

ทฤษฎีการตอบสนองข้อสอบกับข้อสอบแบบทดสอบแบบทดสอบ

Item Response Theory and Testlet Test

พิสิฐ พินิจสกุล¹, สมประสงค์ เสนารัตน์² และเบญจมาภรณ์ เสนารัตน์²

Pisit Pinitakul¹, Somprasong Senarat², and Benjamaporn Senarat²

(วันรับบทความ 15 กรกฎาคม 2563, วันรับแก้ไขบทความ 20 กันยายน 2563, วันตอบรับบทความ 2 พฤศจิกายน 2563)

บทคัดย่อ

บทความนี้เป็นการนำเสนอมุมมองของทฤษฎีการตอบสนองข้อสอบแบบทดสอบแบบทดสอบที่เป็นลักษณะของข้อสอบเชิงสถานการณ์ที่เป็นชุดของข้อสอบหลายข้อในการที่จะตอบสิ่งเร้าเดียว ซึ่งเป็นลักษณะของข้อสอบที่พบได้บ่อยในทางภาษาและทางสังคมศาสตร์ การที่มีข้อคำถามหลายข้อในข้อสอบแบบทดสอบแบบทดสอบจะทำให้ข้อคำถามแต่ละข้อซึ่งนำคำตอบซึ่งกันและกัน ซึ่งเป็นการละเมิดข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบในเรื่องการเป็นอิสระจากกันของข้อสอบ ดังนั้นการประมาณค่าความสามารถของผู้สอบด้วยทฤษฎีการตอบสนองข้อสอบ จึงจำเป็นต้องมีการทบทวนวิธีการประมาณค่าพารามิเตอร์ในโมเดลที่มีข้อสอบประเภทข้อสอบแบบทดสอบ โดยที่ผ่านมามีนักวิชาการเสนอวิธีการแก้ปัญหาไว้หลายวิธี บทความนี้จะนำเสนอประเด็นที่เหมาะสมสำหรับข้อสอบประเภทข้อสอบแบบทดสอบที่สามารถดำเนินการประมาณค่าพารามิเตอร์อย่างไร รวมทั้งนำเสนอตัวอย่างงานวิจัยที่เกี่ยวข้องกับการประมาณค่าพารามิเตอร์ข้อสอบแบบทดสอบ

คำสำคัญ ทฤษฎีการตอบสนองข้อสอบแบบทดสอบ, ทฤษฎีการตอบสนองข้อสอบ, ข้อสอบเชิงสถานการณ์

ABSTRACT

This article presented perspectives of testlet response theory which is characteristics of a situation test comprising a set of items for responding one stimulus. This type of test is commonly used in language and social sciences. Items in testlet could guide the answer among items within the testlet which breaks the local assumptions of item response theory. Consequently, there is necessary to reconsider parameter estimation of examinees for testlet response theory model. The researchers have been proposing multiple ways to solve the problem. This paper presented proper issues for estimating parameter of testlet test as well as research examples on parameter estimation of testlet test.

Keyword testlet item response theory, item response theory, situation test

¹ นักศึกษาหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัยและประเมินผลการศึกษา คณะครุศาสตร์ มหาวิทยาลัยราชภัฏร้อยเอ็ด

Email: tatar2p@gmail.com, pisit.pi@kkumail.com

² ผู้ช่วยศาสตราจารย์ ดร., สาขาวิชาวิจัยและประเมินผลการศึกษา คณะครุศาสตร์ มหาวิทยาลัยราชภัฏร้อยเอ็ด

บทนำ

ข้อสอบลักษณะทดสอบแบบทดสอบที่มีการใช้ข้อสอบอย่างแพร่หลายในการประมาณค่าความสามารถของผู้เรียน ซึ่งลักษณะของข้อสอบชุดทดสอบแบบทดสอบนั้นมีความเป็นมาคือ ในปี 1987 มีนักวิจัยสองท่านคือ Wainer และ Kiely ได้เสนอแฟกเตอร์ของข้อสอบที่ถูกใช้ร่วมกันคือ ทฤษฎีการตอบสนองข้อสอบหรือเรียกได้ว่า Testlet Response Theory (TRT) ที่จะประมาณค่าความสามารถของบุคคลในความเข้าใจสิ่งเร้าเดียวกันหรือสื่อต่าง ๆ เช่น บทความ กราฟข้อมูล บทความ ตารางตัวเลข เป็นต้น โดยในการทำข้อสอบที่ใช้สถานการณ์การอ่านเหล่านี้ทำให้ผู้สอบใช้เวลาค่อนข้างมากถ้าจะตอบคำถามเพียงข้อเดียว ดังนั้นในแต่ละบทความจึงมีข้อคำถามหลายข้อเป็นเซตเดียวกัน จึงเรียกว่ารูปแบบข้อสอบแบบทดสอบแบบทดสอบ (Wainer, Bradlow, & Wang, 2007, p. 60) ทำให้ความเป็นอิสระของข้อสอบแต่ละข้อมีข้อจำกัดเนื่องจากต้องพึ่งพาข้อสอบข้ออื่น โดยในปัจจุบันมีการใช้ทดสอบแบบทดสอบในข้อสอบศาสตร์ต่าง ๆ ซึ่งจะเรียกได้ว่าเป็นข้อสอบเชิงสถานการณ์ที่เป็นแบบทดสอบที่ถามและตอบภายใต้สถานการณ์ที่กำหนดให้ และให้ความสำคัญเรื่องความตรงเชิงเนื้อหา (วิรัช วรรณรัตน์, 2558, น. 4) โดยลักษณะของข้อสอบในชุดของทดสอบนั้นอาจจะมีการขึ้นคำตอบจากการตอบข้อหนึ่งไปขึ้นเพื่อคำตอบอีกข้อหนึ่ง ซึ่งทำให้เกิดการละเมิดข้อตกลงเบื้องต้นที่ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ในประเด็นที่ว่าข้อสอบแต่ละข้อต้องเป็นอิสระจากกัน นอกจากนี้การที่ข้อสอบขึ้นคำตอบซึ่งกันและกันในทฤษฎีการทดสอบแบบดั้งเดิมถือว่าเป็นลักษณะของข้อสอบที่ไม่มีดี อย่างไรก็ตามทฤษฎีการทดสอบแบบดั้งเดิมจะมองความคลาดเคลื่อนเป็นรายฉบับจึงยังไม่สามารถแก้ปัญหาในประเด็นนี้ได้อย่างชัดเจน แต่ในทางตรงข้าม ทฤษฎีการตอบสนองข้อสอบที่ได้รับการพัฒนาภายหลังมุ่งที่จะประมาณค่าพารามิเตอร์ของข้อสอบแต่ละข้อ เช่น ความยาก อำนาจจำแนก โอกาสการเดาได้ จึงถือว่า IRT สามารถวิเคราะห์รายข้อได้ และมีการวัดความคลาดเคลื่อนรายข้อ ทำให้ทฤษฎีการวัดแบบดั้งเดิมมีความคลาดเคลื่อนในการวัดมากกว่าทฤษฎีการตอบสนองข้อสอบ แต่อย่างไรก็ตามข้อสอบในรูปแบบของชุดทดสอบนั้น มีการพึ่งพาคำตอบกันในแต่ละข้อถือว่าเป็นการขึ้นคำตอบซึ่งกันและกันทำให้ข้อสอบแต่ละข้อไม่เป็นอิสระจากกัน และละเมิดข้อตกลงเบื้องต้นที่สำคัญของ IRT สองประการคือ ข้อสอบต้องมุ่งวัดลักษณะเดียวกัน (Unidimensionality) และมีความเป็นอิสระระหว่างข้อคำถาม (Independence) (ศิริชัย กาญจนวาสี, 2555, น. 77) นอกจากนี้ข้อสอบแบบทดสอบแบบทดสอบยังมีการวัดคุณลักษณะหลายด้านในข้อเดียวกัน ซึ่งสอดคล้องกับทฤษฎีการตอบสนองแบบพหุมิติ (Multidimensional Item Response Theory: MIRT) ที่ไม่มีข้อตกลงเบื้องต้นเกี่ยวกับความเป็นเอกมิติและความเป็นอิสระระหว่างข้อสอบ ซึ่งการละเมิดข้อตกลงเบื้องต้นที่กล่าวมานั้นส่งผลต่อการประมาณค่าความเที่ยงของข้อสอบเนื่องจากว่าการตอบถูกของแต่ละข้อไม่เป็นอิสระจากกัน (Sireci et al., 1991, p. 238) จากที่กล่าวมาข้างต้น จะเห็นว่าข้อสอบที่เป็นเชิงสถานการณ์แบบทดสอบได้มีการฝ่าฝืนข้อตกลงเบื้องต้นนี้ของทฤษฎีการวัด ทำให้เกิดช่องว่างที่จะต้องประยุกต์ใช้ทฤษฎีต่าง ๆ เพื่อให้การประมาณค่าพารามิเตอร์ของผู้สอบมีความถูกต้องและมีความเที่ยงยิ่งขึ้น

บทความนี้จะกล่าวถึงทฤษฎีการตอบสนองข้อสอบ ข้อสอบแบบทดสอบ การแก้ปัญหาข้อสอบแบบทดสอบ และงานวิจัยที่เกี่ยวข้อง

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory)

ทฤษฎีการตอบสนองข้อสอบจะแสดงความสัมพันธ์ระหว่างความสามารถหรือคุณลักษณะที่วัดโดยเครื่องมือ และการตอบสนองต่อรายข้อ ซึ่งจะมีการตอบสนองรายข้อแบบสองค่า (Dichotomous) ยกตัวอย่างเช่น ตอบถูก หรือตอบผิด ใช่หรือไม่ใช่ เห็นด้วยหรือไม่เห็นด้วย โดยค่าคะแนนที่ตรวจจะเป็นแค่สองค่า ส่วนการตอบรายข้อที่มากกว่าสองค่า (Polytomous) ยกตัวอย่างเช่น การให้คะแนนข้อสอบแต่ละข้อไม่เท่ากันขึ้นกับความยาก การจัดอันดับการตัดสินใจ หรือผู้ให้คะแนนในแบบมาตรา Likert จากข้อคำถามในแบบสอบถาม (Demars, 2010, p. 3) โดย Fox (2010, p. 6) กล่าวว่า ทฤษฎีการตอบสนองข้อสอบ (IRT) เกี่ยวข้องกับการวัดโครงสร้างเชิงสมมติฐาน ที่มีตัวแปรแฝงอยู่และสามารถวัดได้โดยอ้อมผ่านทางตัวแปรสังเกตได้ตัวอื่น โดยโครงสร้างเชิงสมมติฐานนี้จะเป็นตัวแปรแฝงที่อยู่ในรูปแบบทักษะ ความสามารถ หรือคุณลักษณะของบุคคลที่วัดได้จากข้อสอบแต่ละข้อ ซึ่งอาจจะเป็น การวัดความสามารถทางวิชาการ ความถนัดหรืออาจเป็นทัศนคติ ความเชื่อ

ทฤษฎีการตอบสนองข้อสอบกำหนดอยู่บนแนวคิดสำคัญ 2 ประการ ได้แก่ พฤติกรรมในการตอบข้อสอบของผู้สอบนำไปใช้ทำนายความสามารถ (Ability) หรือลักษณะภายใน (Trait) ของผู้สอบ และพฤติกรรมในการตอบข้อสอบของผู้สอบสัมพันธ์กับความสามารถ หรือลักษณะภายในของผู้สอบ (Hambleton & Swaminathan, 1985, p. 9) โดยมีแนวคิดอยู่บนพื้นฐานความเชื่อว่าลักษณะภายในตัวบุคคลซึ่งไม่สามารถสังเกตได้ แต่สามารถวัดพฤติกรรมหรือคุณลักษณะภายในตัวบุคคลได้โดยอ้อมจากผลการตอบข้อสอบที่สามารถสังเกตได้ ทฤษฎีการตอบสนองข้อสอบยังให้ดัชนีความแม่นยำของคะแนนการทดสอบในรูปแบบความคลาดเคลื่อนมาตรฐานของการวัดของผู้สอบแต่ละคน ผลลัพธ์จากการวัดภายใต้ทฤษฎีการตอบสนองข้อสอบเป็นการจัดวางระดับความสามารถของผู้สอบ แต่อย่างไรก็ตาม ตัวแปรแฝงไม่สามารถวัดได้โดยตรง จึงจำเป็นต้องมีเครื่องมือวัดเพื่อมาช่วยในการแก้ปัญหา โดยทั่วไปสันนิษฐานว่าความสามารถของบุคคลโดยทั่วไปมีจุดกึ่งกลางอยู่ที่ศูนย์ในหน่วยของคะแนนมาตรฐานที่มีค่าเฉลี่ยเท่ากับศูนย์ และความแปรปรวนเท่ากับหนึ่ง และช่วงพิสัยจากอนันต์ลบไปจนถึงอนันต์บวก นอกจากนี้ Demars (2010, p. 3) ยังกล่าวว่า ทฤษฎีการตอบสนองข้อสอบสามารถใช้การทดสอบหรือใช้มาตรวจวัดได้ ทั้งนี้การวิเคราะห์คุณภาพของข้อสอบรายข้อตามทฤษฎีการตอบสนองข้อสอบจะมีพารามิเตอร์ของข้อสอบ (ศิริชัย กาญจนวาสี, 2555, p. 53) เช่น ค่าความยากรายข้อ ค่าอำนาจจำแนกรายข้อ และอาจมีพารามิเตอร์อื่น ๆ ได้ ซึ่งพารามิเตอร์ของข้อสอบรายข้อจะมีประโยชน์ก็ต่อเมื่อทำการสร้างข้อสอบให้ตรงกับระดับคุณลักษณะของกลุ่มประชากรเป้าหมาย เช่น ค่าอำนาจจำแนกรายข้อจะจำแนกความต่างกันระหว่างผู้สอบที่มีระดับความสามารถต่างกันได้ ส่วนค่าความยากรายข้อจะบอกระดับความสามารถของผู้สอบได้ นอกจากนี้ค่าพารามิเตอร์ของข้อสอบสามารถใช้ในการประมาณค่าความคลาดเคลื่อนมาตรฐานของการวัดหรือความน่าเชื่อถือของคะแนน

ลักษณะแฝงภายในตัวบุคคลเป็นตัวกำหนดพฤติกรรมหรือผลการตอบสนองข้อสอบที่สังเกตได้ โดยความสัมพันธ์ดังกล่าวสามารถอธิบายได้ด้วยฟังก์ชันทางคณิตศาสตร์ที่เรียกว่า ฟังก์ชันการตอบข้อสอบ (Item Response Function) มี 2 ลักษณะคือฟังก์ชันโลจิสติก (Logistic Function) และฟังก์ชันปกติสะสม (Normal Ogive Function) ทั้งนี้ Fox (2010, p. 6) ยังได้นำเสนอสมมติฐานหลักสองข้อที่เกี่ยวข้องกับทฤษฎีการตอบสนองข้อสอบ ประการแรกระบุว่า เส้นโค้งลักษณะข้อสอบ (Item Curve Characteristics: ICC) ซึ่งระบุว่าความน่าจะเป็นของการตอบสนองข้อสอบว่าเปลี่ยนแปลงอย่างไรมีผลมาจากการเปลี่ยนแปลงของตัวแปรแฝง รูปแบบทางคณิตศาสตร์ที่แตกต่างกันของเส้นโค้งลักษณะข้อสอบแต่ละข้อจะนำไปสู่โมเดลการตอบสนองต่อข้อสอบที่ต่างกัน ข้อสมมติฐานที่สองระบุว่า การตอบสนองรายข้อแต่ละข้อนั้นจะเป็นอิสระจากกันทางสถิติเมื่อตัวแปรแฝงที่แฝงอยู่คงที่ ซึ่งหมายถึงความเป็นเอกมิติ (Unidimensionality) และความเป็นอิสระจากกันของข้อสอบ

ข้อสอบแบบเทสต์เลท (Testlet Test)

จากการละเมิดข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบในเรื่องความเป็นเอกมิติ และความเป็นอิสระจากกันของข้อคำถาม ซึ่งเกิดจากการใช้ข้อคำถามหลายข้อเป็นหน่วยหลักของการสร้างข้อสอบ ซึ่งลักษณะดังกล่าวได้นิยามว่าเป็น เทสต์เลท หรือกลุ่มของข้อคำถามที่ถูกพัฒนาขึ้นจากหน่วยเดียว ซึ่งหมายถึงจะนำไปดำเนินการสอบด้วยกัน (Wainer et al., 2007, p. 60) โดยในแต่ละการตอบข้อคำถาม ผู้สอบจะได้ทำเทสต์เลทที่มีชุดข้อคำถามเดียวกัน ซึ่งจะมีข้อคำถามหลายข้อ ยกตัวอย่างเช่น ข้อสอบความเข้าใจในการอ่านภาษาอังกฤษ โดยบทความที่มีความยาวมากกว่า 300 คำและนำมาตอบคำถามหลายข้อ (Wheadon, 2010, p. 106) ผู้สอบจะเห็นบทความและมีกลุ่มข้อคำถามที่เกี่ยวข้องกับบทความ ในการเรียงลำดับข้อคำถามจะเรียงจากง่ายไปหายาก ซึ่งการที่มีข้อคำถามเป็นชุดแบบนี้จะสอดคล้องกับลักษณะของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ ที่ระบุว่าข้อสอบแต่ละข้อจะต้องเป็นอิสระจากกัน แต่ในเรื่องของการสอบบทความที่เป็นย่อหน้านั้นจะมีข้อคำถามหลายข้อ ถ้าจะทำการสอบเพียงข้อเดียวก็จะทำให้ผู้สอบใช้เวลามากในการที่จะอ่านบทความยาว ๆ ซึ่งเป็นการเสียเวลาในการทำข้อสอบ มีการแนะนำว่าให้ใช้บทความที่สั้นลงในการตอบคำถาม 1 ข้อ แต่จะเป็นการแก้ปัญหาเพียงชั่วคราวเท่านั้น และไม่เป็นที่ยอมรับเนื่องจากผลสุดท้ายการแก้ปัญหาดังกล่าวจะไปเปลี่ยนโครงสร้างที่ถูกต้อง เนื่องจากโครงสร้างการวัดของข้อสอบเชิงสถานการณ์จะประกอบด้วยชุดข้อคำถามที่มีคำถามหลายข้อ จึงถือว่าเป็นข้อคำถามที่ไม่เป็นอิสระจากกันแต่จะเรียกว่าเป็น เทสต์เลท ซึ่งชุดข้อคำถามจะถูกใช้ด้วยกันทั้งหมด โดย Wainer et al. (2007, p. 54) ได้กล่าวว่า เนื่องจากการที่ไม่มีความเป็นเอกพันธ์ของโครงสร้างข้อสอบ ทำให้ความเป็นอิสระถูกละเมิด ดังนั้นถ้าใช้เทสต์เลทเป็นหน่วยในการวัดความเป็นเอกพันธ์ของโมเดลการตอบสนองข้อสอบจึงอาจจะยังเหมาะสมอยู่ ในส่วนนี้ที่ละเมิดข้อตกลงเบื้องต้นในเรื่องของความเป็นอิสระของข้อคำถาม จึงทำให้ข้อคำถามที่เป็นข้อสอบเชิงสถานการณ์เมื่อทำเป็นคะแนนแบบเทสต์เลทจะสามารถใช้ได้โดยไม่มีข้อกังวลสำหรับผู้สร้างแบบทดสอบ จุดเด่นของผลจากงานของ Rosenbaum (1988, as cited in Wainer et al., 2007, p. 55) แสดงให้เห็นว่าถ้าโมเดลการตอบสนองข้อสอบมีความสอดคล้องกับข้อมูลเชิงประจักษ์ (Model Fit) ถ้าใช้ข้อมูลเทสต์เลททดสอบทางสถิติโดยใช้ตารางการฉกฉวย (Contingency Table) เพื่อที่จะพิจารณาว่าการที่โมเดลสอดคล้องหรือไม่เป็นอิสระจากกันของข้อสอบแบบเทสต์เลท ถ้าเป็นเช่นนั้นความสามารถของผู้สอบจะสรุปออกมาเป็นพารามิเตอร์เดียว และการให้คะแนนระหว่างเทสต์เลทยังคงเป็นไปตามข้อกำหนดของความเป็นเอกมิติ

ในปี ค.ศ. 1987 Wainer และ Kiely ได้นำเสนอประเด็นความเป็นทฤษฎีการตอบสนองข้อสอบเทสต์เลท (Testlet Response Theory: TRT) ซึ่งเป็นรูปแบบเฉพาะในการศึกษาข้อสอบเชิงสถานการณ์แบบนี้ โดยเฉพาะอย่างยิ่งในข้อสอบทางด้านภาษาและสังคมศาสตร์ เช่น ข้อสอบภาษาอังกฤษจะมีข้อสอบการอ่าน การฟัง ซึ่งเป็นรูปแบบที่มาเป็นชุดหรือเรียกว่า เทสต์เลท (Wainer & Wang, 2001, p. 1) โดยจะมีบทความ (Reading Text) หรือบทฟัง (Listening Text) พร้อมกับชุดของข้อคำถามที่จะต้องอ่านหรือฟังจากบทเหล่านี้เพื่อที่จะสามารถตอบคำถามได้ โดยในชุดข้อคำถามอาจจะมีการใช้ข้อมูลจากข้อคำถามอื่นเพื่อที่จะตอบข้อคำถามนั้น (Cross Information) ซึ่งหมายถึง ข้อสอบแต่ละข้อในชุดเทสต์เลทนั้นอาจจะฟังพากันในการที่จะใช้ตอบข้อคำถามอื่นในชุดเทสต์เลท จึงถือว่าละเมิดข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ โดยข้อคำถามแต่ละข้อไม่เป็นอิสระจากกัน และมีข้อคำถามที่อาจวัดหลายองค์ประกอบ ซึ่งอาจเกิดโครงสร้างเป็นพหุมิติ (Multidimensional Structure) โดยในความสามารถทางภาษาศาสตร์นั้นถือว่าเป็นสิ่งที่ผสมผสานความเกี่ยวเนื่องกันของความสามารถหลายด้านที่อย่างซับซ้อน (Wainer et al., 2007, p. 110) ในการที่จะวัดจึงเป็นเรื่องยาก ซึ่งอาจจะทำได้ในรูปแบบของการให้คะแนนแบบหลายค่า (Polytomous) ในโมเดล IRT เฉพาะของข้อสอบที่เป็นเทสต์เลทต่อไป

Wainer et al. (2007, p. 55) กล่าวว่ากลุ่มของข้อคำถามในเทสต์เลทจะมีโครงสร้างซึ่งความน่าจะเป็นของสถานการณ์ที่ไม่พึงประสงค์จะถูกลดลง และจะทำให้โครงสร้างของข้อสอบใกล้เคียงโครงสร้างหลักมากขึ้น ซึ่งประเด็นที่เกี่ยวข้องมี 3 ประเด็นหลักดังนี้

1. ผลกระทบของบริบท (Context Effects)

1.1 การเชื่อมโยงของข้อมูล ในการที่จะขยายหน่วยของข้อสอบจากรายข้อไปสู่เทสต์เลทเป็นการลดขอบเขตของอิทธิพล ในขอบเขตนี้ผู้เชี่ยวชาญมองในประเด็นของการข้ามเรื่องข้อมูลซึ่งจะมีการโยงของข้อมูลในแต่ละข้อ และจะลดขอบเขตของอิทธิพล

1.2 ความไม่สมดุลของเนื้อหา เนื่องจากมีความหลากหลายในชุดข้อคำถามที่ต้องใช้ในข้อสอบแบบเทสต์เลท จึงเกิดความยากที่แตกต่างกันในแต่ละข้อ ในจุดนี้อาจกล่าวได้ว่าชุดข้อสอบเทสต์เลทนี้เป็นการวัดแบบพหุมิติ ยกตัวอย่างเช่น ข้อสอบชุดเทสต์เลทในการอ่านบทความภาษาอังกฤษ ในชุดข้อคำถามอาจจะมีข้อที่วัดทักษะการอ่าน การอ่านเพื่อความเข้าใจ หรืออาจมีเรื่องความรู้คำศัพท์จากในบทความ โดยในชุดข้อสอบที่มีการวัดหลายประเด็นนี้ทำให้ผู้สอบที่มีความสามารถในด้านใดด้านหนึ่งตอบข้อที่ถนัดถูก แต่ก็จะมีด้านที่ตอบไม่ถูกในข้อที่ไม่ถนัด ในส่วนนี้ส่งผลให้เกิดความไม่เป็นเอกมิตภายในชุดข้อสอบและสามารถแก้ปัญหาได้ด้วยวิธีการสอบแบบปรับเหมาะทางคอมพิวเตอร์แบบดั้งเดิม แต่การแก้ปัญหาด้วยชุดข้อสอบเทสต์เลทพิจารณาว่ามีประสิทธิภาพมากกว่า

2. ความแกร่ง (Robustness)

ความแกร่งจะเกี่ยวเนื่องกับความแม่นยำของการวัดด้วยข้อสอบแบบเทสต์เลท โดยในการทดสอบผู้สอบที่มีความสามารถใกล้เคียงกันจะได้รับข้อสอบชุดเดียวกัน ซึ่งจะทำให้ลดผลกระทบของการเปรียบเทียบระหว่างบุคคล ข้อสอบเทสต์เลทถูกสร้างเป็นเซตขึ้นมาโดยสมเหตุสมผลกันในชุดข้อคำถามซึ่งจะผ่านการตรวจสอบคุณภาพจากผู้เชี่ยวชาญและการตรวจสอบคุณภาพด้วยวิธีการทางสถิติจากการทดลองใช้ และเนื่องด้วยชุดข้อสอบแบบเทสต์เลทจะมาเป็นชุดหลายข้อคำถามพร้อมสถานการณ์ที่กำหนด ความคลาดเคลื่อนของผลการทดสอบที่เกิดจากวิธีการสร้างข้อสอบที่มีลักษณะเป็นเทสต์เลทที่อาจจะขึ้นคำตอบซึ่งกันและกันจะถูกชดเชยด้วยข้อดีของความแม่นยำในการวัดโดยใช้ข้อสอบแบบเทสต์เลท

3. อิทธิพลลำดับ (Order Effects)

ในการเรียงลำดับความยากของข้อจะมีการควบคุมจากจุดเริ่มต้นเพื่อให้ได้ความยุติธรรม เนื่องจากว่าคุณลักษณะของเทสต์เลทในการประมาณค่าความสามารถผู้สอบจะใกล้เคียงกับข้อสอบแบบจับคู่ (matching test) ในการที่เลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ CAT จะเริ่มจากข้อยากเพื่อป้องกันปัญหาอิทธิพลลำดับ ดังนั้นในการเลือกข้อจึงจะลำดับจากง่ายไปหายาก

การแก้ปัญหาข้อสอบแบบทดสอบ

ในโมเดลทดสอบสมัยใหม่จะเป็นโครงสร้างแบบพหุมิติ (Multidimensional Structure) โดยในชุดข้อสอบแบบทดสอบจะประกอบด้วยข้อมูลที่เน้นคุณลักษณะที่สนใจจะวัด ซึ่งแต่ละข้อคำถามในชุดทดสอบจะมีความเกี่ยวข้องกับข้ออื่น ๆ ในชุดเดียวกันมากกว่าและเหนือกว่าปัจจัยทั่วไปที่วัด เรียกว่าปัจจัยเฉพาะทดสอบ ดังนั้นจะเป็นปัจจัย $k+1$ คือปัจจัยทั่วไปและมีปัจจัยเพิ่มมาอีกหนึ่งของแต่ละทดสอบ (Wainer et al., 2007, p. 110) โดยในอดีตมีการศึกษาในปี ค.ศ. 1937 โดย Holzinger and Swineford ที่เป็นวิธีการ bi-factor ที่ขยายมาจากวิธี two-factor model ของ Spearman ในส่วนนี้จึงทำให้แต่ละข้อนั้นมี loading ทั้งสองปัจจัย คือปัจจัยทั่วไป และปัจจัยเฉพาะ ในการวิเคราะห์ bi-factor ของ Holzinger and Swineford (1937, as cited in Wainer et al., 2007, p. 110) จึงเป็นแนวทางให้เกิดการวิเคราะห์ในระดับรายข้อ และมีการนำไปใช้ในการวิเคราะห์องค์ประกอบเชิงยืนยัน โดย Joreskog (1969, as cited in Wainer et al., 2007, p. 112) ซึ่งประมาณค่าพารามิเตอร์โดยวิธีการความน่าจะเป็นสูงสุด (Maximum Likelihood)

จากการประมาณค่าของทฤษฎีต่าง ๆ ที่ผ่านมามีการใช้เทคนิคการประมาณค่าของ Monte Carlo ถึงแม้ว่าจะมีข้อจำกัดทางการคำนวณด้วยคอมพิวเตอร์ แต่ยุคปัจจุบันก็ Albert & Chib (1993, p. 678) ที่ใช้เทคโนโลยีในการวิเคราะห์ IRT ที่ประมาณค่ารูปแบบการตอบสนองแบบปกติสะสมภายใต้แนวคิดของเบส์ จึงได้มีการนำวิธีนี้มาใช้ในการพัฒนาการวิเคราะห์โมเดลทดสอบเนื่องจากให้วิธีการที่เน้นโครงสร้างของข้อสอบและวิธีการได้มาซึ่งข้อมูล แต่อย่างไรก็ตามข้อจำกัดในบางจุด ต่อมาเมื่อมีการพัฒนาวิธีการประมาณค่า Monte Carlo และสามารถนำไปใช้ได้จริง โดยวิธีการแบบเบส์ทำให้ได้จุดการประมาณค่าและการแจกแจงภายหลังทั้งหมดอีกด้วย ซึ่งค่าการแจกแจงภายหลังนี้ประกอบด้วยสัดส่วนของผู้สอบที่ผ่านจุดตัด ทำให้สามารถนำไปแก้ปัญหาของบุคคลที่ไม่ผ่านจุดตัดได้โดยการขยายขนาดของตัวอย่าง (Wainer et al., 2007, p. 114)

ในการแก้ปัญหาอิทธิพลของทดสอบเหล่านั้นมีวิธีการอยู่หลายแนวทาง เช่น จงกล บัวแก้ว และคณะ (2561, p. 142) ได้เสนอวิธีการแก้ปัญหาไว้ 3 วิธี คือ 1) การใช้ทฤษฎีการตอบสนองข้อสอบแบบการให้คะแนนสองค่าและเป็นการวัดมิติเดียว (Unidimensional Dichotomous IRT) ซึ่งละเว้นข้อตกลงเบื้องต้นและให้แต่ละข้อเป็นอิสระจากกัน 2) การใช้ทฤษฎีการตอบสนองข้อสอบแบบ Graded Response Model คือเป็นการเปลี่ยนจากการให้คะแนนแบบสองค่าของข้อสอบในทดสอบเหล่านั้น (Dichotomous) ทำให้เป็นการให้คะแนนที่มากกว่าสองค่า (Polytomous) ในชุดข้อสอบทดสอบแต่ละชุด ซึ่งมีผลทำให้ไม่สามารถวิเคราะห์สารสนเทศรายข้อได้ 3) การวิเคราะห์ตามโมเดลการตอบสนองข้อสอบแบบทดสอบที่ใช้อิทธิพลของทดสอบเหล่านั้นมาเกี่ยวข้อง

โดยวิธีการแรกในการวิเคราะห์โมเดลทดสอบแบบใช้ทฤษฎีการตอบสนองข้อสอบแบบการให้คะแนนสองค่าและเป็นการวัดมิติเดียว (Unidimensional Dichotomous IRT) ซึ่งละเมิดข้อตกลงเบื้องต้นความอิสระของข้อคำถาม (Local Independence) จะมีการใช้การประมาณค่าแบบพื้นฐานของ Kernel ที่เป็นการให้คะแนนแบบสองค่า (Dichotomous) แบบ 3 พารามิเตอร์จะมีลักษณะดังนี้ (Wainer et al., 2007, p. 114)

$$P(Y_{ij} = 1) = c_j + (1 - c_j) \logit^{-1}(t_{ij})$$

โดยค่า Y_{ij} คือ การตอบของผู้สอบ i ในข้อที่ j

c_j คือ ค่าการเดา (Lower Asymptote) สำหรับข้อสอบแบบให้คะแนนสองค่า (Dichotomous) ข้อที่ j

$\logit$ คือ $\log(x/(1-x))$

t_{ij} คือ การพยากรณ์เชิงเส้นของตัวแปรแฝงของคะแนน

วิธีการที่สอง จะเป็นการเปลี่ยนการให้คะแนนข้อคำถามให้เป็นแบบหลายค่า (Polytomous IRT) โดยจะ
เป็นการรวมคะแนนกันของข้อคำถามในทดสอบแต่ละชุด โดยมีลักษณะโมเดลดังนี้ (Wainer et al., 2007, p. 114)

$$P(Y_{ij} = r) = \Phi(d_r - t_{ij}) - \Phi(d_{r-1} - t_{ij})$$

โดยค่า Y_{ij} คือ การตอบของผู้สอบ i ในข้อที่ j

d_r และ d_{r-1} คือ จุดตัดแฝงของคะแนนที่ให้คะแนนแบบมากกว่าสองค่า (Polytomous)

Φ คือ ฟังก์ชันแจกแจงสะสมปกติ

t_{ij} คือ การพยากรณ์เชิงเส้นของตัวแปรแฝงของคะแนน

และ Wainer et al. (2007, p. 115) ได้อธิบายโมเดลที่มีความไม่อิสระจากกันด้วยการเป็นทดสอบด้วย
การขยายการพยากรณ์คะแนนเชิงเส้นตรง t_{ij} จากรูปแบบเดิมได้ดังนี้

$$t_{ij} = a_j(\theta_j - b_j)$$

โดยที่ a_j , b_j , และ θ_j แทน ค่าความชันของข้อ ค่าความยาก และความสามารถของผู้สอบ

$$t_{ij} = a_j(\theta_j - b_j - \gamma_{id(j)})$$

โดยมี γ_{id} บ่งบอกถึงผลของทดสอบเชิงความสัมพันธ์ในข้อที่ j กับผู้สอบภายในทดสอบเลข $d(j)$

ผลของทดสอบนี้เป็นความไม่อิสระของข้อสอบที่เพิ่มขึ้นมาในชุดข้อสอบทดสอบเลข ซึ่งอาจเรียกค่าเป็น
อิทธิพลร่วมกัน γ_{id} ในการพยากรณ์คะแนน ดังนั้นพารามิเตอร์อิทธิพลทดสอบเลขนี้จึงแสดงถึงความแตกต่างของ
โมเดลทดสอบเลขกับโมเดลตอบสนองข้อสอบแบบปกติ

โดยอิทธิพลของทดสอบเลขในวิธีการที่ 3 ที่จะวิเคราะห์ในรูปแบบพารามิเตอร์ที่เพิ่มมา คือ ค่าอิทธิพล
ทดสอบเลข $\gamma_{g(i)}$ โมเดลความน่าจะเป็นที่สมบูรณ์ที่มีค่าการกระจายของความน่าจะเป็นร่วมกันนั้นคงที่ ดังโมเดล
ทฤษฎีการตอบสนองทดสอบเลขมีลักษณะดังนี้ (Chen, 2014, p. 20)

$$P_{ij} = c_i + (1 - c_i) \frac{e^{1.7a_j(\theta_j - b_i + \gamma_{g(i)j})}}{1 + e^{1.7a_j(\theta_j - b_i + \gamma_{g(i)j})}}$$

P_{ij} คือ ความน่าจะเป็นในการตอบข้อสอบถูกข้อที่ i ของคนที่ j

θ_j คือ พารามิเตอร์ความสามารถของคนที่ย

a_i คือ พารามิเตอร์อำนาจจำแนกของข้อสอบ i

b_i คือ พารามิเตอร์ความยากของข้อสอบ i

c_i คือ พารามิเตอร์การเดาของข้อสอบ i

$\gamma_{g(i)j}$ คือ อิทธิพลร่วมของข้อสอบ i กับคนที่ j ภายในทดสอบเลข g

ผู้วิจัยจึงนำเสนอวิธีการที่ 4 ในการวิเคราะห์ทฤษฎีการตอบสนองข้อสอบแบบทดสอบแบบทดสอบแบบทดสอบเพื่อแก้ปัญหาความเป็นเอกพันธ์ (Unidimensional) และความเป็นอิสระจากกันของข้อคำถาม (Local Independence) โดยเป็นการใช้รูปแบบโมเดลของทฤษฎีการตอบสนองข้อสอบแบบพหุมิติแบบ 3 พารามิเตอร์ (3PL MIRT) เนื่องจากข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบแบบพหุมิติสามารถที่จะวัดได้พร้อมกันหลายมิติและแต่ละมิติมีความสัมพันธ์กันได้ (Reckase, 2009, p. 91) โดยมีสมการดังนี้

$$P(U_{ij} = 1|\theta_j, a_i, c_i, d_i) = c_i + (1 - c_i) \frac{e^{a_i \theta_j + d_i}}{1 + e^{a_i \theta_j + d_i}}.$$

P คือ ความน่าจะเป็นในการตอบข้อสอบถูก

U_{ij} คือ โอกาสที่จะตอบถูกในข้อที่ i ของคนที่ j

θ_j คือ พารามิเตอร์ความสามารถของคนที่ j

a_j คือ พารามิเตอร์อำนาจจำแนกของข้อสอบ i

c_i คือ พารามิเตอร์การเดาของข้อสอบ i

d_i คือ จุดตัดความยากของข้อสอบ i

โมเดลตอบสนองข้อสอบแบบพหุมิติ เป็นการขยายขีดความสามารถของในการประมาณค่าความสามารถของผู้สอบจากโมเดลการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional IRT) (Rackase, 2009, p. 54) โดยโมเดลตอบสนองข้อสอบแบบพหุมิตียอมให้มีการวัดคุณลักษณะหลายคุณลักษณะพร้อมกันในข้อสอบข้อเดียวและคุณลักษณะมีความสัมพันธ์กันได้ ทำให้ข้อสอบแต่ละข้อให้สารสนเทศสูงกว่าการทดสอบด้วยโมเดลตอบสนองข้อสอบแบบมิติเดียว ทำให้ผลการวัดมีความแม่นยำสูง สามารถลดจำนวนข้อสอบที่ใช้ในการทดสอบได้ประมาณ 30-50 % (Senarat et al., 2013, p. 1010) ซึ่งการลดจำนวนข้อสอบลงก็สามารถลดการขึ้นาคำตอบซึ่งกันและกันของข้อสอบลงด้วย นอกจากนี้เนื่องจากข้อสอบหนึ่งข้อต้องใช้ความสามารถหลายอย่างในการตอบทำให้ข้อสอบมีความยากสูงซึ่งจะลดการขึ้นาคำตอบระหว่างข้อได้อีกระดับหนึ่ง

ตัวอย่างงานวิจัย

Sireci, Thissen, & Wainer (1991, pp. 237-247) ได้ศึกษา On the Reliability of Testlet-Based Tests เป็นการศึกษาการประมาณค่าของข้อสอบที่มีโครงสร้างภายในแบบทดสอบเฉพาะในทางสถิติโดยเน้นในเรื่องความเที่ยง ได้นำเสนอถึงการวิเคราะห์ข้อสอบตามทฤษฎีตอบสนองข้อสอบสำหรับการนำไปใช้ในข้อสอบชุดทดสอบเฉพาะ โดยพบว่าข้อสอบความเข้าใจในการอ่านมีความสัมพันธ์รายข้อภายในบทความเดียวกันสูงกว่าความสัมพันธ์ของข้อคำถามระหว่างชุดทดสอบเฉพาะ

Wainer, Bradlow, & Wang (2001) ได้ศึกษาเรื่อง Using a New Statistical Model for Testlets to Score TOEFL ซึ่งเป็นการศึกษา 3 พารามิเตอร์ของ IRT และพารามิเตอร์ความไม่เป็นอิสระของเทสต์เลท โดยใช้ข้อสอบ TOEFL ชุดเทสต์เลท 86 ชุดที่เป็นชุดเทสต์เลทการอ่าน 50 ชุดมี 699 ข้อและชุดเทสต์เลทข้อสอบการฟัง 36 ชุด มี 290 ข้อ มีผู้สอบ 2,000 คน ผลพบว่าพารามิเตอร์ความยากประมาณค่าได้ดีในข้อสอบเทสต์เลททั้งสองแบบ แต่พารามิเตอร์อำนาจจำแนกและพารามิเตอร์การเดานั้นมีความเอนเอียงเนื่องจากเงื่อนไขความไม่เป็นอิสระจากกันของข้อคำถาม อีกทั้งยังส่งผลให้สารสนเทศของข้อสอบเกิดการประมาณค่าเกินจริง over-estimated อีกด้วย

Paap, Glas, & Veldkamp (2013) ได้เขียนบทความเรื่อง Minimizing the Testlet Effect: Identifying Critical Testlet ซึ่งเป็นงานที่ศึกษาในเรื่องของการลดขนาดอิทธิพลของเทสต์เลทในข้อสอบมาตรฐานที่เป็นชุดเทสต์เลท (standardized test) ซึ่งผลพบว่าคุณลักษณะของข้อสอบเทสต์เลทจะเกี่ยวกับขนาดอิทธิพลของเทสต์เลท ถ้าข้อสอบในชุดเทสต์เลทใดมีความง่ายก็จะลดขนาดอิทธิพลของเทสต์เลท หรือข้อที่มีความยากปานกลางก็จะมีผลของข้อมูลแต่ละข้อในชุดนั้น งานวิจัยชิ้นนี้ได้แนะนำว่าต่อไปควรจะมีการศึกษาความสัมพันธ์ระหว่างคุณลักษณะข้อสอบเทสต์เลทให้เป็นตัวแปรอิสระกับความยากรายข้อโดยเปลี่ยนให้เป็นตัวแปรตาม

Eckes & Baghaei (2015, pp. 85-98) ได้ศึกษาวิจัยเรื่อง “Using Testlet Response Theory to Examine Local Dependence in C-Tests” เป็นการใช้ทฤษฎีตอบสนองข้อสอบเทสต์เลทที่จะศึกษาความไม่เป็นอิสระต่อกันในข้อสอบแบบ C-test หรือเติมในช่องว่างซึ่งเป็นข้อสอบวัดความสามารถทางภาษา อิงตามทฤษฎีตอบสนองข้อสอบเทสต์เลท โดยมีการใช้โปรแกรม SCORIGHT ในการประมาณค่า 2 พารามิเตอร์ (2-PL TRT) โดยใช้กลุ่มตัวอย่าง 4,708 คนในการที่จะทำข้อสอบชุดเทสต์เลท C-test 8 บทความ พบว่าขนาดอิทธิพลของเทสต์เลทนั้นยังไม่มาก มีความแปรปรวนอยู่ระหว่าง 0.09-0.18 และการประมาณค่าพารามิเตอร์ออกมาเป็นรูปแบบการให้คะแนนแบบมากกว่าสองค่า (polytomous IRT) และทฤษฎีการตอบสนองข้อสอบแบบปกติ (IRT) ที่ละเว้นข้อตกลงเบื้องต้นเรื่องความเป็นอิสระของข้อสอบ (local independence) จากการประมาณค่าความเป็นอิสระจากกันแล้ว จึงสรุปได้ว่าข้อตกลงเบื้องต้นนี้ไม่เป็นปัญหาสำหรับข้อสอบชุด C-test แบบเทสต์เลท

จกมล บัวแก้ว, ไพรัตน์ วงษ์นาม และสมพงษ์ ปันหุ่น (2561, น. 139-155) ได้ทำงานวิจัยเรื่อง “การประมาณค่าพารามิเตอร์ของแบบสอบเทสต์เลทตามโมเดลทฤษฎีการตอบสนองข้อสอบและโมเดลทฤษฎีการตอบสนองเทสต์เลท” โดยมีวัตถุประสงค์เพื่อศึกษาการประมาณค่าพารามิเตอร์ข้อสอบประกอบด้วย ค่าอำนาจจำแนก ค่าความยาก ค่าการเดา และพารามิเตอร์ความสามารถของผู้สอบ ที่ประมาณค่าตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (IRT 3PL) และโมเดลทฤษฎีการตอบสนองเทสต์เลทแบบ 3 พารามิเตอร์ (TRT 3PL) ผลที่ได้คือการประมาณค่า 3 พารามิเตอร์ของ IRT ทั้งสองแบบมีค่าสัมประสิทธิ์สหสัมพันธ์สูงมาก และค่าเฉลี่ยของความคลาดเคลื่อนทั้งสามพารามิเตอร์ของทั้งสองแบบมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

สรุป

ทฤษฎีการตอบสนองข้อสอบเทสต์เลทเป็นการวิเคราะห์ข้อสอบรายข้อของชุดข้อสอบเทสต์เลท ซึ่งมีลักษณะเฉพาะเป็นรูปแบบสิ่งเร้าเดียวกันเช่น บทความ ตาราง ข้อมูล บทสนทนา ฯลฯ หรือเรียกว่าข้อสอบเชิงสถานการณ์ ในทางการทดสอบภาษาอังกฤษจะมีด้านความเข้าใจในการอ่าน (Reading Comprehension) และข้อสอบการฟัง (Listening Comprehension) โดยจะมีข้อคำถามหลายข้อในแต่ละชุด จึงเกิดการละเมิดข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบในประเด็นที่ว่าข้อสอบจะต้องเป็นอิสระจากกัน (Local Independence) เนื่องจากว่าข้อสอบแบบเทสต์เลทเป็นข้อสอบเชิงสถานการณ์ที่เป็นชุดข้อคำถาม มีการต้องพึ่งพากันของข้อคำถามแต่ละข้อและอาจมีการชี้นำคำตอบซึ่งกันและกัน จากการศึกษาเอกสารและงานวิจัยที่ผ่านมาได้มีการเสนอการแก้ปัญหาการละเมิดข้อตกลงเบื้องต้นของข้อสอบแบบเทสต์เลทไว้ 4 รูปแบบ โดย 3 รูปแบบแรกเป็นการรวบรวมจากนักวิชาการคนอื่น ๆ และในรูปแบบที่ 4 เป็นรูปแบบที่ผู้เขียนนำเสนอ ดังนี้ รูปแบบที่ 1 ประมาณค่าตามทฤษฎีการตอบสนองข้อสอบโดยละเลยการละเมิดข้อตกลงเบื้องต้น รูปแบบที่ 2 ประมาณค่าตามทฤษฎีการตอบสนองข้อสอบแบบหลายค่า (Polytomous IRT) รูปแบบที่ 3 ประมาณค่าตามทฤษฎีการตอบสนองข้อสอบ ซึ่งมีการประมาณค่าพารามิเตอร์ที่เพิ่มเติมเป็นแบบ 4 พารามิเตอร์ โดยพารามิเตอร์ที่ 4 คืออิทธิพลของเทสต์เลท และรูปแบบที่ 4 เป็นการประมาณค่าทฤษฎีการตอบสนองแบบพหุมิติด้วยโมเดล 3 พารามิเตอร์

บรรณานุกรม

- จกกล บัวแก้ว, ไพรัตน์ วงษ์นาม และสมพงษ์ ปันพูน. (2561). การประมาณค่าพารามิเตอร์ของแบบสอบทดสอบทดสอบตามโมเดลทฤษฎีการตอบสนองข้อสอบและโมเดลทฤษฎีการตอบสนองทดสอบ. *วารสารราชภัฏสุราษฎร์ธานี*, 5(1), 139-155.
- Buakaew, J., Wongnam, P., & Panhoon, S. (2018). Estimated parameters of the testlet-test under item response theory and testlet response theory model. *Suratthani Rajabhat Journal*, 5(1), 139-155. [in Thai].
- ศิริชัย กาญจนวาสิ. (2555). *ทฤษฎีการทดสอบแนวใหม่* (พิมพ์ครั้งที่ 4). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- Kanjanawasee, S. (2012). *Modern test theory* (4th ed). Bangkok: Chulalongkorn University Printing House. [in Thai].
- วิรัช วรรณรัตน์. (2558). หลักและวิธีการสอบวัด. *วารสารมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยราชพฤกษ์*, 1(2), 1-12.
- Wannarat, W. (2015). Principles and measurement methods. *Journal of Humanities and Social Sciences, Rajapruk University*, 1(2), 1-12. [in Thai].
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669–679.
- Chen, J. (2014). *Model selection for IRT equating of Testlet-based tests in the random groups design* (Doctoral dissertation). University of Iowa, United States of America.
- DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. Oxford University Press: Oxford.
- Eckes, T., & Baghaei P. (2015). Using testlet response theory to examine local dependence in C-Tests. *Applied Measurement in Education*, 28(2), 85-98.
- Fox, J.P., (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41–54.
- Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
- Loken, E. & Rulison, K. L. (2010). Estimation of a 4-parameter Item response theory model. *The British Journal of Mathematical and Statistical Psychology*, 63(3), 509-525.
- Paap, M. C. S., Glas, C. A. W., & Veldkamp, B. P. (2013). An overview of research on the testlet effect: Associated features, implications for test assembly, and the impact of model choice on ability estimates, in *LSAC Research Reports Series (LR 13-03)*. Newtown, PA: Law School Admission Council.

- Reckase, M.D., (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Rosenbaum, P. R. (1988). A note on item bundles. *Psychometrika*, 53, 349–360.
- Senarat, S., Tayraukham, S., Piyapimonsit, C., & Tongkhambanjong, S. (2013). Development of a computerized adaptive testing for diagnosing the cognitive process of grade 7 students in learning algebra, using multidimensional item response theory. *Educational Research and Reviews*, 8(13), 1009-1021.
- Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, 28(3), 237-247.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.
- Wainer, H., & Wang, X. (2001). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37(3), 203 – 220.
- Wheadon, B. C. (2010). *An item response theory approach to the maintenance of standards in public examination in England* (Doctoral dissertation). Durham University, England.