพยาบาลศาสตรบัณฑิต

ประชากรทั้งหมดที่เข้าถึงได้ (Accessible population) คือ ข้อสอบ และคะแนนการสอบรายวิชาภาคทฤษฎีของนักศึกษา หลักสูตรพยาบาลศาสตรบัณฑิต มหาวิทยาลัยสวนดุสิต ในภาคการศึกษาที่ 1 ปีการศึกษา 2555 จำนวนทั้งหมด 22 วิชา

กลุ่มตัวอย่าง เลือกจากข้อสอบ และคะแนนการสอบของรายวิชาภาคทฤษฎี ที่มีการวัดและประเมินผลโดยใช้ข้อสอบแบบเลือกตอบ (Multiple choices) 4 ตัวเลือก ในหมวดวิชาเฉพาะ ทั้งกลุ่มวิชาพื้นฐานวิชาชีพและกลุ่มวิชาชีพ และ หมวดวิชาเลือกเสรี ได้จำนวน 11 รายวิชา

คำจำกัดความที่ใช้ในงานวิจัย

1. คุณภาพของข้อสอบ ประกอบด้วย คุณภาพของข้อสอบรายข้อ และ คุณภาพของแบบทดสอบทั้งฉบับ

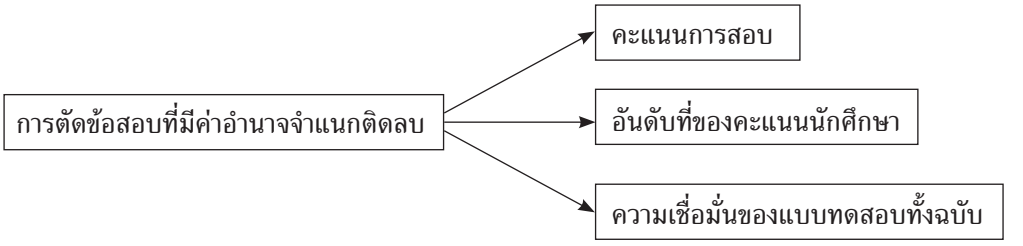
2. คุณภาพของข้อสอบรายข้อพิจารณาจาก ค่าความยากง่าย และ ค่าอำนาจจำแนก โดยข้อสอบที่มีค่าความยากง่ายเหมาะสม (p) อยู่ระหว่าง 0.20-0.80 และ ข้อสอบที่มีค่าอำนาจจำแนก (r) ที่เหมาะสม คือ ≥0.20 ข้อสอบที่ค่าอำนาจจำแนกติดลบ หมายถึง ข้อสอบข้อนั้นจำแนกผู้เรียนกลับด้าน แสดงว่าเป็นข้อสอบที่ไม่ดี เนื่องจาก นักศึกษาที่เรียนเก่ง ส่วนใหญ่ตอบผิด แต่นักศึกษาที่เรียนอ่อนส่วนใหญ่ตอบถูก

3. คุณภาพของแบบทดสอบทั้งฉบับ ในการวิจัยครั้งนี้ หมายถึง ค่าความเชื่อมั่นของแบบทดสอบ แสดงถึง ความคงเส้นคงวาของคะแนนที่ได้จากการใช้แบบทดสอบชุดนั้น สามารถคำนวณโดยใช้ค่า สัมประสิทธิ์ของครอนบัค (Cronbach's Alpha) ใช้กับแบบทดสอบแบบเลือกตอบ ที่มีข้อที่ถูกเพียงข้อเดียว (ตอบถูกได้คะแนน 1 ตอบผิด ได้คะแนน 0) และจะได้ค่าเท่ากับการใช้สูตร KR-20 (Kuder-Richardson Formula 20) ค่าสัมประสิทธิ์ความเชื่อมั่น มีค่าระหว่าง 0-1.00 ค่าสัมประสิทธิ์ความเชื่อมั่นที่มีค่าเข้าใกล้ 1 แสดงว่ายิ่งดี และ หาก <.50 แสดงว่าแบบทดสอบฉบับนั้น ไม่น่าเชื่อถือ จำเป็นต้องปรับปรุง และไม่ควรใช้เป็นหลักในการตัดสินเกรด

4. การตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก หมายถึง การคิดคะแนนที่นักศึกษาได้รับจากข้อสอบทั้งฉบับใหม่ ภายหลังทราบผลการวิเคราะห์ข้อสอบ โดยการไม่ตรวจให้คะแนนข้อสอบที่มีค่าอำนาจจำแนกเป็นลบ และปรับคะแนนรวมของจำนวนข้อสอบที่เหลือให้เป็นร้อยละ 100 เหมือนเดิม

กรอบแนวคิดในการวิจัย

จากแผนภูมิกรอบแนวคิดในการทำวิจัยแสดงให้เห็นถึงตัวแปรที่ใช้ในการศึกษา โดย ตัวแปรอิสระ คือ การตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบ ตัวแปรตาม คือ คะแนนการสอบ อันดับที่ของคะแนนนักศึกษา และความเชื่อมั่นของแบบทดสอบทั้งฉบับ



แผนภูมิที่ 1 กรอบแนวคิดในการวิจัย

เครื่องมือในการวิจัยและการตรวจสอบคุณภาพเครื่องมือ

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ คือ แบบบันทึกข้อมูลมีลักษณะเป็นตารางสำหรับสำหรับบันทึกคะแนนของนักศึกษารายบุคคล ที่ผู้วิจัยใช้สำหรับเก็บข้อมูลแต่ละรายวิชา และโปรแกรมการวิเคราะห์ข้อสอบรายข้อ (Items Analysis) ที่มหาวิทยาลัยใช้ในการวิเคราะห์ข้อสอบ ดำเนินการโดยสำนักส่งเสริมวิชาการและงานทะเบียน

การเก็บรวบรวมข้อมูล

1. เชิญชวนอาจารย์ที่เป็นผู้รับผิดชอบวิชา ในภาคการศึกษาที่ 1 ปีการศึกษา 2555 ที่มีการวัดประเมินผลโดยใช้ข้อสอบแบบหลายตัวเลือก (Multiple choices) เข้าร่วมในการวิจัยในโครงการย่อย

2. ดำเนินการเรียนการสอนตามตามที่กำหนดในการออกแบบการเรียนการสอนรายวิชา (มคอ.3) มีการออกข้อสอบตามตารางวิเคราะห์หลักสูตร (Test blueprint) วิพากษ์ข้อสอบโดยทีมผู้สอน และคณะกรรมการบริหารหลักสูตรของคณะ ดำเนินการสอบกลางภาค และปลายภาค และตัดสินเกรด ตามระเบียบการวัดและประเมินผลของมหาวิทยาลัย

3. ส่งข้อสอบไปวิเคราะห์ที่สำนักส่งเสริมวิชาการและงานทะเบียนของมหาวิทยาลัย ภายหลังการสอบ

4. ส่งตรวจกระดาษคำตอบใหม่ เมื่อพบว่ามีข้อสอบที่มีค่าอำนาจจำแนกติดลบ โดยยกเว้นการตรวจให้คะแนนข้อนั้น ๆ

5. บันทึกคะแนนที่นักศึกษาแต่ละคนได้รับลงในแบบบันทึกข้อมูล ซึ่งมีลักษณะเป็นตาราง Excel ที่เตรียมไว้ ทั้งคะแนนครั้งแรก และคะแนนหลังจากดึงข้อสอบที่มีค่าอำนาจจำแนกติดลบออกเทียบบัญญัติไตรยางศ์เพื่อให้ได้คะแนนครั้งหลังที่นักศึกษาแต่ละคนจะได้รับจากคะแนนเต็มเท่าเดิม หลังจากนั้นผู้วิจัยร่วมในโครงการย่อยทั้ง 11 วิชา นำคะแนนที่ได้ไปวิเคราะห์ตามวัตถุประสงค์ของการวิจัย

6. แต่เนื่องจาก จำนวนข้อสอบในแบบทดสอบแต่ละวิชามีไม่เท่ากัน ผู้วิจัยในโครงการนี้ จึงได้เทียบบัญญัติไตรยางศ์ทำให้คะแนนเต็มของทุกวิชาเป็น 100 คะแนนเท่ากัน เพื่อให้การนำเสนอผลสามารถเปรียบเทียบรายวิชาได้ด้วย

การวิเคราะห์ข้อมูล

1. การวิเคราะห์หาคุณภาพของข้อสอบของ

คณะพยาบาลศาสตร์ โดยการหาค่าความยากง่าย และอำนาจจำแนกเป็นรายข้อ ของแบบทดสอบแต่ละฉบับในทุกวิชา และวิเคราะห์หาค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับโดยใช้โปรแกรมการวิเคราะห์ข้อสอบรายข้อ (Items analysis) ดำเนินการโดยสำนักส่งเสริมวิชาการและงานทะเบียนของมหาวิทยาลัย หลังจากนั้นผู้วิจัยนำผลการวิเคราะห์ มาแจงนับ จัดกลุ่ม และคำนวณหาค่าร้อยละ

2. การวิเคราะห์การเปลี่ยนแปลงของคะแนนสอบรายวิชาและอันดับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก มีการดำเนินการดังนี้

1. การเปรียบเทียบค่าคะแนนเฉลี่ยที่นักศึกษาได้รับ ใน 11 รายวิชา ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Paired t-test

2. การทดสอบการเปลี่ยนแปลงของอันดับคะแนนที่นักศึกษาได้รับ ใน 11 รายวิชา ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Signed rank t est

3. การวิเคราะห์การเปลี่ยนแปลงค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก มีการดำเนินการดังนี้

1. การวิเคราะห์เปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก โดยใช้สถิติ Paired t-test

2. การทดสอบความสัมพันธ์ระหว่างร้อยละของจำนวนข้อสอบที่มีค่าอำนาจจำแนกติดลบ กับค่าความเชื่อมั่นที่เปลี่ยนแปลงไป โดยใช้สถิติ Pearson's Product Moment Correlation

ผลการวิจัย

ผู้วิจัยได้นำเสนอผลการวิจัย ตามวัตถุประสงค์เฉพาะของการวิจัย คือ 1) คุณภาพของข้อสอบ 2) การเปลี่ยนแปลงคะแนนสอบ และอันดับคะแนนที่นักศึกษาได้รับ 3) การเปลี่ยนแปลงค่าความเชื่อมั่น ของแบบทดสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

คุณภาพข้อสอบของคณะพยาบาลศาสตร์

ผลการวิเคราะห์ข้อสอบกลางภาค และปลายภาคของข้อสอบที่ใช้วัดผลการเรียนการสอนในทั้ง 11 วิชา รวม 26 ฉบับ ในปีการศึกษา 2555 ภาคการศึกษาที่ 1 โดยใช้ค่าความยากง่าย และ ค่าอำนาจจำแนก ของข้อสอบ เป็นดัชนีแสดงถึงคุณภาพของข้อสอบรายข้อ และใช้ค่าสัมประสิทธิ์ความเชื่อมั่นแสดงถึงคุณภาพของแบบทดสอบทั้งฉบับ แสดงในตารางที่ 1

ตารางที่ 1 ค่าความยากง่าย (P) และ ค่าอำนาจจำแนก (r) ที่อยู่ในเกณฑ์ที่เหมาะสม และค่าความเชื่อมั่นของแบบทดสอบ ทั้ง 11 รายวิชาที่ใช้ในการศึกษา

รายวิชา	แบบทดสอบ	จำนวนข้อ	จำนวนนักศึกษาที่เข้าสอบ	จำนวนข้อที่ P=.20-.80 (%)	จำนวนข้อที่ r ≥ .20 (%)	จำนวนข้อที่ r<0 (%)	ค่าความเชื่อมั่นของแบบทดสอบ
A	กลางภาค	70	92	62 (88.57)	52 (74.29)	4 (5.71)	0.83
	ปลายภาค	90	92	75 (83.33)	63 (70.00)	6 (6.67)	0.80
B	กลางภาค	70	92	30 (42.86)	30 (42.86)	4 (5.71)	0.50
	ปลายภาค	40	92	10 (25.00)	10 (25.00)	2 (5.00)	0.34
C	ปลายภาค	50	92	27 (54.00)	19 (38.00)	2 (4.00)	0.37
D	ครั้งที่ 1	105	88	49 (46.67)	50 (47.62)	5 (4.76)	0.81
	ครั้งที่ 2	65	88	21 (32.31)	26 (40.00)	3 (4.62)	0.74
	ปลายภาค	105	88	75 (71.43)	55 (52.38)	6 (5.71)	0.81
E	ครั้งที่ 1	70	88	54 (77.14)	29 (41.43)	9 (12.86)	0.43
	ครั้งที่ 2	65	88	30 (46.15)	36 (55.39)	6 (9.23)	0.57
	ปลายภาค	100	88	31 (31.00)	41 (41.00)	13(13.00)	0.57
F	ครั้งที่ 1	75	88	62 (82.67)	38 (50.67)	5 (6.67)	0.60
	ครั้งที่ 2	70	88	60 (85.71)	39 (55.71)	6 (8.57)	0.66
	ครั้งที่ 3	75	88	60 (80.00)	36 (48.00)	4 (5.33)	0.66
	ปลายภาค	58	88	47 (81.03)	27 (46.55)	3 (5.17)	0.50
G	กลางภาค	60	92	33 (55.00)	33 (55.00)	6 (10.00)	0.55
	ปลายภาค	80	92	45 (56.25)	45 (56.25)	7 (8.75)	0.69
H	กลางภาค	120	92	95 (79.17)	75 (62.50)	9 (7.50)	0.80
	ปลายภาค	75	92	58 (73.33)	31 (41.33)	6 (8.00)	0.45
I	ครั้งที่ 1	95	92	37 (38.95)	47 (49.47)	7 (7.37)	0.82
	ครั้งที่ 2	90	92	44 (48.89)	57 (63.33)	3 (3.33)	0.81
J	ปลายภาค	85	92	38 (44.70)	50 (58.82)	3 (3.53)	0.74
	กลางภาค	75	92	29 (38.67)	29 (38.67)	6 (8.00)	0.42
	ปลายภาค	102	92	41 (40.20)	28 (27.45)	12(11.76)	0.71
K	กลางภาค	40	60	30 (75.00)	19 (47.5)	1 (2.50)	0.48
	ปลายภาค	40	60	26 (65.00)	23 (57.50)	1 (2.50)	0.40
11 วิชา	26 ฉบับ	1,970ข้อ (100%)	332 คน	1,169 ข้อ (59.34%)	988 ข้อ (50.15%)	139 ข้อ (7.06%)	0.34-0.83

จากตารางที่ 1 เมื่อพิจารณาคุณภาพเป็นรายข้อของข้อสอบทั้ง 11 วิชา จำนวนทั้งสิ้น 26 ฉบับ จำนวนทั้งสิ้น 1,970 ข้อ โดยอาศัยดัชนีความยากง่าย (P) พบว่า ข้อสอบที่มีค่าความยากง่ายในเกณฑ์ที่เหมาะสม (P อยู่ระหว่าง .20-.80) มีจำนวน 1,169 ข้อ คิดเป็นร้อยละ 59.34 ข้อสอบที่มีค่าอำนาจจำแนกในเกณฑ์ที่เหมาะสม ($r \geq .20$) มีจำนวน 988 ข้อ คิดเป็นร้อยละ 50.15 และมีข้อสอบที่ค่าอำนาจจำแนกติดลบทั้งหมด 139 ข้อ คิดเป็นร้อยละ 7.06 นอกจากนั้นเมื่อพิจารณาคุณภาพของแบบทดสอบทั้งฉบับ พบว่า แบบทดสอบที่มีค่าสัมประสิทธิ์ความเชื่อมั่น .70 ขึ้นไป ซึ่งแสดงว่ามีค่าความเชื่อมั่นอยู่ในเกณฑ์ดี ถึงดีมาก มีจำนวน 10 ฉบับ คิดเป็น ร้อยละ 38.46 และแบบทดสอบที่มีค่าความเชื่อมั่นต่ำกว่า .50 ซึ่งจำเป็นต้องมีการปรับปรุง จำนวน 7 ฉบับ คิดเป็นร้อยละ 26.92

การเปลี่ยนแปลงของคะแนนสอบรายวิชา ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

ภายหลังจากได้รับผลการวิเคราะห์ข้อสอบหากมีข้อสอบที่มีค่าอำนาจจำแนกติดลบ ซึ่งแสดงว่าคุณภาพของข้อสอบข้อนั้น ๆไม่สามารถจำแนกผู้เรียนเก่งและอ่อนได้ ผู้เรียนอ่อนส่วนมากตอบข้อนั้นถูก แต่ผู้เรียนเก่งส่วนมากไม่ได้คะแนนจากข้อนั้น ดังนั้น ในการศึกษาครั้งนี้ ผู้วิจัยจึงได้ดึงข้อสอบที่ที่มีค่าอำนาจจำแนกติดลบออก และส่งตรวจพร้อมทั้งวิเคราะห์ข้อสอบใหม่ หลังจากนั้น จึงนำคะแนนที่ได้ครั้งหลังมาคิดคำนวณเทียบบัญญัติไตรยางศ์เพื่อให้ได้คะแนนที่นักศึกษาแต่ละคนได้รับมาจากคะแนนเต็มเท่าเดิม เพื่อให้สามารถเปรียบเทียบคะแนน และ อันดับของคะแนนของนักศึกษาเป็นรายบุคคลก่อน และหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก ผลการเปรียบเทียบการเปลี่ยนแปลงของคะแนนสอบรายวิชาที่นักศึกษาได้รับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก จากแบบทดสอบทั้ง 11 รายวิชา จำนวน 26 ฉบับ สรุปได้ดังตารางที่ 2 แต่เพื่อให้ผู้อ่านสามารถเปรียบเทียบความแตกต่างระหว่างวิชา ผู้วิจัยจึงได้นำเสนอคะแนนเฉลี่ยของทุกวิชา ที่ได้เทียบคะแนนเต็มให้เป็น 100

ตารางที่ 2 ผลการเปรียบเทียบค่าคะแนนเฉลี่ยคิดเป็นร้อยละ ที่นักศึกษาได้รับ ใน 11 รายวิชา ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Paired t-test (one-tailed)

วิชา	ข้อสอบ	ก่อนตัดข้อที่มีค่า r ติดลบออก		หลังการตัดข้อที่มีค่า r ติดลบออก		คะแนนที่เพิ่มขึ้น (%)	จำนวนข้อที่ r ติดลบ (%)	df= (n-1)	Paired t-test	p-value
		Mean %	S.D. %	Mean %	S.D. %					
A	กลางภาค	48.21	12.55	54.30	13.23	6.09	5.71	91	42.151	<0.01
	ปลายภาค	50.86	10.71	57.52	11.21	6.66	6.67	91	44.195	<0.01
B	กลางภาค	58.74	7.05	60.33	7.65	1.59	5.71	91	10.225	<0.01
	ปลายภาค	70.00	7.15	71.08	7.58	1.08	5.00	91	4.972	<0.01
C	ปลายภาค	55.70	7.01	55.68	7.39	-0.02	4.00	91	-0.129	>0.05

ตารางที่ 2 ผลการเปรียบเทียบค่าคะแนนเฉลี่ยคิดเป็นร้อยละ ที่นักศึกษาได้รับ ใน 11 รายวิชา ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Paired t-test (one-tailed) (ต่อ)

วิชา	ข้อสอบ	ก่อนตัดข้อที่มีค่า r ติดลบออก		หลังการตัดข้อที่มีค่า r ติดลบออก		คะแนนที่เพิ่มขึ้น (%)	จำนวนข้อที่ r ติดลบ (%)	df= (n-1)	Paired t-test	p-value
		Mean %	S.D. %	Mean %	S.D. %					
D	ครั้งที่ 1	69.22	8.04	71.94	8.60	2.72	4.76	87	29.31	<0.01
	ครั้งที่ 2	77.38	8.10	77.75	8.54	0.37	4.62	87	3.53	<0.01
	ปลายภาค	64.02	9.23	66.45	10.09	2.43	5.71	87	17.83	<0.01
E	ครั้งที่ 1	52.60	7.46	55.35	8.49	2.75	12.86	87	11.98	<0.01
	ครั้งที่ 2	56.54	8.39	56.82	9.32	0.28	9.23	87	1.17	>0.05
	ปลายภาค	51.83	6.44	54.92	8.04	3.09	12.00	87	7.12	<0.01
F	ครั้งที่ 1	53.14	8.29	54.50	9.09	1.36	6.67	87	8.714	<0.01
	ครั้งที่ 2	54.42	9.20	54.35	10.56	-0.07	8.57	87	-.306	>0.05
	ครั้งที่ 3	56.83	8.74	58.05	9.35	1.22	5.33	87	8.016	<0.01
	ปลายภาค	53.08	7.80	55.25	8.28	2.17	5.17	87	17.995	<0.01
G	กลางภาค	55.36	8.79	56.68	10.08	1.32	10.00	91	4.75	<0.01
	ปลายภาค	49.76	8.99	51.19	10.04	1.43	8.75	91	7.17	<0.01
H	กลางภาค	55.07	6.98	55.95	7.86	0.88	7.50	91	12.70	<0.01
	ปลายภาค	55.65	6.74	55.66	7.72	0.01	8.00	91	4.88	<0.01
I	ครั้งที่ 1	69.02	7.95	70.19	8.85	1.17	7.37	91	.989	<0.01
	ครั้งที่ 2	63.55	9.91	63.64	10.36	0.09	3.33	91	.997	<0.01
	ปลายภาค	64.64	8.69	64.57	9.16	-0.10	3.53	91	-.997	<0.01
J	กลางภาค	61.03	6.43	60.87	7.33	-0.16	8.00	91	-.961	>0.05
	ปลายภาค	60.96	7.62	62.75	9.14	1.79	11.76	91	8.377	<0.01
K	กลางภาค	56.33	9.74	56.79	10.16	0.46	2.50	59	2.75	<0.01
	ปลายภาค	56.25	8.62	56.95	8.82	0.70	2.50	59	4.75	<0.01

ผลการเปรียบเทียบค่าคะแนนเฉลี่ยที่นักศึกษาได้รับ ใน 11 รายวิชา จากแบบทดสอบสอบทั้ง 26 ฉบับ ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Paired t-test (one-tailed) พบว่า นักศึกษามีคะแนนเฉลี่ยสูงขึ้นอย่างมีนัยสำคัญทางสถิติ (p-value <0.025) จากข้อสอบจำนวน 21 ฉบับ (คิดเป็นร้อยละ 80.77 ของจำนวนชุดข้อสอบทั้งหมด) คะแนนลดลง 1 ฉบับ (คิดเป็นร้อยละ 0.10) ที่เหลืออีก 4 ฉบับ ไม่พบว่าคะแนนเฉลี่ยของนักศึกษามีการเปลี่ยนแปลงอย่างมีนัยสำคัญทางสถิติ ที่ α =0.05

แสดงว่า โดยส่วนใหญ่แล้วการตัดข้อสอบที่มีค่า r ติดลบออกทำให้คะแนนเฉลี่ยของนักศึกษาเพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติ คิดเป็น ร้อยละ 80.77 ของจำนวนชุดข้อสอบทั้งหมด

การเปลี่ยนแปลงอันดับคะแนนที่นักศึกษาได้รับก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

ตารางที่ 3 ผลการทดสอบการเปลี่ยนแปลงของอันดับคะแนนที่นักศึกษาได้รับ ใน 11 รายวิชา ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Signed Rank Test

วิชา	ข้อสอบ	จำนวนนักศึกษาทั้งหมด	จำนวนนักศึกษาที่คะแนนเพิ่มขึ้น		จำนวนนักศึกษาที่คะแนนคงเดิม		จำนวนนักศึกษาที่คะแนนลดลง		Signed Rank Test	p-value
			จำนวน	%	จำนวน	%	จำนวน	%		
A	กลางภาค	92	92	100.00	0	0.00	0	0.00	42.151	<0.01
	ปลายภาค	92	92	100.00	0	0.00	0	0.00	44.195	<0.01
B	กลางภาค	92	83	90.22	1	1.09	8	8.70	7.339	<0.01
	ปลายภาค	92	29	31.52	0	0.00	63	68.48	4.884	<0.01
C	ปลายภาค	92	58	63.04	5	5.43	29	31.52	2.397	<0.01
D	ครั้งที่ 1	88	88	100.00	0	0.00	0	0.00	8.148	<0.01
	ครั้งที่ 2	88	65	73.86	0	0.00	23	26.14	3.342	<0.01
	ปลายภาค	88	86	97.73	0	0.00	2	2.27	8.023	<0.01
E	ครั้งที่ 1	88	78	88.64	0	0.00	10	11.36	7.688	<0.01
	ครั้งที่ 2	88	50	56.82	0	0.00	38	43.18	1.355	>0.05
	ปลายภาค	88	83	94.32	0	0.00	5	5.68	7.925	<0.01
F	ครั้งที่ 1	88	71	80.68	3	3.41	14	15.91	6.504	<0.01
	ครั้งที่ 2	88	41	46.59	2	2.27	45	51.14	.306	>0.05
	ครั้งที่ 3	88	73	82.95	0	0.00	15	17.05	6.315	<0.01
	ปลายภาค	88	84	95.45	0	0.00	4	4.55	7.934	<0.01

ตารางที่ 3 ผลการทดสอบการเปลี่ยนแปลงของอันดับคะแนนที่นักศึกษาได้รับ ใน 11 รายวิชา ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Signed rank test (ต่อ)

วิชา	ข้อสอบ	จำนวนนักศึกษาทั้งหมด	จำนวนนักศึกษาที่คะแนนเพิ่มขึ้น		จำนวนนักศึกษาที่คะแนนคงเดิม		จำนวนนักศึกษาที่คะแนนลดลง		Signed Rank Test	p-value
			จำนวน	%	จำนวน	%	จำนวน	%		
G	กลางภาค	92	65	70.65	2	2.17	25	27.17	4.313	<0.01
	ปลายภาค	92	75	81.52	0	0.00	17	18.48	6.038	<0.01
H	กลางภาค	92	83	90.22	0	0.00	9	9.78	7.83	<0.01
	ปลายภาค	92	64	69.57	1	1.09	27	29.35	4.39	<0.01
I	ครั้งที่ 1	92	74	80.43	0	0.00	18	19.57	6.020	<0.01
	ครั้งที่ 2	92	41	44.57	2	2.17	49	53.26	.577	>0.05
	ปลายภาค	92	39	42.39	0	0.00	53	57.61	1.468	>0.05
J	กลางภาค	92	37	40.22	4	4.35	51	55.43	1.092	>0.05
	ปลายภาค	92	71	77.17	0	0.00	21	22.83	6.464	<0.01
K	กลางภาค	60	37	61.67	0	0.00	23	38.33	000	<0.01
	ปลายภาค	60	43	71.67	0	0.00	17	28.33	000	<0.01

จากตารางที่ 3 ผลการทดสอบการเปลี่ยนแปลงของลำดับคะแนนที่นักศึกษาได้รับ ใน 11 รายวิชา จากข้อสอบทั้งหมด 26 ฉบับ ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก โดยสถิติ Signed rank test พบว่าลำดับคะแนนที่นักศึกษาได้รับมีการเปลี่ยนแปลงลำดับที่อย่างมีนัยสำคัญทางสถิติที่ α =0.05 จากข้อสอบจำนวน 21 ฉบับ คิดเป็น ร้อยละ 80.77 ที่เหลืออีก 5 ฉบับคะแนนของนักศึกษาไม่มีการเปลี่ยนแปลงลำดับที่อย่างมีนัยสำคัญทางสถิติ (p value >0.05)

แสดงว่า โดยส่วนใหญ่แล้วการตัดข้อสอบที่มีค่า r ติดลบออกทำให้มีผลต่อการเปลี่ยนแปลงลำดับที่ของนักศึกษา

การเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

เพื่อศึกษาการเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก ผู้วิจัยได้วิเคราะห์เปรียบเทียบค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก และได้วิเคราะห์ความสัมพันธ์ระหว่างร้อยละของจำนวนข้อสอบที่มีค่าอำนาจจำแนกติดลบ กับการเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบ ดังแสดงในตารางที่ 4

ตารางที่ 4 การเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบทั้ง 26 ฉบับหลังการการตัดข้อสอบที่มีค่า r ติดลบออก และความสัมพันธ์ระหว่างร้อยละของข้อสอบที่ค่า r ติดลบ กับค่าความเชื่อมั่นที่เปลี่ยนแปลงไป

ค่าความเชื่อมั่น	Mean	S.D.	t	df	p-value (2-tailed)
ก่อน	0.62	0.16	4.46	25	.000
หลัง	0.66	0.14			

ผลการวิเคราะห์เปรียบเทียบค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก โดยใช้สถิติ Paired t-test (one-tailed) พบว่า ค่าความเชื่อมั่นของข้อสอบทั้งฉบับหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออกเพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติ ที่ α =0.05 (t = 4.46, df = 25, p value <0.025) และ ผลการทดสอบความสัมพันธ์ระหว่างร้อยละของข้อสอบที่ค่า r ติดลบ กับค่าความเชื่อมั่นที่เปลี่ยนแปลงไป ด้วยสถิติ Pearson's Product moment correlation พบว่า มีความสัมพันธ์กันในเชิงบวก (r= .667) ที่ α =0.05

แสดงว่า ข้อสอบฉบับที่มีจำนวนข้อที่มีค่า r ติดลบจำนวนมาก หากตัดข้อสอบเหล่านี้ออก จะเพิ่มค่าความเชื่อมั่นได้อย่างมีนัยสำคัญ

อภิปรายผล

1. คุณภาพของข้อสอบของคณะพยาบาลศาสตร์

1.1 เมื่อพิจารณาจากดัชนีความยากง่าย (P) พบว่า ข้อสอบส่วนใหญ่ร้อยละ 70.39 มีค่าความยากง่ายในเกณฑ์ที่เหมาะสม (P อยู่ระหว่าง .20-.80) ซึ่งสุมาลี จันทร์ชะลอ (2542: 136) ได้ให้ข้อเสนอแนะว่าเป็นข้อสอบที่ดี ควรเก็บไว้ใช้ ส่วนข้อสอบที่ยากเกินไป (P=.00-.19) และ ข้อสอบที่ง่ายเกินไป (P=.81-1.00) ควรตัดทิ้ง เพราะข้อสอบที่ยากเกินไป แม้นักศึกษาที่เรียนเก่ง ก็ทำข้อสอบข้อนั้นไม่ได้อาจเนื่องจาก วัดในเนื้อหาที่ไม่ได้สอน หรือหากเป็นข้อสอบที่ง่ายเกินไป นักศึกษาไม่ว่าเก่ง หรือ อ่อน ก็ทำได้ ก็แสดงว่าข้อสอบข้อนั้นไม่สามารถวัดความรู้ ของนักศึกษาได้อย่างมีประสิทธิผล

ปัจจัยที่มีผลต่อความยากง่ายของข้อสอบมีหลายประการ เช่น ระดับการวัดของข้อสอบ ซึ่งต้องออกให้สอดคล้องตามตารางวิเคราะห์ข้อสอบ (Test blueprint) ที่กำหนดไว้ ข้อสอบที่มีระดับการวัดในระดับนำไปใช้ วิเคราะห์ สังเคราะห์/ประเมินค่า ย่อมยากกว่าข้อสอบที่วัดในระดับความรู้-จำ และ เข้าใจ ข้อสอบที่วัดในระดับสูง มักอยู่ในรูปสถานการณ์ จึงมีโจทย์ที่ยาว อันอาจเป็นผลให้นักศึกษาต้องใช้เวลาในการอ่านทำความเข้าใจนาน ต้องใช้ความรู้ในระดับสูง นอกจากนั้น ความคล้ายคลึงของตัวลวง ก็มีผลต่อความยากง่ายของข้อสอบ ดังผลการวิจัยของ Ascalon, Meyers, Davis & Smits (2007) ที่ได้ศึกษาผลของความคล้ายคลึงของตัวลวง (Distractor similarity) และโครงสร้างของข้อคำถาม (Item-Stem structure) ต่อความยากของข้อสอบ พบว่า ตัวลวงของข้อคำตอบที่มีความสั้นยาวของตัวอักษรแต่ละข้อตัวลวง ใกล้เคียงกัน เนื้อหาของคำตอบคล้ายคลึงกันกับข้อคำตอบที่ถูกต้อง จะทำให้ข้อสอบข้อนั้นมีความยากมากกว่าข้อสอบที่มีตัวลวงที่มีเนื้อหาของคำตอบแตกต่างกันกับคำตอบที่ถูกต้องอย่างมีนัยสำคัญทางสถิติ ที่ระดับ 0.01 (P value <0.01) เพราะตัวลวงที่มีเนื้อหาแตกต่างกัน จะง่ายต่อการคาดเดา ส่วนลักษณะการตั้งโจทย์คำถามที่เป็นข้อคำถามหรือโจทย์คำถามที่เป็นข้อความปลายเปิด ไม่พบความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ในเรื่องความยากง่ายของข้อสอบ Hamzah & Abdullah (2011) ได้เสนอ

ว่าคุณสมบัติของตัวลวงที่ดี คือ ทำให้นักเรียนที่ไม่มีความรู้ในเรื่องนั้นเลือกคำตอบข้อนั้น และข้อสอบที่มีตัวลวงดี นักเรียนจะเลือกตัวลวงทุกตัวในข้อนั้นในจำนวนที่ใกล้เคียงกัน ข้อสอบที่ยากมักจะมีตัวลวงที่ดี ส่วนข้อสอบที่ง่ายหมายถึงตัวลวงไม่มีประสิทธิภาพ

ปัจจัยที่สำคัญอีกประการหนึ่งที่มีผลต่อความยากง่ายของข้อสอบ คือลักษณะของคำถามที่ใช้ ดังผลการศึกษาของ Caldwell & Pate (2013) ที่ได้ศึกษารูปแบบของคำถามต่อคะแนนการสอบของนักศึกษาและคุณภาพของข้อสอบ โดยมีวัตถุประสงค์เพื่อเปรียบเทียบ การใช้รูปแบบคำถามของข้อสอบที่สร้างตามแนวปฏิบัติที่เป็นมาตรฐานและไม่มีมาตรฐาน ข้อสอบที่สร้างตามแนวปฏิบัติที่เป็นมาตรฐาน จะมีหลักการในการสร้างคำถาม ดังนี้ 1) การใช้คำถามเชิงบวก หลีกเลี่ยงคำถามเชิงลบ เช่น ยกเว้น, ไม่ใช่ หากจำเป็นต้องใช้ต้องขีดเส้นใต้ หรือ ใช้ตัวหนาเพื่อให้เห็นคำที่เป็นเชิงลบอย่างชัดเจน 2) ในแต่ละคำถามสามารถใช้ตัวเลือกได้มากเท่าที่ต้องการ แต่ผลการวิจัยพบว่าการใช้ตัวเลือกเพียง 3 ตัวเลือกก็เพียงพอแล้ว 3) การใช้ตัวเลือก " ไม่มีข้อใดถูกต้อง" ควรใช้ด้วยความระมัดระวัง ผลการศึกษาของ Caldwell & Pate (2013) พบว่า ข้อสอบที่ไม่ได้สร้างตามแนวปฏิบัติที่เป็นมาตรฐาน (Nonstandard scale) มีความยากกว่า ข้อสอบ Standard scale และแนวปฏิบัติที่แนะนำให้หลีกเลี่ยงการใช้ตัวลวง "ไม่มีข้อใดถูกต้อง" ทำให้คะแนนของนักศึกษาในทั้งสองกลุ่มแตกต่างกันอย่างมีนัยสำคัญทางสถิติ โดยคะแนนเฉลี่ยของกลุ่มที่ทำข้อสอบ Standard scale สูงกว่ากลุ่ม ที่ทำข้อสอบ Nonstandard scale จากผลการวิจัยสรุปได้ว่าข้อสอบที่ไม่ได้มาตรฐานจะทำให้ยากต่อการที่นักศึกษาจะตอบถูกมากกว่าข้อสอบที่เป็นมาตรฐาน จึงไม่ช่วยในการแยกแยะกลุ่มที่เก่ง กับกลุ่มที่อ่อนได้ และทำให้นักศึกษาที่เรียนอ่อนมักยิ่งได้คะแนนน้อยลงไปอีก ดังนั้นครูผู้สอนควรได้คำนึงถึงหลักการในการสร้างคำถามตามแนวปฏิบัติที่เป็นมาตรฐานในการสร้างข้อสอบทุกครั้ง

1.2 คุณภาพข้อสอบ ด้านค่าดัชนีอำนาจจำแนก (r) พบว่า ข้อสอบที่มีค่าอำนาจจำแนกในเกณฑ์ที่เหมาะสม ($r \geq .20$) มีเพียงร้อยละ 51.80 ที่สามารถจำแนกนักศึกษาที่เรียนเก่ง และอ่อนได้ (Chase 1978: 140) ข้อสอบส่วนที่เหลือ ร้อยละ 41.19 เป็นข้อสอบที่มีค่าอำนาจจำแนกต่ำ (r=0.01-0.19) และ อีกร้อยละ 7.05 มีค่าอำนาจจำแนกติดลบ (r=-1.00-0.00) ทั้งนี้ข้อสอบที่ค่าดัชนีอำนาจจำแนก ต่ำ มักเกิดจากการใช้คำในข้อคำถาม และตัวเลือกที่กำกวม ไม่ชัดเจน ส่วนข้อที่มีค่าดัชนีอำนาจจำแนกติดลบ ผู้สอนควรได้นำมาพิจารณาว่าเกิดจากสาเหตุใด เช่น การเฉลยผิด ทำให้ผู้เรียนที่เก่งไม่ได้เลือกตัวเลือกนั้น เป็นต้น (Office of Educational Assessment, University of Washington, 2005)

เมื่อพิจารณาจากค่าดัชนีความยากง่าย และค่าดัชนีอำนาจจำแนก จะเห็นว่าจะมีข้อสอบที่ควรตัดทิ้งประมาณ ร้อยละ 20 และอีกร้อยละ 25 ควรต้องนำมาปรับปรุงก่อนนำมาใช้อีก ในแต่ละปีการศึกษา อาจารย์จึงควรนำผลการวิเคราะห์ข้อสอบมาปรับปรุงข้อสอบ และมีการออกข้อสอบใหม่เพื่อทดแทนของเดิม ทั้งนี้จำเป็นอย่างยิ่งที่ต้องออกข้อสอบใหม่จากตารางวิเคราะห์ข้อสอบ (Test blueprint) ดังที่ Usova GM. (1997) ซึ่งได้ศึกษาประสิทธิผลของการใช้ข้อสอบเก่าในคลังข้อสอบร่วมกับข้อสอบใหม่ และได้เสนอวิธีการเลือกข้อสอบให้ประกอบไปด้วย 3 กลุ่มคือ (1) test bank items เป็นข้อสอบที่เคยใช้สอบมาก่อนและผ่านการพิจารณาความตรง และอำนาจจำแนกแล้ว (2) modified items เป็นข้อสอบที่ปรับปรุงจากข้อสอบเดิม และ (3) new items ข้อสอบใหม่ในสัดส่วน 50-40-10

1.3 เมื่อพิจารณาถึงคุณภาพของข้อสอบทั้งฉบับจากค่าสัมประสิทธิ์ความเชื่อมั่น ซึ่งแสดงถึงความคงเส้นคงวาของคะแนนที่ได้จากการใช้ข้อสอบชุดนั้น ทั้งนี้ค่าความเชื่อมั่นของแบบทดสอบ มีผลมาจาก 3 องค์ประกอบ คือ 1) ความสัมพันธ์ระหว่างข้อสอบ

แต่ละข้อ หากข้อสอบแต่ละข้อมีความสัมพันธ์กันในเชิงบวกสูง ค่าสัมประสิทธิ์ความเชื่อมั่นก็จะสูง 2) ความยาวหรือ จำนวนข้อของแบบทดสอบ ข้อสอบที่มีจำนวนข้อมากกว่าจะมีค่าความเชื่อมั่นสูงกว่าข้อสอบที่จำนวนข้อน้อยกว่า ในกรณีที่คุณสมบัติด้านอื่น ๆทัดเทียมกัน 3) เนื้อหาของแบบทดสอบ ข้อสอบที่วัดเนื้อหาแตกต่างกันมาก จะมีค่าความเชื่อมั่นต่ำ จากตารางที่ 1 พบว่าข้อสอบจำนวน 7 ฉบับ (ร้อยละ 26.92) มีค่าความเชื่อมั่นต่ำกว่า .50 สาเหตุประการหนึ่งที่ทำให้แบบทดสอบทั้ง 7 ฉบับ มีค่าความเชื่อมั่นต่ำอาจเนื่องมาจาก มีจำนวนข้อน้อย (40-50 ข้อ ในขณะที่วิชาอื่น ๆมีจำนวนข้อสอบ 70-120 ข้อ)

2. การเปลี่ยนแปลงของคะแนนสอบรายวิชาและอันดับคะแนนที่นักศึกษาได้รับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

2.1 ผลการเปรียบเทียบค่าคะแนนเฉลี่ยที่นักศึกษาได้รับ ใน 11 รายวิชา จากข้อสอบทั้ง 26 ฉบับ ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก พบว่านักศึกษามีคะแนนเฉลี่ยสูงขึ้นอย่างมีนัยสำคัญทางสถิติที่ (α =0.05) จากข้อสอบจำนวน 21 ฉบับ คะแนนลดลง 1 ฉบับ ที่เหลืออีก 4 ฉบับคะแนนเฉลี่ยของนักศึกษาไม่มีการเปลี่ยนแปลงอย่างมีนัยสำคัญทางสถิติ (p value >0.05) แสดงว่า โดยส่วนใหญ่แล้วการตัดข้อสอบที่มีค่า r ติดลบออกทำให้คะแนนเฉลี่ยของนักศึกษาเพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติ ทั้งนี้ อาจอธิบายได้ว่าข้อสอบส่วนที่เหลือ หลังจากตัดข้อที่มีค่า r ติดลบออกแล้วสามารถวัดความรู้ที่แท้จริงของนักศึกษาได้ดีกว่าข้อสอบชุดเดิม หรือกล่าวอีกนัยหนึ่งว่า คะแนนที่ได้จากการทำข้อสอบชุดใหม่นี้ มีค่าเข้าใกล้คะแนนที่แท้จริง (True score) มากกว่าคะแนนที่ได้จากข้อสอบชุดเดิม ซึ่งมีความคลาดเคลื่อน (Error) มากกว่า ดังที่ Tavakol & Dennick (2011) ได้แนะนำให้มีการวิเคราะห์ข้อสอบวัดผลสัมฤทธิ์ภายหลังการสอบ เพราะว่าเป้าหมายที่สำคัญประการหนึ่งของการประเมินในทางแพทยศาสตร์ศึกษา คือการลดความคลาดเคลื่อนของการทดสอบให้มีน้อยที่สุดเท่าที่จะทำได้ เพื่อที่จะให้ผลคะแนนสอบที่วัดได้ใกล้เคียงกับ ความรู้ ความสามารถที่แท้จริงของนักศึกษามากที่สุด

นอกจากการพิจารณาถึงระดับนัยสำคัญของการเปลี่ยนแปลงค่าคะแนนแล้ว ในตารางที่ 2 ยังแสดงถึงขนาดของค่าคะแนนที่เปลี่ยนแปลงไป (Magnitude of changes) ในแต่ละรายวิชา ดังพบว่าในรายวิชา A หลังจากตัดข้อสอบที่มีค่า r ติดลบออก คะแนนเฉลี่ยของนักศึกษาสูงขึ้นสูงสุดคิดเป็นร้อยละ 6.66 (จาก 50.86 เป็น 57.52) ในขณะที่วิชา I คะแนนเฉลี่ยลดลงคิดเป็นร้อยละ 0.10 (จาก 64.66 เป็น 64.57) ผู้วิจัยจึงได้วิเคราะห์ข้อมูลเพิ่มเติมโดยสถิติ Pearson's Product Moment Correlation เพื่อทดสอบว่าคะแนนที่เปลี่ยนแปลงไป (คิดเป็น %) จะมีความสัมพันธ์กับจำนวนข้อสอบที่มีค่าอำนาจจำแนกติดลบ (คิดเป็น %) หรือไม่ ผลการทดสอบพบว่า ไม่มีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติที่ α =0.05 ระหว่างร้อยละของผลต่างของคะแนนก่อนและหลังการตัดข้อสอบที่มีค่า r ติดลบออก กับร้อยละของจำนวนข้อสอบที่มีค่า r ติดลบ ซึ่งแสดงว่า คะแนนของนักศึกษาที่เปลี่ยนแปลงไปไม่ขึ้นอยู่กับ ว่าจำนวนข้อสอบที่มีค่า r ติดลบจะน้อยหรือมาก

ทั้งนี้อาจเนื่องมาจากข้อสอบที่มีค่าอำนาจจำแนกติดลบจำนวนมากนั้นไม่สามารถที่จะลวงได้ทั้งคนเก่ง หรือคนอ่อน กล่าวคือ ทั้งคนเก่ง และคนอ่อนไม่เลือก ถึงแม้ว่าตัดออกไปแล้วคะแนนของนักศึกษาก็ไม่ได้เปลี่ยนไปมากนัก แต่ในข้อสอบชุดที่ดีนั้น แม้จะตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออกจำนวนน้อยข้อ ก็มีผลต่อคะแนนของนักศึกษา เพราะทำให้คนเก่งเลือกผิด ผลการศึกษานี้นับว่ามีความสำคัญมากในเชิงนโยบาย และการนำผลการวิจัยไปใช้ กล่าวคือ ข้อสอบชุดใด ๆ ไม่ว่าจะมีจำนวนข้อที่มีค่าอำนาจจำแนกน้อยหรือมาก ก็ควรตัดออก และคิดคะแนนใหม่ เพราะจะ

ทำให้ผลคะแนนสอบที่วัดได้ครั้งหลังใกล้เคียงกับความรู้ ความสามารถที่แท้จริงของนักศึกษามากกว่าคะแนนชุดเดิม

2.2 ผลการทดสอบการเปลี่ยนแปลงของลำดับคะแนนที่นักศึกษาได้รับ ใน 11 รายวิชา จากข้อสอบทั้งหมด 26 ฉบับ ก่อน และหลังการตัดข้อสอบที่มีค่า r ติดลบออก พบว่า ลำดับคะแนนที่นักศึกษาได้รับมีการเปลี่ยนแปลงลำดับที่อย่างมีนัยสำคัญทางสถิติที่ α =0.05 จากข้อสอบจำนวน 21 ฉบับ ที่เหลืออีก 5 ฉบับคะแนนของนักศึกษาไม่มีการเปลี่ยนแปลงลำดับที่อย่างมีนัยสำคัญทางสถิติ (p-value >0.05) แสดงว่าการตัดข้อสอบที่มีค่า r ติดลบออกทำให้มีผลต่อการเปลี่ยนแปลงลำดับที่ของนักศึกษา ซึ่งจากการทบทวนวรรณกรรมไม่พบงานวิจัย ที่ศึกษาเรื่องนี้ แต่หากพิจารณาว่าในกรณีที่ลำดับที่ของนักศึกษามีการเปลี่ยนแปลงไปภายหลังจากตัดข้อสอบที่มีค่า r ติดลบออก การวัดผลใด ๆที่มีผลในการคัดเลือกผู้เข้าสอบตามลำดับที่สอบได้ จะต้องทำด้วยความระมัดระวังอย่างยิ่ง

3. การเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก

เพื่อศึกษาการเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก ผู้วิจัยได้วิเคราะห์เปรียบเทียบค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก และ ได้วิเคราะห์ความสัมพันธ์ระหว่างร้อยละของจำนวนข้อสอบที่มีค่าอำนาจจำแนกติดลบ กับการเปลี่ยนแปลงค่าความเชื่อมั่นของข้อสอบ ผลการวิเคราะห์เปรียบเทียบค่าความเชื่อมั่นของข้อสอบทั้งฉบับ ก่อนและหลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก โดยใช้สถิติ Paired t-test (one-tailed) พบว่า ค่าความเชื่อมั่นของข้อสอบทั้งฉบับ หลังการตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออก เพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติ ที่ α =0.05 และ ผลการทดสอบความสัมพันธ์ระหว่างร้อยละของข้อสอบที่ค่า r ติดลบ กับค่าความเชื่อมั่นที่เปลี่ยนแปลงไป พบว่า มีความสัมพันธ์กันในเชิงบวก (r= .677) ที่ α =0.05 แสดงว่า การตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออกทำให้ค่าความเชื่อมั่นของข้อสอบทั้งฉบับ เพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติที่ระดับความเชื่อมั่น 95% และ ร้อยละของจำนวนข้อสอบที่มีค่า r ติดลบมีความสัมพันธ์ในทิศทางตรงกันข้ามกับค่าความเชื่อมั่นของข้อสอบ จากการทบทวนวรรณกรรมไม่พบงานวิจัยที่ศึกษาในลักษณะเดียวกัน แต่ผลการศึกษาสอดคล้องกับหลักการที่ว่าค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับมีผลมาจาก 3 องค์ประกอบ คือ 1) ความสัมพันธ์ระหว่างข้อสอบแต่ละข้อ 2) ความยาวหรือ จำนวนข้อของแบบทดสอบ และ 3) เนื้อหาของแบบทดสอบ ข้อสอบที่มีค่าอำนาจจำแนกติดลบมักเป็นข้อที่ใช้คำในข้อคำถาม และตัวเลือกที่กำกวม ไม่ชัดเจน หรือเฉลยผิด ดังนั้น การตัดข้อสอบที่มีค่าอำนาจจำแนกติดลบออกก็จะทำให้ข้อสอบที่เหลือมีความสัมพันธ์กันในเชิงบวกสูงขึ้น ค่าสัมประสิทธิ์ความเชื่อมั่นของแบบทดสอบสอบทั้งฉบับก็จะสูงขึ้นด้วย

จากการศึกษาครั้งนี้ แสดงให้เห็นถึงความสำคัญว่าคุณภาพข้อสอบมีผลต่อคะแนน และอันดับที่ในการสอบของนักศึกษา อาจารย์และผู้ที่รับผิดชอบทางด้านการศึกษา จึงควรพยายามทุกวิถีทางในการสร้างข้อสอบที่มีคุณภาพ ที่สามารถวัดความรู้ความสามารถที่แท้จริงของนักศึกษา ซึ่งกระบวนการในการที่จะได้ข้อสอบที่มีคุณภาพ เริ่มตั้งแต่การที่อาจารย์ผู้สอนมีความรู้ความเข้าใจในการเขียนข้อสอบ และการวัดการประเมินผล มีตารางวิเคราะห์หลักสูตร (Test blueprint) ที่ระบุระดับของความรู้ที่คาดหวังในแต่ละหัวข้อเนื้อหา มีการวิพากษ์ข้อสอบก่อนการนำมาใช้ มีการวิเคราะห์ข้อสอบหลังการสอบทุกครั้ง และ นำผลที่ได้จากการวิเคราะห์มาปรับปรุงคุณภาพของข้อสอบ หรือคัดเลือกข้อสอบที่ดีสำหรับใช้ต่อไป ดังที่ Oluseyi & Olufemi

(2012) ได้ให้ข้อเสนอแนะว่าการสร้างข้อสอบมีความสำคัญอย่างยิ่ง และควรทำตารางแสดงผลการวิเคราะห์ข้อสอบไว้ท้ายข้อเพื่อเลือกข้อที่ดีไว้ในธนาคารข้อสอบ รวมทั้งควรวิพากษ์ข้อสอบก่อนนำมาใช้กับนักศึกษาด้วย ส่วนJozefowicz et al. 2002 ได้แสดงทัศนะว่าข้อสอบแบบเลือกตอบ (Multiple choices) ที่ใช้ในการสอบสำคัญ ๆหากสร้างขึ้นอย่างเหมาะสมจะมีความตรงและมีความเชื่อมั่นสูง สามารถวัดพื้นความรู้และความสามารถของผู้สอบได้ อย่างไรก็ตามงานวิจัยหลายเรื่องพบว่าข้อสอบที่สร้างขึ้นเพื่อใช้เองในสถาบันการศึกษาไม่มีคุณภาพดีพอ หากอาจารย์ผู้ออกข้อสอบไม่ได้รับอบรม หรือเรียนรู้เกี่ยวกับหลักของการวัดผล และการเขียนข้อสอบ (Jozefowicz et al. 2002) เช่นเดียวกับ Burton (2001) ที่ให้ความเห็นว่า แบบทดสอบที่มีการให้คะแนนรายข้อเป็น 0 (ศูนย์) และ 1 จะมีคุณภาพหรือไม่สามารถบอกได้จากการหาดัชนีอำนาจจำแนกของข้อสอบแต่ละข้อ ในบทความนี้ผู้เขียนได้ให้ข้อเสนอแนะ ให้คำนวณหาค่าอำนาจจำแนก จากสหสัมพันธ์ของคะแนนแต่ละข้อและคะแนนรวม แต่ควรใช้เป็นเพียงแนวทางในการตัดสินใจ ร่วมกับการพิจารณาถึงเนื้อหาและถ้อยคำที่ใช้ในข้อคำถามแต่ละข้อ ดังนั้น จึงไม่มีอะไรที่จะมาทดแทนการที่ต้องพิจารณา หรือระมัดระวังเรื่องการใช้ถ้อยคำในการเขียนคำถาม และตัวเลือกของข้อสอบแต่ละข้อ

ข้อเสนอแนะในการนำผลวิจัยไปใช้

1. ควรมีการวิเคราะห์ข้อสอบหลังการสอบทุกครั้ง ไม่ว่าจะเป็นข้อสอบที่ใช้ในการวัดและประเมินผลรายวิชา หรือข้อสอบเพื่อคัดเลือกนักศึกษา หากพบว่ามีข้อสอบที่มีค่าอำนาจจำแนกติดลบ ควรดึงข้อสอบข้อนั้น ๆ ออก และประมวลคะแนนใหม่

2. ควรมีการนำผลการวิเคราะห์ข้อสอบมาประกอบในการพิจารณาคัดเลือกข้อสอบที่มีคุณภาพเก็บไว้ในคลังข้อสอบ และปรับปรุงข้อสอบ

3. สำหรับการวัดและประเมินผลรายวิชาที่ยังไม่มีข้อสอบมาตรฐาน อาจารย์ผู้สอนจำเป็นต้องใช้ข้อสอบที่เขียนขึ้นเอง (Teacher-made test) ควรได้เอาใจใส่ในการวิพากษ์ข้อสอบ และพิจารณาเลือกใช้ข้อสอบที่ผ่านการวิเคราะห์แล้วว่ามีคุณภาพ บวกกับข้อสอบที่ปรับปรุงและนำกลับมาใช้อีก และข้อสอบที่สร้างขึ้นใหม่ โดยอาศัยตารางวิเคราะห์หลักสูตร (Test blueprint) เป็นแนวทางในการคัดเลือกข้อสอบที่มีอยู่และในการสร้างข้อสอบขึ้นใหม่

3. ควรให้ความสำคัญกับกระบวนการในการที่จะได้ข้อสอบที่มีคุณภาพทุกขั้นตอน ตั้งแต่การที่อาจารย์ผู้สอนมีความรู้ความเข้าใจในการเขียนข้อสอบและการวัดการประเมินผล มีตารางวิเคราะห์หลักสูตร (Test blueprint) ที่ระบุระดับของความรู้ที่คาดหวังในแต่ละหัวข้อเนื้อหา มีการวิพากษ์ข้อสอบก่อนการนำมาใช้ มีการวิเคราะห์ข้อสอบหลังการสอบทุกครั้ง และนำผลที่ได้จากการวิเคราะห์มาปรับปรุงคุณภาพของข้อสอบ หรือคัดเลือกข้อสอบที่ดีสำหรับใช้ต่อไป

4. ควรมีการกำหนดจำนวนข้อสอบตามตารางวิเคราะห์หลักสูตร (Test blueprint) เผื่อไว้ เมื่อตัดข้อที่ r ติดลบออกจะได้เหลือจำนวนข้อสอบที่วัดได้ครบทุกประเด็น

ข้อเสนอแนะสำหรับการทำวิจัยครั้งต่อไป

1. ในการเก็บรวบรวมข้อมูลข้อสอบรายวิชา ควรมีการวิเคราะห์ว่า จำนวนข้อสอบที่มีค่า r ติดลบ หรือค่า r ต่ำ ๆ จะมีความสัมพันธ์กับการเปลี่ยนแปลงอย่างมีนัยสำคัญทางสถิติของค่าความเชื่อมั่นของข้อสอบทั้งฉบับอย่างไร และมีผลต่อการเปลี่ยนแปลงคะแนนเฉลี่ยและลำดับที่ของนักศึกษาอย่างไร

2. ควรมีการศึกษาติดตามว่าการนำข้อเสนอแนะในการนำผลวิจัยไปใช้จะทำให้เกิดการเปลี่ยนแปลงในคุณภาพข้อสอบแต่ละรายวิชาของสถาบันหรือไม่อย่างไร

หนังสืออ้างอิง

ปวรส บุตะเขียว. (2555). *การวิเคราะห์ข้อสอบ (Item analysis)*. สืบค้นจาก www.rta.mi.th/630a0u/file/item_analysis.doc

พวงรัตน์ ทวีรัตน์. (2540). *วิธีการวิจัยทางพฤติกรรมศาสตร์และสังคมศาสตร์*. พิมพ์ครั้งที่ 7. กรุงเทพฯ: มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

ศิริชัย กาญจนวาสี. (2552). *ทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory)*. พิมพ์ครั้งที่ 6. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สุมาลี จันทร์ชลอ. (2542). *การวัดและประเมินผล*. กรุงเทพฯ: ศูนย์สื่อเสริมกรุงเทพ.

อุทุมพร จามรมาน.(2535). *ข้อสอบ: การสร้างและการพัฒนา*. กรุงเทพฯ: ฟันนี่พับบลิชชิ่ง.

Ascalon M.E., Meyers L.S., Davis B.W. & Smits N. (2007). Distractor Similarity and Item-Stem Structure: Effects on Item Difficulty. *Applied Measurement in Education. 20* (2), 153-170.

Burton, R.F. (2001). Do Item-discrimination Indices Really Help Us to Improve Our Tests. *Assessment & Evaluation in Higher Education. 26*(3), 213-220.

Caldwell D.J. & Pate A.N. (2013). Effects of Question Formats on Student and Item Performance. *American Journal of Pharmaceutical Education.* 77 (4). Article 71: 1-5.

Chase, C. I. (1978). *Measurement for Educational Evaluation*. 2nd ed. Reading, MA: Addison-Wesley Publishing Company.

DeMars, C.E. (2004). Detection of Item Parameter Drift over Multiple Test Administrations. *Applied Measurement in Education. 17*(3), 265-300.

Hamzah, M.S.G. & Abdullah S.K. (2011). Test Item Analysis: An Educator Professionalism Approach. *US-China Education Review* A3, 307-322.

Jozefowicz, RF, Koeppen BM, Case S, Galbrath R, Swanson D, Glew RH. (2002). The quality of in-house medical school examinations. *Acad Med* 77, 156-161.

Lee H, & Winke, P. (2012). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing. 30*(1), 99-123.

Office of Educational Assessment, University of Washington. (2005). *SCOREPAK®*: Item analysis. Retrieved from http://www.washington.edu/oea/score1.htm

Oluseyi, A.E. & Olufemi A.T. (2012). The Analysis of Multiple Choice Item of the Test of an Introductory Course in Chemistry in a Nigerian University. *The International Journal of Learning. 18*(4), 237-246.

Phipps, S.D., Brackbill, M.L. & Dunn, B.J. (2009). Relationship between Assessment Item Format and Items Performance Characteristics. *American Journal of Pharmaceutical Education. 73*(8), 1-7.

Tasdemir, M. (2010). A Comparison of Multiple-Choice Tests and True-False Tests Used in Evaluating Student Progress. *Journal of Instructional Psychology. 37*(3), 258-266.

Tavakol, M. & Dennick R. (2011). Post-examination analysis of objective tests. *Medical Teacher.* 33: 447-458.

Usova, GM. (1997). Effective test item discrimination using Bloom's taxonomy. *Education. 118* (1), 100-110.

Ware, J. & Vik T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher. 31*, 238-243.