



สรุป อภิปรายผลและข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์หลักเพื่อพัฒนาเกณฑ์ตัดสินข้อสอบล่าเอียง ระหว่างผู้สอบหญิงและชาย ด้วยข้อมูลเชิงประจักษ์ของดัชนีตัดสินข้อสอบล่าเอียง 4 ตัว คือ SA UA α_{MH} และ β_{SIB} ซึ่งเป็นดัชนีจากวิธีการวิเคราะห์ความล่าเอียง 3 วิธี คือ วิธีทฤษฎีการตอบข้อสอบโมเดล 2 พารามิเตอร์ (IRT-2 พารามิเตอร์) วิธีของ Mantel และ Heanszel (MH) และวิธี SIBTEST โดยมีวัตถุประสงค์ เฉพาะ 4 ข้อ คือ

1. เพื่อพัฒนาเกณฑ์ตัดสินข้อสอบล่าเอียงทางเพศด้วยข้อมูลเชิงประจักษ์ในวิชาภาษาอังกฤษและวิชาคณิตศาสตร์จากวิธี IRT-2 พารามิเตอร์ วิธี MH และวิธี SIBTEST
2. เพื่อพิจารณาผลการทดลองใช้เกณฑ์ที่พัฒนาจากข้อมูลเชิงประจักษ์ในประเด็นต่อไปนี้

- 2.1 ผลการตัดสินข้อสอบล่าเอียงทางเพศวิชาภาษาอังกฤษ และวิชาคณิตศาสตร์ เมื่อมีความยาวแบบสอบและจำนวนผู้สอบต่างกัน
- 2.2 ความสอดคล้องในผลการตัดสินข้อสอบล่าเอียงทางเพศในวิชาภาษาอังกฤษ และวิชาคณิตศาสตร์ เมื่อเปรียบเทียบกับผลจากการทดสอบความมีนัยสำคัญของค่าสถิติ
- 2.3 ลักษณะของความล่าเอียงของข้อสอบระหว่างเพศของผู้สอบในวิชาภาษาอังกฤษ และวิชาคณิตศาสตร์

ทั้งนี้วิธีการดำเนินงานและผลการวิจัย โดยสังเขป ดังนี้

1. การพัฒนาเกณฑ์การตัดสินความล่าเอียง
ดัชนีที่นำมาศึกษาเพื่อพัฒนาเกณฑ์ในครั้งนี้ผู้วิจัยได้เลือกดัชนีที่เป็นที่นิยมใช้ 4 ตัว ได้แก่ ดัชนีพื้นที่ความแตกต่างระหว่างโค้งการตอบข้อสอบชนิดมีเครื่องหมาย (SA) และชนิดไม่มีเครื่องหมาย (UA) ซึ่งมีพื้นฐานตามทฤษฎีการตอบข้อสอบ ดัชนี α_{MH} ซึ่งเสนอขึ้นโดย Mantel และ Heanszel และ β_{SIB} จากวิธี SIBTEST เสนอขึ้นโดย Shealy และ Stout ทั้งนี้เพื่อไม่ต้องใช้การทดสอบค่าสถิติซึ่งโดยมากจะมีความไวต่อขนาดของกลุ่มผู้สอบที่นำมาใช้ในการ

วิเคราะห์มีวิธีการกำหนดเกณฑ์การตัดสินใจจากค่าเฉลี่ยของค่าดัชนีที่ได้จากผู้สอบเพศเดียวกัน โดยพิจารณาค่าที่เบี่ยงเบนออกจากค่าดัชนีที่แสดงถึงการไม่มีความลำเอียงระหว่างผู้สอบเพศเดียวกัน ทั้งนี้โดยหวังว่าจำนวนกลุ่มตัวอย่างที่นำมาใช้ในการวิเคราะห์จะช่วยแสดงขอบเขตค่าดัชนีที่แสดงถึงความไม่ลำเอียง เมื่อใช้ข้อมูลจากการตอบข้อสอบจริง และจะสามารถกำหนดค่าดัชนีที่แสดงว่าข้อสอบลำเอียงได้ มีขั้นตอนการวิเคราะห์ ดังนี้

1.1 ข้อมูลและกลุ่มตัวอย่าง

1.1.1 แบบสอบและผลการตอบข้อสอบ

ข้อมูลที่นำมาใช้ในการวิจัยคือ ผลการตอบข้อสอบคัดเลือกบุคคลเพื่อศึกษาต่อในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัย ปีการศึกษา 2535 2 วิชา คือ วิชาภาษาอังกฤษ กข ซึ่งเป็นแบบสอบเลือกตอบ 4 ตัวเลือก ความยาว 100 ข้อ และวิชาคณิตศาสตร์ กข ซึ่งเป็นแบบสอบเลือกตอบ 4 ตัวเลือก จำนวน 41 ข้อ และข้อสอบเติมค่า 6 ข้อ โดยนำมาใช้ในการวิจัยเฉพาะข้อมูลจากข้อสอบเลือกตอบ ผู้วิจัยได้จัดกลุ่มข้อสอบแต่ละวิชาตามเนื้อหาที่วัด และตัดข้อสอบที่มีค่าอำนาจจำแนกต่ำ และมีค่าความยากสูงมากหรือต่ำมากออกจนเหลือข้อสอบภาษาอังกฤษ 80 ข้อ คณิตศาสตร์ 40 ข้อ จากนั้นใช้วิธีสุ่มแบบง่าย สุ่มข้อสอบจากแต่ละกลุ่มเนื้อหาตามสัดส่วนข้อสอบที่มีอยู่ จัดเป็นแบบสอบ 50 60 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ และ 20 30 และ 40 ข้อ สำหรับวิชาคณิตศาสตร์ โดยที่แบบสอบทุกขนาดมีเนื้อหาที่วัดครบตามฉบับสมบูรณ์ และมีค่าเฉลี่ยของความยากและอำนาจจำแนกใกล้เคียงกันทุกความยาวและแบบสอบภายในแต่ละวิชา

1.1.2 กลุ่มผู้สอบ

กลุ่มผู้สอบที่ใช้ในการวิจัยรวม 6 ขนาด ได้แก่ ขนาด 100 200 400 600 800 และ 1,000 คน นั้น ผู้วิจัยสุ่มจากกลุ่มผู้สอบแยกตามเพศรวมเพศละ 101 กลุ่ม นำไปใช้ในการพัฒนาเกณฑ์ 100 กลุ่ม และวิเคราะห์ความลำเอียง 1 กลุ่ม ในแต่ละขนาดผู้สอบและแต่ละวิชา โดยใช้การสุ่มแบบง่ายแบบไม่ใส่คืนด้วยโปรแกรม SAS รวมเป็นแฟ้มข้อมูลที่นำมาใช้ในการศึกษารั้งนี้ 2 เพศ X 101 กลุ่ม X 6 ขนาด ผู้สอบ X 7 ความยาวแบบสอบรวมเป็น 8,484 แฟ้ม

1.2 วิธีการพัฒนาเกณฑ์

1.2.1 การวิเคราะห์ค่าดัชนี SA และ UA

- 1) ใช้โปรแกรม BILOG 3.04 วิเคราะห์ค่าพารามิเตอร์ของข้อสอบแยกแต่ละเพศแต่ละความยาวแบบสอบและแต่ละขนาดผู้สอบ
- 2) สุ่มจับคู่ค่าพารามิเตอร์ที่ได้จากการวิเคราะห์จากผู้สอบเพศเดียวกัน (หญิงกับหญิง และชายกับชาย) ในความยาวแบบสอบและขนาดผู้สอบเดียวกัน จะได้ผลการจับคู่ระหว่างผู้สอบแต่ละเพศ ในแต่ละความยาวแบบสอบและขนาดผู้สอบจำนวน 50 คู่

3) เทียบมาตรฐานของค่าพารามิเตอร์แต่ละคู่ด้วยโปรแกรม

EQUATE

4) คำนวณพื้นที่ระหว่างโค้งการตอบข้อสอบทั้ง SA และ UA ของผู้สอบแต่ละคู่ด้วยโปรแกรม AREAB

1.2.2 การวิเคราะห์ค่าดัชนี α_{MH} และ β_{SIB}

ใช้โปรแกรม SIBIN และ SIBTEST คำนวณค่า α_{MH} และ β_{SIB} ระหว่างข้อมูลการตอบข้อสอบของผู้สอบคู่ที่ใช้ในการวิเคราะห์ค่า SA และ UA

1.2.3 วิเคราะห์ค่าเฉลี่ยของค่า SA UA α_{MH} และ β_{SIB}

ใช้โปรแกรม SPSS-PC วิเคราะห์ค่าเฉลี่ยของ SA UA α_{MH} และ β_{SIB} รายข้อและรวมทุกข้อตามขนาดผู้สอบและความยาวแบบสอบ และรวมทุกข้อทุกขนาดผู้สอบและทุกความยาวแบบสอบ

1.2.4 นำค่าเฉลี่ยที่ได้จากการวิเคราะห์รวมทุกข้อในข้อ 1.2.3 มากำหนดเกณฑ์ตัดสินข้อสอบล่าเอียง

2. การวิเคราะห์ความล่าเอียงของข้อสอบ

2.1 วิเคราะห์ค่าสถิติเบื้องต้นของข้อมูล ได้แก่ ค่าเฉลี่ยของคะแนนดิบ ค่าเฉลี่ยของคะแนนความสามารถ ค่าพารามิเตอร์ความยาก และค่าพารามิเตอร์อำนาจจำแนก

2.2 คำนวณดัชนีแสดงความล่าเอียงของข้อสอบทั้ง 4 ตัว ระหว่างผู้สอบชาย และหญิงด้วยวิธีการเช่นเดียวกับข้อ 1.2.1 และ 1.2.2 จำแนกการวิเคราะห์ตามขนาดผู้สอบและความยาวแบบสอบ

สรุปผลการวิจัย

1. เกณฑ์ที่กำหนดจากค่าเฉลี่ยของดัชนี SA UA α_{MH} และ β_{SIB}

1.1 เกณฑ์ของค่าดัชนี SA

ผลจากการวิเคราะห์ค่าเฉลี่ยของค่า SA วิชาคณิตศาสตร์พบว่า ค่าเฉลี่ยรวมทุกข้อจำแนกตามขนาดผู้สอบมีค่าระหว่าง .47-.58 ส่วนเบี่ยงเบนมาตรฐานระหว่าง .27-.33 ซึ่งมีค่าสูงกว่าค่าที่ได้จากวิชาภาษาอังกฤษ ในขณะที่ค่าเฉลี่ยรวมทุกข้อจำแนกตามความยาวแบบสอบมีค่าระหว่าง .34-.67 ซึ่งแตกต่างกันมากกว่าค่าที่ได้จากการวิเคราะห์ตามขนาดผู้สอบ ค่าเฉลี่ยรวมทุกข้อทุกขนาดผู้สอบและทุกความยาวแบบสอบมีค่าเท่ากับ .467 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ .30 และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยเท่ากับ .01

สำหรับวิชาภาษาอังกฤษพบว่า ค่าเฉลี่ยรวมทุกข้อจำแนกตามขนาดผู้สอบมีค่าใกล้เคียงกันมากคือ มีค่าระหว่าง .21-.27 ส่วนเบี่ยงเบนมาตรฐานระหว่าง .10-.31 ในขณะที่ค่าเฉลี่ยจำแนกตาม ความยาวแบบสอบมีค่าระหว่าง .22-.31 ส่วนเบี่ยงเบนมาตรฐาน .15-.27 โดยมีค่าเฉลี่ยรวม ทุกข้อ ทุกความยาวแบบสอบและทุกขนาดผู้สอบมีค่า .25 ส่วนเบี่ยงเบนมาตรฐาน .041 และ ความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยมีค่าเท่ากับ .02 เกณฑ์ที่กำหนดโดยพิจารณาจากค่าเฉลี่ย รวมทุกข้อทุกขนาดผู้สอบ และทุกความยาวแบบสอบ สำหรับวิชาคณิตศาสตร์และวิชาภาษาอังกฤษมี ค่าต่างกันเนื่องจากผลการวิเคราะห์ต่างกัน ดังนี้

เกณฑ์สำหรับวิชาคณิตศาสตร์ในการตัดสินว่าข้อสอบล่าเอียงเมื่อใช้ดัชนี SA คือ $|SA| > .75$ และเกณฑ์สำหรับวิชาภาษาอังกฤษ คือ $|SA| > .35$ เมื่อพิจารณาใช้ค่าเฉลี่ย รวมทุกข้อทุกความยาวแบบสอบและทุกขนาดผู้สอบ แต่เมื่อพิจารณาความแตกต่างของค่าเฉลี่ยข้าม ความยาวแบบสอบและขนาดผู้สอบแล้วเกณฑ์ที่กำหนดคือ $|SA| > .75$ สำหรับวิชาภาษาอังกฤษ ส่วนวิชาคณิตศาสตร์ที่ความยาวแบบสอบ 20 และ 30 ข้อ $|SA| > .70$ และที่ความยาวแบบ สอบ 40 ข้อ $|SA| > .80$

1.2 เกณฑ์ของดัชนี UA

ค่าเฉลี่ยรวมทุกข้อสำหรับวิชาภาษาอังกฤษจำแนกตามขนาดผู้สอบมีค่า ระหว่าง .403-.422 ส่วนเบี่ยงเบนมาตรฐาน .22-.23 ในขณะที่ค่าเฉลี่ยรวมทุกข้อจำแนกตาม ความยาวแบบสอบมีค่าระหว่าง .40-.45 ส่วนเบี่ยงเบนมาตรฐานมีค่าระหว่าง .22-.23 ซึ่งพบ ว่ามีค่าใกล้เคียงกันมากทั้งผลจากการวิเคราะห์ตามขนาดผู้สอบและตามความยาวแบบสอบ

ส่วนค่าเฉลี่ยรวมทุกข้อทุกขนาดผู้สอบ และทุกความยาวแบบสอบมีค่าเท่ากับ .423 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ .021 และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยมี ค่าเท่ากับ .010

สำหรับวิชาคณิตศาสตร์ค่า UA เฉลี่ยรวมทุกข้อจำแนกตามขนาดผู้สอบมีค่า ระหว่าง .744-.889 ส่วนเบี่ยงเบนมาตรฐานระหว่าง .412-.469 และเมื่อจำแนกตามความ ยาวแบบสอบ พบว่ามีค่าเฉลี่ยแตกต่างกันมากกว่า โดยมีค่าระหว่าง .57-1.07 ส่วนเบี่ยงเบน มาตรฐาน .22-.42 มีค่าเฉลี่ยรวมทุกข้อทุกขนาดผู้สอบ และทุกความยาวแบบสอบเท่ากับ .81 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ .42 และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยมีค่าเท่ากับ .02

เกณฑ์สำหรับค่า UA ที่ตัดสินว่าข้อสอบล่าเอียงในวิชาภาษาอังกฤษ คือ $UA > .50$ และในวิชาคณิตศาสตร์ คือ $UA > 1.20$ เมื่อไม่พิจารณาความแตกต่างของค่าเฉลี่ย ข้ามความยาวแบบสอบและขนาดผู้สอบ แต่เมื่อพิจารณาความแตกต่างของค่าเฉลี่ยในบริบทดังกล่าว เกณฑ์ของ UA ในการตัดสิน ข้อสอบล่าเอียงคือ $UA > .50$ สำหรับวิชาภาษาอังกฤษ วิชา

คณิตศาสตร์ เมื่อความยาวแบบสอบ 20 และ 30 ข้อ เกณฑ์คือ $UA > 1.00$ และเมื่อความยาวแบบสอบ 40 ข้อ เกณฑ์คือ $UA > 1.20$

1.3 เกณฑ์ของดัชนี α_{MH}

ค่าเฉลี่ยของดัชนี α_{MH} รวมทุกข้อ สำหรับวิชาภาษาอังกฤษจำแนกตามขนาดผู้สอบ ตั้งแต่ 100-1,000 คน มีค่าตามลำดับดังนี้ 1.39 1.24 1.15 1.12 1.10 และ 1.09 ตามลำดับ ส่วนเบี่ยงเบนมาตรฐานเท่ากับ .08 .04 .02 .02 .01 และ .01 ส่วนวิชาคณิตศาสตร์ค่าเฉลี่ยตามขนาดผู้สอบมีค่า 1.37 1.23 1.15 1.12 1.10 และ 1.18 มีส่วนเบี่ยงเบนมาตรฐานเท่ากับ .08 .04 .02 .02 .01 และ .11 เมื่อเรียงลำดับตามความยาวแบบสอบ ค่าเฉลี่ยทั้งวิชาภาษาอังกฤษและวิชาคณิตศาสตร์มีค่า 1.18 เท่ากันทุกความยาว โดยมีส่วนเบี่ยงเบนมาตรฐานเท่ากับ .01 เท่ากันหมดเช่นกัน ส่วนค่าเฉลี่ยรวมทุกข้อทุกความยาวแบบสอบและทุกขนาดผู้สอบ สำหรับวิชาภาษาอังกฤษมีค่า 1.18 ส่วนเบี่ยงเบนมาตรฐาน .115 และค่าความคลาดเคลื่อนมาตรฐานเท่ากับ .00 สำหรับวิชาภาษาอังกฤษ ส่วนวิชาคณิตศาสตร์มีค่าเฉลี่ยรวมทุกข้อเท่ากับ .87 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ .058 ค่าความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยมีค่าเท่ากับ .00

เกณฑ์สำหรับค่า α_{MH} ที่จะตัดสินว่าข้อสอบล่าเอียงทั้งในวิชาคณิตศาสตร์และวิชาภาษาอังกฤษใช้ค่าเดียวกันคือ $.70 > \alpha_{MH} > 1.30$ เมื่อไม่ได้พิจารณาค่าเฉลี่ยที่แตกต่างกันข้ามความยาวแบบสอบและขนาดผู้สอบ เมื่อพิจารณาความแตกต่างในบริบทดังกล่าว เกณฑ์สำหรับค่า α_{MH} ทั้งวิชาคณิตศาสตร์และวิชาภาษาอังกฤษใช้ค่าเดียวกันคือ $.50 > \alpha_{MH} > 1.50$ สำหรับขนาดผู้สอบ 100 คน $.60 > \alpha_{MH} > 1.40$ สำหรับขนาดผู้สอบ 200 คน และ $.70 > \alpha_{MH} > 1.30$ สำหรับผู้สอบขนาด 400 คนขึ้นไป

1.4 เกณฑ์ของดัชนี β_{SIB}

ผลการวิเคราะห์ค่าเฉลี่ยของ β_{SIB} สำหรับวิชาภาษาอังกฤษ พบว่าค่าเฉลี่ยรวมทุกข้อจำแนกตามขนาดผู้สอบมีค่าระหว่าง .02-.06 ส่วนเบี่ยงเบนมาตรฐานมีค่าระหว่าง .00-.01 ในขณะที่ค่าเฉลี่ยรวมทุกข้อจำแนกตามขนาดผู้สอบมีค่าเท่ากันคือ .03 และส่วนเบี่ยงเบนมาตรฐานมีค่าเท่ากับเช่นกันคือ .02 ค่าเฉลี่ยรวมทุกข้อทุกขนาดผู้สอบและทุกความยาวแบบสอบมีค่า .032 ส่วนเบี่ยงเบนมาตรฐาน .016 และความคลาดเคลื่อนมาตรฐานเท่ากับ .007

ค่าเฉลี่ยรวมทุกข้อ สำหรับวิชาคณิตศาสตร์จำแนกตามขนาดผู้สอบ มีค่าระหว่าง .02-.07 ส่วนเบี่ยงเบนมาตรฐาน .00-.01 ในขณะที่ค่าเฉลี่ยจำแนกตามความยาวแบบสอบมีค่า .03-.04 ส่วนเบี่ยงเบนมาตรฐานเท่ากันทุกความยาวคือ .02 และมีค่าเฉลี่ยรวมทุกข้อทุกความยาวแบบสอบและทุกขนาดผู้สอบเท่ากับ .04 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ .02 และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยมีค่าเท่ากับ .007

เกณฑ์สำหรับดัชนี β_{SIB} สำหรับข้อสอบล่าเอียงทั้งในวิชาภาษาอังกฤษ และวิชาคณิตศาสตร์ กำหนดเท่ากันคือ $|\beta_{SIB}| > .09$ เมื่อไม่ได้พิจารณาค่าเฉลี่ยที่แตกต่างกันข้ามความยาวแบบสอบและขนาดผู้สอบ แต่เมื่อพิจารณาความแตกต่างในบริบทดังกล่าว เกณฑ์สำหรับตัดสินของดัชนี β_{SIB} สำหรับวิชาคณิตศาสตร์และวิชาภาษาอังกฤษ คือ $|\beta_{SIB}| > .08$ สำหรับขนาดผู้สอบ 100 คน $|\beta_{SIB}| > .06$ สำหรับขนาดผู้สอบ 200 คน และ $|\beta_{SIB}| > .05$ สำหรับขนาดผู้สอบ 400 คนขึ้นไป

2. ผลการตัดสินความล่าเอียงของข้อสอบ

สำหรับข้อมูลจากการตอบข้อสอบจากผู้สอบหญิงและชายที่นำมาใช้ในการวิเคราะห์ค่าดัชนี SA UA α_{MH} และ β_{SIB} นั้น พบว่าค่าเฉลี่ยของคะแนนดิบที่ได้จากการตอบข้อสอบของกลุ่มผู้สอบหญิงและชายมีค่าไม่ต่างกันมากนัก ในแต่ละความยาวแบบสอบและขนาดผู้สอบ โดยคะแนนเฉลี่ยของผู้สอบหญิงมีค่าระหว่าง 23.13-36.23 และของผู้สอบชายมีค่าระหว่าง 22.21-34.95 มีค่าพารามิเตอร์ ความสามารถ (θ) สำหรับผู้สอบหญิงระหว่าง -1.873 และ 3.447 และสำหรับผู้สอบชายระหว่าง -1.721 และ 3.291 และมีค่าเฉลี่ยระหว่าง .02 และ .08 และระหว่าง -.01 และ .00 สำหรับผู้สอบหญิงและชายตามลำดับ นอกจากนี้พบว่า โดยเฉลี่ยแล้วแบบสอบมีความยากสำหรับกลุ่มผู้สอบชายมากกว่ากลุ่มผู้สอบหญิงเล็กน้อย ยกเว้นที่ความยาวแบบสอบ 50 ข้อ ขนาดผู้สอบ 100 คน ที่แบบสอบมีความยากสำหรับผู้สอบหญิงมากกว่าผู้สอบชายเล็กน้อย ในขณะที่มีค่าอำนาจจำแนกเฉลี่ยของแบบสอบใกล้เคียงกันทั้ง 2 เพศ

เมื่อพิจารณาถึงจำนวนข้อสอบที่ได้รับการตัดสินว่าล่าเอียงในกรณีที่ใช้การทดสอบค่าสถิติ พบว่าจำนวนข้อสอบล่าเอียงที่ได้จากวิธีทั้ง 3 วิธี มีจำนวนแตกต่างกันโดยที่วิธี MH และวิธี SIBTEST มีความใกล้เคียงกันมากกว่า และเมื่อขนาดผู้สอบเพิ่มขึ้นภายใต้ความยาวเดียวกัน จำนวนข้อสอบล่าเอียงจะเพิ่มมากขึ้น เมื่อใช้ดัชนี SA และ UA วิธี MH และวิธี SIBTEST จะตรงข้ามคือจำนวนข้อสอบล่าเอียงจะมากขึ้นเมื่อจำนวนผู้สอบน้อยลง

3. ความสอดคล้องในการตัดสินข้อสอบล่าเอียง

สำหรับความสอดคล้องในการตัดสินข้อสอบล่าเอียงระหว่างวิธี IRT-2 พารามิเตอร์ วิธี MH และวิธี SIBTEST ที่พบในวิชาภาษาอังกฤษมีค่าค่อนข้างต่ำ ส่วนใหญ่มีความสอดคล้องร้อยละต่ำกว่า 50.00 และส่วนมากพบว่าความสอดคล้องระหว่างวิธี MH และวิธี SIBTEST จะสูงกว่าวิธี IRT-2 พารามิเตอร์ และวิธี MH หรือ วิธี IRT-2 พารามิเตอร์ และวิธี SIBTEST ทั้งนี้พบทั้งในวิชาภาษาอังกฤษและคณิตศาสตร์ เมื่อพิจารณาถึงจำนวนข้อสอบที่ได้รับการตัดสินว่าล่าเอียงภายในดัชนีเดียวกันแต่ขนาดผู้สอบต่างกันและ/หรือความยาวแบบสอบต่างกัน พบว่ามีความไม่คงที่ในจำนวนข้อสอบล่าเอียงในแต่ละความยาวแบบสอบและขนาดผู้สอบ รูปแบบของความไม่คงที่เท่าที่พบ คือ ในวิชาภาษาอังกฤษ กรณีใช้ดัชนี SA และ UA ส่วนมากแล้วจำนวน

และร้อยละของข้อสอบล่าเอียงจะเพิ่มขึ้นที่ขนาดผู้สอบ 800 และ 1,000 คน และเมื่อความยาวแบบสอบเพิ่มขึ้น ซึ่งมีลักษณะคล้ายกับผลการทดสอบค่า χ^2 ที่มีความไวต่อการเพิ่มของจำนวนผู้สอบและความยาวแบบสอบ แต่ผลจากดัชนี α_{MH} และ β_{SIB} จะตรงกันข้ามนั่นคือ เมื่อขนาดผู้สอบเพิ่มขึ้น จำนวนข้อสอบล่าเอียงลดลงจนถึงขนาดผู้สอบ 600 ลงไปจำนวนข้อสอบล่าเอียงจะค่อนข้างคงที่ และพบว่าร้อยละของแบบสอบล่าเอียงจะเพิ่มขึ้นจากเดิมเพียงเล็กน้อยเมื่อความยาวแบบสอบเพิ่มขึ้น ผลในลักษณะดังกล่าวนี้พบทั้งในกรณีของวิชาภาษาอังกฤษและวิชาคณิตศาสตร์ เมื่อพิจารณาความสอดคล้องในการตัดสินข้อสอบล่าเอียง พบว่าในการใช้ความยาวแบบสอบและขนาดผู้สอบต่างกันภายในดัชนีเดียวกันความสอดคล้องในการตัดสินข้อสอบล่าเอียงด้วยดัชนีทั้ง 4 ไม่ค่อยคงที่ ผลในประเด็นนี้สำหรับดัชนีทุกตัว ไม่สามารถสรุปออกมาได้อย่างเป็นระบบ เนื่องจากลักษณะการลดลงหรือเพิ่มขึ้นของข้อสอบล่าเอียงมีความแตกต่างกันไป ความสอดคล้องที่พบ ไม่ว่าจะเป็นในด้านของขนาดผู้สอบหรือความยาวแบบสอบที่ต่างกันมีค่าค่อนข้างต่ำ ส่วนมากจะต่ำกว่าร้อยละ 50.00

เมื่อวิเคราะห์ความสอดคล้องในการตัดสินข้อสอบล่าเอียง ภายในดัชนีเดียวกันระหว่างความยาวแบบสอบต่างกัน พบว่ามีความไม่คงที่เช่นกัน โดยมีค่ามากบ้างน้อยบ้าง ผลการใช้ดัชนี SA ให้ความสอดคล้องค่อนข้างสูงกว่าดัชนีอื่น ๆ รองลงมาคือ α_{MH} และ β_{SIB} อย่างไรก็ตามข้อสอบส่วนมากจะมีค่าดัชนีที่ต่างกันไปเมื่อปรับด้านความยาวแบบสอบ และขนาดผู้สอบเปลี่ยนไปข้อสอบบางข้อมีค่าดัชนีที่แสดงว่าล่าเอียงในสภาพการณ์หนึ่ง ในขณะที่มีค่าอยู่ในขอบเขตไม่ล่าเอียงในอีกสถานการณ์หนึ่ง ร้อยละของความสอดคล้องที่พบอยู่ระหว่าง 6.06-38.89 ผลของความสอดคล้องที่ค่อนข้างต่ำและการที่มีข้อสอบที่ล่าเอียงในบางบริบท และไม่ล่าเอียงในบางบริบทพบทั้งในวิชาภาษาอังกฤษและคณิตศาสตร์

ในด้านความสอดคล้องของการตัดสินภายในดัชนีเดียวกันในความยาวแบบสอบเดียวกันแต่ขนาดผู้สอบต่างกัน พบความสัมพันธ์ค่อนข้างสูง โดยเฉพาะอย่างยิ่งสำหรับดัชนี SA และ UA เมื่อขนาดผู้สอบเพิ่มขึ้นความสอดคล้องมีแนวโน้มจะสูงขึ้น ทั้งวิชาภาษาอังกฤษและวิชาคณิตศาสตร์ ในขณะที่ความสอดคล้องในดัชนี α_{MH} และ β_{SIB} จะต่ำอย่างมาก และมีความแปรเปลี่ยนไปตามจำนวนผู้สอบ

เมื่อวิเคราะห์ผลการตัดสินข้อสอบล่าเอียงภายในวิธีวิเคราะห์วิธีเดียวกัน โดยพิจารณาระหว่างการใช้การทดสอบค่าสถิติและการใช้เกณฑ์ พบว่าวิธี IRT-2 พารามิเตอร์ ให้ความสอดคล้องต่ำกว่าวิธี MH และวิธี SIBTEST โดยร้อยละของความสอดคล้องระหว่าง UA และ χ^2 สูงกว่า SA และ χ^2 เป็นส่วนมาก ส่วนวิธี MH และวิธี SIBTEST ให้ผลความสอดคล้องในแต่ละวิธีใกล้เคียงกัน ความสอดคล้องที่พบจะมีค่าสูงที่ขนาดผู้สอบ 400 600 คน มากกว่าในขนาดอื่น ๆ

ในขณะที่ตัดสินข้อสอบล่าเอียงด้วยเกณฑ์ที่กำหนดจากค่าเฉลี่ย โดยไม่พิจารณาความแตกต่างในบริบทด้านความยาวของแบบสอบและจำนวนผู้สอบ พบว่าจำนวนข้อสอบล่าเอียงในบริบทต่าง ๆ กัน มีความแตกต่างกันมาก การตัดสินด้วยเกณฑ์ที่กำหนดโดยพิจารณาถึงความแตกต่างในบริบทดังกล่าว พบว่าข้อสอบล่าเอียงในบริบทต่างกันมีจำนวนใกล้เคียงกันมากขึ้น โดยที่ร้อยละของข้อสอบล่าเอียง สำหรับวิชาภาษาอังกฤษมีค่าใกล้เคียงกันระหว่างขนาดผู้สอบต่างกันทั้งนี้ในความยาวแบบสอบ 50-70 ข้อ และ 20-30 ข้อ สำหรับคณิตศาสตร์ ส่วนในความยาว 80 และ 40 ข้อ จะมีจำนวนและร้อยละของข้อสอบล่าเอียงมากกว่าในความยาวอื่น ๆ มากสำหรับดัชนี SA และ UA ในขณะที่การใช้ α_{MH} และ β_{SIB} ให้ผลการตรวจสอบคิดเป็นร้อยละตามความยาวแบบสอบใกล้เคียงกันทั้ง 2 วิชา อย่างไรก็ตามพบว่าไม่ว่าจะใช้เกณฑ์ใด ความสอดคล้องระหว่าง UA และ SA จะค่อนข้างสูงกว่าวิชาอื่น ๆ เป็นส่วนมาก ใกล้เคียงกับความสอดคล้องระหว่าง α_{MH} และ β_{SIB} ส่วนความสอดคล้องในระหว่างดัชนีอื่น ๆ นอกจากนั้นจะมีค่าต่ำกว่า แต่ลักษณะของความสอดคล้องแต่ละขนาดผู้สอบ ความยาวแบบสอบยังไม่คงที่เช่นเดียวกับข้อค้นพบอื่น ๆ ที่ผ่านมา และพบว่าภายในสภาพการณ์เดียวกันผลจากเกณฑ์ที่กำหนดลักษณะให้ร้อยละของความสอดคล้อง ของข้อสอบล่าเอียงสูงกว่าลักษณะหลังเล็กน้อย

4. ทิศทางของความล่าเอียงของข้อสอบที่มีต่อเพศของผู้สอบ

เมื่อพิจารณาทิศทางของความล่าเอียงของข้อสอบ พบว่าข้อสอบล่าเอียงทั้งในวิชาภาษาอังกฤษและคณิตศาสตร์จะล่าเอียงเข้าข้างทั้งกับผู้สอบเพศหญิงและเพศชาย แต่เกินกว่าครึ่งในวิชาภาษาอังกฤษจะล่าเอียงเข้าข้างผู้สอบหญิง ในขณะที่เกินกว่าครึ่งหรือเกือบทั้งหมดของข้อสอบล่าเอียงในวิชาคณิตศาสตร์จะล่าเอียงเข้าข้างเพศชาย

ผลที่พบในวิชาภาษาอังกฤษเมื่อใช้ดัชนี β_{SIB} นั้นค่อนข้างจะตรงข้ามกับที่พบในการใช้ดัชนี SA และ α_{MH} โดยพบว่าส่วนมากข้อสอบล่าเอียงจะล่าเอียงเข้าข้างผู้สอบชายมากกว่าผู้สอบหญิงเล็กน้อย

นอกจากนั้น ยังพบว่าข้อสอบหลายข้อที่ล่าเอียงเข้าข้างผู้สอบเพศหนึ่งเมื่อวิเคราะห์ภายในความยาวแบบสอบ และ/หรือขนาดผู้สอบหนึ่ง และเข้าข้างผู้สอบอีกเพศหนึ่ง เมื่อวิเคราะห์ภายในความยาวแบบสอบและ/หรือขนาดผู้สอบอีกสถานการณ์หนึ่ง

สรุปผลการวิจัยทั้งหมดได้ว่า

1. ค่าเฉลี่ยของดัชนี SA UA α_{MH} และ β_{SIB} ที่ได้จากการเปรียบเทียบระหว่างผู้สอบเพศเดียวกันภายในความยาวแบบสอบ และขนาดผู้สอบต่างกัน มีค่าใกล้เคียงกันในแต่ละวิชา ค่าเฉลี่ยที่ได้จากวิชาภาษาอังกฤษค่อนข้างจะต่ำกว่าค่าเฉลี่ยที่ได้จากวิชาคณิตศาสตร์ สำหรับดัชนี SA และ UA ในขณะที่ค่าเฉลี่ยของ α_{MH} และ β_{SIB} มีค่าใกล้เคียงกันทั้ง 2 วิชา

2. ผลการตัดสินเมื่อใช้ค่าดัชนีตามที่กำหนด พบว่ามีร้อยละของข้อสอบล่าเอียงเมื่อเทียบกับความยาวแบบสอบเปลี่ยนแปรไปเมื่อใช้ขนาดผู้สอบและความยาวแบบสอบต่างกัน ทั้งนี้พบว่าผลจากความแตกต่างของจำนวนผู้สอบมีอิทธิพลต่อจำนวนข้อสอบล่าเอียงมากกว่าความยาวแบบสอบ และที่ส่วนทางกับการทดสอบค่าสถิติคือ ผลการทดสอบค่าสถิติพบว่าจำนวนข้อสอบล่าเอียงจะมากขึ้นเมื่อขนาดผู้สอบเพิ่มขึ้น ในขณะที่เมื่อใช้ดัชนีทั้ง 4 ตัว พบว่ามีข้อสอบล่าเอียงมากขึ้นเมื่อขนาดผู้สอบลดลงผลดังนี้พบเมื่อกำหนดเกณฑ์โดยไม่ว่าค่าถึงถึงขนาดผู้สอบ

ความแตกต่างระหว่างจำนวนข้อสอบล่าเอียง เมื่อขนาดผู้สอบและความยาวแบบสอบเปลี่ยนแปรไปจะลดลง เมื่อกำหนดเกณฑ์โดยพิจารณาความแตกต่างระหว่างขนาดผู้สอบด้วย

3. ความสอดคล้องของผลการตัดสินข้อสอบล่าเอียงในแต่ละสภาพการณ์ที่ศึกษาแตกต่างกันไปโดยที่ไม่สามารถจะวิเคราะห์ออกมาเป็นระบบได้ ส่วนหนึ่งเป็นเพราะค่าดัชนีที่วิเคราะห์ได้ในแต่ละสถานการณ์จะแตกต่างกันไม่คงที่

4. ผลการวิเคราะห์ความล่าเอียงของข้อสอบที่มีต่อเพศผู้สอบ พบว่าข้อสอบล่าเอียงในวิชาภาษาอังกฤษเข้าข้างเพศหญิงมากกว่าชายเป็นส่วนมาก ในกรณีของการใช้ดัชนี SA และ α_{MH} ในขณะที่ผลการใช้ดัชนี α_{SIB} จะให้ผลตรงข้าม ส่วนวิชาคณิตศาสตร์นั้นข้อสอบล่าเอียงที่พบจะล่าเอียงเข้าข้างเพศชายมากกว่าเพศหญิง

อภิปรายผลการวิจัย

1. การพัฒนาเกณฑ์ตัดสินข้อสอบล่าเอียงทางเพศ โดยวิธี IRT-2 พารามิเตอร์ MH และ SIBTEST

เนื่องจากลักษณะการกระจายของค่าดัชนีที่ได้จากวิธีการทั้ง 3 มีลักษณะการกระจายค่อนข้างจะสมมาตร มีฐานนิยมเดียว ค่ามัธยิมเลขคณิต และฐานนิยมมีความแตกต่างกันน้อย ดังนั้นการกำหนดเกณฑ์ที่จะใช้ในการตัดสินความล่าเอียงของข้อสอบจึงใช้ค่าเฉลี่ย และค่าความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยเป็นพื้นฐานในการกำหนด อย่างไรก็ตามลักษณะการกำหนดค่ากระทำใน 2 ลักษณะ คือ ใช้ค่าเดียวกันหมดทุกสถานการณ์ที่ศึกษา และพิจารณาตามความแตกต่างของค่าเฉลี่ย เมื่อขนาดผู้สอบ และความยาวแบบสอบต่างกัน แต่ไม่ว่าจะเป็นในลักษณะใดพบว่ามีค่าคงที่ในจำนวนข้อสอบล่าเอียงกับขนาดผู้สอบ และความยาวแบบสอบทั้งสิ้น ส่วนหนึ่งของความไม่คงที่อาจเป็นทั้งความไม่คงที่ของค่าดัชนีที่ได้จากผู้สอบเพศเดียวกันในขนาดผู้สอบและความยาวแบบสอบต่างกัน ทำให้ไม่สามารถกำหนดเกณฑ์ที่เหมาะสมพอ และเกิดจากความไม่คงที่ของค่าดัชนีที่ได้จากผู้สอบต่างเพศกันในขนาดผู้สอบและความยาวแบบสอบต่างกัน เมื่อ

เป็นดังนี้การกำหนดเกณฑ์ที่เหมาะสมจริง ๆ อาจต้องการการศึกษาอิทธิพลของตัวแปรด้านขนาดแบบสอบและความยาวแบบสอบ ให้ความแตกต่างละเอียดยิ่งขึ้นโดยศึกษาเฉพาะดัชนีแต่ละตัวในแนวลึก

นอกจากนั้นในการวิเคราะห์เพื่อพิจารณาความเป็นเอกมิติของแบบสอบด้วยการวิเคราะห์ตัวประกอบ พบว่าค่าไอเกนของตัวประกอบตัวแรกมีค่าค่อนข้างต่ำ ค่าร้อยละของความแปรปรวนของไอเกนตัวแรกมีค่าต่ำมากเกินกว่าจะน่ายอมรับได้ในความเป็นเอกมิติของแบบสอบ ผลการใช้ IRT โมเดล 3 พารามิเตอร์ พบข้อสอบหลายข้อที่ไม่สามารถหาค่าพารามิเตอร์ได้ และข้อที่เป็นดังนี้แตกต่างกันไปในความแตกต่างของบริบทด้านขนาดผู้สอบ และความยาวแบบสอบ เมื่อเปลี่ยนใช้โมเดล 2 พารามิเตอร์ พบว่าไม่มีปัญหานี้เกิดขึ้น กรณีเช่นนี้อาจเป็นไปได้ว่าข้อตกลงเบื้องต้นในด้านความเป็นเอกมิติ อาจมีระดับความเข้มข้น และระดับการยอมให้มีการฝ่าฝืนระหว่างโมเดลต่าง ๆ โดยที่โมเดลชนิด 3 พารามิเตอร์ อาจมีระดับที่ต้องการความเป็นเอกมิติสูงกว่าโมเดลชนิด 2 พารามิเตอร์ก็ได้ อย่างไรก็ตามนี้อาจเป็นส่วนหนึ่งของความไม่คงที่ของผลการทดลองใช้เกณฑ์

ความไม่คงที่ของค่าที่ได้ อาจมีสาเหตุอีกประการ คือ การใช้ขนาดผู้สอบซึ่งอาจน้อยเกินกว่าจะทำให้เกิดความคงที่ของค่าดัชนีที่ศึกษา โดยเฉพาะอย่างยิ่งในกรณีของค่า SA และ UA ซึ่งเป็นดัชนีที่มีพื้นฐานบนทฤษฎีการตอบข้อสอบ แม้ว่าโปรแกรมที่ใช้จะสามารถคำนวณค่าพารามิเตอร์ของข้อสอบได้ แต่การจะได้ค่าที่คงที่อาจต้องใช้จำนวนผู้สอบมากกว่านี้ อย่างไรก็ตามโปรแกรมที่ใช้ทำให้มีข้อจำกัดในด้านขนาดกลุ่มตัวอย่าง เนื่องจากไม่สามารถใช้กลุ่มตัวอย่างเกิน 1,000 คน

สำหรับข้อจำกัดประการต่อไปของการศึกษาคั้งนี้คือ จำนวนข้อสอบวิชาคณิตศาสตร์มีน้อย ไม่สามารถใช้เปรียบเทียบกับวิชาภาษาอังกฤษในความยาวแบบสอบ 50 ข้อได้ นอกจากนี้ถ้ามีการใช้แบบสอบภาษาอังกฤษความยาว 40 ข้อ ก็อาจให้การค้นพบในด้านเกณฑ์ และผลการวิเคราะห์ความลำเอียงดีกว่านี้ เนื่องจากมีช่วงที่เปรียบเทียบกันได้ใน 2 วิชา

นอกจากนั้นเมื่อวิเคราะห์ค่าความยากของแบบสอบ พบว่าโดยเฉลี่ยแบบสอบที่ใช้เป็นแบบสอบค่อนข้างยาก ยิ่งวัตถุประสงค์ในการนำแบบสอบไปใช้เอื้อต่อการที่ผู้สอบจะใช้เวลาเดาสั่งอยู่ด้วยแล้ว ผลการตอบข้อสอบจึงน่าจะมีการคลาดเคลื่อนสูง และกระทบต่อความคงที่ของค่าดัชนีต่าง ๆ ที่วิเคราะห์ได้

ประการสุดท้ายที่อาจมีผลต่อความไม่คงที่ของเกณฑ์ และดัชนีจากข้อมูลเชิงประจักษ์ที่นำมาใช้คือการที่ใช้คะแนนรวมเป็นเกณฑ์ในการแบ่งความสามารถ ซึ่งถ้ามีจำนวนข้อสอบลำเอียงมาก และคะแนนจากแบบสอบมีความคลาดเคลื่อนมากย่อมมีผลทำให้เกิดความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ของข้อสอบตามทฤษฎีการตอบข้อสอบ และการแบ่งกลุ่มผู้สอบในช่วงความสามารถต่าง ๆ เช่นกัน โดยเฉพาะอย่างยิ่งเมื่อเปรียบเทียบกับผลการศึกษาในต่างประเทศซึ่ง



ส่วนมากใช้แบบสอบถามมาตรฐานหรือไม่ก็ใช้แบบสอบถามการตรวจสอบคุณภาพแบบสอบถามแล้ว

ในด้านเกณฑ์ที่พบจากการวิเคราะห์โดยใช้ผู้สอบเพศเดียวกัน พบว่าเกณฑ์ขั้นต่ำที่ควรใช้ตัดสินข้อสอบล่าเอียงระหว่างเพศของผู้สอบ ควรมีค่าดังตารางที่ 64 ซึ่งเมื่อเปรียบเทียบกับเกณฑ์ที่มีผู้นำมาใช้ ได้แก่ $UA(\phi) > .20$ (ลักษณะ และคอฟแมน, 1980 อ้างถึงใน สุรศักดิ์ อมรรัตนศึกษา 1989, 54) $UA(\phi) > .40$ (ทัศนีย์, 1957 และสพัฒน, 1991) แล้วพบว่าเกณฑ์ที่ใช้กันนั้นต่ำกว่าเกณฑ์ที่พัฒนาขึ้น จากการใช้ข้อมูลเชิงประจักษ์ซึ่งพิจารณาถึงความคลาดเคลื่อนในการสุ่มตัวอย่างที่ผู้วิจัยได้พบในการศึกษาครั้งนี้ และอาจมีผลทำให้มีการระบุข้อสอบล่าเอียงทั้ง ๆ ที่ข้อสอบนั้นไม่ได้ล่าเอียงจริง แต่เกิดความคลาดเคลื่อนโดยบังเอิญจากการสุ่มตัวอย่างได้

นอกจากนี้กรณีของค่าดัชนี α_{HH} ซึ่งไรอันได้ใช้ค่าเพื่อแสดงลักษณะของข้อสอบ 3 ลักษณะคือ

A แสดงว่าเป็นข้อสอบที่ไม่ล่าเอียงโดยมีค่า α_{HH} น้อยกว่า 1.00 และไม่ต่างจาก 0 อย่างมีนัยสำคัญ

B แสดงว่าเป็นข้อสอบพอใช้ได้โดยมีค่า $1 = \alpha_{HH} < 1.5$ และไม่ต่างจาก 1.00 อย่างมีนัยสำคัญ

C แสดงว่าเป็นข้อสอบที่ล่าเอียง ควรพิจารณาในการเลือกมาใช้โดยมีค่า $\alpha_{HH} > 1.5$ และมากกว่า 1.00 อย่างมีนัยสำคัญ

ค่าที่กำหนดดังกล่าวค่อนข้างจะสอดคล้องกับผลการพัฒนาเกณฑ์ จากข้อมูลเชิงประจักษ์ ซึ่งค่าสูงสุด ควรจะมากกว่า 1.5 และไม่ควรถ่ำกว่า 1.30 ในกรณีที่ค่า $\alpha_{HH} > 1.00$ แต่พบกรณีไม่สอดคล้องที่ $\alpha_{HH} < 1.00$ นั่นคือเมื่อค่า α_{HH} ต่ำกว่า 1.00 ในระดับหนึ่ง เช่น ต่ำกว่า .50 หรือ .70 ควรจะตัดสินว่าเป็นข้อสอบล่าเอียงเช่นเดียวกัน เพียงแต่มีทิศทางหรือล่าเอียงเข้าข้างหรือต่อต้านผู้สอบอีกกลุ่มเท่านั้น ซึ่งถ้าพิจารณาตามสูตรการคำนวณค่า α_{HH} แล้วค่าที่ได้ต่ำกว่า 1.00 น่าจะเป็นค่าที่ล่าเอียงต่อกลุ่มอ้างอิง

2. ผลการตัดสินความล่าเอียงของข้อสอบด้วยเกณฑ์ที่พัฒนาจากข้อมูลเชิงประจักษ์

เมื่อมีจำนวนข้อสอบ และขนาดผู้สอบต่างกัน

พบว่าจำนวนข้อสอบและขนาดผู้สอบมีผลต่อค่าดัชนีที่วิเคราะห์ได้ และส่งผลต่อความไม่คงที่ของจำนวนข้อสอบที่ได้รับการตัดสินว่าล่าเอียงในแต่ละความยาวแบบสอบ และขนาดผู้สอบเช่นเดียวกันกับการทดสอบค่าสถิติ ที่เห็นชัด ได้แก่ วิธี IRT ซึ่งเมื่อใช้การทดสอบค่า χ^2 พบว่าในความยาวแบบสอบเดียวกันนั้น เมื่อขนาดผู้สอบมากขึ้นจำนวนข้อสอบล่าเอียงจะสูงขึ้น แต่

การใช้ดัชนีแทนการทดสอบค่าสถิติ พบว่าจำนวนข้อสอบจะลดลงเพื่อจำนวนผู้สอบเพิ่มขึ้น

โดยสรุปแล้วพบว่า การจะกำหนดเกณฑ์การตัดสินข้อสอบล่าเอียง โดยใช้ค่าดัชนี SA UA \propto_{MH} และ β_{SIB} แทนการทดสอบค่าสถิตินั้น พบว่ามีข้อควรระวังในแง่ของขนาดผู้สอบ และความยาวแบบสอบเช่นเดียวกับการทดสอบค่าสถิติ กรณีเช่นนี้อาจใช้จำนวนผู้สอบตามที่เหมาะสมกับทฤษฎีที่นำมาใช้ และพิจารณาจากผลการวิจัยโดยใช้ข้อมูลจำลองประกอบด้วย เช่น ในสถานการณ์ที่ใช้ข้อมูลจำลอง Baghi และ Ferrara (1990) พบว่าเมื่อใช้กลุ่มตัวอย่าง 750 คน วิธี MH ใช้แทน IRT-3 พารามิเตอร์ ในการตรวจค้นความล่าเอียงได้อย่างพอเพียงหรือ Swaminathan และ Rogers (1990) พบว่าเมื่อใช้ตัวอย่าง 250 คน วิธี MH ตรวจค้นได้ถูกต้องร้อยละ 75 และเมื่อใช้ตัวอย่าง 750 คน ตรวจค้นได้ถูกต้องร้อยละ 100.00 ในขณะที่ MAZOR และคณะ (1991) พบว่า ถ้าใช้ผู้สอบ 500 คน หรือต่ำกว่า การใช้วิธี MH ตรวจค้นข้อสอบล่าเอียงมีความผิดพลาดร้อยละ 50.00 และการค้นพบผลการใช้ SIBTEST และ MH ที่พบว่าความแตกต่างในอัตราความคลาดเคลื่อนประเภทที่ 1 ต่างกันอย่างไม่มีนัยสำคัญ เมื่อขนาดผู้สอบเท่ากับ 200 คนขึ้นไป และเมื่อใช้กลุ่มตัวอย่างสูงขึ้น SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 (ระบุว่าข้อสอบล่าเอียงทั้ง ๆ ที่ไม่ล่าเอียง) สูงกว่า MH เล็กน้อย ถ้ากลุ่มตัวอย่างที่ใช้ไม่เกิน 1,000 คน ถ้าใช้กลุ่มตัวอย่างมากกว่านั้นความคลาดเคลื่อนประเภทที่ 1 จะมีค่าสูงขึ้นเป็นต้น

3. ความสอดคล้องของการตัดสินข้อสอบล่าเอียงระหว่างวิธี IRT-2 พารามิเตอร์

กับวิธี MH และวิธี SIBTEST เมื่อใช้เกณฑ์ที่พัฒนาจากข้อมูลเชิงประจักษ์

ผู้วิจัยได้วิเคราะห์ความสอดคล้องระหว่างวิธีทั้ง 3 ทั้งในแง่ผลที่ได้จากการทดสอบนัยสำคัญของค่าสถิติและผลการใช้เกณฑ์ที่พัฒนาขึ้นจากข้อมูลเชิงประจักษ์ พบความสอดคล้องไม่ว่าจะภายในวิธีเดียวกันหรือต่างวิธีกัน มีความไม่คงที่ข้ามขนาดผู้สอบ และความยาวแบบสอบคล้าย ๆ กัน ทั้งนี้ไม่ว่าจะใช้เกณฑ์ที่กำหนดจากค่าเฉลี่ยรวมทุกข้อหรือเกณฑ์ที่พิจารณาจากความแตกต่างในบริบทที่ศึกษา ซึ่งอาจเป็นผลมาจากความไม่คงที่ของค่าดัชนีตามธรรมชาติอยู่แล้วเมื่อใช้ขนาดผู้สอบไม่เหมาะสม ดังได้กล่าวมาแล้วในข้อ 1 ข้อสอบทุกข้อจะมีดัชนีวิเคราะห์ได้ไม่คงที่มีความชันลงไม่เป็นระบบในแต่ละสภาพการณ์ที่ศึกษา ดังนั้นการจะกำหนดเกณฑ์ตัดสินได้โดยให้ผลการตัดสินมีความสอดคล้องกันสูงภายในวิธีเดียวกันกับสภาพการณ์ที่ต่างกัน จึงต้องมีการศึกษาในแนวลึก และละเอียดโดยเน้นที่ดัชนีแต่ละตัวหรือวิธีการแต่ละวิธี

สำหรับความไม่คงที่ของค่าดัชนี และผลการทดสอบนัยสำคัญของค่าสถิตินี้ดังกล่าวพบว่าสอดคล้องกับการศึกษาที่มีผู้ทำมาแล้ว เช่น MAZOR และคณะ (1991) ซึ่งใช้ข้อมูลจำลองพบความผิดพลาดในการตรวจค้นข้อสอบล่าเอียงต่างกัน เมื่อใช้ขนาดผู้สอบต่างกันโดยที่ขนาดผู้สอบ

น้อย (ต่ำกว่า 500 คน) ตรวจสอบความล่าช้าเชิงผิดพลาดถึงร้อยละ 50 ในขณะที่ขนาดผู้สอบมากขึ้น (มากกว่า 2,000 คน) จะตรวจสอบข้อผิดพลาดน้อยกว่า คือ ร้อยละ 30 ในขณะที่ Perlman และคณะ (1988) พบความไม่คงที่ของวิธีการประมาณด้วยวิธี MH และวิธีอื่น ๆ เช่นกัน และพบว่าความเที่ยงของดัชนีจะมีปัญหาถ้าใช้จำนวนตัวอย่างน้อยกว่า 660 คน เช่นเดียวกับการศึกษาของ Hambleton และ Rogers (1989) และ Ryan (1991) ที่พบความไม่คงที่ในการตัดสินความล่าช้าเชิงข้ามขนาดผู้สอบ และความยาวแบบสอบ โดยเฉพาะกรณีของ Ryan ซึ่งได้แบ่งระดับของความล่าช้าเชิง โดยใช้ดัชนี MH เป็น 3 ระดับ ความล่าช้าเชิง คือ

ระดับ A เมื่อ $\alpha_{MH} < 1.00$

B เมื่อ $\alpha_{MH} = 1$ และ < 1.5

C เมื่อ $\alpha_{MH} > 1.5$

และพบว่าข้อสอบล่าช้าเชิงเพียงข้อเดียวจากหลาย ๆ ข้อ ที่ได้รับการตัดสินว่าเป็นข้อล่าช้าเชิงประเภท B ทุกสภาพการณ์ที่ศึกษา นอกนั้นจะมีความไม่คงที่ข้ามสภาพการณ์ที่ศึกษาเช่นกัน

เมื่อพิจารณาถึงทฤษฎีการตอบข้อสอบที่เป็นพื้นฐานของดัชนี SA และ UA รวมทั้งการค้นพบว่าการวิเคราะห์ความล่าช้าเชิง ได้ผลของค่าดัชนีไม่คงที่ข้ามขนาดผู้สอบ อาจเป็นไปได้ 3 กรณี คือ ขนาดกลุ่มตัวอย่างน้อยเกินกว่าจะประมาณค่าพารามิเตอร์ได้คงที่ ทำให้พื้นที่วัดได้มีค่าไม่คงที่เช่นกัน หรืออาจเป็นไปได้ว่า การใช้กลุ่มตัวอย่างวิเคราะห์ความล่าช้าเชิงเพียง 1 คู่ อาจได้กลุ่มตัวอย่างที่มีความคลาดเคลื่อนสูงในการเป็นตัวแทนของกลุ่มประชากร การศึกษาโดยใช้กลุ่มวิเคราะห์ความล่าช้าเชิงมากกว่า 1 คู่ แล้วใช้ค่าเฉลี่ยเช่นเดียวกับการพัฒนาเกณฑ์อาจให้ผลการตัดสินมีความคงที่มากขึ้น และได้ค่าดัชนีที่เหมาะสมที่จะเป็นตัวแทนของกลุ่มประชากรมากขึ้น

นอกจากนั้นกรณีที่เป็นไปได้อีกกรณี คือ ความล่าช้าเชิงที่พบอาจไม่ใช่ความล่าช้าเชิงระหว่างผู้สอบต่างเพศกันอย่างเดียว อาจเป็นความล่าช้าเชิงอันเนื่องมาจากตัวแปรอื่น เช่น ภูมิฐานะที่อยู่ของผู้สอบ การใช้เพศเป็นตัวแปรแบ่งกลุ่มในการศึกษาครั้งนี้ จึงอาจไม่ใช่ภาพของความล่าช้าเชิงที่ถูกต้องทั้งหมด

ข้ออธิบายเพิ่มเติมอีกประการในกรณีของการใช้ α_{MH} และ β_{SIB} ซึ่งใช้คะแนนดิบเป็นตัวแบ่งกลุ่มความสามารถ ในขณะที่ SA และ UA ใช้ค่าพารามิเตอร์ความสามารถ กรณีนี้อาจอธิบายได้ถึงข้อคล่องในการตัดสินข้อสอบล่าช้าเชิงที่ค่อนข้างต่ำ อันเนื่องมาจากความสามารถที่แบ่งด้วยคะแนนดังกล่าวมีความคลาดเคลื่อนมากกว่า

4. ผลการศึกษาลักษณะของความลำเอียงของข้อสอบระหว่างเพศผู้สอบ

ส่วนมากพบว่าข้อสอบลำเอียงในวิชาภาษาอังกฤษลำเอียงเข้าข้างผู้สอบหญิงในจำนวนมากกว่าผู้สอบชาย และข้อสอบลำเอียงในวิชาคณิตศาสตร์จะลำเอียงเข้าข้างผู้สอบชายมากกว่า ทั้งนี้ยกเว้นกรณีที่ใช้ค่าดัชนี β_{S1B} ซึ่งพบว่าข้อสอบลำเอียงในวิชาภาษาอังกฤษมีแนวโน้มจะลำเอียงเข้าข้างผู้สอบชายมากกว่าผู้สอบหญิง

ข้อเสนอแนะในการนำเกณฑ์ไปใช้

1. จากข้อจำกัดที่อาจเป็นไปได้ดังที่อภิปรายมาแล้ว ประกอบกับการค้นพบว่าความสอดคล้องของการตัดสินใจในวิธีเดียวกันเมื่อจำนวนผู้สอบต่างกันมีแนวโน้มสูงขึ้นระหว่างผู้สอบ 600 คนขึ้นไป โดยเฉพาะอย่างยิ่งสำหรับดัชนี SA และ UA เช่น ที่พบในตารางที่ 31-36 สำหรับดัชนี α_{MH} และ ซึ่งผลจากการวิจัยครั้งนี้พบว่าความสอดคล้องภายในวิธี α_{MH} เมื่อกลุ่มผู้สอบมีขนาดต่างกันค่อนข้างต่ำ และต่ำกว่าที่พบในกรณีที่ใช้ค่า SA และ UA และผลจาก β_{S1B} ก็พบในลักษณะคล้ายคลึงกัน ผู้วิจัยจึงเสนอให้ใช้จำนวนผู้สอบที่พบให้ความถูกต้องมากที่สุดเมื่อใช้ข้อมูลจำลอง คือ จำนวนประมาณ 600 คน ขึ้นไป และสำหรับดัชนี SA และ UA ซึ่งมีรากฐานบนทฤษฎีการตอบข้อสอบซึ่งปกติควรจะใช้จำนวนผู้สอบ 1,000 คนขึ้นไป ผู้วิจัยเสนอให้ใช้จำนวนระหว่างที่เสนอตามทฤษฎี และจำนวนที่พบว่ามีความสอดคล้องในการตัดสินใจข้อสอบลำเอียงข้ามขนาดผู้สอบในระดับที่ค่อนข้างน่าพอใจ คือ 800 คน โดยที่เกณฑ์ของดัชนีแต่ละตัวตามที่ได้จากข้อมูลเชิงประจักษ์ที่เสนอให้ใช้ คือ

1.1 ค่าดัชนี SA

- 1) กรณีความยาวแบบสอบต่ำกว่า 50 ข้อ $|SA| > .80$
- 2) กรณีความยาวแบบสอบ 50 ข้อขึ้นไป $|SA| > .40$

1.2 ดัชนี UA

- 1) กรณีความยาวแบบสอบต่ำกว่า 50 ข้อ UA > .50
- 2) กรณีความยาวแบบสอบ 50 ข้อ ขึ้นไป UA > 1.20

1.3 ดัชนี α_{MH}

$$.60 > \alpha_{MH} > 1.40$$

1.4 ดัชนี β_{S1B}

$$\beta_{S1B} > .06$$

2. กรณีของการกำหนดเกณฑ์ตัดสินความล่าช้าของข้อสอบ ควรต้องใช้วิจารณญาณของผู้นำไปใช้โดยพิจารณาร่วมกับวัตถุประสงค์ของการใช้แบบสอบว่าการมีข้อสอบล่าช้าอยู่ด้วย จะก่อให้เกิดความยุติธรรมหรือยุติธรรมต่อกลุ่มผู้สอบข้อสอบเพียงไร กรณีที่ไม่แน่ใจในผลการตัดสินข้อสอบล่าช้า และไม่ต้องการตัดข้อสอบที่พบว่าล่าช้าออกจากแบบสอบอาจพิจารณาใช้ข้อสอบล่าช้าเข้าข้างผู้สอบแต่ละกลุ่มในแบบสอบเป็นจำนวนเท่า ๆ กัน เพื่อให้เกิดความยุติธรรมต่อผู้สอบ โดยให้มีคุณสมบัติอื่น ๆ ของข้อสอบใกล้เคียงกันมากที่สุด ประเด็นที่น่าจะความเข้าใจก็คือข้อสอบที่มีค่าดัชนีอยู่ในช่วงที่แสดงถึงความล่าช้าไม่ได้หมายความว่า เป็นข้อสอบที่ไม่ดี ข้อสอบล่าช้าเป็นเพียงกระแงเงาที่สะท้อนให้เห็นถึงผลโดยส่วนรวม (totality) ที่สังคมได้จัดการกระทำ (treat) ต่อเด็กหรือบุคคลในกลุ่มข้อสอบต่าง ๆ ต่างกัน โดยที่สังคมในที่นี้หมายถึงความรวมถึงสังคมทั้งภายใน และนอกครอบครัว เช่นกรณีของกลุ่มเพศหญิงและเพศชายอาจรวมถึงการกระทำต่อเด็กทั้ง 2 เพศของครอบครัว ครู อาจารย์ สังคมโดยทั่ว ๆ ไป รวมถึงหลักสูตร อุปกรณ์การสอน ของเล่นและอื่น ๆ

3. การใช้ IRT กับแบบสอบควรพิจารณาจุดมุ่งหมายที่นำแบบสอบไปใช้ แบบสอบคัดเลือกอาจเป็นตัวอย่างที่ดีของความไม่เหมาะสมของการวิเคราะห์ด้วยทฤษฎี IRT เพราะลักษณะของข้อสอบจะกระจายวัดเนื้อหาที่ไม่ใช่เนื้อหาเดียว ควรพิจารณานำทฤษฎีการตอบข้อสอบชนิดพหุมิติมาใช้จะดีกว่า หรืออาจวิเคราะห์ตัวประกอบแล้วจัดกลุ่มข้อสอบเพื่อการวิเคราะห์ตามโครงสร้างเนื้อหา

4. วิธีการวิเคราะห์ตามทฤษฎี IRT-2 พารามิเตอร์ ในด้านความสอดคล้องของการค้นพบข้อสอบล่าช้า เมื่อขนาดผู้สอบต่างกันสูงกว่าวิธีอื่น โดยเฉพาะอย่างยิ่งเมื่อขนาดผู้สอบ 600 คนขึ้นไป ดังนั้นกรณีใช้ผู้สอบขนาด 600 คนขึ้นไป การเลือกใช้วิธี IRT-2 พารามิเตอร์ น่าจะให้ผลการตัดสินที่ดีกว่า แต่ในกรณีตัวอย่างน้อยกว่านี้การเลือกใช้วิธี MH หรือ SIBTEST น่าจะให้ผลไม่ต่างกันนักและประหยัดกว่า

ข้อเสนอแนะในการวิจัย

จากประเด็นต่างๆ ที่พบในการวิจัยครั้งนี้ ผู้วิจัยขอเสนอแนะหัวข้อที่น่าจะได้รับการศึกษาต่อไปดังนี้

1. ศึกษาการเบี่ยงเบนของค่าดัชนีจากผู้สอบกลุ่มเดียวกัน โดยใช้ตัวแปรอื่น ๆ (เช่น ที่ตั้งของสถานศึกษา) ในสภาพของความแตกต่างระหว่างขนาดผู้สอบและความยาวแบบสอบ เพื่อเปรียบเทียบกับผลในการศึกษาคั้งนี้

2. ศึกษาถึงการเบี่ยงเบนของค่าดัชนีโดยจำแนกขนาดผู้สอบให้ละเอียดกว่านี้ ควรใช้แบบสอบวิชาเดียวกันที่ความยาวต่างกัน โดยพิจารณาว่าในสภาพใดที่ทำให้เกิดความคงที่และความสอดคล้องในจำนวนข้อสอบล่าเอียงมากที่สุด

3. น่าจะมีการศึกษาถึงผลการใช้แบบสอบที่มีข้อสอบล่าเอียงทางเพศใดเพศหนึ่งแบบมีปฏิสัมพันธ์ในแบบสอบเพื่อผลที่จะเกิดขึ้นต่อการใช้คะแนนที่ได้จากแบบสอบในการตัดสินใจเกี่ยวกับผู้สอบ ว่าแตกต่างจากการตัดข้อสอบออกหรือไม่

4. ควรมีการศึกษาความคงที่ในการตัดสินข้อสอบล่าเอียงเมื่อมีการเปรียบเทียบกลุ่มผู้สอบกลุ่มย่อยมากกว่า 1 ครั้ง