



การเปรียบเทียบดัชนีความเป็นคู่ขนานของข้อสอบและแบบสอบที่ได้จากการคัดเลือกข้อสอบระหว่าง
กลุ่มผู้เชี่ยวชาญที่มีลักษณะแตกต่างกัน: การประยุกต์ใช้ทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ

Comparison of Item and Test Parallel Index between Test Forms among
Different Characteristics of experts: Application of Formal Cultural Consensus Theory

ร.อ. อนูปพงษ์ กันธิวงศ์ *

Anupong Kantiwong

รศ.ดร.โชติกา ภาชีผล **

Assoc. Prof. Shotiga Pasiphol Ph.D.

บทคัดย่อ

การวิจัยในครั้งนี้มีจุดประสงค์ 1) เพื่อศึกษาระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญ 2) เพื่อเปรียบเทียบดัชนีความเป็นคู่ขนานของแบบสอบที่ได้จากการคัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีลักษณะแตกต่างกันและ 3) เพื่อเปรียบเทียบดัชนีความเป็นคู่ขนานของข้อสอบที่ได้จากการคัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีลักษณะแตกต่างกัน ตัวอย่างที่ใช้ในการศึกษาค้นคว้าครั้งนี้ คือ ผู้เชี่ยวชาญที่มีประสบการณ์สอนนักศึกษาแพทยศาสตร์แพทยศาสตรพระมงกุฎเกล้าจากการเลือกแบบเจาะจง 100 คน โดยจัดกลุ่มเพื่อคัดเลือกข้อสอบสำหรับสร้างแบบสอบคู่ขนานตามระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบและแบบสอบ เครื่องมือที่ใช้ในการวิจัยประกอบด้วย แบบวัดระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบและข้อสอบสำหรับคัดเลือกเพื่อสร้างแบบสอบคู่ขนานสำหรับกลุ่มผู้เชี่ยวชาญ ผลการวิจัยสรุปได้ดังนี้ 1) ผู้เชี่ยวชาญที่มีประสบการณ์สอนน้อยกว่า 10 ปี ส่วนใหญ่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบในช่วง 0.40-0.59, 0.60-0.79 และ 0.80-1.00 ตามลำดับ แตกต่างจากผู้เชี่ยวชาญที่มีประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไป 0.60-0.79, 0.40-0.59 และ 0.80-1.00 ตามลำดับ 2) แบบสอบคู่ขนานที่สร้างขึ้นจากการคัดเลือกข้อสอบโดยกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบระดับ 0.80-1.00 และมีประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไปมีความเป็นคู่ขนานของแบบสอบสูงกว่ากลุ่มที่อยู่ในระดับต่ำกว่า ($RMSD = 0.283$ [$MSG = 0.020$] และ $MRD = 0.042$ [$MRIG = 0.003$]) และ 3) ในทุกกลุ่มผู้เชี่ยวชาญที่ผ่านเกณฑ์ขั้นต่ำของการตรวจสอบความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบเป็นโดยอาศัยทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบมีค่าเฉลี่ยของ $RMSD$ ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F=0.582$, $p=0.755$)

* นิสิตมหาบัณฑิตสาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

E-mail Address: anupongpcm31@gmail.com

** อาจารย์ประจำสาขาวิชา การวัดและประเมินผลการศึกษา ภาควิชา จิตวิทยาและสถิติ

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

E-mail Address: aimornj@hotmail.com

ISSN 1905-4491

Abstract

The purposes of this research were 1) to determine consensus competence of the expert about item parallelism 2) to comparison of test parallel index between tests form that selected by different characteristics of experts 3) to comparison of item parallel index between tests form that selected by different characteristics of experts. The sample consisted of 100 instructors of Phramongkutklo College of Medicine using purposive sampling for devised the expert group by level of consensus competence. The research instruments were consensus competence test for the expert about the item parallelism and item pool for constructed parallel test forms. The research findings were as follows: 1) Experts with less than 10 years of teaching experience are most likely to have the consensus competence of the expert about item parallelism of the test in the range of 0.40-0.59, 0.60-0.79 and 0.80-1.00, respectively, unlike experts with higher 10 years of teaching experience to have in the range of 0.60-0.79, 0.40-0.59 and 0.80-1.00, respectively. 2) The parallel tests created by the selection of the items by a group of experts with the consensus competence of the expert about item parallelism of the test in the range of 0.80-1.00 and teaching experience above 10 years has the parallelism index of the test more than lower competences (RMSD = 0.283 [MSG = 0.020] and MRD = 0.042 [MRIG = 0.003]) and 3) it was found that all groups of experts who passed the minimum criterion of the consensus competence of the expert about item parallelism of the test were not significantly different of the RMSD means ($F=0.582$, $p=0.755$).

คำสำคัญ: แบบสอบคู่ขนาน / ทฤษฎีฉันทามติทางวัฒนธรรม

KEYWORDS: PARALLEL TEST/ CULTURAL CONSENSUS THEORY

บทนำ

จากพระราชบัญญัติการศึกษาแห่งชาติ ในหมวด 4 มาตรา 26 ได้กำหนดให้สถานศึกษาจัดการประเมินผู้เรียนโดยพิจารณาจากพัฒนาการของผู้เรียน ความประพฤติ การสังเกตพฤติกรรมการเรียน การร่วมกิจกรรมและการทดสอบควบคู่ไปในกระบวนการเรียนการสอนตามความเหมาะสมของแต่ละระดับและรูปแบบการศึกษา แสดงให้เห็นถึงความสำคัญของการวัดและประเมินผลระหว่างเรียน (Formative evaluation) ซึ่งเป็นการตรวจสอบความก้าวหน้าของผู้เรียนว่าบรรลุวัตถุประสงค์การเรียนรู้หรือไม่ เพื่อพิจารณาปรับปรุงข้อบกพร่องโดยการให้ข้อมูลย้อนกลับเพื่อส่งเสริมการเรียนรู้อย่างมีประสิทธิภาพ (โชติกา ภาชีผล, 2559; Hattie, 2009) สอดคล้องกับจุดมุ่งหมายของการวัดและประเมินผลเพื่อเปรียบเทียบ ซึ่งเป็นการวัดและประเมินผลเพื่อเปรียบเทียบพัฒนาการของผู้เรียน ว่ามีความเจริญงอกงามเพียงใด โดยเปรียบเทียบเป็นระยะๆ คนละช่วงเวลา อาจจะเป็นการวัดและประเมินผลช่วงก่อนเรียน ระหว่างเรียน หรือหลังเรียน แต่ต้องคำนึงถึงเนื้อหาที่วัดเพื่อเปรียบเทียบพัฒนาการของแต่ละบุคคลควรเป็นเนื้อหาเดียวกัน (โชติกา ภาชีผล, 2559; Gronlund, 2009)

แบบสอบคู่ขนาน ได้ถูกนำมาใช้ในการสอบต่างๆ อย่างแพร่หลาย โดยมีเป้าหมายให้แบบสอบทุกฉบับมีความเป็นคู่ขนานกันทั้งในด้านเนื้อหาและการตรวจสอบโดยอาศัยค่าทางสถิติ ทั้งนี้การออกแบบการสร้างแบบสอบคู่ขนาน จะต้องมีการกำหนดรูปแบบของแบบสอบ สร้างแผนผังการสอบ (Test blueprint) และสร้างตารางวิเคราะห์ข้อสอบ (Table of specifications) ซึ่งเป็นขั้นตอนที่สำคัญที่ทำให้ได้ข้อสอบที่วัดได้ตรงตามจุดประสงค์และพฤติกรรมการเรียนรู้ และสามารถใช้เป็นแนวทางในการพัฒนาแบบสอบคู่ขนานได้ (โชติกา ภาชีผล, 2559) แบบสอบคู่ขนาน ได้ถูกนำมาใช้ในการสอบทั้งการสอบขนาดใหญ่ ที่มีการสอบหลายครั้ง เช่น การทดสอบวิชาความถนัดทั่วไป (General Aptitude Test) การทดสอบวิชาความถนัดทางวิชาการและ

วิชาชีพ (Professional and Academic Aptitude Test) ในวิชาชีพแพทยมีการใช้แบบสอบคู่ขนานในการสอบเพื่อประเมินและรับรองความรู้ความสามารถในการประกอบวิชาชีพเวชกรรม สอดคล้องกับการใช้แบบสอบคู่ขนานในการสอบประเมินเพื่อรับใบอนุญาตเป็นผู้ประกอบวิชาชีพเวชกรรมของประเทศแคนาดาซึ่งประกอบด้วยข้อสอบจำนวน 168 ข้อและสร้างข้อสอบคู่ขนานจากคลังข้อสอบโดยอาศัยทฤษฎีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ นอกจากนี้ มีการใช้แบบสอบคู่ขนานเพื่อการประเมินผลความก้าวหน้าและประเมินผลรวบยอดในรายวิชาต่างๆ รวมถึงแบบสอบก่อนและหลังเรียนในหลายรายวิชา เช่น เกสซ์วิทยา (Keikesers, 2014) และสรีรวิทยา (Hettiaratchi, 1978) เป็นต้น โดยพบว่า การตรวจสอบความเป็นคู่ขนานของแบบสอบทั้งฉบับโดยอาศัยทฤษฎีการทดสอบแบบดั้งเดิมมีหลักการว่า ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของข้อสอบคู่ขนานแต่ละคู่มิค่าใกล้เคียงกัน แต่ค่าเฉลี่ยรวมของแบบสอบคู่ขนานทั้งฉบับอาจจะแตกต่างกัน ซึ่งเกิดจากข้อจำกัดของคะแนนความคลาดเคลื่อนและค่าพารามิเตอร์ของข้อสอบและแบบสอบมีลักษณะเฉพาะขึ้นอยู่กับกลุ่มผู้ตอบและเปลี่ยนไปตามการสอบแต่ละครั้ง (ศิริชัย กาญจนวาสี, 2555)

การสร้างแบบสอบคู่ขนาน แบ่งตามการบริหารการสอบได้เป็น 2 รูปแบบ รูปแบบที่ 1 การสร้างแบบสอบคู่ขนานหลังการสอบ เป็นการสร้างข้อสอบจากคลังข้อสอบที่มีการเก็บรวบรวมข้อมูลค่าพารามิเตอร์หรือค่าสถิติของข้อสอบรายข้อจากการสอบครั้งก่อนเพื่อใช้เป็นเงื่อนไขหรือสารสนเทศในการคัดเลือกข้อสอบเพื่อสร้างแบบสอบคู่ขนาน สามารถสร้างแบบสอบคู่ขนานโดยการใช้โปรแกรมคอมพิวเตอร์เพื่อสร้างชุดข้อสอบตามเงื่อนไขที่กำหนดขึ้นตามค่าสถิติที่ใช้ตรวจสอบความเป็นคู่ขนาน โดยกำหนดเนื้อหา จำนวนข้อสอบ และโครงสร้างในการวัดของข้อสอบตามผังข้อสอบ และทำการตรวจสอบความเป็นคู่ขนานที่สร้างขึ้นโดยอาศัยทฤษฎีการทดสอบแบบดั้งเดิมหรือทฤษฎีการตอบสนองข้อสอบ ข้อดีของแบบสอบคู่ขนานหลังการสอบ คือ มีการตรวจสอบความเป็นคู่ขนานโดยอาศัยค่าทางสถิติ ทำให้มีความน่าเชื่อถือและมีความเป็นคู่ขนานของข้อสอบและแบบสอบตามเกณฑ์ความเป็นคู่ขนานที่ชัดเจน สามารถสร้างชุดแบบสอบคู่ขนานได้จำนวนหลายชุดขึ้นกับขนาดของคลังข้อสอบและไม่ขึ้นกับการตัดสินใจโดยผู้เชี่ยวชาญ แต่ข้อเสียของการสร้างแบบสอบคู่ขนานวิธีนี้ คือ ต้องมีคลังข้อสอบและการวิเคราะห์ข้อสอบทางสถิติเพื่อหาค่าพารามิเตอร์ โดยต้องมีจำนวนผู้สอบเพียงพอสำหรับตรวจสอบความเป็นคู่ขนานโดยเฉพาะการหาค่าพารามิเตอร์และการตรวจสอบความเป็นคู่ขนานของแบบสอบโดยอาศัยทฤษฎีการตอบสนองข้อสอบ นอกจากนี้การประยุกต์ใช้ในสภาพจริงค่อนข้างยากเนื่องจากต้องอาศัยโปรแกรมคอมพิวเตอร์ที่ต้องการความเชี่ยวชาญในการใช้งานโปรแกรมสร้างชุดข้อสอบ รูปแบบที่ 2 การสร้างแบบสอบคู่ขนานก่อนการสอบ เป็นการสร้างแบบสอบคู่ขนานโดยผู้สอบเขียนข้อคำถามและตัวเลือกตามแผนผังการสอบและตารางวิเคราะห์ข้อสอบ ที่กำหนดจุดประสงค์การเรียนรู้และโครงสร้างการวัดเหมือนกัน และตรวจสอบความเป็นคู่ขนานโดยอาศัยผู้เชี่ยวชาญ ทั้งนี้สามารถตรวจสอบความเป็นคู่ขนานของแบบสอบรายข้อโดยประยุกต์ใช้การตรวจสอบความตรงเชิงเนื้อหา ซึ่งเป็นวิธีที่ให้ผู้เชี่ยวชาญตัดสินใจข้อคำถามที่สร้างขึ้นว่าเป็นไปตามเนื้อหาละวัตถุประสงค์ที่ต้องการวัดหรือไม่ มีความสอดคล้องของข้อคำถามกับเนื้อหาและพฤติกรรมการเรียนรู้ และพิจารณาถึงความเป็นคู่ขนานของแบบสอบเป็นรายข้อคำถาม ข้อดีของการสร้างแบบสอบคู่ขนานก่อนการสอบและการตรวจสอบโดยผู้เชี่ยวชาญ คือ ในการสร้างข้อสอบในสภาพจริงได้ง่ายเนื่องจากใช้ผู้เชี่ยวชาญที่มีอยู่ในสาขาที่เชี่ยวชาญนั้นๆ และสามารถสร้างข้อสอบใหม่ได้ตามโครงสร้างและจุดประสงค์ที่กำหนด โดยไม่ต้องอาศัยค่าสถิติหรือพารามิเตอร์ของข้อสอบเดิมจากคลังข้อสอบ แต่ข้อเสียของวิธีนี้ คือ การพิจารณาความเป็นคู่ขนานของแบบสอบโดยผู้เชี่ยวชาญ ขึ้นอยู่กับระดับความสามารถในการตัดสินใจรายบุคคลของผู้เชี่ยวชาญที่แตกต่างกัน (โชติกา ภาชีผล, 2559) รวมถึงประสบการณ์และความเชี่ยวชาญในการออกข้อสอบ คัดเลือกข้อสอบ รวมถึงการตัดสินใจตรงตามเนื้อหา

ของข้อสอบตามจุดประสงค์การเรียนรู้รายวิชานั้นของแต่ละผู้เชี่ยวชาญด้วย (Weller, 2007; Romney, Batchelder and Weller, 1986)

การตรวจสอบความเป็นคู่ขนานของแบบสอบโดยอาศัยข้อตกลงเบื้องต้นของทฤษฎีการทดสอบแบบดั้งเดิม พบว่าแบบสอบคู่ขนาน 2 ฉบับจะต้องมีคะแนนจริงของผู้สอบแต่ละคนและค่าความคลาดเคลื่อนมาตรฐานของประชากรที่สอบทั้ง 2 ฉบับเท่ากัน (Lord & Novik, 1968; Lord, 1980) โดยคะแนนจริง มีค่าเท่ากับคะแนนเฉลี่ยของคะแนนที่ได้จากการวัดซ้ำหลายๆ ครั้ง ซึ่งเป็นการพิจารณาค่าสถิติหลังจากการสอบ (ค่าเฉลี่ยรวม ค่าความคลาดเคลื่อนมาตรฐานและค่าความเที่ยงของแบบสอบ) หากจะนำแบบสอบคู่ขนานนี้ไปใช้ในการสอบครั้งต่อไป ต้องใช้แบบสอบซ้ำและความเที่ยงของแบบสอบคู่ขนานอาจเปลี่ยนไปตามประชากรที่แตกต่างกัน ซึ่งเป็นจุดอ่อนของการตรวจสอบความเป็นคู่ขนานของแบบสอบโดยอาศัยทฤษฎีการทดสอบแบบดั้งเดิม ซึ่งขาดสารสนเทศในการพิจารณาความเป็นคู่ขนานของข้อสอบรายข้อด้วย (Doran, 2007) ต่อมาจึงมีการนำทฤษฎีการตอบสนองข้อสอบมาประยุกต์ใช้ในการสร้างแบบสอบคู่ขนานโดยอาศัยค่าพารามิเตอร์ของข้อสอบที่ได้จากการสอบ โดยสามารถจัดชุดแบบสอบคู่ขนานโดยกำหนดเงื่อนไขของพารามิเตอร์ข้อสอบที่แตกต่างกัน (Lin, 2008; Ali & Rijn, 2016) และตรวจสอบความเป็นคู่ขนานของแบบสอบ โดยอาศัยโค้งลักษณะข้อสอบ (Item characteristic curve; ICC) พบว่า แบบสอบคู่ขนาน 2 ฉบับจะต้องมีความน่าจะเป็นที่ผู้ตอบข้อสอบถูก ($P_i(\theta)$) เท่ากันในทุกความสามารถที่มีอยู่ในผู้ตอบ (θ) และมีรูปแบบการกระจายไม่แตกต่างกัน (Linden & Luecht, 1998; Spray & Miller, 1992) นอกจากนี้ มีการใช้ฟังก์ชันสารสนเทศของแบบสอบ (Test information; TIF) โดยแบบสอบคู่ขนานจะต้องมีฟังก์ชันสารสนเทศของแบบสอบเหมือนกัน (Linden & Adema, 1998; Veldkamp, Matteucci & Jong, 2013) สอดคล้องกับการศึกษาของ Sanders & Verschoor (1998) และ Ali and Rijn (2016) พบว่า การวิเคราะห์ดัชนีความเป็นคู่ขนานของแบบสอบโดยอาศัยทฤษฎีการตอบสนองข้อสอบ จะพิจารณาจากฟังก์ชันสารสนเทศของแบบสอบ โค้งลักษณะของแบบสอบซึ่งให้สารสนเทศที่ถูกต้องและน่าเชื่อถือมากขึ้น

จากการศึกษาของ Wanichsan (2012) และ Panjaburee, Hwang, Triampo and Shih (2010) พบว่าการสร้างแบบสอบโดยกลุ่มผู้เชี่ยวชาญ สามารถเพิ่มระดับความสัมพันธ์ระหว่างเนื้อหาหรือจุดประสงค์การเรียนรู้กับแบบสอบได้ดีขึ้น ซึ่งลดจุดอ่อนของการสร้างแบบสอบโดยใช้โปรแกรมคอมพิวเตอร์เนื่องจากมีความซับซ้อนของรายละเอียดในแต่ละเนื้อหา ทำให้การประมาณค่าความสามารถของผู้สอบมีความตรงมากขึ้น และสามารถให้ข้อมูลย้อนกลับได้อย่างมีประสิทธิภาพสำหรับข้อสอบวินิจฉัยที่ต้องการความสัมพันธ์ระหว่างข้อสอบกับเนื้อหาค่อนข้างสูง สอดคล้องกับสภาพจริงที่การคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างเป็นแบบสอบคู่ขนาน จะใช้กลุ่มที่ประกอบด้วยผู้เชี่ยวชาญอย่างน้อย 3 คน โดยพิจารณาจากเกณฑ์มาตรฐานจุดประสงค์การเรียนรู้ และเนื้อหาที่กำหนดตามแผนผังการสอบและตารางวิเคราะห์ข้อสอบเพื่อให้แบบสอบที่สร้างขึ้นมีความเที่ยงและความตรงสูง แต่พบว่า ปัญหาของการตัดสินใจโดยอาศัยผู้เชี่ยวชาญ คือ ระดับความสามารถในการตัดสินใจของผู้เชี่ยวชาญแต่ละคนมีความแตกต่างกัน จึงส่งผลให้ความตรงและความถูกต้องของผลการตัดสินใจที่เบี่ยงเบนไปจากค่าที่แท้จริง (Weller, 2007; Romney, Batchelder & Weller, 1986) จึงเป็นที่มาของการนำทฤษฎีฉันทามติทางวัฒนธรรม (Cultural consensus theory) มาประยุกต์ใช้เพื่อตรวจสอบความสามารถในการเห็นพ้องร่วมกันของผู้เชี่ยวชาญ ทั้งนี้ ฉันทามติเป็นกระบวนการตัดสินใจที่กลุ่มคนต้องการให้เกิดการมีส่วนร่วมและความเห็นพ้องร่วมกันมากที่สุด การสรุปผลการตัดสินใจขึ้นกับข้อกำหนดที่ตกลงกันได้ ตามน้ำหนักและขอบเขตของการเห็นพ้องร่วมกัน โดยมีการประยุกต์ใช้ทฤษฎีฉันทามติทางวัฒนธรรม สำหรับวิเคราะห์ผลจากการตัดสินใจหรือการลงฉันทามติ เพื่อประมาณค่าระดับความสามารถในการ

เห็นพ้องร่วมกันของแต่ละบุคคลว่ามีความสามารถในการตัดสินใจในเรื่องนั้นๆ มากน้อยเพียงใด โดยการหาความสอดคล้องของคำตอบระหว่างผู้เชี่ยวชาญในกลุ่มของแต่ละข้อคำถาม ซึ่งเป็นประมาณการระดับความสามารถในการเห็นพ้องร่วมกันว่าคำตอบของผู้เชี่ยวชาญมีความถูกต้องมากน้อยเพียงใด และสามารถตรวจสอบคำตอบที่ถูกต้องของข้อคำถามนั้นๆ โดยอาศัยทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ ทำให้เพิ่มความเชื่อมั่นของการลงฉันทามติโดยผู้เชี่ยวชาญมีความน่าเชื่อถือมากขึ้นและสามารถถ่วงน้ำหนักระดับความสามารถในการเห็นพ้องร่วมกันของผู้เชี่ยวชาญแต่ละคนในกลุ่มได้

โมเดลฉันทามติทางวัฒนธรรมแบ่งตามรูปแบบการประมาณค่าความสามารถในการเห็นพ้องร่วมกันได้ 2 โมเดล (Romney, 1986) ได้แก่ 1. โมเดลฉันทามติทางวัฒนธรรมแบบไม่มีรูปแบบ (Informal cultural consensus model) เป็นวิธีการวิเคราะห์เพื่อระบุระดับความสอดคล้องของการตอบข้อคำถามของผู้เชี่ยวชาญแต่ละคนกับฉันทามติของกลุ่ม แต่จะไม่เน้นการหาคำตอบที่ถูกต้องของข้อคำถาม แต่จะเน้นความสอดคล้องของฉันทามติของบุคคลต่อฉันทามติของกลุ่ม เพราะมีแนวคิดที่ว่าฉันทามติเกิดขึ้นมาจากข้อคำถามหรือเนื้อหาที่ยังไม่ทราบคำตอบที่แน่ชัดและคำตอบที่แท้จริงของข้อคำถามทางสังคมวิทยามักจะแตกต่างกันตามความหลากหลายทางวัฒนธรรม ดังนั้น จึงเป็นโมเดลที่เน้นการหาค่าถ่วงน้ำหนักฉันทามติของผู้เชี่ยวชาญต่อฉันทามติของกลุ่มว่าควรถ่วงน้ำหนักแต่ละมากน้อยเพียงใดตามความสามารถในการเห็นพ้องร่วมกัน 2. โมเดลฉันทามติทางวัฒนธรรมแบบมีรูปแบบ (Formal cultural consensus model) เป็นวิธีการประมาณค่าความสามารถในการเห็นพ้องร่วมกันของแต่ละผู้เชี่ยวชาญ โดยได้จากความเห็นพ้องร่วมกันของคำตอบแต่ละข้อคำถามจากฉันทามติของผู้เชี่ยวชาญ โดยอาศัยเมทริกซ์ความสัมพันธ์หรือเมทริกซ์ของความแปรปรวนร่วมของฉันทามติของผู้เชี่ยวชาญทุกคู่ในกลุ่มผู้เชี่ยวชาญทั้งหมดที่ลงฉันทามติหรือตอบข้อคำถามนั้นๆ และสามารถคำนวณความน่าจะเป็นของความถูกต้องสำหรับคำตอบที่ได้จากฉันทามติของผู้เชี่ยวชาญโดยอาศัยทฤษฎีสถิติของเบย์มาใช้ในการคำนวณ ภายใต้ข้อตกลงเบื้องต้นอย่างเคร่งครัดตามทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ คือ ผู้เชี่ยวชาญแต่ละคนต้องเป็นอิสระต่อกัน ไม่มีการปรึกษากันและต้องไม่ใช้การสัมภาษณ์กลุ่มย่อย ข้อคำถามที่ใช้เป็นเอกมิตี โดยต้องมีความเท่าเทียมกันและเป็นข้อคำถามที่คำตอบมีความสอดคล้องกันสูง จึงจะสามารถนำมาใช้ตรวจสอบความสามารถในการเห็นพ้องร่วมกันของผู้เชี่ยวชาญต่อความเป็นคู่ขนานของแบบสอบถามได้

ผู้วิจัยจึงมีความสนใจศึกษาเปรียบเทียบดัชนีความเป็นคู่ขนานของแบบสอบถามคู่ขนานที่ระดับความสามารถในการตัดสินใจความเป็นคู่ขนานของข้อสอบถามของผู้เชี่ยวชาญและประสบการณ์สอนที่แตกต่างกัน เพื่อให้ได้แบบสอบถามที่มีคุณภาพสูงและนำไปใช้ได้จริง การวิจัยครั้งนี้จึงใช้วิธีการสร้างแบบสอบถาม โดยอาศัยผู้เชี่ยวชาญในการคัดเลือกข้อสอบถามที่มีตัวแปรอิสระ คือ ระดับความสามารถในการตัดสินใจความเป็นคู่ขนานของข้อสอบถามของผู้เชี่ยวชาญเพื่อตรวจสอบระดับความสามารถของผู้เชี่ยวชาญ มุ่งเน้นลดจุดอ่อนในเรื่องความเป็นปัจเจกบุคคลและระดับความถูกต้องในการตัดสินใจที่แตกต่างกันของผู้เชี่ยวชาญแต่ละบุคคล และประสบการณ์สอนเพื่อเป็นสารสนเทศในการตัดสินใจคัดเลือกผู้เชี่ยวชาญเพื่อสร้างแบบสอบถาม ซึ่งการใช้การคัดเลือกข้อสอบถามคู่ขนานโดยกลุ่มผู้เชี่ยวชาญสามารถนำไปประยุกต์ใช้ในการสร้างแบบสอบถามในสภาพจริงได้ สามารถตรวจสอบความเป็นคู่ขนานของแบบสอบถามเมื่อนำไปใช้ในการสอบโดยอาศัยการทดสอบทางสถิติและนำข้อสอบไปใช้ในการสร้างคลังข้อสอบสำหรับคัดเลือกข้อสอบเพื่อสร้างแบบสอบถามที่มีคุณภาพสูงในครั้งต่อไป

วัตถุประสงค์

1. เพื่อศึกษาระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญ
2. เพื่อเปรียบเทียบดัชนีความเป็นคู่ขนานของแบบสอบคู่ขนานที่ได้จากการคัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบและประสบการณ์สอนที่แตกต่างกัน
3. เพื่อเปรียบเทียบดัชนีความเป็นคู่ขนานของข้อสอบคู่ขนานที่ได้จากการคัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบและประสบการณ์สอนที่แตกต่างกัน

วิธีดำเนินการวิจัย

การดำเนินการวิจัยเพื่อเปรียบเทียบดัชนีความเป็นคู่ขนานของข้อสอบและแบบสอบที่ได้จากการคัดเลือกข้อสอบโดยกลุ่มผู้เชี่ยวชาญที่มีลักษณะแตกต่างกันแบ่งเป็น 3 ส่วน ได้แก่

ส่วนที่ 1 ศึกษาความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญ เป็นการประมาณค่าระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบซึ่งประยุกต์ใช้ทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบและการวิเคราะห์ฉันทามติ แหล่งข้อมูลสำหรับการวิจัยในส่วนนี้เป็นตัวอย่างผู้เชี่ยวชาญที่มีประสบการณ์สอนนักศึกษาแพทย์ วิทยาลัยแพทยศาสตร์พระมงกุฎเกล้าจำนวน 63 คนที่ได้จากการเลือกแบบเจาะจงและตอบรับในการให้ความร่วมมือในการเก็บรวบรวมข้อมูล เครื่องมือที่ใช้ ได้แก่ แบบวัดระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบที่ประกอบด้วย ข้อสอบจำนวน 20 คู่ โดยให้ผู้เชี่ยวชาญพิจารณาตัดสินใจเป็นคู่ขนานของข้อสอบแต่ละข้ออย่างอิสระโดยไม่ถามความคิดเห็นระหว่างกัน และไม่มีการประชุมกลุ่มย่อยตามข้อตกลงเบื้องต้นของทฤษฎีฉันทามติทางวัฒนธรรม การวิเคราะห์ข้อมูลแบ่งออกเป็น การตรวจสอบระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญโดยอาศัยทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ โดยตรวจสอบข้อตกลงเบื้องต้นของทฤษฎีฉันทามติทางวัฒนธรรม 3 ด้าน คือ ความเป็นเอกมิติของแบบวัด ความเป็นอิสระต่อกันของผู้ตอบและความถูกต้อง และประมาณค่าความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญจากการวิเคราะห์ฉันทามติโดยอาศัยโปรแกรม UCINET เพื่อแบ่งกลุ่มผู้เชี่ยวชาญจำแนกตามระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบและประสบการณ์สอนนักศึกษาแพทย์

ส่วนที่ 2 เปรียบเทียบดัชนีความเป็นคู่ขนานของแบบสอบคู่ขนานที่ได้จากการคัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบและประสบการณ์สอนที่แตกต่างกัน ตัวอย่างที่ใช้ คือ ผู้เชี่ยวชาญจำนวน 18 คน แบ่งออกเป็น 6 กลุ่มละ 3 คนตามระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญ (ระดับที่ 1, 2 และ 3 โดยมีค่าความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบในช่วง 0.80-1.00, 0.60-0.79 และ 0.40-0.59 ตามลำดับ) และประสบการณ์สอนนักศึกษาแพทย์ (ตั้งแต่ 10 ปีขึ้นไปและน้อยกว่า 10 ปี) โดยให้กลุ่มผู้เชี่ยวชาญคัดเลือกข้อสอบเพื่อสร้างแบบสอบคู่ขนาน 1 ชุดละ 2 แบบสอบละ 25 ข้อ โดยใช้เนื้อหาการเลือกใช้ใน 3 ระบบ ประกอบด้วย ยา ในระบบหลอดเลือดและหัวใจ ระบบทางเดินอาหาร และระบบต่อมไร้ท่อ โดยจำแนกเป็น 5 จุดประสงค์ละ 5 ข้อ ซึ่งในแต่ละจุดประสงค์ประกอบด้วยข้อสอบจำนวนทั้งหมด 40 ข้อ กลุ่มผู้ให้ข้อมูลสำหรับตรวจสอบความเป็นคู่ขนานของแบบสอบและข้อสอบ ได้แก่ นักศึกษาแพทย์วิทยาลัยแพทยศาสตร์พระมงกุฎเกล้า ชั้นปีที่ 3-6 ปีการศึกษา 2560 จำนวน 400 คน โดยทำการสอบแบบสอบคู่ขนานละ 200 คน วิเคราะห์ข้อสอบโดยใช้

โปรแกรม IRTpro เพื่อหาค่าสารสนเทศของแบบสอบและพารามิเตอร์ของข้อสอบโดยอาศัยโมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ และตรวจสอบความสอดคล้องกลมกลืนของข้อสอบกับโมเดล 2PL โดยอาศัยสถิติทดสอบไคสแควร์ และเปรียบเทียบค่าดัชนีความเป็นคู่ขนานของแบบสอบที่สร้างขึ้น โดยพิจารณาค่า $RMSD_{TIF}$ และ MRD จากนั้นจึงเปรียบเทียบค่าเฉลี่ยของ $RMSD_{TIF}$ และ MRD โดยใช้สถิติทดสอบ one-way ANOVA มีรายละเอียดดังต่อไปนี้

1) ดัชนีความเป็นคู่ขนานของแบบสอบโดยอาศัยค่า $RMSD_{TIF}$ หมายถึง ค่าสถิติที่บ่งชี้ว่าแบบสอบคู่ขนาน 2 ฉบับมีส่วนเบี่ยงเบนค่าเฉลี่ยกำลังสองของสารสนเทศของแบบสอบ (Root mean square deviation of test information; $RMSD_{TIF}$) สำหรับตรวจสอบการรูปแบบและการกระจายของฟังก์ชันสารสนเทศของแบบสอบ 2 ฉบับที่วัดโครงสร้างทางเนื้อหาเดียวกันในระดับความสามารถของผู้ตอบข้อสอบทุกระดับแตกต่างกันไม่เกิน 0.50 ซึ่งคำนวณได้จากสมการ ดังต่อไปนี้

$$RMSD_{TIF} = \sqrt{\frac{\sum_{j=1}^q [I(\theta_j) - I'(\theta_j)]^2}{q}}$$

เมื่อ q = จำนวนระดับความสามารถที่ใช้ในการตรวจสอบฟังก์ชัน

$I(\theta)$ = ค่าสารสนเทศของแบบสอบที่ระดับความสามารถของผู้สอบ θ

ทั้งนี้ การพิจารณาดัชนีความเป็นคู่ขนานของแบบสอบโดยอาศัยค่า $RMSD_{TIF}$ ควรพิจารณาร่วมกับค่าผลต่างระหว่างสารสนเทศของแบบสอบสูงสุดที่ระดับความสามารถของผู้สอบเดียวกัน (Maximum score gap; MSG) โดยค่า MSG ของแบบสอบคู่ขนานจะไม่เกิน 0.50 ซึ่งคำนวณได้จากสมการ ดังต่อไปนี้

$$MSG = \max(\theta) \{ \max I(\theta) - \min I'(\theta) \}$$

2) ดัชนีความเป็นคู่ขนานของแบบสอบโดยอาศัยค่า MRD หมายถึง ค่าสถิติที่บ่งชี้ว่าแบบสอบคู่ขนาน 2 ฉบับมีค่าเฉลี่ยสัดส่วนที่แตกต่างกันของสารสนเทศของแบบสอบ (MRD) สำหรับตรวจสอบรูปแบบและการกระจายของฟังก์ชันสารสนเทศของแบบสอบ 2 ฉบับที่วัดโครงสร้างทางเนื้อหาเดียวกันในระดับความสามารถของผู้ตอบข้อสอบทุกระดับแตกต่างกันไม่เกิน 0.10 ซึ่งคำนวณได้จากสมการ ดังต่อไปนี้

$$MRD = \frac{\sum_{j=1}^q \left| \frac{I(\theta_j) - I'(\theta_j)}{I'(\theta_j)} \right|}{q}$$

ทั้งนี้ การพิจารณาดัชนีความเป็นคู่ขนานของแบบสอบโดยอาศัยค่า MRD ควรพิจารณาร่วมกับค่าผลต่างสัมพัทธ์ระหว่างสารสนเทศของแบบสอบสูงสุดที่ระดับความสามารถของผู้สอบเดียวกัน (Maximum relative information gap; MRIG) โดยค่า MRIG ของแบบสอบคู่ขนานจะไม่เกิน 0.10 ซึ่งคำนวณได้ดังนี้

$$MRIG = \max(\theta) \left\{ \frac{\max I(\theta) - \min I'(\theta)}{I'(\theta)} \right\}$$

ส่วนที่ 3 เปรียบเทียบดัชนีความเป็นคู่ขนานของข้อสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบและประสบการณ์สอนนักศึกษาแพทย์ที่แตกต่างกัน โดยพิจารณาค่า $RMSD_{IF}$ และ ค่าเฉลี่ย $RMSD_{IF}$ ซึ่งเป็นค่าสถิติที่บ่งชี้ว่าข้อสอบคู่ขนาน 2 ข้อมีส่วนเบี่ยงเบนค่าเฉลี่ยกำลังสองของสารสนเทศของข้อสอบ สำหรับตรวจสอบการรูปแบบและการกระจายของฟังก์ชันสารสนเทศของข้อสอบ 2 ข้อที่วัดโครงสร้างทางเนื้อหาเดียวกันในระดับความสามารถของผู้ตอบข้อสอบทุกระดับ โดยกำหนดให้ข้อสอบคู่ขนานจะต้องมีค่า $RMSD_{IF}$ แตกต่างกันไปไม่เกิน 0.50 ทั้งนี้สามารถคำนวณได้จากสมการดังต่อไปนี้

$$RMSD_{IF} = \sqrt{\frac{\sum_{j=1}^q [i(\theta_j) - i'(\theta_j)]^2}{q}}$$

เมื่อ $i(\theta)$ = ค่าสารสนเทศของข้อสอบที่ระดับความสามารถของผู้สอบ θ

ผลการวิจัย

การตรวจสอบระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญ โดยใช้แบบวัดระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญ พบว่า แบบวัดนี้มีความเป็นอิสระต่อกันของผู้ตอบ (Conditional independence) และความถูกต้อง (Common truth) ตามข้อตกลงเบื้องต้นของทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ แต่ไม่เป็นไปตามข้อตกลงเบื้องต้นในด้านความเป็นเอกมิตินี้ของแบบวัด (Homogeneity) โดยพบว่า สัดส่วนของ Eigenvalue สูงสุดต่อลำดับที่ 2 พบว่า มีสัดส่วนน้อยกว่า 3 เท่า จึงพิจารณาตัดผู้เชี่ยวชาญที่มีความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบในระดับต่ำออกเพราะจะทำให้การประมาณค่าระดับความสามารถของผู้เชี่ยวชาญมีความคลาดเคลื่อน ผู้เชี่ยวชาญจะต้องมีความสามารถในการตัดสินใจไปในทิศทางเดียวกันและมีความสอดคล้องกันสูงในกลุ่ม เมื่อวิเคราะห์ฉันทามติโดยไม่พิจารณาผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบต่ำแล้ว พบว่าแบบวัดระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญมีคุณสมบัติด้านความเป็นเอกมิตินี้ของแบบวัดเป็นไปตามข้อตกลงเบื้องต้นของทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ พบว่าผู้เชี่ยวชาญส่วนใหญ่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบอยู่ในระดับที่ 3 มากกว่าระดับที่ 2 และ 1 ตามลำดับ ทั้งนี้จำนวนตัวอย่างผู้เชี่ยวชาญในแต่ละกลุ่มเพียงพอสำหรับการจำแนกผู้เชี่ยวชาญสำหรับการคัดเลือกข้อสอบเพื่อสร้างแบบสอบคู่ขนานได้ นอกจากนี้พบว่า ระดับความสามารถในการตรวจสอบความเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญมีค่าสูงสุดเท่ากับ 0.942 และ ค่าต่ำสุดเท่ากับ 0.401 โดยมีค่าเฉลี่ยเท่ากับ 0.637 และพิจารณาจำแนกกลุ่มผู้เชี่ยวชาญตามประสบการณ์สอนออกเป็น 2 กลุ่ม ได้แก่ ประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไปและน้อยกว่า 10 ปี จึงสามารถแบ่งกลุ่มผู้เชี่ยวชาญที่มีลักษณะแตกต่างกันเป็น 6 กลุ่ม โดยพบว่า ผู้เชี่ยวชาญส่วนใหญ่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบอยู่ในระดับที่ 3 มากกว่าระดับที่ 2 และ 1 โดยเฉพาะในกลุ่มผู้เชี่ยวชาญที่มีประสบการณ์สอนนักศึกษาแพทย์น้อยกว่า 10 ปี แต่ในกลุ่มผู้เชี่ยวชาญที่มีประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไปจะมีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบในระดับที่ 2 จำนวนสูงสุด ดังแสดงในตาราง 1

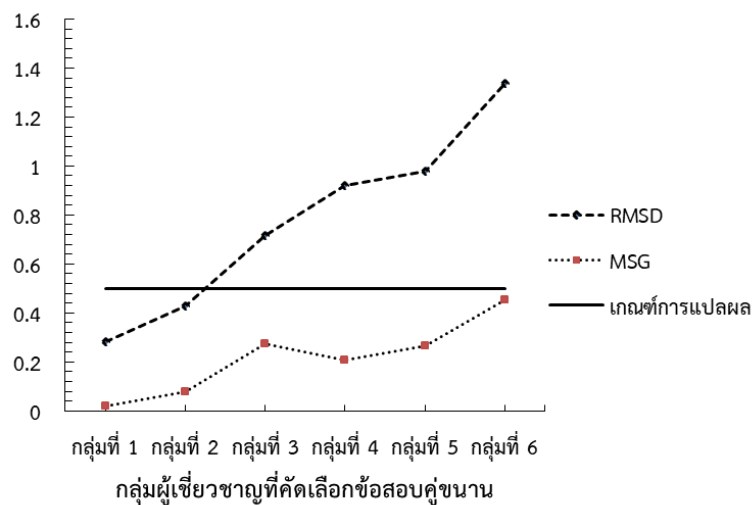
ตาราง 1 การแบ่งกลุ่มผู้เชี่ยวชาญจำแนกตามระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบ

ลักษณะของผู้เชี่ยวชาญ		ระดับความสามารถในการตัดสิน ความเป็นคู่ขนานของข้อสอบ			รวม
		ระดับที่ 1 (0.80-1.00)	ระดับที่ 2 (0.60-0.79)	ระดับที่ 3 (0.40-0.59)	
ประสบการณ์สอน	ตั้งแต่ 10 ปีขึ้นไป	4	10	5	19
	น้อยกว่า 10 ปี	3	6	15	24
รวม		7	16	20	43

เมื่อกลุ่มผู้เชี่ยวชาญแต่ละกลุ่มทำการคัดเลือกข้อสอบเพื่อสร้างแบบสอบคู่ขนานแล้วจึงดำเนินการเก็บข้อมูลกับกลุ่มนักศึกษาแพทย์ผู้ให้ข้อมูล จากนั้นจึงวิเคราะห์ทางสถิติโดยใช้โปรแกรม IRTpro เพื่อเปรียบเทียบดัชนีความเป็นคู่ขนานของแบบสอบ พบว่า เมื่อกำหนดให้ดัชนีความเป็นคู่ขนานของแบบสอบโดยใช้ RMSD และ MSG สำหรับแบบสอบคู่ขนานต้องมีค่า RMSD และ MSG ต่ำกว่า 0.50 พบว่า แบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบในระดับที่ 1 ทั้งกลุ่มที่มีประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไป (กลุ่มที่ 1) และประสบการณ์สอนน้อยกว่า 10 ปี (กลุ่มที่ 2) มีความเป็นคู่ขนานของแบบสอบ ส่วนแบบสอบที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญกลุ่มอื่นๆ ไม่มีความเป็นคู่ขนานของแบบสอบ โดยมีแนวโน้มของค่า RMSD และ MSG สูงขึ้นตามระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบของกลุ่มผู้เชี่ยวชาญที่ลดลงและประสบการณ์สอนน้อยกว่า 10 ปี (ภาพที่ 1)

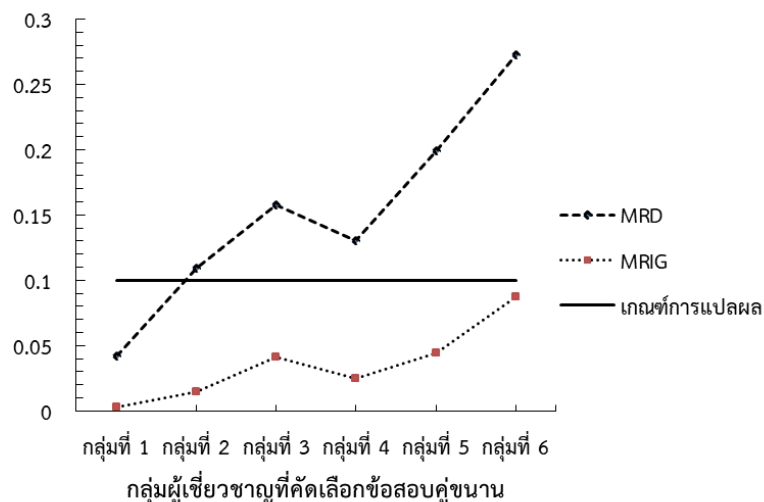
ตาราง 2 ผลการตรวจสอบความเป็นคู่ขนานของแบบสอบโดยอาศัยทฤษฎีการตอบสนองข้อสอบ

กลุ่ม	1		2		3		4		5		6	
ระดับความสามารถในการ ตัดสินความเป็นคู่ขนาน ของข้อสอบ	ระดับที่ 1		ระดับที่ 1		ระดับที่ 2		ระดับที่ 2		ระดับที่ 3		ระดับที่ 3	
ประสบการณ์สอน	≥ 10 ปี		< 10 ปี		≥ 10 ปี		< 10 ปี		≥ 10 ปี		< 10 ปี	
ชุดข้อสอบ	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B
RMSD	0.283		0.430		0.714		0.922		0.980		1.339	
MSG	0.020		0.080		0.275		0.208		0.266		0.452	
MRD	0.042		0.109		0.158		0.130		0.199		0.273	
MRIG	0.003		0.015		0.041		0.025		0.044		0.087	



ภาพ 1 เปรียบเทียบค่า RMSD และ MSG ระหว่างแบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้สูงอายุ (กำหนดเกณฑ์การแปลผลแบบสอบคู่ขนาน; ค่า RMSD และ MSG ต่ำกว่า 0.50)

เมื่อกำหนดเกณฑ์สำหรับแบบสอบคู่ขนานต้องมีค่า MRD และ MRIG ต่ำกว่า 0.10 (น้อยกว่าร้อยละ 10) พบว่า แบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้สูงอายุที่มีความสามารถในการตัดสินใจความเป็นคู่ขนานของข้อสอบในระดับที่ 1 ที่มีประสบการณ์สอน ≥ 10 ปี มีความเป็นคู่ขนานของแบบสอบ ส่วนแบบสอบที่คัดเลือกโดยกลุ่มผู้สูงอายุกลุ่มอื่นๆ ไม่มีความเป็นคู่ขนานของแบบสอบ โดยมีแนวโน้มของค่า MRD และ MRIG มีค่าสูงขึ้นตามระดับความสามารถในการตัดสินใจความเป็นคู่ขนานของข้อสอบของกลุ่มผู้สูงอายุที่ลดลงและมีประสบการณ์สอนนักศึกษาแพทย์น้อยกว่า 10 ปี (ภาพ 2)



ภาพ 2 เปรียบเทียบค่า MRD และ MRIG ระหว่างแบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้สูงอายุ (กำหนดเกณฑ์การแปลผลแบบสอบคู่ขนาน; ค่า MRD และ MRIG ต่ำกว่า 0.10)

เมื่อเปรียบเทียบค่าเฉลี่ยของ MRD ที่ได้จากแบบสอบที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญทั้ง 6 กลุ่มมีอย่างน้อย 1 คู่ที่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เมื่อทำการวิเคราะห์เปรียบเทียบรายคู่ พบว่า กลุ่มผู้เชี่ยวชาญในระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบที่ 1 และประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไปมีค่าเฉลี่ยของ MRD ต่ำกว่ากลุ่มผู้เชี่ยวชาญอื่นๆทุกกลุ่ม และกลุ่มผู้เชี่ยวชาญในระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบที่ 1 และประสบการณ์สอนน้อยกว่า 10 ปี มีค่าเฉลี่ยของ MRD ต่ำกว่ากลุ่มผู้เชี่ยวชาญในระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบที่ 3 และประสบการณ์สอนน้อยกว่า 10 ปี และพบว่าแบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญส่วนใหญ่มีความเป็นคู่ขนานกันของข้อสอบสอดคล้องกับการตรวจสอบความเป็นคู่ขนานของแบบสอบโดยใช้ค่าเฉลี่ย RMSD แบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญทุกกลุ่มมีค่าเฉลี่ย RMSD ต่ำกว่า 0.50 จึงสรุปได้ว่า ข้อสอบที่ถูกคัดเลือกโดยกลุ่มผู้เชี่ยวชาญเมื่อนำมาสร้างเป็นแบบสอบแล้วในภาพรวมมีความเป็นคู่ขนานกันของข้อสอบในทุกแบบสอบ (ตาราง 3) เมื่อวิเคราะห์เปรียบเทียบค่าเฉลี่ยของ RMSD ที่ได้จากแบบสอบจำแนกตามกลุ่มผู้เชี่ยวชาญที่มีลักษณะแตกต่างกัน 6 กลุ่ม พบว่า แบบสอบที่ได้จากการคัดเลือกข้อสอบโดยกลุ่มผู้เชี่ยวชาญทั้ง 6 กลุ่มมีค่าเฉลี่ยของ RMSD ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ตาราง 3 ผลการตรวจสอบความเป็นคู่ขนานของข้อสอบและแบบสอบโดยอาศัยทฤษฎีการตอบสนองข้อสอบ

กลุ่ม	1	2	3	4	5	6
ชุดแบบสอบ	1A-1B	2A-2B	3A-3B	4A-4B	5A-5B	6A-6B
ระดับความสามารถในการตัดสิน ความเป็นคู่ขนานของข้อสอบ	ระดับที่ 1	ระดับที่ 1	ระดับที่ 2	ระดับที่ 2	ระดับที่ 3	ระดับที่ 3
ประสบการณ์สอน	≥ 10 ปี	< 10 ปี	≥ 10 ปี	< 10 ปี	≥ 10 ปี	< 10 ปี
ร้อยละของข้อสอบคู่ขนานโดย พิจารณาค่า RMSD	100.00	100.00	98.00	100.00	100.00	100.00
ค่าเฉลี่ย RMSD ($\overline{RMSD}$)	0.095	0.084	0.123	0.091	0.079	0.096

อภิปรายผล

ในการวิจัยครั้งนี้พบว่า เมื่อวิเคราะห์ระดับความสามารถในการตัดสินความเป็นคู่ขนานของแบบสอบของผู้เชี่ยวชาญที่เป็นตัวอย่างทั้งหมดไม่เป็นไปตามข้อตกลงเบื้องต้นด้านความเป็นเอกมิติของแบบวัด โดยพบว่ามีกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบค่อนข้างต่ำถึงต่ำมาก (0.00 ถึง 0.40) ซึ่งอาจเกิดจากการขาดองค์ความรู้ในด้านความเป็นคู่ขนานของข้อสอบหรือให้ความร่วมมือในการทำแบบวัดค่อนข้างน้อย ทำให้ฉันทามตินั้นแตกต่างจากกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบสูงกว่า เมื่อพิจารณาจากผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบต่ำกว่าเกณฑ์ออกและวิเคราะห์ฉันทามติใหม่พบว่า แบบวัดเป็นไปตามข้อตกลงเบื้องต้นของทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ สอดคล้องกับงานวิจัยของ Krackhardt (1987); Johnson (1990) และ Weller (2007) ที่วิเคราะห์ฉันทามติหลังจากตัดกลุ่มที่มีระดับความเห็นพ้องร่วมกันต่ำออกและแสดงถึงความเป็นเอกมิติของกลุ่มที่อยู่ในวัฒนธรรมเดียวกันมากขึ้นอย่างชัดเจน การตรวจสอบข้อตกลงเบื้องต้นของทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ มีความสำคัญที่แบบสอบหรือแบบวัดจะต้องมีความเป็นเอกมิติเพราะคำตอบที่ต้องการทราบเกิดจากฉันทามติรวมของกลุ่มที่มีคุณลักษณะเดียวกันจึงจะมี

ความน่าเชื่อถือและเป็นประโยชน์ในการเป็นสารสนเทศในการตัดสินใจในเรื่องที่วัดและประเมินผลนั้นๆ หากพบว่าการละเมิดข้อตกลงเบื้องต้นข้อนี้ควรพิจารณาวิเคราะห์ฉันทามติเฉพาะกลุ่มตัวอย่างหรือประชากรที่มีความสอดคล้องกลมกลืนเป็นฉันทามติไปในทิศทางเดียวกัน หรือวิเคราะห์ฉันทามติแยกกลุ่มวัฒนธรรม (Cross-cultural consensus analysis) (Borgatti and Halgin, 2011; Weller, 2007) ส่วนข้อตกลงเบื้องต้นด้านความเป็นอิสระต่อกันของผู้ตอบหรือลงฉันทามติและด้านความถูกต้องของคำตอบ จะต้องวางแผนออกแบบและบริหารจัดการวัดหรือการสอบให้รัดกุมและเป็นไปตามข้อตกลงเบื้องต้นอย่างเคร่งครัด

จากการประมาณค่าระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบของผู้เชี่ยวชาญโดยอาศัยทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบ พบว่า ผู้เชี่ยวชาญส่วนใหญ่มีระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบอยู่ในระดับที่ 3 มากกว่าระดับที่ 2 และ 1 โดยเฉพาะในกลุ่มผู้เชี่ยวชาญที่มีประสบการณ์สอนนักศึกษาแพทย์น้อยกว่า 10 ปี แต่ในกลุ่มผู้เชี่ยวชาญที่มีประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไปจะมีระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบในระดับที่ 2 มากที่สุด ซึ่งแสดงให้เห็นว่าประสบการณ์ในการสอนนักศึกษาแพทย์มีส่วนช่วยในการตัดสินใจเป็นคู่ขนานของแบบสอบได้ดีขึ้น โดยคุณสมบัติที่สำคัญของผู้เชี่ยวชาญที่ควรพิจารณาสำหรับคัดเลือกผู้เชี่ยวชาญในการสร้างแบบสอบคู่ขนานควรมีประสบการณ์ในการสอนรายวิชานั้นๆ รวมถึงประสบการณ์และความเชี่ยวชาญในการออกข้อสอบ คัดเลือกข้อสอบ รวมถึงการตัดสินใจตรงตามเนื้อหาของข้อสอบตามจุดประสงค์การเรียนรู้รายวิชานั้นของแต่ละผู้เชี่ยวชาญด้วย (Romney, Batchelder and Weller, 1986; Weller, 2007) ทั้งนี้การตรวจสอบระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบทำให้ได้สารสนเทศในการคัดเลือกผู้เชี่ยวชาญที่เหมาะสมในการสร้างแบบสอบคู่ขนานให้มีคุณภาพความตรงและความถูกต้องสูงขึ้น รวมถึงเพิ่มระดับความสัมพันธ์ระหว่างเนื้อหาหรือจุดประสงค์การเรียนรู้กับแบบสอบได้ดีขึ้น (Wanichsan D., 2012 และ Panjaburee, Hwang, Triampo and Shih, 2010)

จากการวิเคราะห์เปรียบเทียบค่าเฉลี่ยของ MRD ที่ได้จากแบบสอบทั้งฉบับจำแนกตามกลุ่มผู้เชี่ยวชาญที่มีลักษณะแตกต่างกัน 6 กลุ่ม พบว่า ค่าเฉลี่ยของ MRD ที่ได้จากแบบสอบที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญทั้ง 6 กลุ่มพบว่า กลุ่มผู้เชี่ยวชาญในระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบที่ 1 และประสบการณ์สอนตั้งแต่ 10 ปีขึ้นไป มีค่าเฉลี่ยของ MRD ต่ำกว่ากลุ่มผู้เชี่ยวชาญอื่นๆทุกกลุ่ม และกลุ่มผู้เชี่ยวชาญในระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบที่ 1 และประสบการณ์สอนน้อยกว่า 10 ปี มีค่าเฉลี่ยของ MRD ต่ำกว่ากลุ่มผู้เชี่ยวชาญในระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบที่ 3 และประสบการณ์สอนน้อยกว่า 10 ปี ซึ่งแสดงให้เห็นว่า กลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินใจระดับความเป็นคู่ขนานของข้อสอบที่สูงสามารถคัดเลือกข้อสอบให้มีดัชนีความเป็นคู่ขนานของแบบสอบได้ดีกว่ากลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินใจระดับความเป็นคู่ขนานของข้อสอบที่ต่ำกว่า รวมถึงประสบการณ์การสอนนักศึกษาแพทย์ที่มากขึ้นก็มีส่วนช่วยให้การคัดเลือกข้อสอบให้มีดัชนีความเป็นคู่ขนานของแบบสอบได้ดีขึ้นเช่นกัน ดังนั้น การตรวจสอบระดับความสามารถในการตัดสินใจเป็นคู่ขนานของข้อสอบ จึงให้สารสนเทศที่สำคัญต่อการคัดเลือกกลุ่มผู้เชี่ยวชาญที่ทำหน้าที่พิจารณาคัดเลือกข้อสอบเพื่อสร้างแบบสอบคู่ขนานในสภาพจริง นอกจากนี้ พบว่า แบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญส่วนใหญ่มีความเป็นคู่ขนานกันของข้อสอบสอดคล้องกับการตรวจสอบความเป็นคู่ขนานของแบบสอบโดยใช้ $RMSD_F$ และ ค่าเฉลี่ย $RMSD_F$ แบบสอบคู่ขนานที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญทุกกลุ่มให้ผลสอดคล้องกัน คือ ข้อสอบที่ถูกคัดเลือกโดยกลุ่มผู้เชี่ยวชาญเมื่อนำมาสร้างเป็นแบบสอบแล้วในภาพรวมมีความเป็นคู่ขนานกันของข้อสอบในทุกแบบสอบ และเมื่อวิเคราะห์เปรียบเทียบค่าเฉลี่ยของ $RMSD_F$ พบว่า

แบบสอบที่ได้จากการคัดเลือกข้อสอบโดยกลุ่มผู้เชี่ยวชาญทั้ง 6 กลุ่มมีค่าเฉลี่ยของ $RMSD_F$ ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 ทั้งนี้ เมื่อพิจารณาระดับความสามารถของผู้เชี่ยวชาญในการตัดสินความเป็นคู่ขนานของข้อสอบพบว่า กลุ่มผู้เชี่ยวชาญมีความเป็นเอกมิตีและอยู่ในกลุ่มฉันทามติที่มีทิศทางการตัดสินใจอยู่ในทิศทางเดียวกัน จากการวิเคราะห์สถิติตามทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบหลังจากตัดกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบต่ำกว่า 0.40 ดังนั้นผู้เชี่ยวชาญส่วนใหญ่ จึงมีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบผ่านเกณฑ์ที่กำหนดซึ่งเป็นผลให้การคัดเลือกข้อสอบมีดัชนีความเป็นคู่ขนานของข้อสอบมีประสิทธิภาพ สอดคล้องกับงานวิจัยของ Spray and Miller (1992) และ Debeer, Ali and Rijn (2017) ที่ให้ความสำคัญต่อความเป็นคู่ขนานรายข้อ เพราะจะนำไปสู่ผลรวมของความเป็นคู่ขนานของแบบสอบทั้งฉบับ ซึ่งจะพบว่า แบบสอบ 2 ฉบับที่ประกอบด้วยข้อสอบคู่ขนานรายข้อในสัดส่วนที่สูงมีโอกาสที่แบบสอบคู่นั้นจะคู่ขนานกันสูงขึ้น ทั้งนี้การตรวจสอบความเป็นคู่ขนานของข้อสอบควรใช้ระดับความสามารถของผู้สอบอย่างน้อย 6 ระดับ จึงจะทำให้การตรวจสอบมีความน่าเชื่อถือและคลาดเคลื่อนน้อย (Debeer, Ali and Rijn, 2017) โดยงานวิจัยในครั้งนี้ใช้ระดับความสามารถของผู้สอบ 15 ระดับซึ่งให้ผลการตรวจสอบความเป็นคู่ขนานของข้อสอบสอดคล้องกันดี

ข้อเสนอแนะในการนำผลวิจัยไปใช้

กลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบโดยอาศัยทฤษฎีฉันทามติทางวัฒนธรรมแบบมีรูปแบบอยู่ในระดับสูง สามารถคัดเลือกข้อสอบเพื่อสร้างแบบสอบคู่ขนานให้มีดัชนีความเป็นคู่ขนานของแบบสอบได้สูงกว่ากลุ่มผู้เชี่ยวชาญที่อยู่ในระดับต่ำกว่า ดังนั้น ควรมีการตรวจสอบระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบก่อนเพื่อให้ได้แบบสอบคู่ขนานที่มีคุณภาพทั้งความตรงและดัชนีความเป็นคู่ขนานของแบบสอบสูงขึ้น รวมถึงประสบการณ์ในการสอนนักศึกษาแพทย์มีส่วนช่วยเพิ่มค่าดัชนีความเป็นคู่ขนานของแบบสอบที่คัดเลือกโดยกลุ่มผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบในระดับสูง ดังนั้น การคัดเลือกกลุ่มผู้เชี่ยวชาญในการคัดเลือกข้อสอบเพื่อสร้างแบบสอบคู่ขนานควรพิจารณาประสบการณ์การสอนนักศึกษาในรายวิชาหรือเนื้อหาที่เกี่ยวข้องกับข้อสอบร่วมด้วย โดยการตรวจสอบระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบจะให้สารสนเทศที่สำคัญถึงคุณสมบัติขั้นต่ำที่ใช้เป็นการคัดกรองผู้เชี่ยวชาญเบื้องต้นสำหรับคัดเลือกผู้เชี่ยวชาญในการสร้างแบบสอบคู่ขนานในสภาพจริง ทั้งนี้ การตรวจสอบความเป็นคู่ขนานของข้อสอบและแบบสอบโดยพิจารณาระดับความสามารถของผู้สอบจำนวนทั้งสิ้น 15 ระดับ แสดงให้เห็นว่าข้อสอบและแบบสอบมีความเป็นคู่ขนานตลอดไค้สารสนเทศของข้อสอบและแบบสอบ

ข้อเสนอแนะในการวิจัยครั้งต่อไป

จากผลการวิจัยในครั้งนี้ พบว่า ทุกกลุ่มผู้เชี่ยวชาญที่ผ่านเกณฑ์ขั้นต่ำของการตรวจสอบความเป็นคู่ขนานของข้อสอบโดยอาศัยทฤษฎีฉันทามติทางวัฒนธรรมสามารถคัดเลือกข้อสอบให้มีความเป็นคู่ขนานของข้อสอบรายข้อได้ดีไม่แตกต่างกัน ดังนั้นจึงควรศึกษาการจัดกลุ่มผู้เชี่ยวชาญที่มีสัดส่วนระดับความสามารถในการตัดสินความเป็นคู่ขนานของแบบสอบที่แตกต่างกันส่งผลต่อดัชนีความเป็นคู่ขนานของข้อสอบและแบบสอบหรือไม่ เพื่อเป็นสารสนเทศสำหรับการจัดกลุ่มผู้เชี่ยวชาญในสภาพจริงซึ่งอาจจะมีจำนวนผู้เชี่ยวชาญในการคัดเลือกข้อสอบคู่ขนานผู้เชี่ยวชาญที่มีระดับความสามารถในการตัดสินความเป็นคู่ขนานของข้อสอบ

ระดับสูงค่อนข้างจำกัด ดังแสดงค่าสถิติที่ได้จากการวิเคราะห์ภูมิหลังของผู้เชี่ยวชาญในงานวิจัยครั้งนี้ ควรประยุกต์ใช้ดัชนีการตรวจสอบความเป็นคู่ขนานของข้อสอบและแบบสอบนี้สำหรับตรวจสอบความเป็นคู่ขนานของแบบสอบในรูปแบบอื่นๆ เช่น ข้อสอบเติมคำตอบ หรือรูปแบบผสมที่มีทั้งข้อสอบแบบเลือกตอบหลายตัวเลือกและเติมคำตอบในสัดส่วนที่ต่างกัน โดยพิจารณาจากค่าสารสนเทศของแบบสอบตามโมเดลต่างๆ ที่เหมาะสมโดยอาศัยทฤษฎีการตอบสนองข้อสอบ รวมถึงการตรวจสอบความเป็นคู่ขนานของข้อสอบและแบบสอบโดยใช้ทั้งค่าสารสนเทศและโครงสร้างของข้อสอบและแบบสอบหากทำการวิจัยสำหรับการสอบขนาดใหญ่และมีผู้เข้าสอบเป็นจำนวนมากต่อไป

รายการอ้างอิง

ภาษาไทย

- โชติกา ภาชีผล. (2559). *การวัดและประเมินผลการเรียนรู้*. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่* (พิมพ์ครั้งที่ 4). กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

- Ali, U. S., & Rijn, P. W. (2016). *An evaluation of different statistical target for assembling parallel form in item response theory.*, Applied psychological measurement.
- Borgatti, S. P., & Halgin, D. S. (2011). Consensus Analysis. In D. B. Kronenfeld, G. Bennardo, V. C. DeMunck, & M. D. Fischer (Eds.), *A companion to cognitive anthropology* (pp.171-190). West Sussex, UK: Blackell Publishing.
- Chen, P. H., Chang, H. H., & Wu, H. (2012). Item selection for the development of *parallel forms from an IRT-based seed test using a sampling and classification approach*. Education and psychological measurement, 72(6), 933-953.
- Debeer, D., Ali, U. S., & Rijn, P. W. (2017). Evaluating Statistical Targets for Assembling Parallel Mixed-Format Test Forms. *Journal of Educational Measurement*, 54(2), 218-242.
- Gronlund, N. E. (2009). *Assessment of Student Achievement* (9th Ed.). London: Pearson.
- Hattie, J. (2005). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Hettiaratchi, E. S. G. (1978). A comparison of student performance in two parallel physiology tests in multiple choice and short answer forms. *Medical education*, 12(4), 290-296.
- Johnson, J. C. (1990). *Selecting Ethnographic Informants*. Newbury Park: Sage Publications.
- Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, 9(2), 109-134.

- Keikesers, C. J. P., Brouwers, J. R. B., Wildt, D. J., & Custers, E. J. F., Cate, O. Th., Hazen, A. C. M., & Jansen, P. A. (2014). A comparison of medical and pharmacy students' knowledge and skills of pharmacology and pharmacotherapy. *British Journal of Clinical Pharmacology*, 78(4), 781-788.
- Lin, C. J. (2008). Comparisons between classical theory and item response theory in automated assembly of parallel test forms. *The journal of technology, learning and assessment*, 6(8). Retrieved from <https://ejournals.bc.edu/index.php/jtla/article/view/1638>
- Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, 35(3), 185-198.
- Linden, W. J., & Luecht, R. M. (1998). Observed-score equating as a test assembly problem. *Psychometrika*, 63(4), 401-418.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Massachusetts: Addison-Wesley.
- Panjaburee, P., Hwang, G. j., Triampo, W., & Shih, B. Y. (2010). A multi-expert approach for developing testing and diagnostic system based on the concept-effect model. *Computer and education*, 55(2), 527-540.
- Romney, A. K., Batchelder W. H., & Weller S. C. (1986). Culture as Consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88(2), 313-338.
- Sanders, P. F., & Verschoor, A. J. (1998). Parallel test construction using classical item Parameters. *Applied Psychological Measurement*, 22(3), 212-223.
- Spray, J.A., & Miller, T.R. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportions correct when population ability distributions are incongruent* (Research Report). Retrieved from <https://eric.ed.gov/?id=ED352388>.
- Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, 37(2), 123-139. doi:<https://doi.org/10.1177/0146621612469825>
- Wanichsan D., Panjaburee P., Laosinchai P., Triampo W., & Chookaew S. (2012). A majority-density approach to developing testing and diagnostic systems with the cooperation of multiple experts based on an enhanced concept-effect relationship mode. *Expert systems with application*, 39, 8380-8388.
- Weller S. C. (2007). Cultural consensus theory applications and frequently asked questions. *Field methods*, 19(4), 339-368.