# บทที่ 2

## เอกสารและผลงานวิจัยที่เกี่ยวข้อง

เนื้อหาในบทนี้จะแบ่งออกเป็น 5 ส่วนคือ

1. ความเป็นมาและวิธีการตรวจสอบความลำเอียงของข้อสอบ
2. วิธีตรวจสอบความลำเอียงของข้อสอบตามแนวคิดทฤษฎีการตอบข้อสอบ
3. วิธีตรวจสอบความลำเอียงของข้อสอบตามวิธีของแมนเทลและแฮนส์เซล
4. วิธีตรวจสอบความลำเอียงของข้อสอบตามการทดสอบ SIB (SIBTEST)
5. งานวิจัยที่เกี่ยวข้องกับการตรวจสอบความลำเอียงของข้อสอบ

### 1. ความเป็นมาและวิธีการตรวจสอบความลำเอียงของข้อสอบ

การศึกษาความลำเอียงของข้อสอบหรือแบบสอบนั้น เริ่มมาจากการสังเกตปรากฏการณ์ของผลการสอบเพื่อการคัดเลือก (Selection bias) ซึ่งพบว่าไม่เป็นไปตามสัดส่วน ระดับสติปัญญาหรือโครงสร้างของประชากร ไม่ว่าจะเป็นการคัดเลือกเพื่อการศึกษาต่อ เพื่อการบรรจุเข้าทำงาน เพื่อเลื่อนตำแหน่งหรืออื่น ๆ ส่งผลให้เกิดการศึกษาความลำเอียงของแบบสอบ และข้อสอบรายข้อในแบบสอบที่ใช้ (test item bias) ในระยะต่อ ๆ มา

อย่างไรก็ตามแม้แนวคิดในการตรวจสอบความยุติธรรมของการใช้ข้อสอบ เพื่อการคัดเลือกบุคคลจะมีมาตั้งแต่ปี 1951 ดังได้กล่าวมาแล้ว แต่เหตุการณ์ที่ทำให้มีการตื่นตัวเป็นอันมากในเรื่องของความลำเอียงของข้อสอบ ได้แก่ กรณีที่เรียกว่า Defunis and Odegaard ซึ่งเกิดขึ้นในปี 1971 (Breland และ Ironson, 1976: 89)เป็นกรณีที่ Marco Defunis และคณะ ซึ่งถูกปฏิเสธจากโรงเรียนกฎหมายของมหาวิทยาลัยแห่งวอชิงตันได้ฟ้องร้องว่า เขาได้คะแนนการสอบสูงกว่าผู้ได้รับการคัดเลือกบางคนที่มหาวิทยาลัยรับเข้าศึกษา และยื่นฎีกาฟ้องร้อง Charles Odegaard อธิการบดีมหาวิทยาลัยดังกล่าวและคณะเป็นจำเลย เพื่อให้พิจารณาทบทวนการคัดเลือกนักศึกษาใหม่และจากนั้นเป็นต้นมา การพิจารณาตรวจสอบความลำเอียงของข้อสอบระหว่างผู้สอบกลุ่มย่อยเป็นสิ่งที่มีการปฏิบัติกันจนปัจจุบันเป็นเสมือนส่วนหนึ่งของขั้นตอนการ

พัฒนาแบบทดสอบโดยทั่ว ๆ ไป โดยเฉพาะอย่างยิ่งสำหรับการพัฒนาแบบทดสอบมาตรฐาน ตัวอย่างเช่น ศูนย์บริการการทดสอบทางการศึกษา (Educational Testing Service) ของประเทศสหรัฐอเมริกา ซึ่งให้บริการด้านการทดสอบด้วยแบบสอบมาตรฐานทั่วประเทศสหรัฐอเมริกา ได้มีการตรวจสอบความลำเอียงทั้งก่อนและหลังจากการนำแบบทดสอบไปทดลองใช้การตรวจสอบความลำเอียงของข้อสอบก่อนนำไปทดลองใช้จะใช้การตรวจสอบโดยผู้เชี่ยวชาญ ซึ่งประกอบด้วยคณะบุคคลหลายฝ่ายตรวจสอบความไว (sensitivity) ของแบบสอบ โดยพิจารณาถึงรูปแบบของข้อสอบ เนื้อหา ภาพประกอบคำที่ใช้และอื่น ๆ เพื่อไม่ให้มีความลำเอียงเป็นประโยชน์หรือเสียประโยชน์สำหรับผู้สอบกลุ่มย่อยใด ๆ เช่น ในกลุ่มผู้สอบที่ต่างเพศ ต่างเชื้อชาติ และวัฒนธรรม เป็นต้น

เท่าที่ผ่านมาจนถึงปัจจุบัน มีผู้เสนอวิธีการตรวจสอบความลำเอียงของข้อสอบโดยพิจารณาจากค่าสถิติอยู่หลายวิธีดังนี้

1) การพิจารณาความแตกต่างกันในค่าเฉลี่ยระหว่างกลุ่ม กรณีเช่นนี้อาจไม่ได้แสดงถึงความลำเอียงของข้อสอบหรือแบบสอบ เพราะกลุ่มย่อยทั้งสองกลุ่มอาจมีความแตกต่างกันอยู่แล้วในสิ่งที่วัด

2) การวิเคราะห์ตัวประกอบ เป็นการดูความสัมพันธ์ระหว่างคะแนนย่อย ๆ ในแบบสอบสำหรับกลุ่มย่อย ๆ โดยมากใช้เทคนิควิเคราะห์แบบ maximum likelihood ปัญหาอยู่ที่การเปรียบเทียบโครงสร้างระหว่างกลุ่มว่าจะใช้เกณฑ์ใดดีที่สุดรวมทั้งความซับซ้อนของวิธีการและการวิเคราะห์

3) การเปรียบเทียบค่าความยาก เป็นการพิจารณาลำดับที่สัมพัทธ์ของค่าความยากที่ได้จากการวิเคราะห์แยกกลุ่ม เป็นการพิจารณาปฏิสัมพันธ์ระหว่างข้อกระทงและกลุ่มผู้สอบ ข้อกระทงใดที่มีลักษณะของความสัมพันธ์สำหรับกลุ่มบางกลุ่มแตกต่างไปจากข้ออื่น ๆ ในแบบสอบเดียวกันจะเป็นข้อสอบที่ถือว่ามีความลำเอียง วิธีการทำได้ทั้งการวิเคราะห์ความแปรปรวนและการใช้แผนภูมิเส้น

ผู้ที่พัฒนาวิธีการนี้ได้สมบูรณ์ที่สุด คือ Angoff ซึ่งแปลงค่า p ให้เป็น ค่า z มาตรฐานเพื่อขจัดความสัมพันธ์เชิงเส้นโค้งระหว่างค่า P ของข้อสอบ 2 ชุด แล้วแปลงค่า z เป็นเดลต้า ($\Delta$) มีค่าเฉลี่ย = 13 และค่าส่วนเบี่ยงเบนมาตรฐาน = 4 เพื่อกำจัดค่าที่เป็นลบแล้วพิจารณาค่าเดลต้าของข้อสอบข้อเดียวกันจากกลุ่มที่ต่างกันโดยใช้แนวเส้นทะแยงที่เป็นเส้นแกนหลัก ข้อสอบจะลำเอียงหรือไม่พิจารณาได้จากระยะทางจากจุดค่าเดลต้าที่ลากไปตั้งฉากกับเส้นแกนหลัก

4) การใช้ทฤษฎีการตอบข้อสอบ เป็นการเปรียบเทียบความน่าจะเป็นในการตอบข้อสอบถูกของกลุ่มผู้สอบย่อย โดยพิจารณาถึงค่าพารามิเตอร์ของข้อสอบ และผู้สอบตามแบบจำลองที่เลือกใช้ซึ่งมีทั้งชนิด 1 2 และ 3 พารามิเตอร์

5) การวิเคราะห์ด้วยค่าไคสแควร์ ซึ่งมีผู้เสนอหลายคน เช่น วิธีของ Scheuneman (1979) เป็นวิธีการที่แปลงค่าความสามารถของสมาชิกในกลุ่มเป็นช่วง ๆ อาจเป็น 3 หรือ 5 ช่วง แล้วใส่ข้อมูลในตาราง 2 ทาง (จำนวนกลุ่ม X การตอบถูก - ผิด) จำนวนตารางเท่ากับจำนวนช่วงคะแนนจากนั้นคำนวณค่าไคสแควร์แต่ละตาราง และเมื่อรวมค่าทุกตาราง จะได้ค่าไคสแควร์ที่มีชั้นแห่งความเป็นอิสระเท่ากับจำนวนกลุ่ม การจัดกลุ่มตามระดับคะแนนเช่นนี้ ถ้ากลุ่มผู้สอบ 2 กลุ่ม มีคะแนนเฉลี่ยต่างกันแล้วจะเกิดผลที่ดูเป็นความลำเอียงได้

6) ที่แตกแยกออกมาจาก Loglinear ได้แก่วิธีของ Mantel-Haenszel ที่ศึกษาตัวแปรประเภทแบ่งสองในกลุ่มที่แตกต่างกันด้วยตัวแปรบล็อค และนำมาใช้เป็นตัวบ่งชี้ความลำเอียงของข้อสอบ

7) Linn และ Harnisch, (1981) ได้เสนอวิธีวิเคราะห์ในลักษณะทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ โดยรวมกลุ่มย่อยเข้าด้วยกัน แล้วแบ่งความสามารถเป็นช่วงดูความแตกต่างระหว่างความน่าจะเป็นของการตอบถูกที่คาดหวัง ซึ่งประมาณได้ด้วยทฤษฎีการตอบข้อสอบกับความน่าจะเป็นของการตอบถูกจากข้อมูลเชิงประจักษ์ พิจารณาเป็นช่วงความสามารถแล้วนำมารวมกันเป็นดัชนีรวมของความลำเอียงของแต่ละข้อ วิธีนี้เรียกว่าวิธีทฤษฎีการตอบข้อสอบแบบเทียม และใช้ได้ในกรณีที่มีกลุ่มตัวอย่างน้อยทำให้ใช้ทฤษฎีการตอบข้อสอบแบบเต็มรูปไม่ได้

8) วิธีล่าสุดที่เสนอโดย Shealy and Stout (1989) ได้แก่ การทดสอบ SIB (หรือ SIBTEST) เป็นวิธีที่ดัดแปลงจากวิธีทฤษฎีการตอบข้อสอบ สามารถใช้ได้ทั้งเพื่อการตรวจสอบความลำเอียงของข้อสอบรายข้อการตรวจสอบความลำเอียงของกลุ่มข้อสอบหรือแบบสอบทั้งฉบับ

การวิจัยครั้งนี้วิธีการตรวจสอบความลำเอียงของข้อสอบที่ผู้วิจัยนำมาใช้ในการศึกษาได้แก่วิธีการใช้ทฤษฎีการตอบข้อสอบ วิธีของ Mantel และ Haenszel และวิธีการทดสอบ SIB ซึ่งจะได้กล่าวอย่างละเอียดต่อไป

## 2. วิธีตรวจสอบความลำเอียงของข้อสอบตามแนวคิดทฤษฎีการตอบข้อสอบ

### 2.1 ทฤษฎีการตอบข้อสอบ (Item Response Theory)

ทฤษฎีการตอบข้อสอบเป็นทฤษฎีที่มีกำเนิดมาจากการที่นักวัดผลและนักการศึกษาได้พยายามคิดค้นหาวิธีการแก้ไขข้อบกพร่องของทฤษฎีการทดสอบแบบดั้งเดิม (classical test theory) โดยเริ่มที่งานของ Binet และ Simon ในปี 1916 ที่ได้เริ่มทำกราฟของความสัมพันธ์ระหว่างผลการสอบและตัวแปรอิสระอื่น ๆ และใช้กราฟนั้นในการพัฒนาแบบสอบ และในปี 1936 Richardson ได้แสดงความสัมพันธ์ระหว่างพารามิเตอร์ตามโมเดลการตอบข้อสอบและพารามิเตอร์ ตามทฤษฎีการวัดดั้งเดิม ซึ่งเป็นวิธีเริ่มแรกในการประมาณค่าพารามิเตอร์ตามทฤษฎีการตอบข้อสอบจนกระทั่งในปี 1943-4 Lawley ได้คิดวิธีการใหม่ในการประมาณค่าพารามิเตอร์ และในปี 1952 Lordได้เสนอโมเดลปกติสะสม normal ogive model ชนิด 2 พารามิเตอร์ การประมาณค่าพารามิเตอร์ และการนำโมเดลไปใช้ในทางปฏิบัติ ต่อมา Birnbaum (1957-8) ได้เสนอให้ใช้logistic modelแทนและพัฒนาพื้นฐานด้านสถิติสำหรับโมเดลที่เสนอใหม่

ปี 1960 Rasch ได้พัฒนาโมเดลการตอบข้อสอบขึ้น 3 โมเดล ซึ่งได้มีผู้นำไปใช้ในการศึกษาและพัฒนาต่อในภายหลัง เช่น Wright ซึ่งนอกจากจะนำไปใช้ในการศึกษาวิจัยแล้ว ยังได้ร่วมกับ Panchapakesan (1969) ในการเสนอวิธีการประมาณค่าพารามิเตอร์ และโปรแกรมคอมพิวเตอร์ BICAL เพื่อที่ใช้ในการคำนวณตามโมเดลของ Rasch ด้วย

และหลังจากนั้น มาก็ได้มีนักการศึกษาหลายท่านได้ช่วยกันพัฒนาโปรแกรมคอมพิวเตอร์สำหรับการวิเคราะห์ตามทฤษฎีนี้ขึ้นมาอีกมากมาย เช่น Lord (1982) ได้พัฒนาโปรแกรมการประมาณค่าพารามิเตอร์ ชื่อ LOGIST และ Mislevy และ Bock (1984) พัฒนาโปรแกรม BILOG เป็นต้น

โมเดลตามทฤษฎีการตอบข้อสอบแสดงถึง ความสัมพันธ์ระหว่างผลการตอบข้อสอบกับลักษณะแฝงหรือความสามารถที่ไม่สามารถสังเกตได้โดยตรง และอธิบายได้ในรูปของฟังชั่นทางคณิตศาสตร์ โดยมีข้อสรุปว่า (1) ผลการตอบข้อสอบของผู้สอบสามารถทำนาย หรืออธิบายได้ด้วยองค์ประกอบที่เรียกว่า ลักษณะแฝง หรือความสามารถที่มีอยู่ในตัวบุคคล และ (2) ความสัมพันธ์ระหว่างผลการตอบข้อสอบ และความสามารถที่มีอิทธิพลต่อการตอบข้อสอบสามารถอธิบายได้ด้วยฟังชั่นที่มีลักษณะเป็นโค้งสะสมเพิ่มขึ้นทางเดียว (monotonically increasing function) ซึ่งเรียกว่า ฟังชั่นลักษณะข้อสอบอันเป็นฟังชั่นที่แสดงว่า ผู้สอบที่มีคะแนนในลักษณะแฝงที่วัดสูงกว่าจะมีความน่าจะเป็นที่คาดหวังในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้ที่ได้คะแนนการวัดลักษณะแฝงนั้นต่ำกว่า

## 2.2 ข้อตกลงเบื้องต้นในการใช้ทฤษฎีการตอบข้อสอบ

การวิเคราะห์ข้อสอบโดยใช้ทฤษฎีการตอบข้อสอบ มีข้อตกลงเบื้องต้นเกี่ยวกับแบบสอบดังนี้

2.2.1 ความเป็นเอกมิติ (Unidimensionality) หมายความว่าการตอบข้อสอบทุกข้อได้ถูกต้อง ต้องใช้ความสามารถหรือลักษณะแฝงภายในเพียงลักษณะเดียว หรือข้อสอบทุกข้อในแบบสอบวัดความสามารถแฝงลักษณะเดียว หรือกลุ่มลักษณะแฝงกลุ่มเดียว

2.2.2 Monotonicity

ผู้มีความสามารถในลักษณะแฝงที่วัดสูงกว่า จะมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องมากกว่า

2.2.3 Probabilistic models

หมายความว่าเป็นแบบจำลองที่ฟังชั่นของลักษณะข้อสอบแสดงถึงความสัมพันธ์ระหว่างตัวแปรความสามารถที่สังเกตไม่ได้และตัวแปรการตอบข้อสอบที่สังเกตได้ (ผลการตอบข้อสอบ 0 หรือ 1)

2.2.4 Local independence

การตอบข้อสอบแต่ละข้อเป็นไปอย่างเป็นอิสระต่อกัน การตอบข้อสอบข้อหนึ่งไม่มีอิทธิพลต่อการตอบข้อสอบอีกข้อหนึ่ง นั่นคือ

1) การตอบข้อสอบทุกข้อของผู้สอบคนหนึ่ง หรือการตอบข้อสอบข้อหนึ่ง ๆ ของผู้สอบทุกคนมีความเป็นอิสระต่อกัน

2) ในแง่ของสถิติ หมายความว่า ถ้าให้ $\theta$ เป็นความสามารถที่มีอิทธิพลต่อการตอบข้อสอบของผู้สอบ $U_i$ เป็นผลการตอบข้อสอบข้อ i ของผู้สอบที่สุ่มเลือกได้ (i มีค่า = 1,2...n)และ $P(U_i|\theta)$ เป็นความน่าจะเป็นในการตอบข้อสอบของผู้สอบที่มีความสามารถ $\theta$ $P(U_i = 1|\theta)$ คือ ความน่าจะเป็นของการตอบข้อสอบถูก และ $P(U_i = 0|\theta)$ เป็นความน่าจะเป็นของการตอบข้อสอบผิด คุณสมบัติของ local independence แสดงได้ด้วย

$$\text{Prob}(U_1, U_2, \ldots\ldots U_n|\theta) = P(U_1|\theta)\,P(U_2|\theta)\,P(U_n|\theta)$$

$$= \prod_{i=1}^{n} P(U_i|\theta)$$

ซึ่งให้ความหมายว่า สำหรับผู้สอบใดๆ (หรือผู้สอบที่มีความสามารถตามที่กำหนด) ความน่าจะเป็นของการตอบแบบสอบจะเท่ากับผลคูณของความน่าจะเป็นในการตอบข้อสอบแต่ละข้อ เช่น ถ้ารูปแบบผลการตอบข้อสอบ 4 ข้อ เป็น (0, 1, 1, 1) หรือ $U_1 = 0$ $U_2 = 1$ $U_3 = 1$ และ $U_4 = 1$ ตามข้อตกลงเกี่ยวกับ local independence จะได้ว่า

$$\begin{aligned} P(U_1 = 0, U_2=1, U_3=1, U_4=1 \mid \theta) \\ &= P(U_1 = 0 \mid \theta)\, P(U_2 = 1 \mid \theta)\, P(U_3 = 1 \mid \theta)\, P(U_4 = 1 \mid \theta) \\ &= P_1 P_2 P_3 P_4 \end{aligned}$$

โดยที่ $P_i = P(U_i = 1 \mid \theta)$ และ $Q_i = 1 - P_i$

3) หลังจากควบคุมค่า $\theta$ แล้ว จะได้ความเป็นอิสระอย่างมีเงื่อนไขว่า

$$\rho_{i,j} \mid \theta = 0$$

นั่นคือ ผลการตอบข้อสอบ 2 ข้อใด ๆ ของผู้สอบ ไม่มีความสัมพันธ์กัน หรือ Stout(1987) ใช้คำว่า essential independence คือ มีค่าใกล้กับศูนย์

Lord (1980: 19) กล่าวว่า "Local independence เป็นคุณสมบัติที่เกิดขึ้นโดยอัตโนมัติจากคุณสมบัติของความเป็นเอกมิติ ไม่ใช่เป็นข้อตกลงเบื้องต้นที่เพิ่มขึ้นมา" ในทางปฏิบัติส่วนมากก่อนจะวิเคราะห์ข้อมูลการทดสอบด้วยโมเดลตามทฤษฎีการตอบข้อสอบ จึงมีการตรวจสอบความเป็นเอกมิติของแบบสอบก่อน

## 2.3 โมเดลที่เป็นที่นิยมในทฤษฎีการตอบข้อสอบ

ลักษณะทั่วไปของทฤษฎีการตอบข้อสอบ คือ โมเดลการตอบข้อสอบซึ่งในทางคณิตศาสตร์แสดงถึงความน่าจะเป็นในการตอบข้อสอบถูกในรูปฟังชั่นของความสามารถของผู้สอบ การตอบข้อสอบแบบสองประเภท ข้อ j แสดงได้ในรูปของคะแนน

$$X_j = \begin{cases} 1 \text{ ถ้าตอบถูก} \\ 0 \text{ ถ้าตอบผิด} \end{cases}$$

ความสามารถของผู้สอบใช้แทนด้วย $\theta$ และความน่าจะเป็นของการตอบข้อสอบข้อ j ถูก คือ

$$P(X_j = 1 | \theta) = P_j(\theta)$$

ดังนั้น ความน่าจะเป็นในการตอบข้อสอบผิด จึงเป็น

$$P(X_j = 0 | \theta) = 1-P_j(\theta)$$

โมเดลทางทฤษฎีการตอบข้อสอบที่เป็นที่นิยมใช้ในปัจจุบันมีความแตกต่างที่สำคัญ คือ จำนวนพารามิเตอร์ที่ใช้ในการบรรยายถึงข้อสอบ ซึ่งมีอยู่ 3 โมเดล คือ โมเดลชนิด 1, 2 และ 3 พารามิเตอร์ และเป็นโมเดลที่ใช้กับข้อสอบชนิดมีการให้คะแนน 2 ค่า (dichotomous) (Hambleton และคณะ 1991: 12) ดังนี้

1. โมเดลโลจิสติคชนิด 1 พารามิเตอร์ (Rasch, 1960)

โมเดลนี้เป็นโมเดลที่นิยมใช้กันกว้างขวาง โค้งลักษณะข้อสอบสำหรับโมเดลนี้แสดงได้ดังสมการ

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

โดยที่

$P_i(\theta)$ คือ ความน่าจะเป็นของผู้สอบที่สุ่มขึ้นมาที่มีความสามารถ $\theta$ จะตอบข้อสอบข้อ i ได้ถูกต้อง

$b_i$ คือ ค่าพารามิเตอร์ความยากของข้อสอบ

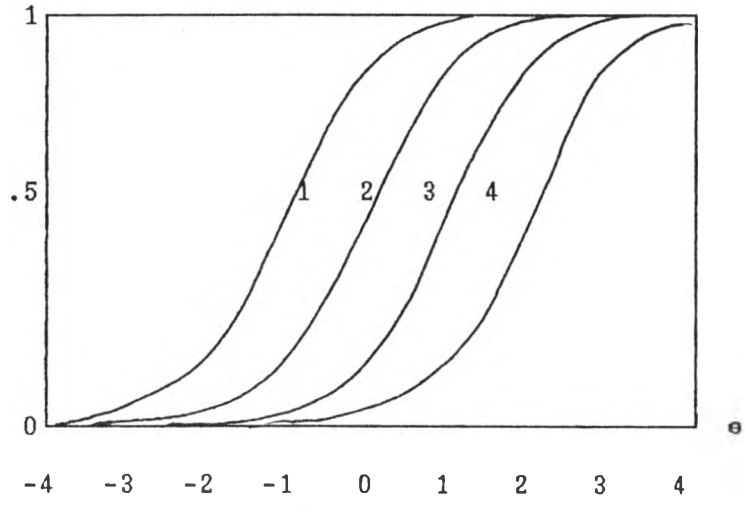
n คือ จำนวนข้อสอบในแบบสอบ

e คือ ฐาน (base) ของลอการิทึมธรรมชาติมีค่า 2.718

ค่า $b_i$ คือ ค่าบนมาตรความสามารถที่มีความน่าจะเป็นในการตอบข้อสอบถูกเป็น 0.5 ถ้า b มีค่าสูง แสดงว่า ข้อสอบนั้นยาก และผู้มีโอกาสจะตอบข้อสอบได้ถูกต้อง 50% ต้องมีความสามารถสูงขึ้น

ถ้าแปลงค่า θ ให้มีค่าเฉลี่ยเป็น 0 และค่าความเบี่ยงเบนมาตรฐานเป็น 1 แล้ว b จะมีค่าประมาณ -2.0 ถึง +2.0 (Hambleton และคณะ, 1991: 13) โมเดลนี้ถือว่า ความยากของข้อกระทงเป็นลักษณะข้อสอบลักษณะเดียวที่มีอิทธิพลต่อการตอบข้อสอบ

ภาพที่ 2 เป็นภาพโค้งลักษณะข้อสอบตามโมเดล 1 พารามิเตอร์ 4 ข้อ ซึ่งมีค่า b เป็น -1, 0 1 และ 2 ตามลำดับ สังเกตได้ว่า โค้งตัดแกน P(θ) ที่ 0 และทุกโค้งมีค่าอำนาจจำแนกเท่ากัน โมเดล 1 พารามิเตอร์นี้ในทางคณิตศาสตร์เทียบได้กับโมเดลของRasch




ภาพที่ 2: โค้งลักษณะข้อสอบตามโมเดล 1 พารามิเตอร์

2. โมเดลโลจิสติค ชนิด 2 พารามิเตอร์

Lord เป็นคนแรกที่พัฒนาโมเดลการตอบข้อสอบ ชนิด 2 พารามิเตอร์ ขึ้นในปี 1952 โดยใช้การกระจายแบบโค้งปกติสะสม และ Birnbaum เป็นผู้เสนอให้ใช้ฟังชั่นโลจิสติคซึ่งมีความสะดวกและง่ายในการคำนวณมากกว่าแทนฟังชั่นโค้งปกติสะสม

โค้งลักษณะข้อสอบสำหรับโมเดลโลจิสติคแบบ 2 พารามิเตอร์ แสดงได้ในรูปสมการดังนี้

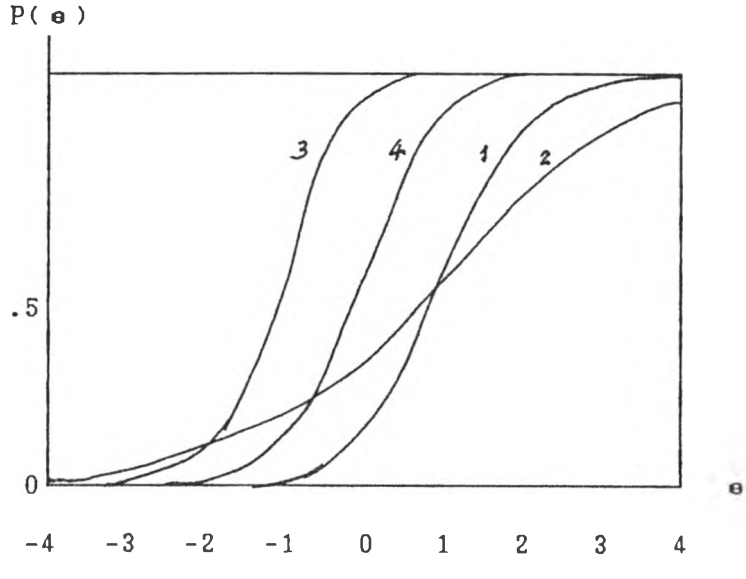
$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

สิ่งที่เพิ่มเข้ามาในสมการคือ D ซึ่งได้แก่ องค์ประกอบของการปรับมาตรที่ทำให้ฟังชั่นโลจิสติคและฟังชั่นปกติสะสมมีค่าสัมบูรณ์ของความแตกต่างน้อยกว่า 0.01 สำหรับทุกค่าของ $\theta$

สิ่งที่เพิ่มเข้ามาอีกตัว คือ ค่า $a_i$ ซึ่งเป็นค่าพารามิเตอร์อำนาจจำแนก เป็นค่าความชันของโค้งลักษณะข้อสอบที่จุดเปลี่ยนโค้งหรือที่จุด $b_i$ บนมาตรความสามารถ

ค่า $a_i$ มีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ ในทางทฤษฎี แต่ในทางปฏิบัติจะมีค่าอยู่ระหว่าง -2 ถึง 2 ข้อสอบที่มีค่า a สูง จะเป็นข้อสอบที่มีโค้งการตอบข้อสอบที่ชันมากและสามารถจำแนกคนได้ดีกว่าข้อสอบที่มีค่า a ต่ำกว่า

ภาพที่ 3 เป็นภาพโค้งลักษณะข้อสอบโมเดล 2 พารามิเตอร์ ซึ่งข้อ 1 มีค่า $b_1$=1.0 $a_1$=1.0 ข้อ 2 มีค่า $b_2$=1.0, $a_2$=0.5 ข้อ 3 มีค่า $b_3$=-1.0, $a_3$=1.5 และข้อ 4 มีค่า $b_4$=0.0, $a_4$=1.2




ภาพที่ 3: โค้งลักษณะข้อสอบตามโมเดล 2 พารามิเตอร์

(Hambleton, Swaminathan and Rogers, 1991: 16)

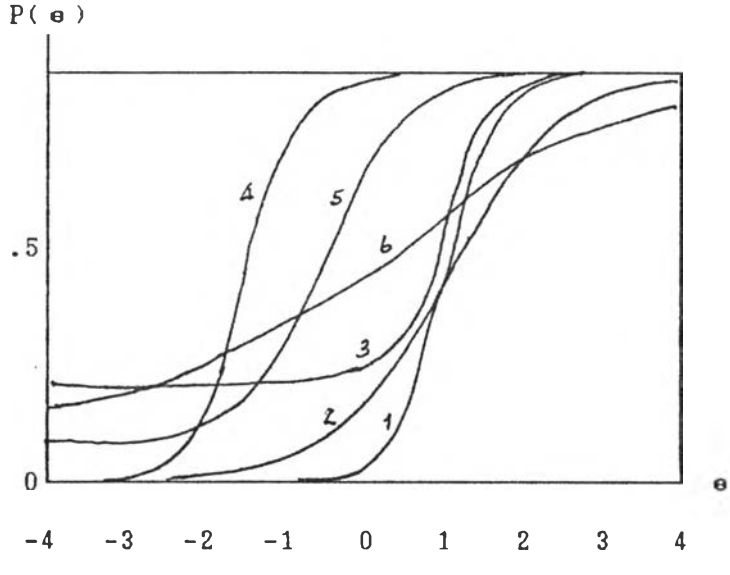
3. โมเดลโลจิสติคแบบ 3 พารามิเตอร์

สมการของฟังชั่นโลจิสติคแบบ 3 พารามิเตอร์ คือ

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

เมื่อ $c_i$ คือ ความน่าจะเป็นในการตอบข้อสอบเลือกตอบได้ถูกต้องอันเนื่องมาจากการเดา ถ้าตัวเลือกในแบบสอบเรียงไว้แบบสุ่ม และผู้สอบทุกคนเดาตอบ ค่าของ $c_i$ จะเท่ากับ 1/A เมื่อ A เป็นจำนวนตัวเลือกในแต่ละข้อ ถ้าผู้สอบแต่ละคนเลือกตอบ โดยมีการกำจัดตัวเลือก 1 ตัว หรือมากกว่าออกก่อน ค่า $c_i$ จะมีค่ามากกว่า 1/A

ภาพที่ 4 เป็นภาพโค้งลักษณะข้อสอบชนิด 3 พารามิเตอร์ จะเห็นว่า ข้อสอบข้อ 1-3 ยากกว่า ข้อสอบข้อที่ 4-6 และข้อสอบข้อที่ 1, 3 และ 4 มีค่าอำนาจจำแนกสูงกว่าข้อที่ 2, 3 และ 6 และข้อสอบข้อที่ 3, 5 และ 6 แสดงถึงอิทธิพลของค่าการเดา ($c_i$) ต่อโค้งลักษณะข้อสอบ



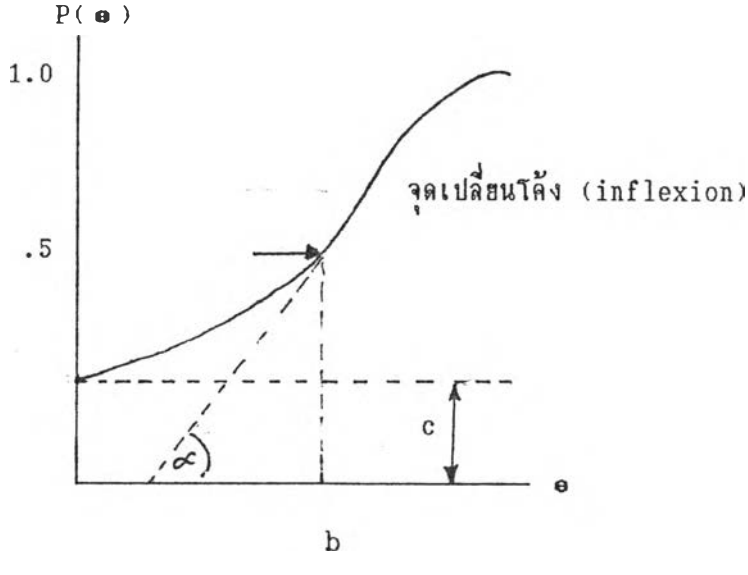

ภาพที่ 4: โค้งลักษณะข้อสอบตามโมเดล 3 พารามิเตอร์

(Hambleton, Swaminathan and Rogers, 1991: 18)

ภาพที่ 5 เป็นภาพแสดงความหมายของค่าพารามิเตอร์ข้อสอบ (Lord, 1980: 13-14) ค่าพารามิเตอร์ c คือ ความน่าจะเป็นที่ผู้สอบที่ไม่มีความสามารถในสิ่งที่วัดโดยสิ้นเชิง ( $\theta = -\infty$ ) จะตอบข้อสอบได้ถูกต้อง เรียกว่า เป็นค่าพารามิเตอร์การเดา หรือระดับคะแนนการตอบถูกโดยบังเอิญ (guessing parameter or pseudo-chance score level) กรณีที่ข้อสอบใดไม่สามารถตอบได้ถูกโดยการเดาแล้ว ค่า c = 0

พารามิเตอร์ b เป็นพารามิเตอร์แสดงตำแหน่ง (location parameter) เพราะเป็นตัวแสดงตำแหน่งของโค้งบนมาตรความสามารถเรียกว่า ค่าความยากของข้อสอบ ยิ่งข้อสอบยากมากโค้งจะยิ่งชิดมาทางขวามากจุดเปลี่ยนโค้งของโค้งโลจิสติคอยู่ที่จุดที่ค่า $\theta$ = b กรณีไม่มีการเดา b คือ ระดับความสามารถที่มีความน่าจะเป็นในการตอบข้อสอบถูก = .5 แต่ ถ้ามีค่าการเดา b จะเป็นค่า $\theta$ ที่มีความน่าจะเป็นในการตอบถูกอยู่ตรงกึ่งกลางระหว่างค่า c และ 1.00 (=(1-c)/2)

ค่าพารามิเตอร์ a เป็นสัดส่วนต่อความชันของโค้งที่จุดเปลี่ยนโค้ง ซึ่งมีค่าเท่ากับ .425(1-c) ดังนั้น a จึงเป็นอำนาจจำแนกของข้อสอบแสดงระดับการตอบข้อสอบที่แปรเปลี่ยนไปภายในระดับความสามารถต่าง ๆ




ภาพที่ 5: ความหมายของค่าพารามิเตอร์ข้อสอบ (Lord, 1980: 14)

## 2.4 การประมาณค่าความสามารถและพารามิเตอร์ข้อสอบ

การประมาณค่าความสามารถมีวิธีการประมาณที่ใช้กันอยู่หลายวิธี เช่น (Hambleton, Swaminathan และ Rogers, 1991: 46)

1) Joint maximun likelihood procedure พัฒนาขึ้นโดย Lord ในปี 1974 และ 1980 สามารถใช้ประมาณค่าพารามิเตอร์ข้อสอบและความสามารถพร้อม ๆ กัน และสามารถใช้ได้กับทั้งโมเดล 1 2 และ 3 ค่าพารามิเตอร์

2) Marginal maximum likelihood procedure โดย Bock และ Aitkin ปี 1981 มีทั้งชนิดโมเดล 1 2 และ 3 พารามิเตอร์ ประมาณค่าพารามิเตอร์ของข้อสอบ แล้วจึงประมาณค่าพารามิเตอร์ของความสามารถ

3) Conditional maximum likelihood procedure โดย Andersen ปี 1972-3 และ Rasch ปี 1960 ใช้ได้เฉพาะโมเดล 1 พารามิเตอร์ เป็นฟังชั่นของความเป็นไปได้ โดยมีเงื่อนไขที่คะแนนที่ได้จากการทำแบบสอบถูก

4) Joint and marginal Bayesian estimation procedures โดยมิสเลวี Mislevy ปี 1986 Swaminathan และ Gifford ปี 1982, 1985 และ 1986 ใช้ได้ทั้งกับโมเดล 1 2 และ 3 พารามิเตอร์

5) Heuristic estimation procedure โดย Urry ปี 1974 และ 1978 ใช้ได้ทั้งกับโมเดล 1 2 และ 3 พารามิเตอร์

6) Method based on nonlinear factor analysis procedure โดย McDonald ปี 1967 และ 1989 ใช้ได้กับโมเดล 2 และ 3 พารามิเตอร์ ที่กำหนดค่า c ให้เป็นค่าคงที่

ในที่นี้ผู้วิจัยขอนำเสนอวิธีการประมาณค่าความสามารถและค่าพารามิเตอร์ของข้อสอบด้วยวิธีที่นิยมใช้ 3 วิธี ได้แก่ วิธี Joint maximum likelihood procedure ซึ่ง Wingersky, Barton และ Lord (1982) ได้ใช้เป็นฐานในการพัฒนาโปรแกรม LOGIST สำหรับ mainframe computer ซึ่งขณะนี้ยังนิยมใช้กันอยู่และในปัจจุบันศูนย์บริการทางการศึกษา (ETS-Educational Testing Service) กำลังพัฒนารูปแบบใหม่ เพื่อให้สามารถใช้กับคอมพิวเตอร์ส่วนบุคคล (PC-Computer) วิธีที่ 2 ที่จะได้นำเสนอคือ วิธี Joint and marginal Bayesian estimate procedure ซึ่งเป็นส่วนหนึ่งที่ Mislevy และ Bock ได้นำไปพัฒนาโปรแกรม BILOG ในปี 1984 โดยมีทั้งโปรแกรมที่ใช้กับคอมพิวเตอร์ชนิด mainframe และ PC และวิธีสุดท้าย คือ วิธี Marginal Maximum Likelihood (MML) ซึ่งเป็นวิธีที่นำไปใช้ในโปรแกรม BILOG เช่นกัน

2.4.1 วิธีการประมาณด้วย Maximum Likelihood

สมมุติให้ $(U_1, U_2, U_3 \ldots U_n)$ เป็นรูปแบบการตอบข้อสอบ n ข้อ ของผู้สอบ โดยที่ $U_j$ มีค่าเป็น 1 ถ้าตอบข้อสอบถูก และ 0 ถ้าตอบข้อสอบผิด

ภายใต้ข้อตกลงในเรื่องความเป็นอิสระในการตอบข้อสอบแต่ละข้อ (Local independence) จะได้ว่า

$$P(U_1, U_2, \ldots, U_j, \ldots, U_n \mid \theta) = P(U_1 \mid \theta)\, P(U_2 \mid \theta) \ldots P(U_j \mid \theta) \ldots P(U_n \mid \theta)$$

หรือเขียนใหม่ได้ในรูป

$$P(U_1, U_2, \ldots, U_n \mid \theta) = \prod_{j=1}^{n} P(U_j \mid \theta)$$

เนื่องจาก $U_j$ มีค่าเป็น 1 หรือ 0 จึงเขียนได้ในรูปฟังชั่นความเป็นไปได้ ดังนี้

$$P(U_1, U_2, \ldots, U_n \mid \theta) = \prod_{j=1}^{n} P(U_j \mid \theta)^{U_j} \, [1-P(U_j \mid \theta)]^{(1-U_j)}$$

หรือ

$$= \prod_{j=1}^{n} P_j^{U_j} \, Q_j^{(1-U_j)}$$

โดยที่ $P_j = P(U_j \mid \theta)$ และ $Q_j = 1 - P(U_j \mid \theta)$

และเมื่อแสดงในรูปฟังชั่นความน่าจะเป็น (likelihood function) แล้วจะได้ว่า

$$L(u_1, u_2, \ldots, u_n \mid \theta) = \prod_{j=1}^{n} P_j^{u_j} \, Q_j^{(1-u_j)}$$

และเนื่องจาก $P_j$ และ $Q_j$ เป็นฟังชั่นของ $\theta$ และพารามิเตอร์ของข้อสอบ ฟังชั่นข้างบนจึงเป็นฟังชั่นของพารามิเตอร์เหล่านี้ด้วย และจากคุณสมบัติของล็อคการิทึม ที่ว่า

$$\ln xy = \ln x + \ln y \quad \text{และ}$$

$$\ln x^a = a \ln x$$

ฟังชั่นข้างบนจึงแสดงได้ในรูปของ log-likelihood ดังนี้

$$\ln L(U|\theta) = \sum_{j=1}^{n} [u_j \ln P_j + (1-u_j) \ln (1-P_j)$$

โดยที่ u คือ เวคเตอร์ของการ์ดอบข้อสอบ

ฟังชั่นความเป็นไปได้ข้างบนมีค่าสูงสุดได้ เมื่อสามารถประมาณค่า $\theta$ ที่ใช้แทนค่าในสมการข้างล่าง ซึ่งเรียกว่า สมการ likelihood ได้

$$\frac{d}{d\theta} \ln L(U|\theta) = 0$$

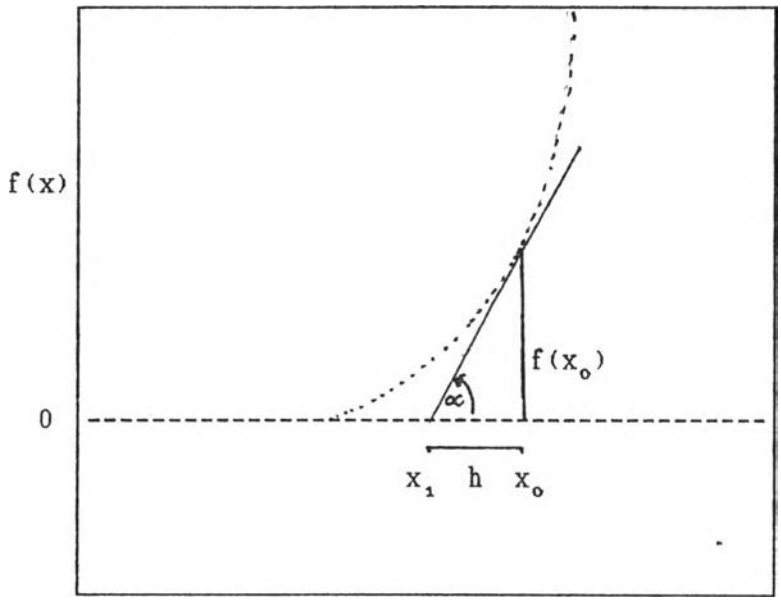
การแก้สมการ likelihood ต้องอาศัยวิธีการของ Newton-Raphson ซึ่งสมมุติว่า ถ้าสมการที่ต้องการแก้คือ

$$f(x) = 0$$

และค่าที่ประมาณได้จากสมการคือ $x_0$ ในขณะที่ค่าที่ถูกต้องมากกว่าคือ $x_1$ สามารถแสดงได้ในรูป

$$x_1 = x_0 - h$$

และจากภาพที่ 6 $h = \dfrac{f(x_o)}{\tan\alpha}$




ภาพที่ 6: ภาพแสดงวิธีการของ Newton - Raphson (Hambleton และ Swaminathan, 1985: 80)

และเนื่องจาก $\tan\alpha$ เป็นความชันของ $f(x)$ ที่ $x_o$ ดังนั้น $\tan\alpha = f'(x_o)$ เมื่อ $f'(x_o)$ เป็น derivative ของฟังชั่นที่ประเมินได้ที่ $x_o$ ดังนั้น

$$h = \frac{f(x_o)}{f'(x_o)} \qquad \text{และ}$$

$$X_1 = X_o - \frac{f(x_o)}{f'(x_o)}$$

และเมื่อได้ค่า $x_1$ แล้ว การประมาณค่า $x$ อื่น ๆ ที่มีความถูกต้องแม่นยำกว่า $x_1$ ก็ดำเนินไปตามที่ได้ประมาณค่า $x_1$ การประมาณค่า $x$ จะมีการทำซ้ำจนกระทั่งความแตกต่างระหว่าง

ค่า $x_m$ และค่า $x_{m+1}(= X_{m+1} - X_m)$ ต่ำกว่าค่าที่กำหนดไว้และค่า $X_m$ ถือเป็นค่าที่ได้จากสมการ likelihood

จากหลักดังกล่าวของ Newton-Raphson

$$\text{เมื่อ} \quad f(x) = \frac{d}{d\theta} \ln L(U|\theta)$$

$$\text{ดังนั้น} \quad f'(X) = \frac{d^2}{d\theta^2} \ln L(U|\theta)$$

และถ้า $\theta_m$ เป็นการประมาณค่ารอบที่ m ของค่า $\theta$ แล้ว การประมาณค่า $\theta$ ในรอบที่ $m + 1$ คือ

$$\theta_{m+1} - \theta_m = \left| \frac{d}{d\theta} \ln L(U|\theta) \right|_m \Big/ \left| \frac{d^2}{d\theta^2} \ln L(U|\theta) \right|_m$$

และการประมาณซ้ำ จะกระทำต่อไปจนกว่าจะถึงเกณฑ์ที่กำหนด ดังการประมาณค่า X ที่กล่าวข้างบน

2.4.2 วิธี Joint and Marginal Bayesian estimation

วิธี Bayesian เป็นวิธีที่ได้ค่าประมาณของความสามารถที่ถูกต้อง ถ้าทราบข้อมูลเกี่ยวกับความสามารถของผู้สอบมาบ้างแล้ว หรือมีการระบุรูปแบบการกระจายของค่าความสามารถผู้สอบได้ เช่น ระบุว่าผู้สอบเป็นตัวอย่างที่สุ่มมาจากประชากรที่มีการกระจายเป็นปกติ และมีการกระจายของความสามารถเป็นปกติ

$$\theta_a \sim N(\mu, \sigma)$$

ค่า $\mu$ และ $\sigma$ ต้องมีการระบุค่าว่าเท่าไร เช่น Owen (1975 อ้างใน Hambleton และ Swaminathan, 1985: 92) กำหนดให้ $\mu = 0$ และ $\sigma = 1$ ในการศึกษาแบบสอบคัดแปลง (adaptive testing) การระบุการกระจายของความสามารถเป็นการกระจายปกตินับว่าเป็น

วิธีที่ค่อนข้างสะดวก แต่ Birnbaum (1969, cited in Hambleton และSwaminathan หน้าเดียวกัน) ระบุการกระจายค่าความสามารถเดิมของผู้สอบในรูปของฟังชั่น $f(\theta)$ โดยที่

$$f(\theta) = \exp(\theta) / [1+\exp(\theta)]^2$$

ในบางกรณี การระบุการกระจายของความสามารถเดิม โดยพิจารณาจากข้อมูลเชิงประจักษ์อาจเหมาะสมกว่า เช่น อาจพิจารณาจากการกระจายของคะแนนดิบ หรือคะแนนดิบแปลง

ตามทฤษฎีของเบย์ส ซึ่งได้กำหนดความสัมพันธ์ของความน่าจะเป็นเชิงเงื่อนไข และความน่าจะเป็นอิสระ (conditional and marginal probabilities) ในรูปของสมการดังนี้

$$P(B|A) = P(A|B)\,P(B)/P(A) \qquad (1)$$

ในบริบทของการประมาณค่าความสามารถ A คือ $\theta_a$ และ B คือ U ซึ่งได้แก่ ชุดการตอบแบบสอบถาม n ข้อ ดังนั้น สมการ (1) สามารถเขียนใหม่เป็น

$$P(\theta_a|U) = P(U|\theta_a)\,P(\theta_a)/P(U) \qquad (2)$$

และเนื่อง $\theta_a$ เป็นตัวแปรต่อเนื่อง สมการ (2) จึงเป็นฟังชั่นความหนาแน่น $P(\theta_a)$ ซึ่งในที่นี้คือ การกระจายเดิมของความสามารถซึ่งในสมการอาจสับสนกับ function การตอบข้อสอบ ดังนั้น ฟังชั่นความหนาแน่นใน (2) จึงระบุได้ในรูปของ f(U) นั่นคือ

$$f(\theta_a|U) = f(U|\theta_a)\,f(\theta_a)/f(U) \qquad (3)$$

และสำหรับการตอบข้อสอบชุดหนึ่ง ๆ f(U) คือ ค่าคงที่ $f(\theta_a|U)$ เป็นความหนาแน่นภายหลังของ $\theta_a$ และ $f(\theta_a)$ เป็นความสามารถเดิม $f(U|\theta_a)$ สามารถแสดงได้ในรูปของฟังชั่น likelihood ของค่าที่สังเกตได้ ดังนั้น สมการ (3) จึงเขียนใหม่ได้เป็น

$$f(\theta_a|U) \propto L(U|\theta_a)\,f(\theta_a) \qquad (4)$$

เมื่อ $L(U|\theta_a)$ คือ ฟังชั่น likelihood ซึ่งสามารถแปลความหมายได้เป็น ฟังชั่นภายหลัง $\propto$ ความเป็นไปได้ x ฟังชั่นเดิม

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \qquad (5)$$

เมื่อจำนวนผู้สอบมี N คน ความหนาแน่นทั้งก่อนและหลังจะเป็นความหนาแน่นร่วมของ $\theta_1, \theta_2, \ldots, \theta_n$ ดังนั้น สมการที่ (5) จะขยายออกเป็น

$$f(\theta_1, \theta_2, \ldots, \theta_n \mid U_1, \ldots, U_n)$$

$$\propto L(U_1, U_2 .., U_n \mid \theta_1, \theta_2, \ldots, \theta_n)\, f(\theta_1, \theta_2, \ldots, \theta_n) \qquad (6)$$

ฟังชั่น likelihood จะได้แก่

$$L(U \mid \theta) = L(U_1, U_2 .., U_n \mid \theta_1, \ldots, \theta_n) = \prod_{a=1}^{N} L(U_a \mid \theta_a)$$

$$= \prod_{a=1}^{N} \prod_{i=1}^{n} L(U_{ia} \mid \theta) = \prod_{a=1}^{N} \prod_{i=1}^{n} P_{ia}^{\,U_{ia}}\, I^{\,1-U_{ia}} \qquad (7)$$

เมื่อได้ค่า $f(\theta_a)$ ออกมาเป็นค่าคงที่แล้วสมการ (4) จะลดลงเป็น

$$f(\theta_a \mid U) \propto L(U \mid \theta_a)$$

ซึ่งได้แก่ ฟังชั่น likelihood ในกรณีนี้การประมาณค่าของเบย์สจะได้ค่าเท่ากับการประมาณค่าด้วย maximum likelihood ตามวิธีของเบย์สแล้ว อาจเริ่มโดยการระบุว่า การกระจายเดิมของ $\theta$ เป็นการกระจายปกติมีค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น 1

$$\theta_a \quad N(0, 1) \quad \text{หรือ}$$

$$f(\theta_a) \propto \exp(-1/2\ \theta_a)^2 \qquad (8)$$

ถ้าเรากำหนดให้ความสามารถมีการแจกแจงอย่างเป็นอิสระ ซึ่งเป็นข้อตกลงที่มีเหตุผลเป็นไปได้ แล้วการแจกแจงภายหลังจะเป็นดังนี้

$$f(\theta_1, \theta_2, \ldots, \theta_n \mid U) \propto L(U \mid \theta_1, \ldots, \theta_n)\, f(\theta_1, \ldots, \theta_n) \qquad (9)$$

และเนื่องจาก

$$f(\theta_1, \ldots, \theta_n) = f(\theta_1)\, f(\theta_2), \ldots, f(\theta_n)$$

$$\propto \prod_{a=1}^{N} \exp((-1/2)\,\theta_a)$$

$$= \exp((-1/2)\,\theta_a) \qquad (10)$$

ดังนั้น การกระจายภายหลังจึงเป็น

$$f(\theta_1, \theta_2, \ldots, \theta_n \mid U) \propto L(U \mid \theta_1, \theta_2 \ldots, \theta_n)\,[\exp(-1/2\ \theta_a)] \qquad (11)$$

จากนั้นประมาณค่าสูงสุดของค่าพารามิเตอร์ที่สอดคล้องกับค่าสูงสุดของฟังชั่นความหนาแน่นภายหลัง โดยการใช้ค่าล็อกการิทึมเข้าช่วย

$$\ln f(\theta \mid U) = \text{ค่าคงที่} + \ln L(U \mid \theta) - \sum_{a=1}^{N} \theta_a \qquad (12)$$

ผลของการแก้สมการ

$$\frac{\partial}{\partial \theta_a} \ln f(\theta \mid U) = 0 \quad a=1,\ldots,n \qquad (13)$$

ได้แก่ตัวประมาณค่าฐานนิยมของ $\theta_1, \ldots, \theta_n$ ของเบย์ส สมการ (13) สามารถเขียนได้ในรูป

$$\sum_{i=1}^{n} k_i(U_{ia}-P_{ia}) - \theta_a = 0 \qquad (14)$$

เมื่อ $k_i = D_{a1}(P_{ia}-c_i) / P_{ia}(1-c_i)$ (15)

สำหรับโมเดล 3 พารามิเตอร์

สมการที่ 14 เป็นตัวอธิบายถึงความแตกต่างประการแรกระหว่างวิธีประมาณค่า MLE และ Bayesian และสามารถประมาณค่า $\theta_a$ สำหรับผู้ที่ทำข้อสอบผิดหมดหรือถูกหมด ได้จากสมการ

$$\sum_{i=1}^{n} k_iP_{ia} = \sum_{i=1}^{n} k_iU_{ia} - \theta_a \qquad (16)$$

$$\sum k_iP_{ia} = -\theta_a \qquad (17)$$

และเนื่องจากค่าเฉลี่ยของ $\theta = 0$ $\theta_a$ ที่สอดคล้องกับคะแนนรวม = 0 จึงมีค่าเป็นลบ เมื่อพิจารณาสมการ (17) และการพิจารณาค่า $\theta$ สำหรับผู้ได้คะแนนเต็มก็มีวิธีการเช่นเดียวกัน สมการฐานนิยมซึ่งได้แก่ สมการที่ (14) แก้ได้โดยใช้วิธีการของ Newton-Raphson เช่นเดียวกับที่ใช้ใน MLE

3. วิธี Marginal Maximum Likelihood (MML)

การประมาณค่าพารามิเตอร์ด้วย MML สามารถใช้ได้กับแบบสอบทั้งยาวและสั้น และเริ่มจากรูปแบบการตอบข้อสอบของผู้สอบ

$$X = (x_1, x_2, \ldots, x_n)$$

และสำหรับผู้มีความสามารถ $\theta$ ค่าความเป็นไปได้แสดงได้ดังนี้

$$P(X \mid \theta) = \prod_{j=1}^{n} [P_j(\theta)]^{x_j} \; [1-P_j(\theta)]^{1-x_j}$$

ค่าความน่าจะเป็นที่ได้เป็นความน่าจะเป็นที่มีเงื่อนไขว่าทราบค่าความสามารถ = $\theta$ ซึ่งต่างจากความน่าจะเป็นของรูปแบบ x จากผู้ไม่ทราบค่า $\theta$ ซึ่งสุ่มมาจากประชากรที่มีการกระจายของค่า $\theta$ อยู่ในรูปฟังชั่นความหนาแน่นต่อเนื่อง (continuous density $g(\theta)$) อยู่ในรูปของความน่าจะเป็นที่ไม่มีเงื่อนไข และแสดงได้ในรูปของการอินทิเกรท

$$P(x) = \int_{-\infty}^{+\infty} P(x \mid \theta)\, g(\theta)\, d\theta$$

ค่านี้เรียกว่าค่าความน่าจะเป็นโดยปราศจากเงื่อนไขของ X และเนื่องจากค่า $\theta$ ถูก integrate ออกไป ค่านี้จึงเป็นฟังชั่นของพารามิเตอร์ข้อสอบเท่านั้น และสามารถหาค่าได้จากสูตร gaussian quadrature ดังนี้

$$\bar{P}(x) = \sum_{k=1}^{q} p(x \mid x_k)\, A(X_k)$$

เมื่อ X คือ quadrature point

$A(x_k)$ คือ น้ำหนักทางบวกซึ่งสอดคล้องกับฟังชั่น $g(x)$ ซึ่งหาค่าได้จากตารางใน Stroud และ Sechrest,1966 (อ้างถึงใน Mislevy และ Bock,1990 : 1- 7)

และในวิธีของ MML ค่าพารามิเตอร์ของข้อสอบจะพิจารณาใช้ค่าที่ทำให้สมการข้างล่างมีค่าสูงสุด

$$\log L_M = \sum_{l=1}^{s} r_l \log_e \bar{P}(X_l)$$

เมื่อ $\bar{P}(X_l)$ ประมาณได้ดังสูตร gaussian guadrature $r_l$ คือ ความถี่ของรูปแบบการตอบข้อสอบรูปแบบ $X_l$ จากจำนวนผู้เข้าสอบทั้งหมด (N) และ S เป็นจำนวนรูปแบบที่เป็นไปได้ทั้งหมดจากข้อมูลการตอบข้อสอบ

กรณี 3 พารามิเตอร์ ค่าสูงสุดของสมการข้างบนสำหรับข้อสอบข้อ j หาได้จากสมการ likelihood

$$\sum_{k=1}^{q} \frac{\bar{r}_{jk} - \bar{N}_k P_j(X_k)}{P_j(X_k)\,[1 - P_j(X_k)]} \cdot \frac{\partial P_j(X_k)}{\partial \begin{bmatrix} c_j \\ a_j \\ b_j \end{bmatrix}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

เมื่อ $$\bar{r}_{jk} = \sum_{l}^{s} r_l x_{lj} P(X_l \mid X_k) A(X_k) / \bar{P}(X_l)$$

$$\bar{N}_k = \sum_{l}^{s} r_l P(X_l \mid X_k) A(X_k) / \bar{P}(X_l)$$

เป็นค่าคาดหวังภายหลังของจำนวนข้อสอบที่ถูกและจำนวนข้อสอบที่ทำที่จุด $X_k$ ($x_{lj}$ คือ คะแนน 0, 1 สำหรับข้อสอบ j ในรูปแบบการตอบ l)

จากนั้นใช้ขั้นตอนการคำนวณ EM และวิธีของ Newton-Gauss (= Fisher-scoring) แก้สมการ

## 2.5 การตรวจสอบความลำเอียงของข้อสอบตามทฤษฎีการตอบข้อสอบ

นิยามของความลำเอียงของข้อสอบในแง่ของทฤษฎีการตอบข้อสอบก็คือ ข้อสอบจะลำเอียงถ้าฟังชั่นการตอบข้อสอบข้อนั้น ๆ ของประชากรย่อย 2 กลุ่ม ไม่เป็นเช่นเดียวกัน

จากนิยามดังกล่าวการตรวจสอบความลำเอียงของข้อสอบตามแนวความคิดของทฤษฎีการตอบข้อสอบสามารถกระทำได้โดยเปรียบเทียบฟังชั่นการตอบข้อสอบของผู้สอบกลุ่มย่อยที่ต้องการศึกษา และการเปรียบเทียบอาจกระทำได้ดังนี้ (Hulin Drasgow และ Parsons, 1983 )

1) ทฤษฎีการตอบข้อสอบโมเดล 1 พารามิเตอร์ ซึ่งกำหนดให้ค่าการเดา = 0 และค่าอำนาจจำแนกของทุกข้อ = 1.00 มีค่าความยาก (b) แปรผันไปตามกลุ่ม

เริ่มด้วยการทดสอบความเหมาะสมของแบบจำลองกับข้อมูลทั้ง 2 กลุ่ม เป็นรายข้อด้วย $H_i$ ซึ่งมีการกระจายแบบ $X^2$ มี df = N-1 คัดข้อที่ไม่เหมาะสมออก แล้ววิเคราะห์ความลำเอียงรายข้อด้วย $Z_i$

$$H_i = \sum_{j=1}^{N} \frac{[U_{ij} - \hat{P}_i(\theta_j)]^2}{[\hat{P}_i(\theta_j)\,\hat{Q}_i(\theta_j)]}$$

$U_{ij}$ = คะแนนจากการตอบข้อสอบ แบบ 0-1 ของคนที่ j ในข้อที่ i

$$P_i(\theta) = \frac{1}{1 + \exp[-D(\theta - b_i)]}$$

$$Z_i = \frac{\hat{b}_{iA} - \hat{b}_{iB}}{SE^2 b_{iA} + SE^2 b_{iB}}$$

$$SE^2_{bi} = \sqrt{I_{ai} / (I_{ai} I_{bi} \; I^2_{aibi})}$$

ค่า $Z_i$ สูงแสดงว่า ข้อสอบนั้นยากสำหรับกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่ง

หรืออาจทดสอบด้วยค่า t ซึ่งคำนวณได้จากสูตร

$$t = \frac{\hat{b}_{iA} - \hat{b}_{iB}}{SE^{2b}_{iA} + SE^{2b}_{iB}}$$

ในกรณีนี้ Draba (Cited in Berk, 1982 : 138) เสนอให้ใช้ค่า t > 2.4 เป็นเกณฑ์การตัดสินความลำเอียงของข้อสอบ

2) ทฤษฎีการตอบข้อสอบโมเดล 2 พารามิเตอร์ กำหนดค่าการเดา = 0 ส่วนค่าความยากและค่าอำนาจจำแนกผันแปรไปตามกลุ่มย่อย วิธีการเริ่มด้วยการประมาณค่าพารามิเตอร์แยกตามกลุ่มย่อย A และ B ด้วยวิธี maximum likelihood แล้วเปรียบเทียบโค้งการตอบข้อสอบ ด้วยค่า F

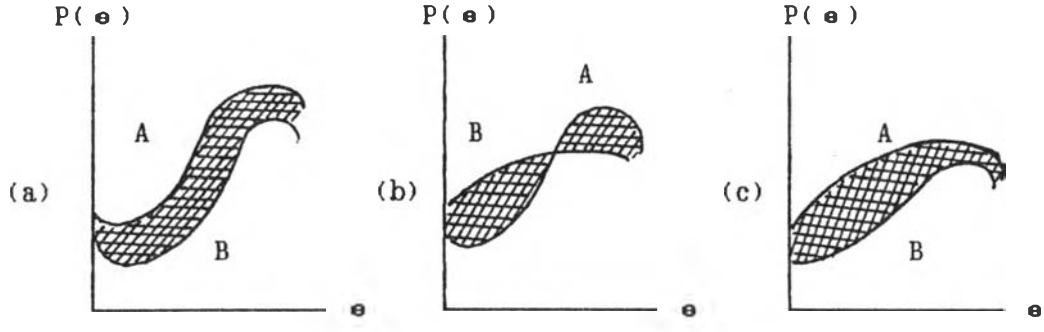
$$F = \frac{SSE(pooled) - [SSE(A)-SSE(B)]}{SSE(A) + SSE(B)} \cdot \frac{[J_A - J_B - 4]}{2}$$

ค่าชั้นแห่งความเป็นอิสระ คือ 2, $(J_A + J_B - 4)$ เมื่อ J เป็นจำนวนช่วงความสามารถ การทดสอบค่า F จะบอกว่าเส้นกดถอยรวมสามารถอธิบายความแปรปรวนของสัดส่วนการตอบของเส้นกดถอยจากแต่ละกลุ่มได้เท่ากันหรือไม่ ถ้าเท่ากัน (ไม่มีนัยสำคัญ) ก็ไม่จำเป็นต้องใช้เส้นกดถอยแยกกันเพราะความสัมพันธ์ระหว่างความสามารถ และการตอบข้อสอบของทั้งสองกลุ่มเป็นเช่นเดียวกัน และข้อสอบทำหน้าที่ในการวัดความสามารถของคน 2 กลุ่ม นี้ได้เหมือนกัน

3) ทฤษฎีการตอบขอสอบโมเดล 3 พารามิเตอร์ เป็นแบบที่ปล่อยให้ค่าการเดาค่าความยาก และค่าอำนาจจำแนกของข้อสอบแปรเปลี่ยนไปตามกลุ่ม

วิธีนี้เริ่มจากการประมาณค่าพารามิเตอร์รวมของทุกคนทุกกลุ่ม แล้วแปลงเป็นค่ามาตรฐานบนค่าประมาณความยาก จากนั้นแยกประมาณค่าความยากและอำนาจจำแนกตามกลุ่มย่อย โดยใช้ค่าการเดาที่ประมาณได้จากการรวมทุกกลุ่ม การทำค่าความยากให้เป็นมาตรฐานต้องตัดข้อที่มีค่าความยากสูงมากหรือต่ำมากออก

การวิเคราะห์ความลำเอียงของข้อสอบตามโมเดล 3 พารามิเตอร์นั้น Linn, Levine, Hastings และ Wardrop,(1981 cited in Hulin, Drasgows และ Parsons, 1983: 176) ได้พัฒนาการทดสอบความลำเอียงจากความแตกต่างระหว่างโค้งลักษณะข้อสอบโดยตรง โดยพิจารณาจากพื้นที่ระหว่างโค้ง 2 โค้งเป็นตัวแสดงความลำเอียง โดยมีลักษณะของพื้นที่อยู่ 3 ลักษณะ ดังภาพที่ 7




ภาพที่ 7: โค้งการตอบข้อสอบที่มีความลำเอียง 3 ลักษณะสำหรับกลุ่ม A และ B

รูป (a) แสดงถึงความลำเอียงที่คงที่ (consistent bias) ข้อสอบมีความลำเอียงต่อกลุ่ม B อย่างสม่ำเสมอตลอดความต่อเนื่องของความสามารถเป็นความลำเอียงอย่างสม่ำเสมอ (uniform bias)

รูป (b) แสดงว่า ข้อสอบมีความยากคล้ายคลึงกันสำหรับทั้ง 2 กลุ่ม แต่มีอำนาจจำแนกสำหรับกลุ่ม B น้อยกว่ากลุ่ม A ทำให้มีความลำเอียงต่อกลุ่ม B ในพวกที่มีความสามารถสูงและลำเอียงต่อกลุ่ม A ที่มีความสามารถต่ำ

รูป (c) มีจุดตัดของโค้งและแกนความเป็นไปได้ในการตอบถูกไม่เท่ากัน สำหรับ 2 กลุ่ม แต่ความยากและอำนาจจำแนกใกล้เคียงกัน มีความลำเอียงต่อกลุ่ม B ในผู้มีความสามารถต่ำ แต่จะลดลงในผู้มีความสามารถสูงขึ้น ลักษณะความลำเอียงเช่นนี้อาจเป็นเพราะสมาชิกในกลุ่ม A ทำข้อสอบถูกเพราะการเดาก็เป็นได้

Shepard และคณะ (1985 : 81) กล่าวว่า เมื่อพบว่าโค้งลักษณะข้อสอบของกลุ่มอ้างอิงและกลุ่มเป้าหมายต่างกันมากพอที่จะเป็นหลักฐานว่า ข้อสอบที่ศึกษาอยู่ไม่ได้วัดลักษณะแฝงเดียวกันสำหรับกลุ่มทั้งสองแล้ว จะใช้ดัชนีชี้ถึงระดับของความลำเอียง ดังนี้

1. พื้นที่ ชนิดไม่มีเครื่องหมาย (unsigned Area) เป็นค่าสัมบูรณ์ของพื้นที่ระหว่างโค้งทั้งสอง

2. ผลบวกกำลังสองแบบที่ 1 (SOS1) เป็นค่าผลบวกของความแตกต่างกำลังสองทุกค่าของความสามารถที่เป็นไปได้

3. ผลบวกกำลังสองแบบที่ 2 (SOS2) คล้าย SOS1 แต่ให้น้ำหนักความแตกต่างกำลังสองด้วยส่วนกลับของความคลาดเคลื่อนของความแปรปรวน ในการประมาณค่าความเป็นไปได้ ดังนั้นถ้ามีการประมาณโค้งอย่างไม่ดีในส่วนใดส่วนหนึ่งแล้ว จะทำให้ลดประโยชน์ต่อความแตกต่างระหว่างกลุ่มสอง

4. ไคสแควร์ (IRT-$\chi^2$) เป็นการทดสอบนัยสำคัญโดยการเปรียบเทียบความแตกต่างของค่าความยาก และอำนาจจำแนกระหว่างกลุ่มพร้อม ๆ กัน โดย Lord (1980)

5. พื้นที่ชนิดมีเครื่องหมาย เหมือนพื้นที่ชนิดไม่มีเครื่องหมาย แต่มีการให้เครื่องหมายไว้เพื่อแสดงให้เห็นว่า กลุ่มใดเสียประโยชน์ และกลุ่มใดได้ประโยชน์ ถ้าโค้งการตอบข้อสอบตัดกัน แสดงว่ามีความลำเอียงต่างกัน สำหรับผู้มีความสามารถต่างกัน และค่าสัมบูรณ์ของพื้นที่ชนิดมีเครื่องหมาย จะน้อยกว่าชนิดไม่มีเครื่องหมาย

6. ผลบวกกำลังสอง แบบที่ 3 (SOS3) เป็นผลบวกกำลังสองของความแตกต่างชนิดมีเครื่องหมายของ SOS1 โดยการใช้เครื่องหมายของความแตกต่างแต่ละตัวที่ยังไม่ยกกำลังสอง

7. ผลบวกกำลังสอง แบบที่ 4 (SOS4) เป็นดัชนีคู่ขนานกับ SOS2 แต่เป็นผลบวกที่มีการให้น้ำหนัก และให้เครื่องหมาย

การวัดพื้นที่ของความแตกต่างระหว่างโค้งลักษณะข้อสอบของกลุ่มตัวอย่าง เพื่อพิจารณาถึงขนาดของความลำเอียงของข้อสอบนั้นมีอยู่ 2 ลักษณะ คือ การวัดช่วงเปิด และช่วงปิด (open and closed intervals) กรณีของการวัดช่วงปิดนั้น จะกำหนดขอบเขตของการวัดให้อยู่ในช่วงความสามารถที่กำหนด ในขณะที่การวัดช่วงเปิดนั้นวัดภายในช่วงความสามารถทั้งหมด เพื่อให้ได้พื้นที่ที่แน่นอน (exact area) ระหว่างโค้ง 2 โค้ง

ภายใต้การวัดแบบช่วงปิดนั้น สามารถวัดพื้นที่ได้ 2 ลักษณะ คือ พื้นที่ชนิดมีเครื่องหมายและไม่มีเครื่องหมาย (signed and unsigned areas) โดยมีสูตรการคำนวณทั่ว ๆ ไป ดังนี้

จากฟังชั่นการตอบข้อสอบ แบบ 3 พารามิเตอร์

$$P(\theta) = c + (1-c)\, P^{*}(\theta)$$

โดยที่ $P^{*}(\theta) = \{1+\exp[+Da(\theta-b)]\}^{-1}$

a, b และ c เป็นพารามิเตอร์ที่แสดงลักษณะของข้อสอบ และ D เป็นตัวคงที่ปกติกำหนดให้ = 1.7

พื้นที่ระหว่างความสามารถที่กำหนดช่วงไว้ 2 จุด บนมาตรความสามารถได้จาก

$$S(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} P(\theta)\,d\theta = c(\theta_2-\theta_1)+(1-c)(Da)^{-1}\ln\frac{\{1-\exp[Da(\theta_2-b)]\}}{\{1+\exp[Da(\theta_1-b)]\}}$$

ในการศึกษาความลำเอียงของข้อสอบนั้น จะมีพารามิเตอร์ของ 2 กลุ่ม คือ กลุ่มอ้างอิง ได้แก่ $a_R$, $b_R$, และ $c_R$ และกลุ่มเปรียบเทียบ ได้แก่ $a_F$, $b_F$, และ $c_F$ พื้นที่ชนิดมีเครื่องหมาย (Closed-interval Signed Area-CSA) ได้จาก

$$CSA = \int_{\theta_1}^{\theta_2} [P_R(\theta) - P_F(\theta)]d\theta = S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)$$

และพื้นที่ชนิดไม่มีเครื่องหมาย (Closed-interval Unsigned Area-CUA) หรือ $\Phi$ ได้จาก

$$CUA = \int_{\theta_1}^{\theta_2} P_R(\theta) - P_F(\theta)d\theta$$

Laksana และ Coffman (1980: 11) เสนอให้ใช้ค่า CUA > .20 เป็นเกณฑ์ชี้ความลำเอียงของข้อสอบ

การวิเคราะห์ความลำเอียงของข้อสอบด้วยทฤษฎีการตอบข้อสอบนั้น Subkoviak, Mack, Ironson และ Craig 1984) พบว่าจากการเปรียบเทียบกับวิธีอื่น ๆ ได้แก่ ค่าความยากแปลง และไคสแควร์แบบ 5 ช่วง พบว่าการวิเคราะห์ด้วยทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ ให้ความถูกต้องมากที่สุด Shepard Camilli และWilliams (1985) ค้นพบสอดคล้องกันในแง่ที่ว่าเป็นวิธีที่ดีที่สุดกว่าวิธี $\chi^2$ และค่าความยากแปลงแต่ 3 คนหลังนี้ ใช้ทฤษฎีการตอบข้อสอบเทียมเข้าไปประยุกต์ใช้ Baghi และ Ferrara (1990) พบว่าในขนาดตัวอย่าง 750 คน สามารถใช้ MH แทน IRT ชนิด 3 พารามิเตอร์ ได้ Kim และ Kohen (1991) พบว่าการเปรียบเทียบความแตกต่างของพื้นที่ใต้โค้งด้วยดัชนีไม่มีเครื่องหมายดีกว่าแบบมีเครื่องหมายรวม ทั้งมีความคลาดเคลื่อนต่ำกว่าแบบมีเครื่องหมาย และระบุว่าลำเอียงโดยที่ไม่ลำเอียงจริงน้อยกว่าวิธีอื่น ๆ คิดเป็นร้อยละ 5 ข้อ Lautenschlager และ Park (1988) พบว่าวิธีของ Linn (1980) ดีกว่าวิธีของ Warm (1978) Rudner Getson และ Knight 1980) พบว่าวิธีการตรวจสอบความลำเอียงด้วยทฤษฎีการตอบข้อสอบโมเดล 3 พารามิเตอร์ ไม่มีผลกระทบจากความยาวของข้อสอบ

การใช้โปรแกรมการถดถอยแบบโลจิสต์ โดย Swaminathan และ Rogers (1990) จากข้อมูลจำลอง พบว่า สามารถตรวจค้นความลำเอียงของข้อสอบได้ร้อยละ 50-75 แล้วแต่ความยาวของข้อสอบ และขนาดของกลุ่มตัวอย่าง ในกรณีที่เป็นการลำเอียงอย่างไม่

สม่ำเสมอ แต่สามารถตรวจค้นความลำเอียงอย่างสม่ำเสมอได้ใกล้เคียงกับวิธี MH แต่ค่าใช้จ่ายมากกว่าถึง 3-4 เท่า Rogers (1989) ได้ใช้การถดถอยแบบโลจิสติ์เพื่อตรวจค้นความลำเอียงของข้อสอบแบบไม่สม่ำเสมอ พบว่า ใช้ได้ดีกับกลุ่มตัวอย่างขนาด 250 คนขึ้นไป และจะตรวจค้นความลำเอียงอย่างสม่ำเสมอได้เท่ากับ MH แต่ตรวจค้นได้ดีกว่าในกรณีของความลำเอียงแบบไม่สม่ำเสมอ

สำหรับการวิจัยครั้งนี้ ผู้วิจัยได้เลือกศึกษา วิธี IRT-2 พารามิเตอร์ โดยเลือกพัฒนาเกณฑ์ของ พื้นที่ความแตกต่างแบบเปิด (open interal) เพื่อให้ได้พื้นที่แน่นอน (exact area) โดยใช้ทั้งพื้นที่ชนิดที่มีเครื่องหมาย (SA) และพื้นที่ชนิดไม่มีเครื่องหมาย (UA)

## 3. วิธีตรวจสอบความลำเอียงของข้อสอบด้วยวิธีของแมนเทลและแฮนส์เซล

เทคนิคแมนเทล-แฮนส์เซล (MH) เป็นเทคนิคที่ Mantel และ Haenszel ได้เสนอขึ้นใช้ตั้งแต่ปี 1959 แต่ Holland และ Thayer เพิ่งนำเสนอมาใช้เพื่อการตรวจสอบความลำเอียงของข้อสอบในปี 1986 และในปีนั้นก็ได้มีการตรวจสอบความลำเอียงของข้อสอบด้วยเทคนิค MH โดย Dorans Hambleton Rogers และ Arrasmith Holland และ Thayer, Wright Raju Bode และ Larsen (1989 : 2)

เทคนิค MH เป็นเทคนิคที่มีความคล้ายคลึงกับวิธีไคสแควร์ที่เสนอโดย Scheuneman (1979) Marascuilo และ Slaughter (1981) และMellenberg (1982) โดยเป็นการเปรียบเทียบผลการตอบข้อสอบของกลุ่มผู้สอบ 2 กลุ่ม หรือที่ฮอลแลนด์เรียกว่ากลุ่มอ้างอิงและกลุ่มเปรียบเทียบ (reference and focal group) โดยที่กลุ่มแรกจะใช้อ้างถึงกลุ่มที่คาดว่าจะได้ประโยชน์จากข้อสอบและกลุ่มหลังเป็นกลุ่มเสียประโยชน์จากข้อสอบในกรณีที่ข้อสอบมีความลำเอียง โดยมีการตรวจสอบทุก ๆ ระดับคะแนนรวมจากการสอบ ข้อสอบใดที่ผู้สอบทั้ง 2 กลุ่ม ทำได้เท่า ๆ กัน จะเป็นข้อสอบที่ถือได้ว่าไม่มีความลำเอียงต่อกลุ่มใดกลุ่มหนึ่ง วิธีการในการตรวจสอบมีดังนี้

ถ้าให้ $N_{Rj}$ และ $N_{Fj}$ แทนจำนวนผู้สอบในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีคะแนนรวมในช่วงคะแนน j และให้ $N_j$ ซึ่ง = $N_{Rj} + N_{Fj}$ เป็นจำนวนผู้สอบรวมทั้ง 2 กลุ่ม ที่ได้คะแนนรวมจากการสอบอยู่ในช่วงคะแนน j แล้วนำมาเขียนในรูปตาราง 2 ทาง แสดงผลการตอบถูก (1) และผิด (0) ดังตารางที่ 1

ตารางที่ 1 ข้อมูลผลการสอบของกลุ่มตัวอย่างกลุ่ม R และ F ที่มีคะแนนรวมอยู่ในช่วงคะแนน j

| | คะแนนที่ได้จากข้อสอบที่ต้องการตรวจสอบความลำเอียง | | | |
|---|---|---|---|---|
| กลุ่มผู้สอบ | | 1 | 0 | รวม |
| | R | $A_j$ | $B_j$ | $N_{Rj}$ |
| | F | $C_{Fj}$ | $D_{Fj}$ | $N_{Fj}$ |
| | รวม | $N_{1j}$ | $N_{0j}$ | $N_j$ |

เมื่อ $A_j$, $B_j$, $C_j$ และ $D_j$ เป็นความถี่ที่สังเกตได้ของการตอบถูก (1) และผิด (0) ในช่วงคะแนน j และแสดงในรูปสัดส่วนได้ ดังตารางที่ 2

ตารางที่ 2 สัดส่วนการตอบข้อสอบของกลุ่มประชากร 2 กลุ่ม

| ผลการตอบ / กลุ่ม | 1 | 0 | รวม |
|---|---|---|---|
| R | $P_{Rj}$ | $Q_{Rj}$ | 1 |
| F | $P_{Fj}$ | $Q_{Fj}$ | 1 |

โดยที่ $P_{Rj}$ คือ สัดส่วนของกลุ่มอ้างอิงที่อยู่ในช่วงความสามารถ j ที่ตอบข้อสอบถูก

$P_{Fj}$ คือ สัดส่วนของกลุ่มสนใจที่อยู่ในช่วงความสามารถ j ที่ตอบข้อสอบถูก

$Q_{Rj}$ คือ $1 - P_{Rj}$

$Q_{Fj}$ คือ $1 - P_{Fj}$

จากตารางที่ 1 จะได้ว่า ในช่วงระดับคะแนน j $A_j$ และ $B_j$ คือ จำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบข้อที่ถูกต้อง และไม่ถูกต้องตามลำดับ ในขณะที่ $C_j$ และ $D_j$ คือ จำนวนผู้สอบกลุ่มเปรียบเทียบที่ตอบข้อสอบข้อที่ถูกต้องและไม่ถูกต้องตามลำดับเช่นกัน

ตามเทคนิควิธี MH ที่เสนอโดย Holland และThayer นั้น ข้อมูลการตอบข้อสอบแต่ละข้อจะถูกนำมาแสดงในรูปตาราง 2 ทางนี้ เป็นจำนวนตารางเท่ากับคะแนนรวมที่มีผู้สอบทำได้บวกด้วย 1 จากนั้นจะคำนวณค่าความเป็นไปได้ในรูปของค่าสัดส่วนการตอบข้อสอบถูก-ผิดระหว่างกลุ่มโดยกำหนดให้เป็นค่า $\alpha$ ดังนี้

$$\alpha_{MH} = \frac{\Sigma\ A_j D_j / N_j}{\Sigma\ B_j C_j / N_j}$$

ค่า $\alpha_{MH}$ จะมีค่าระหว่าง 0 และ $\alpha$ โดยที่ $\alpha_{MH} = 1$ จะแสดงถึงสมมุติฐานสูญที่ว่าไม่มีความแตกต่างหรือไม่มีความลำเอียง ซึ่งแสดงว่ากลุ่มตัวอย่างทั้ง 2 กลุ่ม ทำข้อสอบได้ผลเช่นเดียวกันในแต่ละระดับคะแนนรวม ค่า $A_j D_j / N_j$ และ $B_j C_j / N_j$ มีค่าเท่ากันทั้ง 2 กลุ่มทำให้ $\alpha_{MH} = 1$

อย่างไรก็ตามในทางปฏิบัติแล้ว แม้ว่าจะมีการควบคุมให้ผู้สอบ 2 กลุ่ม มีความสามารถในสิ่งที่ต้องการวัดเท่ากัน แต่เป็นไปไม่ได้ที่ผลการตอบข้อสอบของผู้สอบทั้ง 2 กลุ่ม จะเป็นเช่นเดียวกันดังกล่าว Mantel และHaenszel (1959) จึงเสนอค่าสถิติไคสแควร์เพื่อทดสอบค่าที่ได้ว่าจะมีความแตกต่างจาก 1.00 อย่างมีนัยสำคัญหรือไม่ที่ระดับชั้นแห่งความเป็นอิสระ = 1 ดังนี้

$$\text{MH-CHISQ} = \frac{\Sigma A_j - E(A_j) - 0.5)^2}{\Sigma\ \text{Var}(A_j)}$$

โดยที่ค่าผลรวมเป็นค่าที่ได้จากการรวมทุกชั้นของคะแนนรวม และค่า $E(A_j)$ มีค่าดังนี้

$$E(A_j) = (N_{Rj})\ (N_{1j}) / N_j$$

$$Var(A_j) = \frac{N_{Rj} N_{Ej} N_{1j} N_{0j}}{N_j (N_j - 1)}$$

และสมมุติฐานสูญแสดงว่าข้อสอบไม่ลำเอียง ได้แก่

$H_o$ : $P_{Rj} = P_{Fj}$ สำหรับทุกชั้นคะแนน j

สมมุติฐานศูนย์นี้เป็นสมมุติฐานของความเป็นอิสระอย่างมีเงื่อนไขของสมาชิกกลุ่ม และคะแนนที่ได้จากการตอบข้อสอบที่ต้องการตรวจสอบความลำเอียง และภายใต้สมมุติฐานศูนย์จะได้ค่าคาดหวังของการตอบข้อสอบ ดังนี้

$$E(A_j) = n_{Rj} m_{1j}) / T_j$$
$$E(B_j) = n_{Rj} m_{0j}) / T_j$$
$$E(C_j) = n_{Fj} m_{1j}) / T_j$$
$$E(D_j) = n_{Fj} m_{0j}) / T_j$$

สมมุติฐานอื่นของแมนเทลและแฮนส์เซล ได้แก่

$$H_1: \frac{p_{Rj}}{q_{Rj}} = \alpha \frac{p_{Fj}}{q_{Fj}} \quad j = 1,....K \text{ เมื่อ } \alpha \text{ มีค่า} \neq 1$$

เมื่อ $\alpha$ มีค่า = 1 ซึ่งสอดคล้องกับสมมุติฐานศูนย์ จะได้ว่า

$$H_o: \frac{p_{Rj}}{q_{Rj}} = \frac{p_{Fj}}{q_{Fj}}$$

และค่าประมาณของ $\alpha_{MH} = \frac{p_{Rj}}{q_{Rj}} \cdot \frac{p_{Fj}}{q_{Fj}} = \frac{p_{Rj} q_{Fj}}{p_{Fj} q_{Rj}}$ สำหรับทุก j = 1,...,k

นอกจากนั้น Holland และThayer ได้เสนอแนะให้แปลงค่า $\alpha_{MH}$ ให้เป็นเดลต้า ($\Delta$) เพื่อให้เป็นค่าที่มีค่าเฉลี่ยเป็น 0 ดังนี้

$$\Delta_{MH} = \frac{-4}{1.7} \ln(\alpha_{MH}) = -2.35 \ln(\alpha_{MH})$$

ค่า $\Delta_{MH}$ มีค่าระหว่าง -2.6 ถึง 2.6 และใช้เป็นดัชนีแสดงความลำเอียง ข้อสอบที่ไม่ลำเอียง จะมีค่า $\alpha_{MH} = 1$ หรือ $\Delta_{MH} = 0$

เท่าที่ผ่านมามีผู้นำวิธี MH มาศึกษา เช่น Clauser และคณะ (1991) ใช้ MH วิเคราะห์ความลำเอียงอย่างสม่ำเสมอ พบว่า MH จะใช้ได้ดีในกรณีที่ข้อสอบมีค่าอำนาจจำแนกสูง แต่จะไม่สามารถวิเคราะห์ข้อสอบที่มีค่าความยากสูงมากได้ ซึ่งสอดคล้องกับข้อค้นพบของ Mazor และคณะ (1991) Sudweeks และ Tolman (1990) พบว่า ข้อสอบที่ลำเอียงมักเป็นข้อสอบยาก Baghi และ Ferrara (1990) พบว่า เมื่อใช้ขนาดตัวอย่าง 750 คนขึ้นไป MH ใช้แทน IRT-3 ได้ Swaminathan และ Rogers (1990) ซึ่งศึกษาจากข้อมูลจำลองพบว่า MH วิเคราะห์ได้ดีกว่าการถดถอยแบบโลจิสต์เล็กน้อย โดยตรวจค้นได้ถูกต้อง ร้อยละ 75 ในกรณีใช้กลุ่มตัวอย่าง 250 คน และตรวจค้นได้ ร้อยละ 100 กรณีกลุ่มตัวอย่าง 500 คน กรณีที่ลำเอียงอย่างสม่ำเสมอและกรณีไม่สม่ำเสมอที่ตัดกันที่ปลายข้างใดข้างหนึ่ง แต่ MH มีค่าใช้จ่ายน้อยกว่าโลจิสต์ประมาณ 3-4 เท่า Hambleton และคณะ (1986) พบว่า MH ให้ค่าใกล้เคียงกับ IRT ทั้งที่ใช้ค่าความแตกต่างของค่าเฉลี่ยกำลังสองและการตรวจสอบความแตกต่างของพื้นที่รวมใต้โค้ง แต่ MH มีค่าใช้จ่ายต่ำกว่าและใช้เวลาน้อยกว่า Lincre (1986) พบว่า MH ใช้ได้ดีในสถานการณ์จำลองทุกสถานการณ์พอ ๆ กับโปรแกรม PROX ของ Rasch แต่ควรใช้เกณฑ์ในการคำนวณความลำเอียงและความคลาดเคลื่อนหลาย ๆ เกณฑ์ และใช้ตัวประมาณค่าความคลาดเคลื่อนมาตรฐานมากกว่า 1 ตัว Perlman และคณะ (1988) พบว่า MH มีปัญหาด้านความเที่ยง เมื่อจำนวนกลุ่มตัวอย่างน้อยกว่า 660 คน Thissen Steinberg และ Wainer (1988) พบว่า MH ให้ผลการวิเคราะห์คล้าย IRT-LR และอาจใช้วิเคราะห์ความลำเอียงก่อนใช้ IRT-LR

ในการศึกษาครั้งนี้ ผู้วิจัยได้นำดัชนี $\alpha_{MH}$ มาใช้เพื่อพัฒนาเกณฑ์ตัดสินใจข้อสอบลำเอียง

## 4. วิธีตรวจสอบความลำเอียงของข้อสอบตามการทดสอบ SIB (SIBTEST)

การทดสอบ SIB เป็นวิธีการที่มีแนวคิดบนการตรวจสอบความลำเอียงของแบบสอบชนิดพหุมิติ (multidimensional) โดยมีพื้นฐานบนทฤษฎีการตอบข้อสอบแบบพหุมิติ แต่การทดสอบ SIB แตกต่างจาก IRT ตรงที่เป็นการใช้การทดสอบค่าสถิติไร้พารามิเตอร์(non-parametric) และมีการคำนวณที่ง่ายไม่ซับซ้อน ใช้ได้ดีสำหรับการตรวจสอบความลำเอียงอย่างสม่ำเสมอและมีทิศทางเดียว (unidirectional)

การทดสอบ SIB เป็นวิธีการตรวจสอบความลำเอียงของข้อสอบที่เสนอโดย Shealy และ Stout (Shealy, 1989; Shealy และ Stout, 1991 cited in Shealy และ Stout, 1992: 2) และรูปแบบอยู่ที่การวิเคราะห์ข้อมูลที่มีข้อตกลงว่ามีมิติการวัด 2 มิติ มิติหนึ่งเป็นความสามารถหรือลักษณะแฝงเป้าหมายที่ต้องการวัด และอีกมิติเป็นลักษณะแฝงแทรกซ้อนที่ไม่ต้องการวัด

Shealy และ Stout (1992) เริ่มจากกลุ่มผู้สอบ 2 กลุ่ม คือ กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ การตอบข้อสอบของผู้สอบที่สุ่มมาจากกลุ่มผู้สอบคือ U โดยที่ U = $(U_1,...,U_n)$ เมื่อ $U_i$ มีค่าเป็น 0 ถ้าตอบข้อสอบข้อ i ผิดและเป็น 1 เมื่อตอบถูก

สำหรับโมเดล IRT แล้ว โดยทั่ว ๆ ไป U จะเกิดได้จาก 2 องค์ประกอบ คือ (1) พารามิเตอร์ความสามารถของผู้สอบจำนวน d มิติ และ (2) ฟังชั่นการตอบข้อสอบแต่ละข้อ ซึ่งเป็นตัวกำหนดความน่าจะเป็นของการตอบข้อสอบได้ถูก

ในที่นี้ d คือ 2 เพราะเราจะพิจารณาความสามารถแทรกซ้อนเพิ่มเข้ามาจากความสามารถเป้าหมายที่ต้องการวัดเวคเตอร์ความสามารถในที่นี้คือ ( $\theta$, $\eta$) เมื่อ $\theta$ คือ ความสามารถเป้าหมายและ $\eta$ คือ ความสามารถสอดแทรก ส่วนฟังชั่นการตอบข้อสอบสำหรับข้อ i คือ $P_i(\theta, \eta)$ โดยที่ข้อสอบทุกข้อจะประกอบด้วย $\theta$ ในขณะที่บางข้อจะมี $\eta$ อยู่ด้วย สำหรับข้อที่ประกอบด้วย $\theta$ อย่างเดียว ฟังชั่นการตอบข้อสอบ คือ $P_i(\theta)$ ค่า $P_i(\theta, \eta)$ จะสูงขึ้นใน $\theta$ และใน $\eta$ เมื่อข้อสอบข้อ i ต้องการความสามารถทั้ง 2 อย่าง และจะสูงขึ้นใน $\theta$ ถ้าต้องการเฉพาะความสามารถ $\theta$

นิยามของความลำเอียงของข้อสอบในการทดสอบ SIB คือ "ความลำเอียงของข้อสอบจะเกิดขึ้นต่อกลุ่มเปรียบเทียบที่ระดับความสามารถ $\theta$ ถ้าฟังชั่นการตอบข้อสอบสำหรับความสามารถเป้าหมาย (target alility) ของกลุ่มอ้างอิงมีค่ามากกว่าฟังชั่นการตอบข้อสอบของกลุ่มเปรียบเทียบ" (Stout และShealy,1993) ซึ่งแสดงได้ในรูปของ

$$T_R(\theta) > T_F(\theta)$$

เมื่อ $T_R(\theta) = E_R[P(\theta, \eta) \mid \theta]$

$$= \int p(\theta, \eta) f_R(\eta \mid \theta) d\eta$$

$TF(\theta) = E_F[P(\theta, \eta) \mid \theta]$

$$= T_F(\theta)$$

$\theta = (\theta, \eta)$

โดยที่ R คือ กลุ่มอ้างอิง F คือ กลุ่มเปรียบเทียบ

$\theta$ = ความสามารถเป้าหมาย

$\eta$ = ความสามารถสอดแทรกซึ่งอาจเป็นพหุมิติก็ได้

$T_R(\theta)$ = ฟังชั่นการตอบข้อสอบของกลุ่มอ้างอิง

$T_F(\theta)$ = ฟังชั่นการตอบข้อสอบของกลุ่มเปรียบเทียบ

ความลำเอียงของข้อสอบจะเกิดขึ้นเมื่อ

$$\eta_F \mid \theta = \eta_R \mid \theta$$

นิยามความลำเอียงในแง่ของแบบสอบทั้งฉบับ เมื่อให้ U เป็นการตอบข้อสอบแบบสุ่มของผู้สอบคนหนึ่ง h(U) คือ คะแนนจากแบบสอบ โดยที่ $h(\underline{U}) = \Sigma Ui$ แล้วจะได้ว่าความลำเอียงของข้อสอบจะเกิดต่อกลุ่มเปรียบเทียบในระดับความสามารถ $\theta$ ถ้า

$$T_R(\theta) > T_F(\theta)$$

เมื่อ $T_R(\theta) \underline{\underline{\text{def}}} E_R[h(\underline{U}) \mid \theta]$

$T_F(\theta) \underline{\underline{\text{def}}} E_F[h(\underline{U}) \mid \theta]$

โดยที่ปริมาณของความลำเอียงที่ระดับ $\theta$ คือ $\beta(\theta) \underline{\underline{\text{def}}} T_R(\theta) - T_F(\theta)$ และดัชนีของความลำเอียง คือ

$\beta_u = \int B(\theta)\, f_F(\theta)\, d\theta$ ซึ่งเป็นความลำเอียงเฉลี่ยต่อผู้สอบแต่ละคนในกลุ่มที่สนใจศึกษา

สมมุติฐานเพื่อการทดสอบความลำเอียงของข้อสอบหรือแบบสอบของ SIBTEST

$$H_0: \beta = 0 \qquad H_1: |\beta| > 0$$

การทดสอบทำโดยแบ่งแบบสอบออกเป็นแบบสอบย่อย 2 ฉบับ คือ V เป็นแบบสอบที่มีความตรง และ S แบบสอบที่ต้องการศึกษา และคะแนนของแต่ละแบบสอบย่อยคือ

$$Y = \sum_{S} U_i$$

$$X = \sum_{V} U_i$$

สำหรับการตรวจสอบความลำเอียงของข้อสอบรายข้อ S คือ ข้อสอบแต่ละข้อที่ตรวจสอบ และ V คือ ข้อสอบที่เหลือในแต่ละรอบของการตรวจสอบ

$\bar{Y}_{Rk}$ คือ คะแนนเฉลี่ยที่ได้จากแบบสอบย่อยที่ศึกษาสำหรับผู้เข้าสอบทุกคนที่มีคะแนนรวมในแบบสอบย่อยที่ถือว่าเป็นแบบสอบที่มีความตรง (X) = k

และความลำเอียงวัดได้จาก $(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ S ทั้งนี้ เพราะ $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ เป็นความแตกต่างของการตอบของกลุ่มย่อยที่มีความสามารถระดับเดียวกัน (= k) บน ความสามารถเป้าหมาย และกรณีที่ข้อสอบไม่มีความลำเอียงค่า $\bar{Y}_{Rk} - \bar{Y}_{Fk} = 0$

สำหรับค่าสถิติที่ประมาณค่าความลำเอียง $\beta$ ได้จาก

$$\beta = \sum_{k=0}^{n} P_k(\bar{Y}_{Rk} - \bar{Y}_{Fk})$$

เมื่อ $P_k$ = สัดส่วนของผู้สอบกลุ่มที่สนใจศึกษาที่ตอบแบบสอบย่อยที่ถือว่าไม่ลำเอียงได้ถูก k ข้อ

สถิติทดสอบของ SIBTEST คือ

$$\beta = \frac{\hat{\beta}}{\sigma(\hat{\beta})} \quad \text{ซึ่ง} = N(0,1) \quad \text{เมื่อ} \quad \beta = 0$$

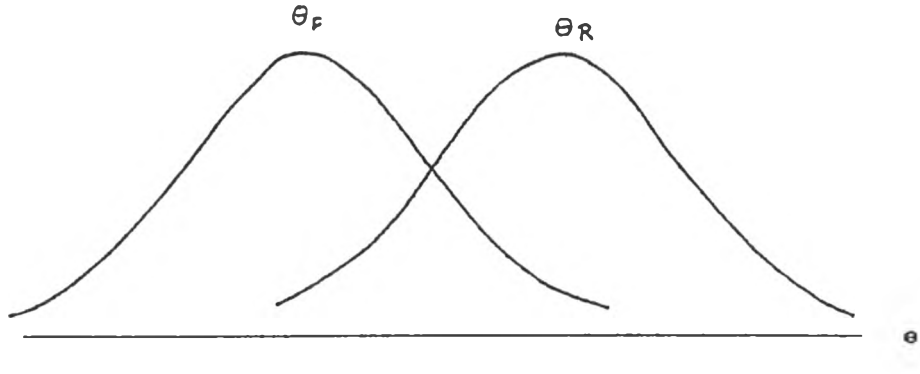
โดยที่

$$\sigma(\hat{\beta}) = \left[ \sum_{k=0}^{n} P_k \left[ (1/J_{Rk}) \sigma^2(y|k,R) + (1/J_{Fk}) \sigma^2(y|k,F) \right] \right]^{1/2}$$

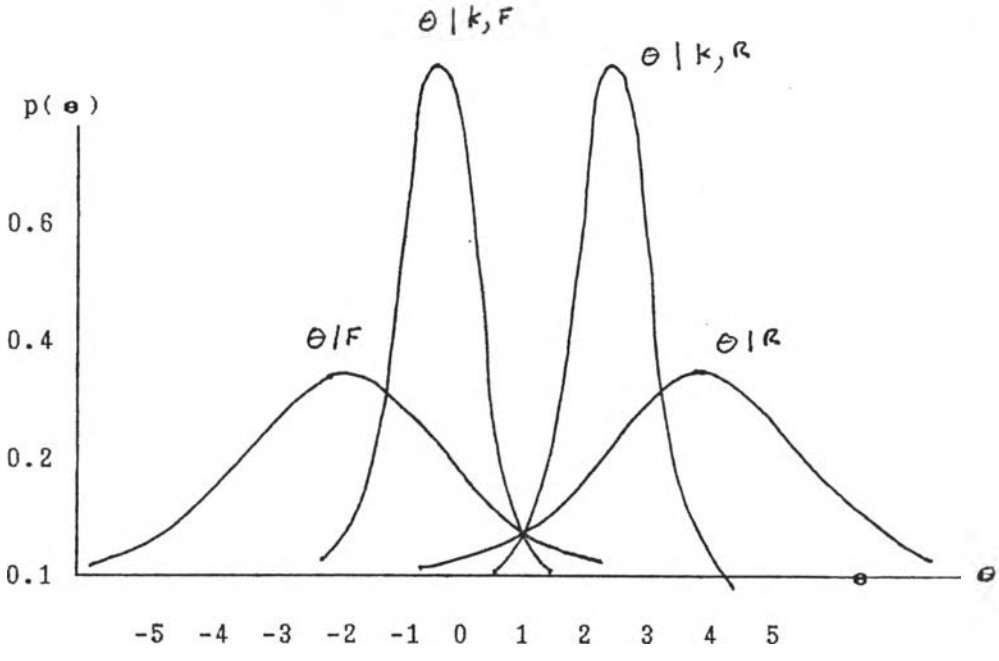
เมื่อ $\sigma^2(Y|k,g)$ เป็นความแปรปรวนของคะแนนของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบในช่วงคะแนน k

อย่างไรก็ตามเมื่อมีความแตกต่างระหว่างความสามารถเป้าหมาย ค่าสถิติ $\beta$ จะสูงกว่าปกติ ทำให้เกิดการระบุว่าข้อสอบลำเอียงทั้ง ๆ ที่ไม่ลำเอียง

จากภาพข้างล่าง ซึ่งแสดงความแตกต่างระหว่างกลุ่มประชากรย่อยในความสามารถเป้าหมาย ( $\theta$ )




เมื่อแบ่งระดับความสามารถตามแบบสอบย่อยที่ถือว่าไม่ลำเอียงจะปรากฏปัญหาการมีค่าสถิติสูงกว่าปกติ ดังภาพ




นั่นคือ ที่ระดับความสามารถ k เราคาดได้ว่า แม้ข้อสอบไม่ลำเอียง ผู้สอบในกลุ่มสนใจศึกษา จะมีความสามารถต่ำกว่าผู้สอบในกลุ่มอ้างอิง นั่นคือ

$$\bar{Y}_{Rk} - \bar{Y}_{Fk} > 0 \quad \text{ทุก ๆ ค่าของ } k$$

ดังนั้น จึงใช้ค่าแก้ด้วยการใช้ค่าถดถอยตามทฤษฎีการตอบข้อสอบชนิด 3 พารามิเตอร์ เพื่อกำจัดค่าเอ้อของค่าสถิติ โดยการแปลงค่า

$$\bar{Y}_{Rk} - \bar{Y}_{Fk} \quad \text{ให้เป็น} \quad \bar{Y}^{*}_{Rk} - \bar{Y}^{*}_{Fk}$$

โดยที่

$$Y_{gk} = \bar{Y}_{gk} + \hat{M}_{gk}\,[\hat{V}(k) - \hat{V}_{g}(k)]; \; g \text{ คือ } R \text{ หรือ } F$$

และ
$$M_{gk} = \frac{\bar{Y}_{g,(k+1)} - \bar{Y}_{g,(k-1)}}{V_{g}(k+1) - V_{g}(k-1)}$$

$$V(k) = (1/2)[V_{R}(k) + V_{F}(k)] = (1/2)(T_{v}(\theta_{Rk}) + T_{v}(\theta_{Fk}))$$

$$V(k) = T_{v}(\theta_{k})$$

การทดสอบนัยสำคัญทางสถิติ เมื่อ $H_0$: $\beta = 0$ ได้แก่

$$Z_{\beta_{IB}} = \frac{\beta}{SE(\beta)}$$

และจะปฏิเสธสมมุติฐานศูนย์ที่ว่า $\beta = 0$

เมื่อ $\beta > Z_\alpha$ โดยที่ค่า $Z_\alpha$ กำหนดไว้จาก $P[N(0,1) > Z_\alpha = \alpha$

การศึกษาวิจัยครั้งนี้ ผู้วิจัยได้เลือกพัฒนาเกณฑ์ การตัดสินข้อสอบลำเอียงของดัชนี $B_{SIB}$ จากข้อมูลเชิงประจักษ์ เพื่อนำไปใช้แทนการทดสอบนัยสำคัญ ของค่า Z ที่ได้จากโปรแกรมการวิเคราะห์

## 5. งานวิจัยที่เกี่ยวข้องกับการตรวจสอบความลำเอียงของข้อสอบ

งานวิจัยเกี่ยวกับการตรวจค้นความลำเอียงของข้อสอบเท่าที่มีผู้ทำมานั้น แบ่งเป็นงานวิจัยภายในประเทศไทยและงานวิจัยต่างประเทศ ดังนี้

### 5.1 งานวิจัยในประเทศ

ชัชชัย เผ่าพงษ์ (2526) ได้ศึกษาความลำเอียงระหว่างผู้สอบเพศชายและหญิงของแบบทดสอบมาตรฐานวัดความถนัดทางการเรียนด้านคณิตศาสตร์ และภาษาในระดับมัธยมต้น จำนวน 2 ฉบับ ด้วยวิธีวิเคราะห์ความแตกต่างของโค้งลักษณะข้อสอบ จากทฤษฎีการตอบข้อสอบ 3 พารามิเตอร์ กลุ่มตัวอย่างที่ใช้เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในปีการศึกษา 2524 จากทุกภาคภูมิศาสตร์ของประเทศไทย ได้แก่ ภาคเหนือ ภาคกลาง ภาคใต้ ภาคตะวันออก และภาคตะวันออกเฉียงเหนือ

ผลการวิจัยพบว่า แบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์ มีข้อกระทงที่ลำเอียงต่อกลุ่มนักเรียนชายและหญิงในระดับปานกลางขึ้นไป จำนวน 5 ข้อ เนื้อหาของข้อกระทงเป็นเรื่องเกี่ยวกับร้อยละ การหาปริมาตรและการหาความยาวเส้นรอบรูปสามเหลี่ยมอย่างละ 1 ข้อ และเป็นเนื้อหาเกี่ยวกับโจทย์ปัญหาทางคณิตศาสตร์ 2 ข้อ ข้อสอบที่พบว่าลำเอียงทั้ง 5 ข้อ ลำเอียงต่อนักเรียนชาย 4 ข้อ อีก 1 ข้อ ลำเอียงต่อนักเรียนหญิงในช่วงความสามารถแรก และลำเอียงต่อนักเรียนชายในช่วงความสามารถต่อมา

สำหรับการวิเคราะห์ความลำเอียงในแบบทดสอบวัดความถนัดทางภาษาเกี่ยวกับการอ่านเข้าใจ มีข้อสอบที่พบว่าลำเอียงในระดับปานกลางขึ้นไป 9 ข้อ เป็นข้อที่มีเนื้อหาวัดความเข้าใจเกี่ยวกับการอ่านคำประพันธ์ และบทร้อยกรองอย่างละ 1 ข้อ อีก 7 ข้อ เป็นเนื้อหาเกี่ยวกับความเข้าใจในการอ่าน ในจำนวนข้อสอบทั้ง 9 ข้อนี้ เป็นข้อสอบที่มีความลำเอียงต่อกลุ่มนักเรียนชาย 1 ข้อ ต่อนักเรียนหญิง 6 ข้อ และอีก 2 ข้อ มีความลำเอียงต่อนักเรียนชายในระดับความสามารถแรก ๆ และลำเอียงต่อนักเรียนหญิงในระดับความสามารถที่สูงขึ้น

ทัศนีย์ นีรมนตรี (2530) ได้ศึกษาถึงความลำเอียงของแบบสอบวิชาคณิตศาสตร์ของโครงการตรวจสอบคุณภาพการศึกษานักเรียนมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526 ด้วยวิธีวิเคราะห์ 3 วิธีคือ วิธีกำหนดจุดค่าเดลต้า ( $D \pm .75\ Z_g$ ) (2) วิธีการตอบข้อสอบแบบ 3 พารามิเตอร์ ($\Phi$ >.40) และ (3) วิธีการทดสอบความแตกต่างระหว่างกลุ่มด้วยสถิติไคสแควร์ในโมเดลล็อกลิเนีย 2 โมเดล คือ โมเดลที่ไม่มีพารามิเตอร์ผลร่วมระหว่างระดับคะแนนกับกลุ่ม และโมเดลที่ไม่มีพารามิเตอร์ของผลหลักที่เกิดจากกลุ่ม (เปรียบเทียบค่า $\chi^2$ ที่คำนวณได้กับค่า $\chi^2$ ในตาราง) โดยเปรียบเทียบจำนวนข้อกระทงที่มีความลำเอียงระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในภาคภูมิศาสตร์ 5 ภาค ได้แก่ ภาคกลาง ภาคเหนือ ภาคใต้ ภาคตะวันออก และภาคตะวันออกเฉียงเหนือ

ผลการศึกษาพบว่าเมื่อแยกข้อสอบออกตามค่าความยากที่วิเคราะห์ด้วยวิธีดั้งเดิม พบว่า ข้อที่ลำเอียงมีจำนวน 43 ข้อ และไม่ลำเอียง 17 ข้อ เหมือนกันทุกคู่ทุกภาคและเมื่อเปรียบเทียบจำนวนข้อที่ลำเอียงระหว่างกลุ่มนักเรียนในกรุงเทพมหานคร และกลุ่มนักเรียนภาคอื่น ๆ พบว่าในการใช้ IRT-3 พารามิเตอร์ พบข้อกระทงที่มีความลำเอียงจำนวนมากที่สุด มีข้อกระทงที่ลำเอียงซ้ำกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับทุก ๆ ภาค แต่จำนวนไม่เท่ากันในแต่ละวิธี วิธีที่ 1 และ 3 มีข้อลำเอียงซ้ำกันมากสุดระหว่างกรุงเทพมหานครและภาคตะวันออกเฉียงเหนือ ส่วนวิธีที่ 2 มีจำนวนข้อที่ลำเอียงซ้ำกันมากที่สุดระหว่างกรุงเทพมหานครและภาคตะวันออก

การเปรียบเทียบผลการวิเคราะห์ทั้ง 3 วิธี กับความลำเอียงที่เกิดขึ้นในภาคเดียวกันพบข้อกระทงที่ลำเอียงซ้ำกัน ข้อกระทงที่ลำเอียงส่วนใหญ่เป็นข้อที่ง่ายสำหรับนักเรียนในกรุงเทพมหานครมากกว่ากลุ่มนักเรียนในภาคอื่น ๆ สำหรับวิธีวิเคราะห์ที่ 1 แต่เป็นข้อที่มีความลำเอียงในเกณฑ์ต่ำสำหรับวิธีที่ 2 และเป็นข้อที่ลำเอียงอย่างสม่ำเสมอในการวิเคราะห์ด้วยวิธีที่ 3

สุรศักดิ์ อมรรัตนศักดิ์ (2531) ได้ศึกษาดัชนีความลำเอียงของข้อสอบ โดยพิจารณาถึงสัมประสิทธิ์สหสัมพันธ์ระหว่าง วิธีวิเคราะห์ความลำเอียงของข้อสอบ 4 วิธี คือ วิธีวิเคราะห์ความแปรปรวน วิธีแปลงค่าความยากของข้อสอบ ($D \pm 3\ Sd$) วิธีโค้งลักษณะข้อสอบ

โมเดล 1 พารามิเตอร์ (ทดสอบความแตกต่างระหว่างค่าความยากด้วย t เกณฑ์ที่แสดงว่าข้อสอบลำเอียง คือ $t \geq 2.4$) และ 3 พารามิเตอร์ ( $\Phi_1 \geq .20$) และเปรียบเทียบความแตกต่างของผลการคัดเลือกก่อนและหลังการศึกษา ความลำเอียงของข้อสอบในด้านจำนวนผู้ได้รับการคัดเลือกสัดส่วนชาย : หญิงที่ได้รับการคัดเลือกและความเที่ยงของแบบสอบ ผลการวิจัยสรุปได้ว่า

1. วิธีโค้งลักษณะข้อสอบโมเดล 3 พารามิเตอร์ ค้นพบข้อสอบลำเอียงได้มากที่สุด รองลงมา ได้แก่ วิธีวิเคราะห์ความแปรปรวน วิธีที่ตรวจค้นข้อสอบที่ลำเอียงได้น้อยที่สุด ได้แก่ วิธีแปลงค่าความยากของข้อสอบ

2. วิธีทั้ง 4 วิธี มีความสัมพันธ์กันทางบวกอย่างมีนัยสำคัญที่ระดับ .001 โดยมีค่าระหว่าง .7535-.9921

3. การใช้คะแนนดิบและคะแนนรวมแบบอื่น ๆ อีก 5 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันประมาณร้อยละ 4 ถึง 24 ส่วนการใช้คะแนนมาตรฐานที่ที่ปกติรวมกับคะแนนแปลงแบบอื่น ๆ อีก 4 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันร้อยละ 4-23

4. เมื่อตัดข้อสอบที่มีความลำเอียงออก พบว่าสัดส่วนหญิงและชายที่ได้รับการคัดเลือกมีความใกล้เคียงกัน และค่าความเที่ยงของแบบสอบลดลงเล็กน้อย

สุพัฒน์ สุกมลสันต์ (2534) ได้วิเคราะห์ความลำเอียงของข้อสอบภาษาอังกฤษเพื่อคัดเลือกเข้าศึกษาในมหาวิทยาลัย โดยใช้ข้อมูลการตอบแบบทดสอบภาษาอังกฤษเข้ามหาวิทยาลัยชุด กข และ ชุด กขค ปี 2531-2533 ซึ่งมีข้อสอบชุดละ 100 ข้อ ตัวแปรที่ศึกษาความลำเอียงคือ เพศ และภาคภูมิศาสตร์ของผู้สอบ ซึ่งแยกออกเป็น 5 ภาคตามภูมิลำเนาของผู้สอบ ได้แก่ ภาคกลาง ภาคตะวันออก ภาคตะวันออกเฉียงเหนือ ภาคเหนือ และภาคใต้ เทคนิควิธีที่นำมาวิเคราะห์ความลำเอียงมี 3 วิธี ได้แก่ วิธีกำหนดจุดค่าเดลต้า (Delta-Plot Method) วิธีไคสแควร์ชนิดที่แบ่งความสามารถของผู้สอบเป็น 3 ระดับ ได้แก่ กลุ่มความสามารถระดับต่ำ (ผู้ได้คะแนนรวม 0-40 คะแนน) กลุ่มความสามารถระดับปานกลาง (ผู้ได้คะแนนรวม 41-70 คะแนน) และระดับสูง (ผู้ได้คะแนนรวม 71-100 คะแนน) วิธีที่ 3 ที่นำมาใช้วิเคราะห์ความลำเอียงของข้อสอบได้แก่ วิธีการวัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบที่วิเคราะห์ตามทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์

เกณฑ์การตัดสินความลำเอียงของข้อสอบสำหรับวิธีต่าง ๆ ที่ สุพัฒน์ สุกมลสันต์ กำหนดขึ้นคือ

1. วิธีกำหนดจุดค่าเดลต้า ข้อสอบลำเอียงคือ ข้อที่มีระยะตั้งฉาก จากค่าคู่อันดับเดลต้าไปยังเส้นแกนหลัก (d) มากกว่า $= \pm 1.964$ Sd.

2. วิธีไคสแควร์ ข้อสอบลำเอียง คือ ข้อสอบที่ผู้สอบต่างกลุ่มที่อยู่ในระดับคะแนนเดียวกัน มีสัดส่วนการตอบถูกหรือผิดแตกต่างกันอย่างมีนัยสำคัญที่ระดับ $\alpha = 0.017$

3. วิธีการวัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะของข้อสอบ ข้อสอบลำเอียง คือ ข้อที่มีพื้นที่ความแตกต่างมากกว่า 0.40 โดยแบ่งเป็น 3 ระดับ

3.1 พื้นที่ > 0.70 แสดงว่าลำเอียงมาก

3.2 พื้นที่ระหว่าง 0.40-0.70 แสดงว่าลำเอียงปานกลาง

3.3 พื้นที่ที่มีค่ามากกว่า 0.00 แต่น้อยกว่า 0.40 แสดงว่าลำเอียงน้อย

ผลการวิเคราะห์พบว่าแบบทดสอบภาษาอังกฤษ ฉบับ กข และ กขค ปี 2531-2533 มีความลำเอียงต่อเพศ 7-28 ข้อ และ 4-41 ข้อ โดยมีแนวโน้มที่ลำเอียงต่อเพศชายมากกว่าเพศหญิง และลำเอียงต่อภาคภูมิศาสตร์ 6-45 ข้อ และ 5-43 ข้อ โดยมีความลำเอียงต่อผู้สอบจากภาคอื่นมากกว่าภาคกลางประมาณ 2-3 เท่า

การวิเคราะห์เปรียบเทียบผลของการวิเคราะห์โดยใช้วิธีวิเคราะห์ต่างกัน พบว่า มีจำนวนข้อสอบที่ระบุว่าลำเอียงต่างกันอย่างมีนัยสำคัญ แต่ละวิธีให้ผลที่มีความสัมพันธ์กันอย่างไม่มีนัยสำคัญ ผลการวิเคราะห์ด้วยวิธีทฤษฎีการตอบข้อสอบ เมื่อใช้เกณฑ์พื้นที่มากกว่า 0.40 แสดงความลำเอียงของข้อสอบ พบว่ามีจำนวนข้อสอบที่ลำเอียงมากที่สุด รองลงมา ได้แก่ วิธีไคสแควร์ และวิธีกำหนดจุดเดลต้า ซึ่งวิธีสุดท้ายนี้พบข้อสอบลำเอียงน้อยกว่า 2 วิธีแรกประมาณ 3-4 เท่า

จะเห็นว่า การศึกษาถึงความลำเอียงของข้อสอบในประเทศไทยมีการศึกษาอยู่ 2 ลักษณะ คือ การศึกษาความลำเอียงระหว่างผู้สอบเพศชายและหญิง และระหว่างผู้สอบที่มีภูมิลำเนาอยู่ต่างภาคภูมิศาสตร์ วิธีที่นำมาใช้ ได้แก่ วิธีทฤษฎีการตอบข้อสอบโมเดล 1 และ 3 พารามิเตอร์ วิธีกำหนดจุดค่าเดลต้า วิธีไคสแควร์ วิธีวิเคราะห์ความแปรปรวน และวิธีแปลงค่าความยากของข้อสอบ ผลการศึกษาพบว่า ในขณะที่ ชัชชัย เผ่าพงษ์ พบว่าข้อสอบวัดความถนัดด้านคณิตศาสตร์ที่ลำเอียงส่วนมากลำเอียงต่อเพศชาย และข้อสอบวัดความถนัดด้านภาษาที่ลำเอียงส่วนมากลำเอียงต่อเพศหญิงนั้น สุพัฒน์ สุกมลสันต์ พบว่าข้อสอบลำเอียงด้านภาษามีแนวโน้มลำเอียงต่อเพศชายมากกว่าหญิง

ในด้านความลำเอียงระหว่างภาคพบว่ามีความสอดคล้องกันระหว่างการศึกษาของทัศนีย์ พีรมนตรี ที่พบว่าข้อสอบลำเอียงมักเป็นข้อสอบที่ง่ายสำหรับผู้สอบในกรุงเทพมหานคร ในขณะที่ สุพัฒน์ สุกมลสันต์ พบว่าข้อสอบที่ลำเอียงมีความลำเอียงต่อผู้สอบภาคอื่นมากกว่าภาคกลาง โดยเฉลี่ยแล้วลำเอียงต่อผู้สอบภาคตะวันออกเฉียงเหนือ ภาคเหนือ และภาคใต้เป็นจำนวนใกล้เคียงกันและมากกว่าที่ลำเอียงต่อผู้สอบภาคกลางโดยเฉลี่ยแล้วลำเอียงต่อผู้สอบภาคตะวันออกเฉียงเหนือ ภาคเหนือและภาคใต้เป็นจำนวนใกล้เคียงกันและมากกว่าที่ลำเอียงต่อผู้สอบภาคกลาง

ข้อค้นพบอีกประการคือ นอกจากแต่ละวิธีที่ใช้จะให้ผลการค้นพบจำนวนข้อสอบลำเอียง ไม่เท่ากันและผลการศึกษาของสุพัฒน์ สุกมลสันต์ พบว่า มีความสัมพันธ์กันระหว่างการค้นพบแต่ละวิธีอย่างไม่มีนัยสำคัญ นอกจากนั้นเกณฑ์ตัดสินความลำเอียง ด้วยดัชนี $\Phi$ หรือ (SA) ยังมีความแตกต่างกันไปด้วย

กล่าวได้ว่า การศึกษาเรื่องความลำเอียงของข้อสอบในประเทศไทยยังมีน้อยมาก วิธีการ MH และ SIBTEST ยังไม่มีการนำมาศึกษาวิจัย และยังไม่พบว่ามีการศึกษาเปรียบเทียบด้านความยาวของข้อสอบ และการพัฒนาเกณฑ์การตัดสินความลำเอียง นอกจากนั้นข้อค้นพบความลำเอียงของข้อสอบคณิตศาสตร์และภาษาอังกฤษระหว่างผู้สอบต่างเพศยังไม่สอดคล้องกัน ผู้วิจัยจึงหวังว่าการศึกษาของผู้วิจัยในครั้งนี้จะช่วยให้ข้อมูลเพิ่มเติมมากขึ้นในส่วนที่กล่าวมานี้ และจะช่วยให้ได้มีการศึกษาวิจัยในเรื่องนี้มากขึ้นในอนาคต

## 5.2 งานวิจัยต่างประเทศ

งานวิจัยเกี่ยวกับความลำเอียงของข้อสอบในต่างประเทศนั้น ผู้วิจัยจะนำเสนอเป็น 2 ประเด็น คือ (1) การตรวจค้นข้อสอบลำเอียงของข้อมูลจำลอง และ (2) การตรวจค้นข้อสอบลำเอียงจากข้อมูลจริง

### 5.2.1 การตรวจค้นข้อสอบลำเอียงจากข้อมูลจำลอง

การตรวจค้นข้อสอบลำเอียงจากข้อมูลจำลองส่วนมากเป็นวิธีการศึกษาเปรียบเทียบเทคนิคการตรวจสอบต่าง ๆ ที่มีผู้เสนอขึ้นมาใช้ดังนี้

Rudner Getson และ Knight (1980) ได้เปรียบเทียบเทคนิคการตรวจค้นความลำเอียงของข้อสอบจากข้อมูล 3 วิธี ได้แก่เทคนิคความยากแปลงชนิดแกนหลัก (Transformed Item Difficulties Major Axis = TID - MA) เทคนิคความยากแปลงเส้นแกน 45 องศา (Transformed Item Difficulties-45 Line = TID-45) เทคนิคทฤษฎีการตอบข้อสอบ 3 พารามิเตอร์ (ICC-3) และ 1 พารามิเตอร์ ชนิดค่าสถิติความเหมาะสม (ICC-1F) ชนิดความแตกต่างในความง่ายของข้อสอบ (ICC-1E) เทคนิคไคสแควร์แบบ 5 ช่วง (CHI-5) และเทคนิคไคแสควร์ชนิดหลายช่วง (CHI-N) ผลจากการวิเคราะห์ข้อมูลพบว่า เมื่อใช้ความยาวของข้อสอบหลายขนาดไม่พบว่ามีเทคนิคใดที่มีผลจากความยาวของข้อสอบเป็นพิเศษ เมื่อพิจารณาจากค่าสหสัมพันธ์ของผลการตรวจค้นกับข้อมูลที่สร้างขึ้น โดยพิจารณาจากปริมาณของความลำเอียงของข้อสอบ พบว่า วิธีที่ถูกต้องมากที่สุด ได้แก่ ICC-3 CHI-5 TID-45($r$ = .80 .73 และ .68) และ ที่ถูกต้องน้อยที่สุด ได้แก่ ICC-1F ($r$ = .55) และทุกวิธียกเว้น ICC-1F สามารถตรวจค้นความลำเอียงได้ในกรณีที่มีความยากต่างกันเท่านั้น มีเพียง ICC-1F ที่มีความไวต่อความลำเอียงในอำนาจจำแนกของข้อสอบด้วย

เทคนิค CHI-5 มีประสิทธิภาพเท่า ๆ กับ ICC-3 และพบว่าถ้าเพิ่มจำนวนช่วงความสามารถมากขึ้น จะลดความถูกต้องในการตรวจค้นลง ซึ่งอาจเป็นเพราะช่วงความสามารถในแต่ละข้อนั้นต่างกันไปก็ได้

สรุปว่า ถ้าไม่คำนึงถึงความยาวของแบบสอบ และขนาดหรือธรรมชาติของความลำเอียงแล้ว ICC-3 และ ไคสแควร์แบบ 5 ช่วง เป็นวิธีที่ควรใช้วิธีค่าความยากแปลงก็อาจนำมาใช้แทนได้ ถ้าจะตรวจค้นในด้านความต่างของค่าความยากของข้อสอบ

Van Der Flier Mellenbergh Ader และ Win (1984) ได้เปรียบเทียบวิธีการตรวจค้นความลำเอียงของข้อสอบจากข้อมูลจำลองด้วยวิธีที่มีเงื่อนไข วิธีโลจิทแบบดัดแปลงจากวิธีของ Lordที่เสนอการใช้ทฤษฎีการตอบข้อสอบ 3 พารามิเตอร์ แบบเป็นขั้นตอนโดย

ขั้นที่ 1 ประมาณค่าพารามิเตอร์ แล้วคำนวณค่าสถิติที่แสดงความลำเอียงของข้อสอบทุกข้อ

ขั้นที่ 2 คัดข้อสอบที่ลำเอียงออก แล้วประมาณค่าพารามิเตอร์ใหม่จากนั้นจึงนำค่าพารามิเตอร์ที่ได้คำนวณค่าสถิติความลำเอียงของข้อสอบทุกข้อ (รวมทั้งข้อสอบที่ลำเอียงด้วย) ข้อที่มีค่าไคสแควร์สูงสุดในรอบนั้นและมีนัยสำคัญ ถือว่าเป็นข้อสอบที่ลำเอียง ข้ออื่น ๆ นอกนั้น ถือว่าเป็นข้อที่ไม่ลำเอียงแม้ว่า ค่าไคสแควร์ จะมีนัยสำคัญ

ขั้นที่ 3 ทำขั้นที่ 1 และ 2 อีกครั้ง การทำซ้ำจะจบลงถ้าทำซ้ำเท่ากับจำนวนรอบที่กำหนดไว้ หรือค่าไคสแควร์สูงสุดของรอบนั้นไม่มีนัยสำคัญ

ในแต่ละรอบของการคำนวณ คำนวณค่าไคสแควร์จากข้อสอบทุกข้อรวมทั้งข้อที่ตรวจค้นว่ามีความลำเอียงในรอบก่อน ๆ

การคำนวณแต่ละรอบ จำนวนข้อสอบที่ขจัดออกไปจะมากกว่าจำนวนข้อสอบในรอบที่แล้ว 1 ข้อ

ข้อสอบที่ตัดออกจากการคำนวณคะแนนรวมในการคำนวณรอบที่แล้ว ไม่จำเป็นต้องถูกตัดออกไปในรอบต่อไป ถ้าค่าไคสแควร์ ของค่านั้นยอมรับได้ในรอบต่อมา ด้วยวิธีนี้คะแนนรวมจะเป็นอิสระจากข้อสอบลำเอียง และข้อสอบทุกข้อจะได้รับการตรวจสอบโดยมีระดับความสามารถที่ได้จากแบบสอบที่ไม่ลำเอียง

ข้อดีของการทำซ้ำคือจะแน่ใจว่าได้ข้อสอบที่ลำเอียงจริง ๆ จากการตรวจสอบเพราะจะระบุเฉพาะข้อที่มีค่าไคสแควร์ที่สูงสุด และมีนัยสำคัญเท่านั้น

Lautenschlager และ Park (1988) ได้ใช้ข้อมูลจำลองศึกษาเปรียบเทียบวิธีการเชื่อมโยงค่าพารามิเตอร์ (parameter linking method) ซึ่งมีการแปลงค่าพารามิเตอร์ที่ประมาณได้จากทฤษฎีการตอบข้อสอบ ให้เป็นค่ามาตรฐานเพื่อเปรียบเทียบกันได้

ข้ามกลุ่ม โดยมีวิธีที่ Linn และคณะ (1980 ; cited in Lautenschlager และ Park, 1988) และวิธีของ Warm (1978) cited in Lautenschlager และ Park 1988)

วิธีของ Warm ง่ายที่สุดจากการใช้ความคิดรวบยอดที่ว่า ความยากของข้อสอบของทุกกลุ่มจะเหมือนกันได้ด้วยการแปลงค่าเชิงเส้นตรง โดยได้แปลงค่าพารามิเตอร์สัมพันธ์โดยใช้สูตรดังนี้

$$b_1 = \frac{SD_{b1}}{SD_{b2}} b_2 + (M_{b1} - \frac{SD_{b1}}{SD_{b2}} M_{b2})$$

$$a_1 = a_2 \frac{SD_{b2}}{SD_{b1}}$$

$b_1$ คือ ค่าพารามิเตอร์ b บนมาตรของกลุ่ม 1

$b_2$ คือ ค่าพารามิเตอร์ b บนมาตรของกลุ่ม 2

$a_1$ คือ ค่าพารามิเตอร์ a บนมาตรของกลุ่ม 1

$a_1$ คือ ค่าพารามิเตอร์ a บนมาตรของกลุ่ม 2

$M_{b1}$, $SD_{b1}$ คือ ค่าเฉลี่ยและค่าความเบี่ยงเบนมาตรฐานของค่า b บนมาตรของกลุ่ม 1

$M_{b2}$, $SD_{b2}$ คือ ค่าเฉลี่ยและค่าความเบี่ยงเบนมาตรฐานของค่า b บนมาตรของกลุ่ม 2

วิธีการของ Linn นั้น จะประมาณความถูกต้องของ $b_1$ ด้วยค่าเฉลี่ยและค่าความเบี่ยงเบนมาตรฐานที่ได้ให้น้ำหนักไว้ น้ำหนักของแต่ละข้อจะได้แก่ส่วนกลับของความแปรปรวนในการสุ่มของ $b_1$ ที่ประมาณค่าได้สูงสุด ดังนั้นข้อที่มีค่าความแปรปรวนในการสุ่มของ $b_1$ มากที่สุดจะมีน้ำหนักน้อยที่สุดในการพิจารณาค่าคงที่ของการเชื่อมโยงค่าพารามิเตอร์

จากการวิเคราะห์ข้อมูลจำลอง ด้วยการทดสอบค่า $\chi^2$ ของ Lord พบว่าภายใต้สถานการณ์ลำเอียงแบบทิศทางเดียว (แต่ละข้อลำเอียงเข้าข้างกลุ่มใดกลุ่มหนึ่งเพียงกลุ่มเดียว) นั้น ยากที่จะตรวจค้นจำนวนข้อสอบที่ถูกระบุตรงข้ามกับความเป็นจริง (ข้อสอบลำเอียงแต่ตรวจพบว่าเป็นข้อสอบไม่ลำเอียงหรือข้อสอบไม่ลำเอียงแต่ตรวจพบว่าเป็นข้อที่ลำเอียง)

ทำให้ยากที่จะเลือกว่าควรใช้วิธีใด แต่ในการศึกษาครั้งนี้ พอจะพูดได้ว่าวิธีของ Linn และคณะดูจะดีกว่า และยังพบว่าข้อสอบที่มีระดับความลำเอียงน้อย มักจะถูกระบุจากการตรวจค้นว่าเป็นข้อสอบไม่ลำเอียง

Lautenschlager และ Park กล่าวว่าการตรวจสอบข้อสอบลำเอียงในอนาคต น่าจะปรับปรุงโดยใช้ทฤษฎีการตอบข้อสอบแบบพหุมิติ (multidimensional IRT) แทนเมื่อแบบสอบขาดคุณสมบัติด้านเอกมิติ

Linacre (1988) ได้ตรวจสอบความถูกต้องของวิธี MH และวิธี PROX ของ Rasch ในการตรวจสอบความลำเอียงของข้อสอบที่ได้จากการจำลองข้อมูลขึ้นโดยใช้ตัวประมาณค่าความคลาดเคลื่อนมาตรฐานหลาย ๆ วิธี สำหรับวิธี MH ผลการวิเคราะห์ พบว่าทั้งสองวิธีใช้ได้ดีในสถานการณ์จำลองทุกสถานการณ์ ส่วนตัวประมาณค่าความคลาดเคลื่อนมาตรฐานของ MH มีคุณสมบัติแตกต่างกันไปและควรมีการแปลผลอย่างระมัดระวัง เมื่อใช้วิธี MH ผู้วิเคราะห์ควรจะคำนวณขนาดของความลำเอียงของข้อสอบ และความแปรปรวนของความคลาดเคลื่อนด้วยการใช้เกณฑ์หลาย ๆ เกณฑ์ และใช้ตัวประมาณค่าความคลาดเคลื่อนมาตรฐานมากกว่า 1 ตัว สำหรับวิธี PROX จะให้การตรวจสอบการประมาณที่ได้จากวิธี MH

Thissen, Steinberg และ Wainer (1988) ได้ศึกษาความแตกต่างของโค้งการตอบข้อสอบระหว่างกลุ่ม ด้วยทฤษฎีการตอบข้อสอบ เพื่อศึกษาถึงสภาพการณ์ในการทดสอบสมมุติฐาน ความเท่ากันของค่าพารามิเตอร์ด้วยค่าไคสแควร์แบบอัตราส่วน likelihood (IRT-LR) โดยใช้ข้อมูลจำลอง และได้ทำการเปรียบเทียบการทดสอบสมมุติฐานความเท่ากันของค่าพารามิเตอร์ ด้วย IRT-LR และวิธีของ MH พบว่า IRT-LR สามารถตรวจค้นความลำเอียงได้ในเมื่อค่าความแตกต่างของความยากระหว่างกลุ่มมีค่า 0.3 และความแตกต่างในด้านการเดามีค่า 0.1 จำนวนตัวอย่างในกลุ่มน้อย (=500คน) โดยจำนวนข้อสอบร่วมจะน้อย หรือมากก็ไม่มีผลกระทบ ค่าความแตกต่างที่น้อยกว่านี้ไม่สามารถใช้ IRT-LR ตรวจค้นได้ว่าลำเอียงหรือไม่ แต่อาจทำได้ถ้าจำนวนตัวอย่างมากกว่านี้

การระบุความลำเอียงของข้อสอบ ใช้วิธีเปรียบเทียบความแตกต่างระหว่างกลุ่มในข้อสอบที่ต้องการตรวจค้นความลำเอียงกับข้อสอบร่วม ซึ่งมีข้อตกลงว่าไม่มีความลำเอียงอย่างไรก็ตามควรได้ศึกษาเพื่อหาวิธีเลือกข้อสอบร่วม และการทดสอบความลำเอียงของแบบสอบทั้งฉบับ

สำหรับผลการตรวจค้นความลำเอียงของข้อสอบด้วยวิธี IRT-LR และ MH ให้ผลคล้ายคลึงกัน ข้อเสนอแนะ คือ MH อาจเป็นประโยชน์ในการประเมินแบบสอบยาว ๆ และอาจใช้เป็นเครื่องมือกลั่นกรองก่อนจะใช้ IRT-LR ต่อไป

Oshima (1989) ได้ศึกษาผลของข้อมูลหลายมิติที่มีต่อการตรวจค้นความลำเอียง โดยจำลองข้อมูล 42 ชุดจากข้อสอบ 40 ข้อ มีโครงสร้างความเป็นพหุมิติ 2 ชนิด คือ ชนิดที่มีลักษณะแฝงที่เด่นอยู่ลักษณะเดียว และที่มีลักษณะแฝงเด่น 2 ลักษณะ เกณฑ์ที่ใช้ในการตัดสินได้จากการคำนวณค่าดัชนี SA UA SSOS (ผลบวกกำลังสองที่คิดเครื่องหมาย) USOS (ผลบวกกำลังสองที่ไม่คิดเครื่องหมาย) เกณฑ์การตัดสินความลำเอียงคือ ค่าที่มากกว่า 2 ส่วนเบี่ยงเบนมาตรฐานของ SA UA SSOS และ USOS ที่คำนวณได้จากกลุ่มอ้างอิง 2 กลุ่ม ผลจากการศึกษาครั้งนี้พบความสอดคล้องกับผลการศึกษาของ Oshima (1989) ที่พบว่าค่าที่ได้จากวิธีนี้ไม่คงที่ ขึ้นอยู่กับกลุ่มผู้สอบ ค่าที่นำมาใช้ในครั้งนี้จึงใช้ค่าเฉลี่ยจากการทำซ้ำ 5 ครั้ง

อย่างไรก็ตาม เมื่อจำนวนข้อสอบที่มีลักษณะพหุมิติ (ข้อสอบที่วัดลักษณะแฝงเด่นมากกว่า 1 ลักษณะ) ในแบบสอบมีจำนวนไม่เท่ากัน โดยเมื่อมีจำนวนข้อสอบที่มีลักษณะพหุมิติอยู่ในแบบสอบร้อยละ 5 10 และ 15 ข้อ ดัชนี SA ที่ใช้เป็นเกณฑ์จะมีค่าเท่ากับ +.173 +.205 +.201 ดัชนี UA เท่ากับ .262 .277 .293 ดัชนี SSOS เท่ากับ +1.13 +1.23 +1.55 และดัชนี USOS เท่ากับ 1.56 1.60 และ 2.03 ตามลำดับ

ผลการศึกษาพบว่า เมื่อผู้สอบ 2 กลุ่ม มีค่าเฉลี่ยของความสามารถแทรกซ้อน (Nuisance ability - $\theta_2$) เท่ากัน พบว่ามีการระบุว่าข้อสอบลำเอียงทั้ง ๆ ที่จำลองขึ้นให้ไม่ลำเอียง ทั้งนี้พบทั้งในข้อสอบลักษณะเอกมิติ และพหุมิติ และเมื่อค่าเฉลี่ยของ $\theta_2$ ต่างกัน จะสามารถระบุข้อสอบที่ลำเอียงได้ถูกต้องในอัตราสูง

การค้นพบข้อสอบลำเอียงร้อยละ 100 สำหรับทุกดัชนี เกิดขึ้นเมื่อค่าเฉลี่ยของ $\theta_2$ ของกลุ่มผู้สอบทั้ง 2 กลุ่มมีค่าเท่ากัน และมีจำนวนข้อสอบพหุมิติ อยู่ร้อยละ 5 ข้อ (มีความเป็นเอกมิติมากกว่า) อัตราการค้นพบสูงเพราะค่าเฉลี่ยของดัชนีมีค่าสูง ทั้งสำหรับข้อสอบพหุมิติและข้อสอบที่จำลองให้ลำเอียง อย่างไรก็ตามเมื่อจำนวนข้อสอบลำเอียง ในแบบสอบเพิ่มขึ้น สัดส่วนของข้อสอบลำเอียงที่ค้นพบลดลง

โดยสรุปแล้วดัชนีจาก IRT สามารถตรวจสอบข้อสอบพหุมิติที่สร้างให้ลำเอียง และที่สร้างให้ไม่ลำเอียงได้อย่างถูกต้อง แต่อำนาจในการตรวจสอบลดลง ถ้าจำนวนข้อสอบลำเอียง ในแบบสอบมีมากขึ้น

Oshima เสนอว่า เนื่องจากพบว่าเมื่อใส่ข้อสอบพหุมิติเข้าไว้ในแบบสอบร้อยละ 20 ทำให้การทดสอบความเป็นเอกมิติถูกปฏิเสธมากกว่าครึ่งหนึ่งของการทดสอบ ดังนั้นจึงควรทดสอบ ความเป็นเอกมิติของแบบสอบทุกครั้ง เมื่อจะวิเคราะห์ด้วย IRT แบบสอบที่ไม่เป็นเอกมิติ ไม่ควรใช้ IRT แบบเอกมิติ แต่อาจพิจารณาใช้ IRT แบบพหุมิติแทน

Swaminathan และ Rogers (1990) ได้ตรวจค้นความลำเอียงของข้อสอบด้วยวิธีการถดถอยแบบโลจิส และวิธี MH โดยมีขนาดของตัวอย่าง 2 ขนาด ได้แก่ 250 และ 500 คน และความยาวของแบบสอบ 3 ขนาด ได้แก่ 40 60 และ 80 ข้อ จำลองข้อมูลการตอบข้อสอบด้วยโปรแกรม DATAGEN แบบ 3 พารามิเตอร์ ข้อมูลการจำลองความลำเอียงใช้ 2 พารามิเตอร์โดยให้อำนาจจำแนกของ 2 กลุ่มเท่ากัน ส่วนค่าความยากแปรผันไปเพื่อให้เกิดระดับความลำเอียงต่าง ๆ ตามต้องการ แต่ในลักษณะการลำเอียงแบบไม่สม่ำเสมอ (Nonuniform)นั้นให้ค่าความยากของ 2 กลุ่มเท่ากัน แต่ค่าอำนาจจำแนกแปรไปผลการวิเคราะห์พบว่าการตรวจค้นความลำเอียงของข้อสอบ 2 วิธี โดยใช้การทดสอบ $\chi^2$ มีขั้นความเป็นอิสระ(df) เท่ากับ 2 สำหรับวิธีทดสอบโลจิสและการทดสอบ MH-$\chi^2$ ให้ผลใกล้เคียงกัน โดย MH ดีกว่าเล็กน้อย คือ มีการตรวจค้นได้ถูกต้องร้อยละ 75 กรณีใช้ตัวอย่าง 250 คน และตรวจค้นได้ร้อยละ 100 ในตัวอย่าง 500 คน ในทุกความยาวแบบสอบและในกรณีความลำเอียงแบบสม่ำเสมอ (Uniform) ในกรณีของการลำเอียงแบบไม่สม่ำเสมอนั้น MH ไม่สามารถตรวจค้นได้ ส่วนการถดถอยแบบโลจิสตรวจค้นได้ถูกต้องประมาณร้อยละ 50 ในกรณีตัวอย่างน้อยและข้อสอบสั้น และถูกต้องประมาณร้อยละ 75 ในแบบสอบยาวและตัวอย่างขนาดใหญ่แต่กรณีมีความลำเอียงแบบไม่สม่ำเสมอ โดยที่โค้งตัดกันที่ปลายสุดของโค้งไม่ว่าจะเป็นด้านความสามารถสูงหรือต่ำก็ตามMH จะดีกว่า

ในการตรวจค้นผิดคือระบุว่าลำเอียงทั้ง ๆ ที่ไม่ลำเอียงนั้น MH ดีกว่า โดยตรวจค้นผิดประมาณร้อยละ 1 ส่วนการถดถอยแบบโลจิสผิดประมาณร้อยละ 1-6

วิธีการถดถอยแบบโลจิสแม้จะยืดหยุ่นกว่า และมีความเป็นทั่วไปมากกว่า แต่ค่าใช้จ่ายมากกว่าประมาณ 3-4 เท่า

ประโยชน์ของการถดถอยแบบโลจิส อยู่ตรงที่เป็นวิธีการที่มีพื้นฐานเป็นแบบจำลองที่สามารถเพิ่มตัวแปรความสามารถเข้าไปในแบบจำลองได้ เช่นเดียวกับองค์ประกอบอื่น ๆ ที่มีความเกี่ยวข้องและสำคัญซึ่งทำให้เข้าใจในธรรมชาติของความลำเอียงได้ดีกว่า

Clauser และคณะ (1991) ได้ศึกษาถึงอิทธิพลต่าง ๆ ที่มีผลต่อวิธีการตรวจค้นความลำเอียงของข้อสอบด้วย MH ข้อมูลที่ใช้เป็นข้อมูลจำลองขึ้นเพื่อใช้ในการศึกษาการตรวจค้นความลำเอียงแบบสม่ำเสมอ ข้อมูลที่จำลองขึ้นเป็นแบบสอบจำนวน 75 ข้อ มีความลำเอียงอยู่ 16 ข้อ มีระดับของความแตกต่างในค่าความยากหลายระดับระหว่างกลุ่มอ้างอิงและกลุ่มเป้าหมายซึ่งมีจำนวนกลุ่มละ 1,500 คน มีการวิเคราะห์ระดับความยากที่ต่างกันหลายระดับกับระดับความยากของแบบสอบ 5 ระดับ อำนาจจำแนก 4 ระดับ และการกระจายของความสามารถ 2 ลักษณะ ผลการวิเคราะห์แสดงอิทธิพลจากความแตกต่างในระดับความยากรวมทั้งค่าอำนาจจำแนกและค่าความยาก ซึ่งทำให้ได้ข้อเสนอแนะว่าค่าสถิติ Mantel Haenszel

นี้ใช้ได้ดี สำหรับข้อกระทงที่มีค่าอำนาจจำแนกสูงและมีแนวโน้มจะไม่สามารถตรวจค้นความลำเอียงในข้อสอบที่มีค่าความยากสูง

Mazor และคณะ (1991) ได้ใช้วิธีการตรวจค้นความลำเอียงแบบ MH ศึกษาข้อมูลที่จำลองขึ้นโดยใช้กลุ่มตัวอย่างขนาด 2,000 1,000 500 200 และ 100 คน เพื่อเปรียบเทียบอัตราการตรวจค้นโดยใช้ข้อสอบที่จำลองขึ้น 5 ชุด ชุดละ 75 ข้อ ตรวจสอบความลำเอียงด้วยการทดสอบค่า MH-$\chi^2$ df = 1 ผลการศึกษาพบว่าเมื่อใช้จำนวนผู้สอบ 2,000 คน วิธีนี้ตรวจค้นความลำเอียงได้ผิดพลาดร้อยละ 30 และเมื่อใช้จำนวนผู้สอบ 500 คนลงไป พบว่าตรวจค้นความลำเอียงผิดพลาดร้อยละ 50 ข้อ กระทงที่ไม่สามารถตรวจค้นความลำเอียงได้ได้แก่ข้อกระทงที่ยากมากข้อกระทงที่มีค่าความยากต่างกันเล็กน้อยสำหรับคน 2 กลุ่ม และเป็นข้อกระทงที่มีค่าอำนาจจำแนกต่ำ

Park และ Lautenschlager (1990) ได้ตรวจค้นความลำเอียงของข้อสอบโดยใช้ทฤษฎีการตอบข้อสอบที่ได้ดัดแปลงจากวิธีของ Lord (1980) ที่เรียกว่า วิธี 'ปลอดจากลำเอียง' (Purification) ซึ่งมีวิธีการดังนี้

1) วิเคราะห์แบบสอบทั้งฉบับ โดยการประมาณค่าพารามิเตอร์รวมแล้วแปลงค่าความยาก (b) ให้เป็นค่ามาตรฐาน จากนั้นกำหนดค่าการเดา ($c_i$) ตามที่ได้จากการวิเคราะห์แล้วประมาณค่าความยากและอำนาจจำแนก ($a_i$) ใหม่อีกครั้ง โดยแยกตามกลุ่มที่ศึกษาแล้วแปลงค่า $b_i$ เป็นค่ามาตรฐานอีกครั้ง จากนั้นเปรียบเทียบฟังชั่นการตอบข้อสอบเพื่อดูความลำเอียง

2) ประมาณค่าระดับความสามารถโดยรวมทุกกลุ่มเข้าด้วยกันใหม่ใช้แต่เฉพาะข้อสอบที่ไม่ลำเอียง จะได้ข้อสอบที่มีความเป็นเอกพันธ์

3) กำหนดค่า $\theta$ เท่ากับที่ประมาณได้ในข้อ 2 แล้วประมาณค่า $a_i$ และ $b_i$ ของข้อสอบทุกข้อรวมข้อที่มีความลำเอียงไว้ด้วย

4) เปรียบเทียบฟังชั่นการตอบข้อสอบ หรือค่าพารามิเตอร์อีกครั้งด้วยค่าไคสแควร์

วิธีนี้เรียกว่าปลอดจากลำเอียง เพราะพยายามที่จะประมาณค่าพารามิเตอร์โดยตัดข้อสอบลำเอียงออกจากแบบสอบก่อน

วิธีที่นำมาใช้ในการศึกษาครั้งนี้เป็นวิธีที่ปรับปรุงจากวิธีของลอร์ดได้แก่

1. วิธีดัดแปลงจาก Drasgow ที่ได้เสนอไว้ เป็นการตรวจค้นความลำเอียงโดยการทำซ้ำ ๆ หลายรอบ จนกว่าจะได้ชุดข้อสอบที่ลำเอียงที่ซ้ำกันในการทำซ้ำ 2 รอบสุดท้าย วิธีนี้ใช้ค่าไคสแควร์ต่ำสุดของ Divgi (1985) แทนวิธีของ Stocking และ Lord (1983)วิธีการนี้เรียกว่า M-DIL ( Modified-Drasgow's Iterative Linking Procedure) มีวิธีดังนี้

1) ใช้วิธีการเชื่อมโยงมาตรของ สต๊อคกิ้งและลอร์ด ประมาณค่าพารามิเตอร์บนมาตรที่เท่ากันโดยประมาณ

2) คำนวณค่าความลำเอียง

3) เชื่อมโยงค่าพารามิเตอร์ที่ประมาณได้อีกครั้ง โดยใช้แค่ข้อที่ไม่ลำเอียง

4) คำนวณค่าความลำเอียงของทุกข้ออีกครั้ง

5) ทำซ้ำ ๆ จนกว่าจะได้ข้อสอบลำเอียงชุดเดียวกันใน 2 รอบสุดท้าย

2. วิธีดัดแปลงจากวิธี 'ปลอดลำเอียง' ของ Lord (1980) ซึ่ง Park (1988) ได้เสนอขึ้นเรียกว่า M-LTP (Modified-Lord Test Purification) ซึ่งมีขั้นตอนดังนี้

1) คำนวณค่า $\theta$ จากทุกกลุ่มรวมกัน

2) จากค่า $\theta$ ของแต่ละคนที่คำนวณได้ใน 1) นำมาคำนวณค่าพารามิเตอร์ของข้อสอบ แล้วคำนวณค่าสถิติความลำเอียง

3) คำนวณค่า $\theta$ ใหม่อีกครั้ง โดยใช้เฉพาะข้อสอบที่ไม่ลำเอียง

4) คำนวณค่าพารามิเตอร์ของข้อสอบอีกครั้งโดยใช้ค่า $\theta$ ที่คำนวณได้จาก 3)

5) คำนวณค่าสถิติความลำเอียง

6) ทำซ้ำจนกว่าจะได้ข้อสอบซ้ำกัน ในการทำซ้ำ 2 รอบสุดท้าย

3. ใช้วิธีผสมระหว่าง M-DIL และวิธีปลอดลำเอียงเรียกว่า ILAP (Iterative Parameter Linking and $\theta$ Scale Purification) ซึ่งจะเริ่มด้วย M-DIL จนถึงรอบสุดท้ายแล้วเริ่มประมาณค่า $\theta$ ใหม่อีกครั้งแยกตามกลุ่มเฉพาะข้อสอบไม่ลำเอียง จากนั้น จึงคำนวณค่าพารามิเตอร์ข้อสอบแยกตามกลุ่ม แล้วใช้วิธีการของ Divgi ประมาณค่าคงที่ของการเชื่อมโยง โดยใช้เฉพาะข้อที่ไม่ลำเอียง แล้วคำนวณค่าสถิติความลำเอียงใหม่ แล้วตามด้วย M-DIL ต่อ

ข้อมูลที่ศึกษา เป็นข้อมูลจำลองด้วยวิธีการจำลองของ Sympson (1978) ภายใต้สภาพการณ์ลำเอียงชนิดทิศทางเดียว และทิศทางผสม พบว่าการวิเคราะห์ 2 วิธีแรก ให้ผลเท่ากัน แต่ M-DIL ต้องทำซ้ำในจำนวนรอบที่มากกว่า M-LTP และถ้าใช้วิธีสุดท้ายแล้วจะสามารถลดจำนวนข้อสอบที่ลำเอียง แต่ถูกระบุจากการตรวจค้นว่าไม่ ลำเอียงลงได้มากกว่า2 วิธีแรก

Cohen และ Kim (1993) ศึกษาเปรียบเทียบการใช้ทฤษฎีการตอบข้อสอบ 2 พารามิเตอร์ โดยใช้สถิติทดสอบ $\chi^2$ ของลอร์ดและสถิติทดสอบพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบชนิดคิดเครื่องหมายและไม่คิดเครื่องหมาย อันได้แก่ สถิติ Z ของราชู โดยใช้ข้อมูลจำลองตามทฤษฎีการตอบข้อสอบ 2 พารามิเตอร์ จำลองการตอบข้อสอบ 2 ขนาด ข้อ 20 และ 60 ข้อ จำนวนผู้สอบ 100 และ 500 คน และมีจำนวนข้อสอบลำเอียงในแบบทดสอบจำนวนร้อยละ 0 10 และ 20

ผลการวิเคราะห์พบว่า

1. แบบสอบยิ่งยาวการระบุข้อสอบลำเอียงผิดพลาด (FP) จะมากขึ้น และถ้ามีจำนวนข้อสอบลำเอียงอยู่ในแบบสอบน้อย จำนวนข้อสอบที่ถูกระบุว่าลำเอียงผิดพลาดจะยิ่งมากขึ้น

องค์ประกอบที่มีอิทธิพลต่อการระบุว่าข้อสอบลำเอียงผิดพลาด (FP) อีกประการ คือ ระดับนัยสำคัญ ($\alpha$) ค่า $\alpha$ ยิ่งมาก FP จะเกิดมาก (ที่ .05 จะเกิด FP มากกว่าที่ .01) ผลดังนี้เกิดในทุกสถานการณ์ที่ศึกษา สำหรับ Z แต่เกิดในบางสถานการณ์สำหรับ $\chi^2$

2. แบบสอบที่ยาวกว่า มีร้อยละของข้อสอบลำเอียงมากกว่า จะมีข้อสอบที่ถูกระบุว่าไม่ลำเอียงผิดพลาด (FN) มากกว่า แต่จะเกิดขึ้นกับการทดสอบด้วย Z มากกว่า

กรณีที่จำนวนผู้สอบเท่ากับ 100 คน พบว่าความแตกต่างในการระบุว่าข้อสอบไม่ลำเอียงผิดพลาด สำหรับ Z และ $\chi^2$ ไม่แตกต่างกันมากเท่ากรณีใช้ตัวอย่าง 500 คนซึ่งมีการระบุว่าไม่ลำเอียงผิดพลาด สำหรับ $\chi^2$ น้อยกว่า Z โดยเฉพาะอย่างยิ่งในกรณีที่มีข้อสอบลำเอียงในแบบสอบมากที่สุดคือ ร้อยละ 20

Cohen และ Kim วิจารณ์ว่า การที่ค่าสถิติทดสอบ $\chi^2$ ให้ผลการตรวจสอบข้อสอบลำเอียงดีกว่า Z ของพื้นที่ที่คิดเครื่องหมายอาจเป็นเพราะ Z (ESA) ใช้การเปรียบเทียบเฉพาะค่าความยาก (b) ในขณะที่ $\chi^2$ ใช้ทั้งค่าความยากและอำนาจจำแนก (b และ a) แต่ไม่สามารถอธิบายความแตกต่างได้ในกรณีใช้ Z(H) ซึ่งทดสอบพื้นที่ที่ไม่คิดเครื่องหมายเพราะ Z(H) ก็ใช้การทดสอบค่าพารามิเตอร์ทั้ง 2 ตัวเช่นกัน

5.2.2 การตรวจค้นข้อสอบลำเอียงจากข้อมูลจริง

การตรวจค้นข้อสอบลำเอียงจากข้อมูลจริงเท่าที่ผ่านมาใช้ข้อมูลทั้งที่ได้จากแบบสอบมาตรฐานและแบบสอบที่ผู้วิจัยแต่ละคนพัฒนาขึ้นเองดังนี้

Shoener (1984) ได้เปรียบเทียบการตรวจค้นความลำเอียงของข้อสอบด้วยวิธีการทางสถิติ ได้แก่ IRT 3 พารามิเตอร์ ซึ่งใช้การทดสอบ $\chi^2$ และการตัดสินด้วยผู้ตัดสินจำนวน 24 คน โดยใช้แบบสอบอิงเกณฑ์กับนักเรียนชั้นมัธยมศึกษาชาย และหญิงจำนวน 1,064 คน ที่มีภูมิหลังทางวัฒนธรรมต่าง ๆ กัน การตัดสินใช้ผู้ตัดสินที่มีความรู้ทางด้านคณิตศาสตร์และหลักสูตรคณิตศาสตร์เป็นอย่างดีและใช้แบบฟอร์มการให้คะแนนเป็นช่วง (rating form) ผู้ตัดสินเป็นผิวดำ 8 คน ผิวขาว 8 คน และเชื้อสายสเปน 8 คน เป็นชาย และหญิงเท่า ๆ กัน

การวิเคราะห์ได้พิจารณา ถึงความสอดคล้องของการตัดสินของกลุ่มผู้ตัดสิน 3 กลุ่ม และความสอดคล้องระหว่างวิธีการทางสถิติและการตัดสินของผู้ตัดสิน โดยใช้ค่าสถิติของ Kappa ผลจากการวิเคราะห์พบว่าความสอดคล้องระหว่างการตัดสินโดยผู้เชี่ยวชาญต่างกลุ่มไม่มีนัยสำคัญสำหรับทั้งความลำเอียงด้านวัฒนธรรมและทางเพศ พบความสอดคล้องกันอย่างมีนัยสำคัญในตัวบ่งชี้บางตัวที่คำนวณขึ้นเพื่อรวมการให้คะแนนของผู้ตัดสิน แต่ไม่พบความสอดคล้องระหว่างวิธีการตรวจค้นความลำเอียงทั้ง 2 วิธี การตัดข้อกระทงที่พบทางสถิติว่าลำเอียงออกไปไม่ทำให้ลำดับของคนในกลุ่มย่อยและรวมทั้งหมดเปลี่ยนไป และทั้งแบบสอบฉบับเดิมและแบบสอบที่ตัดข้อกระทงที่ลำเอียงออกแล้วมีความสัมพันธ์อย่างมีนัยสำคัญกับแบบสอบวิธีคณิตศาสตร์อิงกลุ่มที่เป็นแบบสอบมาตรฐาน

Subkoviak Mack Ironson และ Craig (1984) ได้เปรียบเทียบวิธีการตรวจค้นความลำเอียงของข้อสอบ 3 วิธีคือ (1) ICC ชนิด 3 พารามิเตอร์ (2) ค่าความยากแปลง (TID) และ (3) ไคสแควร์ซึ่งมี 2 วิธี ได้แก่ วิธีที่เสนอโดย Scheuneman (1979) พิจารณาเฉพาะตัวเลือกถูก (CHIS) และที่เสนอโดย Camilli (1979) พิจารณาทั้งตัวเลือกถูกและผิด (CHIC)

วิธีการแปลงค่าความยาก TID ทำโดยคำนวณค่าความยากของข้อสอบตามกลุ่มผิวดำและผิวขาว จากนั้นแปลงค่าความยาก (p) เป็นค่า z ซึ่งมีค่าเท่ากับเปอร์เซ็นไทล์ที่ 1-p ในการกระจายแบบปกติ แล้วแปลงค่า z เป็นค่าเดลต้าโดยที่ = 4Z + 13 ค่าระยะทางระหว่างจุดจบของค่า z ทั้งสองกลุ่ม และเส้นแกนหลักเป็นค่าแสดงความลำเอียงและจะใช้เครื่องหมายบวก ถ้าเป็นการลำเอียงเข้าข้างผิวดำ และใช้ค่าลบถ้าเข้าข้างผิวขาว

วิธี CHIS จะแบ่งคะแนนรวมเป็น 5 ช่วง คำนวณความถี่คาดหวัง (E) ในการตอบถูกในแต่ละช่วงตามกลุ่ม แล้วเปรียบเทียบความถี่คาดหวัง กับความถี่จากการสังเกต (O) ด้วยค่าสถิติไคสแควร์ รวมทุกช่วงคะแนนค่าที่ได้จะเป็นการวัดความลำเอียงโดยไม่มี

เครื่องหมาย การวัดโดยคิดเครื่องหมายหรือมีทิศทางทำโดย ให้เครื่องหมายบวกกรณีลำเอียงเข้าข้างกลุ่มผิวดำ และลบเมื่อเข้าข้างกลุ่มผิวขาว โดยให้เครื่องหมายในแต่ละช่วงคะแนน

วิธี CHIC ก็เช่นเดียวกับ CHIS แต่จะวิเคราะห์ตัวเลือกผิดด้วย เครื่องมือที่ใช้เป็นแบบสอบเลือกคำเหมือนแบบเลือกตอบ 4 ตัวเลือกจำนวน 50 ข้อมีอยู่ 10 ข้อที่เป็นข้อสอบวัดค่าสแลงสำหรับผิวดำอีก 40 ข้อ เป็นคำศัพท์มาตรฐาน ให้รหัส 10 ข้อแรก เป็น 1 คือข้อที่ลำเอียงอีก 40 ข้อให้รหัส 0 เป็นข้อไม่ลำเอียงผลการวิเคราะห์มีดังนี้

ค่าสหสัมพันธ์ระหว่างผลการตรวจค้นของแต่ละวิธีการ กับข้อสอบที่สร้างขึ้นที่ให้รหัสว่าลำเอียงหรือไม่ลำเอียง พบว่าค่าสหสัมพันธ์ระหว่าง CHIS และ CHIC มีค่าสูงสุด (ชนิดไม่คิดเครื่องหมาย = .978 ชนิดคิดเครื่องหมาย = .970) รองลงมาคือค่าสหสัมพันธ์ระหว่าง TID และ CHIC ถ้ามีจำนวนตัวอย่างน้อย เช่น 300 คน หรือน้อยกว่า แทนที่จะใช้ IRT ควรเลือกค่า z และที่ให้ค่าถูกต้องใกล้เคียงกับ z คือ ไคสแควร์ส่วนค่าเดลต้านั้น ในกรณีสุดโต่งจะมีการชี้ถึงความลำเอียงตรงข้ามกับความเป็นจริง ดัชนีเดลต้าที่เหลือ (residaul delta index) ซึ่งได้ลองนำมาศึกษานั้นพบว่าใช้ได้ดีเกือบเท่า ๆ ไคสแควร์ ซึ่งควรมีการแก้ไขต่อไปเพื่อให้นำมาใช้แทน ไคสแควร์ได้

Shepard Camilli และ Williams (1985) ได้ศึกษาความตรงของเทคนิคการตรวจค้นความลำเอียงด้วยวิธีไคสแควร์ ดัชนีเดลต้าของแองกอฟ และทฤษฎีการตอบข้อสอบแบบเทียม โดยใช้ทั้งข้อมูลจริงและข้อมูลจำลองพบว่าผลจากการศึกษาด้วยข้อมูล 2 ชนิด มีความสอดคล้องกัน คือ พบว่าทฤษฎีการตอบข้อสอบแบบเทียมเป็นวิธีดีที่สุด โดยวิเคราะห์จากทั้งค่าสัมประสิทธิ์สหสัมพันธ์ และร้อยละของความสอดคล้อง ไคสแควร์ให้ความถูกต้องใกล้เคียงกัน แต่เดลต้าใช้ได้ไม่พอเพียง และในกรณีที่ลำเอียงมาก ๆ ยังระบุได้ตรงข้ามกับความเป็นจริง เช่นระบุว่าเข้าข้างกลุ่มผิวดำ ทั้ง ๆ ที่ที่จริงเป็นข้อลำเอียงเข้าข้างผิวขาว แต่อย่างไรก็ตามดัชนีดัดแปลงของ Angoff (Modified Angoff Index) ซึ่งนำค่าความแตกต่างของค่าความยาก มาถดถอยบนค่าสหสัมพันธ์แบบพอยท์ไบซีเรียลของข้อกระทงและใช้ค่าที่เหลือเป็นดัชนีให้การตรวจค้นดีเกือบเท่ากับไคสแควร์ แต่ก็ควรระวังถ้าจะใช้เพื่อการตัดข้อกระทงทิ้ง

Hambleton และคณะ (1986) ได้เปรียบเทียบวิธีการตรวจค้นความลำเอียงของข้อสอบ 4 วิธี ได้แก่ วิธีของ Mantel-Haenszel วิธีการลงจุดค่าความยาก วิธีค่าความแตกต่างของค่าเฉลี่ยกำลังสอง (the root meansquared difference) และวิธีของพื้นที่รวม ซึ่ง 2 วิธีหลังนี้เป็นวิธีการที่ใช้ทฤษฎีการตอบข้อสอบ

แบบสอบที่ใช้ในการศึกษา เป็นแบบสอบความสามารถในการอ่านของคลีฟแลนด์มีข้อสอบจำนวน 75 ข้อ โดยใช้กลุ่มตัวอย่างชาย 451 คน หญิง 486 คน คะแนนจุดตัดในการแปลความหมายค่าสถิติที่แสดงความลำเอียงได้จากการจำลองข้อมูลขึ้น ผลปรากฏว่าวิธีทั้ง 4 วิธี ให้ผลการตรวจค้นข้อสอบลำเอียงใกล้เคียงกัน ปัญหาด้านวิธีวิทยาอยู่ที่ความไม่แจ่มชัดใน

การกำหนดจุดตัด ความคลาดเคลื่อนชนิดที่หนึ่ง และ การประมาณค่าพารามิเตอร์ที่ไม่ดี ข้อค้นพบเด่น ๆ อยู่ที่ความสำคัญของการเลือกใช้ช่วงคะแนนความสามารถในการวัดความลำเอียง และพบว่า วิธี MH เป็นวิธีเลือกใช้แทนวิธีทฤษฎีการตอบข้อสอบได้อย่างประหยัดกว่าทั้งเงินและเวลา

Doolitte และ Cleary (1987) ได้ตรวจค้นความลำเอียงระหว่างเพศของข้อสอบในแบบวัดผลสัมฤทธิ์ทางคณิตศาสตร์โดยใช้ตัวอย่างจำนวน 8 กลุ่ม ที่ทำแบบสอบ ACT Assessment Mathematic Test รวม 8 ฟอร์ม แต่ละกลุ่มมีความเท่าเทียมกันและมีจำนวนประมาณ 1,300-1,400 คน ตัวอย่างเป็นหญิงประมาณร้อยละ 55 วิเคราะห์ด้วยดัชนีของ ลินน์ และ ฮาร์นิสช์ (Linn and Harnisch, 1981) ซึ่งเป็นโมเดลโลจิสติคแบบ 3 พารามิเตอร์ ถ้าดัชนี Z ที่ได้เป็นบวกแสดงว่าง่าย สำหรับกลุ่มเปรียบเทียบถ้าเป็นลบแสดงว่ายาก สำหรับกลุ่มเปรียบเทียบ ผลการวิเคราะห์พบว่าดัชนี Z มีค่าเป็นลบ สำหรับข้อกระทงที่วัดด้านเรขาคณิตและการให้เหตุผลเชิงพีชคณิต และเลขคณิตในทุกฟอร์มแสดงว่าง่ายสำหรับชายมากกว่าหญิงที่เหลือส่วนมากเป็นบวก

ผลจากการวิเคราะห์ด้วยการวิเคราะห์ความแปรปรวน พบผลเช่นเดียวกับการคำนวณค่า Z

Perlman และคณะ (1988) ได้ศึกษาความคงที่ของวิธีการประมาณค่าความลำเอียงของข้อสอบ 4 วิธี ได้แก่ วิธีค่าความยากแปลงวิธีค่าความยากแปลงที่นำมาดัดแปลงโดย Shepard วิธีการวิเคราะห์ค่าที่เหลือแบบ 1 พารามิเตอร์ของ Rasch และวิธี MH

กลุ่มตัวอย่างที่ใช้มี 30 กลุ่ม โดยมีขนาด 600 จนถึง 2,000 คน ซึ่งสุ่มมาจากนักเรียนชั้น 9 จำนวน 54,896 คน ซึ่งมีทั้งผิวขาว ผิวดำ และชนชาติที่พูดภาษาสเปนที่ทำแบบทดสอบทักษะความสามารถขั้นต่ำของรัฐชิคาโก ซึ่งประกอบด้วยข้อสอบเลือกตอบ 46 ข้อ จากธนาคารข้อสอบจำนวน 1,000 ข้อ ผลจากการศึกษาพบว่าความเที่ยงของตัวบ่งชี้ความลำเอียงจะมีปัญหา ถ้าใช้จำนวนตัวอย่างน้อยกว่า 666 คน ไม่มีวิธีใดที่ให้ผลการตรวจค้นความลำเอียงได้มากกว่า หรือน้อยกว่าวิธีอื่น ๆ อย่างคงที่ข้ามขนาดตัวอย่าง

Baeza (1989) ได้ศึกษาพฤติกรรมการตอบข้อสอบของชาวอเมริกันอินเดียนและอเมริกันคอเคเซี่ยน เพื่อพิจารณาระดับของความลำเอียงทั้งภายในและภายนอกข้อสอบโดยใช้ตัวแปรด้านเศรษฐกิจ สังคม และระดับคะแนนเฉลี่ยในชั้นมัธยมปลายเป็นตัวแปรจับคู่กลุ่มตัวอย่างนอกเหนือไปจากเพศ โรงเรียนที่เรียนอยู่ และปีการศึกษาที่เรียน

วิธีการที่ใช้ในการตรวจค้นความลำเอียง ได้แก่ วิธี MH และวิธีการจับคู่กลุ่มตัวอย่างของ McNemar ตัวแปรจับคู่สำหรับ MH ได้แก่ คะแนนรวมจากแบบสอบ ส่วนตัวแปรจับคู่สำหรับ McNemar ได้แก่ ตัวแปรด้านเศรษฐกิจสังคมหรือระดับคะแนนเฉลี่ย

ผลการวิเคราะห์พบว่าวิธี MH ซึ่งใช้เกณฑ์ภายในตรวจพบข้อสอบที่ลำเอียงเพียงเล็กน้อยส่วน McNemar ซึ่งใช้เกณฑ์ภายนอกนั้นพบว่าเมื่อใช้ตัวแปรจับคู่เป็นสถานภาพทางเศรษฐกิจสังคม พบว่าข้อสอบมีระดับความลำเอียงต่ำกว่าการจับคู่ด้วยระดับคะแนนเฉลี่ย แสดงว่าระดับคะแนนเฉลี่ยของคน 2 กลุ่มนี้มีระดับของความหมายไม่เท่ากัน

Baghi และFerrara (1989) ได้เปรียบเทียบผลการสืบค้นความลำเอียงด้วยวิธีการแตกต่างกัน 3 วิธี ได้แก่ การใช้ทฤษฎีการตอบข้อสอบ วิธีการใช้ค่าเดลต้า (ค่าความยากแปลง) และวิธี MH โดยศึกษาในแง่ของผลกระทบจากขนาดของกลุ่มตัวอย่างในผลการตรวจค้นความลำเอียง และระดับของความสัมพันธ์ระหว่างค่าสถิติจากทั้ง 3 วิธีนี้

ข้อมูลที่ใช้ในการวิเคราะห์ เป็นข้อมูลที่สุ่มจากนักเรียนระดับ 9 จำนวน 50,000 คน ที่ตอบแบบสอบทักษะความเป็นพลเมืองดี ประกอบด้วยข้อสอบแบบเลือกตอบ 45 ข้อ ประเมินความรู้ และ ทักษะ รวม 3 ด้าน ได้แก่ด้านรัฐบาลประชาธิปไตย ด้านการเมือง และ พฤติกรรมทางการเมือง และ ด้านหลักการ สิทธิ และความรับผิดชอบ จำนวนกลุ่มตัวอย่างที่สุ่มมาใช้มี 4 ขนาด ได้แก่ ขนาด 1,000 คน 750 คน 500 คน และ 200 คน

ผลจากการวิเคราะห์ไม่พบว่า มีข้อสอบข้อใดที่มีความลำเอียงทั้งในการเปรียบเทียบระหว่างผู้สอบผิวดำและผิวขาว และผู้สอบหญิง และชาย การแสดงค่าประมาณความยากด้วยแผนภูมิเส้นแสดงความสัมพันธ์ระหว่างค่าความยากต่างกลุ่มอยู่ในรูปเส้นตรงเกือบสมบูรณ์ ความสอดคล้องกันระหว่างวิธีการตรวจค้นความลำเอียงของ Rasch และค่าความยากแปลงมีความสัมพันธ์กันสูงมากทุกขนาดของกลุ่มตัวอย่างจากค่าสัมประสิทธิ์สหสัมพันธ์แบบลำดับ และทั้ง 2 วิธี มีความสอดคล้องสูงสุดกับวิธีการตอบข้อสอบ 3 พารามิเตอร์ ในการตรวจค้นความลำเอียงและไม่ลำเอียงของข้อสอบ

Hambleton และ Rogers (1989)ได้ศึกษาเปรียบเทียบวิธีการตรวจค้นข้อสอบลำเอียง 2 วิธี ได้แก่ MH และ IRT โดยใช้ข้อมูลการตอบแบบสอบ New Mexico High School Proficiency Exam (NMHSPE) จำนวน 150 ข้อ ของนักเรียนชาวอเมริกันผิวขาว 8,000 คน และชาวอเมริกันพื้นเมือง 2,600 คน จากผู้สอบทั้งหมด 23,000 คน

NMHSPE เป็นแบบสอบชนิดเลือกตอบ 4 ตัวเลือก วัดทักษะการดำเนินชีวิต 5 ด้าน คือ (1) ความรู้เกี่ยวกับแหล่งทรัพยากรของชุมชน (2) ความรู้ด้านผู้บริโภค (3) ความรู้เกี่ยวกับรัฐบาลและกฎหมาย (4) ความรู้ด้านสุขภาพทางกายและจิต และ (5) ความรู้ด้านอาชีพ

กลุ่มตัวอย่างที่นำมาศึกษา ซึ่งได้แก่ ชาวอเมริกันผิวขาวและอเมริกันพื้นเมืองนั้นได้สุ่มมาศึกษาอย่างละ 2,000 คน และแบ่งเป็น 2 กลุ่ม ๆ ละ 1,000 คน ตามลำดับคู่-คี่ การศึกษาความลำเอียง 2 คู่ เช่นนี้เพื่อพิจารณาถึงความคงที่ในการระบุข้อสอบลำเอียง

ในการศึกษาความลำเอียง ผู้วิจัยได้ใช้ข้อสอบเพียง 75 ข้อ โดยตัดข้อสอบที่ง่ายมาก ($p > .70$) และค่าอำนาจจำแนกต่ำ ($r < .10$) วิธี MH วิเคราะห์โดยใช้ 76 ระดับคะแนน ส่วนวิธีทฤษฎีการตอบข้อสอบใช้การวัดพื้นที่ระหว่างโค้งลักษณะข้อสอบ (ICC) พิสัยของความสามารถที่นำมาใช้ในการคำนวณพื้นที่กำหนดใช้ 2 ค่าเบี่ยงเบนมาตรฐานสูงกว่าความสามารถเฉลี่ยของชาวอเมริกันผิวขาว และ 2 ค่าเบี่ยงเบนมาตรฐานต่ำกว่าความสามารถเฉลี่ยของชาวอเมริกันพื้นเมือง ความคงที่ของการตรวจสอบความลำเอียง ได้แก่ ร้อยละของข้อสอบที่ถูกระบุว่าลำเอียงหรือไม่ลำเอียงซ้ำกันในการวิเคราะห์ 2 ครั้ง

เกณฑ์ตัดสินความลำเอียง สำหรับวิธี IRT - 3 พารามิเตอร์ได้จากการวิเคราะห์ดัชนี SA ระหว่างผู้สอบที่เป็นอเมริกันพื้นเมือง 2 กลุ่ม และใช้ค่าสูงสุดที่ได้เป็นเกณฑ์ซึ่งมีค่า = .468 และด้วยวิธีเดียวกันเกณฑ์ตัดสินความลำเอียง สำหรับวิธี MH คือ MH-$\chi^2$ = 6.64

ผลการวิเคราะห์พบว่า วิธี MH มีความสอดคล้องของการตรวจสอบร้อยละ 80 ส่วน ICC ร้อยละ 73 และพบว่า ร้อยละ 47 ของข้อสอบที่ถูกระบุว่าลำเอียงในการวิเคราะห์คู่ที่ 1 เป็นข้อสอบลำเอียงในการวิเคราะห์คู่ที่ 2 ด้วย สำหรับ MH และร้อยละ 61 สำหรับ ICC ในทางกลับกันพบว่า ร้อยละ 64 ของข้อสอบลำเอียงในการวิเคราะห์คู่ที่ 2 เป็นข้อสอบลำเอียงในการวิเคราะห์คู่ที่ 1 สำหรับ MH และร้อยละ 56 สำหรับ ICC ผลดังกล่าวแสดงให้เห็นว่าความคงที่ของการวิเคราะห์ทั้ง MH และ ICC อยู่ในระดับปานกลางซึ่งอาจเป็นเพราะจำนวนตัวอย่างค่อนข้างน้อย (MH มีข้อสอบที่ระบุว่าลำเอียงในการวิเคราะห์ทั้ง 2 ครั้ง 7 ข้อ ICC 14 ข้อ และพื้นที่แสดงความแตกต่างของ ICC มีค่าระหว่าง-2.7 และ 1.5)

สรุปได้ว่า ทั้ง MH และ ICC มีความไม่เที่ยงในระดับหนึ่งในการใช้ตรวจสอบข้อสอบลำเอียงโดยที่ MH มีความคงที่ประมาณร้อยละ 80 และ ICC ประมาณ 73

Reynold (1989) ได้ใช้วิธีการ 4 วิธีในการตรวจค้นความลำเอียงของข้อสอบ CTBS/S ฉบับภาษาสเปน โดยใช้ข้อมูลจากนักเรียน 895 และ 911 คน วิธีการที่ใช้ในการตรวจค้นมี 4 วิธี ได้แก่

1. ค่าความยากแปลงของแองกอฟฟ์

2. การทดสอบไคสแควร์

3. เปรียบเทียบความแตกต่างในค่าความยาก และการทดสอบความเหมาะสมด้วยโมเดลของ Rasch

4. ทฤษฎีการตอบข้อสอบ 3 พารามิเตอร์ (พิจารณาจากพื้นที่และความแตกต่างในค่าสถิติที่ใช้ทดสอบความเหมาะสม)

ตรวจสอบความตรงเชิงโครงสร้างในการวัดความลำเอียงของข้อสอบโดยการพิจารณาค่าสหสัมพันธ์ระหว่างวิธีต่าง ๆ ร้อยละของความสอดคล้องระหว่างวิธีต่าง ๆ ที่ใช้และการวิเคราะห์ตัวประกอบ

ผลจากการวิเคราะห์พบว่าเคลด้า และไคสแควร์ มีความสัมพันธ์เกือบจะสมบูรณ์วิธีที่ 1 และ 2 กับวิธีที่ 3 และ 4 มีความสัมพันธ์กันน้อย สนับสนุนโดยผลจากการวิเคราะห์ร้อยละของความสอดคล้อง การวิเคราะห์ตัวประกอบแสดงว่า เคลด้า ไคสแควร์ และ IRT แบบ 3 พารามิเตอร์ วัดโครงสร้าง (construct) เดียวกัน คือ ความลำเอียงด้านภาษา และความสัมพันธ์ระหว่างเคลด้าและไคสแควร์นั้นเห็นได้ชัดแม้แต่ในกรณีที่มีความลำเอียงน้อย

Baghi และ Ferrara (1990) ได้เปรียบเทียบผลจากการสืบค้นความลำเอียงของข้อสอบด้วยทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ (IRT) และสถิติไคสแควร์แบบ MH (MHCS) โดยใช้ข้อมูลจากการบริหารแบบสอบทักษะความเป็นพลเมืองดีของรัฐแมรี่แลนด์ ให้แก่นักเรียนระดับ 9 จำนวน 50,000 คน ในเดือนมกราคม-กุมภาพันธ์ 1988 จำนวนตัวอย่างที่ใช้ในการสืบค้นด้วย MHCS มี 4 ขนาด ได้แก่1,000 คน 750 คน 500 คน และ 200 คน ส่วนจำนวนตัวอย่างในการสืบค้นด้วย IRT ใช้กลุ่มละ 1,000 คน ในแต่ละกลุ่มเปรียบเทียบ การวิเคราะห์ประกอบด้วย (1) การพิจารณาความคงที่ของค่าสถิติ MHCS เมื่อใช้กลุ่มตัวอย่างขนาดต่าง ๆ (2) ความคงที่ของดัชนีความลำเอียง(MH Alpha)ในวิธี MHCS ข้ามช่วงคะแนน (3) ความสัมพันธ์ระหว่างดัชนีความลำเอียงแบบ IRT และค่า MH Alpha และ (4) ความสอดคล้องระหว่างวิธีทั้งหมดในการสืบค้นข้อกระทงที่ลำเอียง

ผลปรากฏว่าเมื่อใช้วิธี IRT สืบค้นความลำเอียงของข้อสอบระหว่างกลุ่มหญิง และชายพบข้อสอบลำเอียง 4 ข้อ และพบข้อสอบลำเอียง 3 ข้อในการเปรียบเทียบระหว่างผู้สอบผิวขาวและผิวดำ ซึ่งมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องไม่เท่ากันอย่างมีนัยสำคัญ และในขณะที่ผลจากการหาค่าสหสัมพันธ์ระหว่างดัชนีทั้งสองตัวแสดงความไม่สอดคล้องกันในการบ่งชี้ข้อสอบที่ลำเอียงนั้น ความสอดคล้องในผลรวมแสดงว่าเทคนิค MHCS เป็นวิธีที่ใช้แทน IRT แบบ 3 พารามิเตอร์ได้อย่างเพียงพอถ้าจำนวนตัวอย่างมีขนาดอย่างน้อย 750 คน

Sudweeks และ Tolman (1990) ได้เปรียบเทียบผลจากการใช้วิธี MH และการตัดสินด้วยผู้เชี่ยวชาญในการตรวจค้นความลำเอียงของข้อสอบวิชาวิทยาศาสตร์ระหว่างนักเรียนชายและหญิงระดับ 5 จำนวน 926 คน เครื่องมือที่เก็บรวบรวมข้อมูลเป็นแบบสอบอิงเกณฑ์ชนิดเลือกตอบที่สร้างขึ้นตามมาตรฐาน และวัตถุประสงค์ของหลักสูตรวิทยาศาสตร์ร่วมในรัฐฮาห์ ผลจากการวิเคราะห์พบว่า วิธีการทั้งสองให้ผลต่างกัน และข้อสอบหลายข้อที่ถูกระบุด้วยวิธีการ 2 วิธีนี้ว่า ลำเอียง จะเป็นข้อสอบที่ยาก

Cohen Kim และ Subkoviak (1991) ได้ศึกษาอิทธิพลของการทราบการกระจายในการตรวจค้นการทำหน้าที่เบี่ยงเบนของข้อสอบสำหรับกลุ่มผู้สอบผิวดำ และผิวขาวที่สร้างขึ้นอย่างมีเจตนาให้ลำเอียง โดยใช้โปรแกรม LOGIST ซึ่งใช้วิธี Joint Maximum Likelihood Estimation (JMLE) และโปรแกรม BILOG โดยที่โปรแกรม ไบล็อก ใช้การประมาณค่าพารามิเตอร์ 3 สถานการณ์ คือ

1. MMLE (Marginal Maximum Likelihood Estimation) เป็น BILOG ที่ไม่มีการกำหนดค่าพารามิเตอร์ ค่าอำนาจจำแนก (a) จะอยู่ในช่วง 0-$M_a$=1.13 a = .604 เป็น BILOG A

2. MMAPE(Marginal Maximum a Posteriori Estimation) กำหนดค่า a ในรูป lognormal ค่าเฉลี่ย = 0 (ไม่มี FLOAT option) เป็น BILOG B

3. MMAPE กำหนดค่า a ในรูป lognormal เช่นกัน แต่มีข้อตกลงว่าการกระจายของค่า (ค่าแปลงของ a) มีการกระจายเป็นปกติ มีค่าความเบี่ยงเบนมาตรฐาน = 0.5 (มี FLOAT option) เป็น BILOG C

ส่วนการวัดความแตกต่างของพื้นที่ระหว่างโค้งการตอบข้อสอบนั้น ทำ 2 ลักษณะ คือ อย่างคิดเครื่องหมาย(SA) และไม่คิดเครื่องหมาย(UA) การประมาณค่าพารามิเตอร์ใช้แบบ 2 พารามิเตอร์ โดยการกำหนดค่า c ให้เป็นค่าคงที่

พบจากค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันว่า การวัดพื้นที่ โดยใช้ BILOG B และ BILOG C มีความสัมพันธ์กันสูงมาก ทั้งในดัชนีแบบคิดเครื่องหมาย และไม่มีเครื่องหมาย ส่วนระหว่าง LOGIST และ BILOG ก็สูงเช่นกัน แต่ความคงที่ในรูปแบบของความคล้ายคลึงจะน้อยกว่า ส่วนความคล้ายคลึงกันระหว่าง LOGIST กับ BILOG B และ BILOG C ลดลงกว่าเล็กน้อย เป็นผลที่สอดคล้องกับการศึกษาที่คนอื่น ๆ เคยพบ ในการวัดพื้นที่ของความเบี่ยงเบนของข้อสอบ

BILOG B และ C จะดีกว่าวิธีอื่น ๆ และมีความไวในการตรวจค้นมากเมื่อค่า b อยู่ในช่วง +5 มากกว่าในช่วง +3 ยกเว้น BILOG B ที่เท่า ๆ กันทั้ง 2 ช่วง ส่วน MMLE จะดีกว่า JMLE มาก ในการตรวจค้นความเบี่ยงเบน ในกรณีที่ค่าสัมบูรณ์ของ b อยู่

ในช่วง +3 และ ดีกว่าเล็กน้อยในกรณีค่าสัมบูรณ์ของ b น้อยกว่า หรือเท่ากับ 5 ในการวัดแบบมีเครื่องหมาย จำนวนตัวอย่างที่ใช้ก็ไม่มีอิทธิพลมากนักต่อการตรวจค้นความเบี่ยงเบนของข้อสอบ สรุปแล้ว การใช้ BILOG แบบ B และ C ในการตรวจค้นความลำเอียงของข้อสอบที่สร้างขึ้นมาโดยทราบค่า หรือ กำหนดค่าความลำเอียงไว้ก่อนแล้วดีกว่าวิธีอื่น

Kim และ Cohen (1991) ได้ศึกษาความลำเอียงของข้อสอบที่สร้างขึ้นอย่างจงใจให้ลำเอียง โดยการเปรียบเทียบผลการคำนวณพื้นที่ความแตกต่างระหว่างโค้งที่ได้จากทฤษฎีการตอบข้อสอบของกลุ่มตัวอย่างผิวขาว และผิวดำ 2 วิธี ได้แก่ การคำนวณพื้นที่ภายใต้ความสามารถที่มีขอบเขตจำกัด และที่ไม่มีขอบเขตจำกัด

จำนวนกลุ่มตัวอย่างนี้ใช้กลุ่มละ 1,000 คน สุ่มมาจากเดิมที่มีผิวขาว 1,021 คน และผิวดำ 1008 คน ทั้งนี้เพราะขีดจำกัดในการใช้โปรแกรม BILOG ด้วยไมโครคอมพิวเตอร์ เครื่องมือเก็บรวบรวมข้อมูล คือ แบบสอบคำศัพท์แบบเลือกตอบ ชนิด 4 ตัวเลือก จำนวน 50 ข้อ เป็นคำศัพท์ภาษาอังกฤษมาตรฐาน 40 ข้อ อีก 10 ข้อ เป็นคำศัพท์ที่มีความง่ายสำหรับกลุ่มผิวดำ มากกว่ากลุ่มผิวขาว

การวิเคราะห์ความเป็นเอกมิติ (Unidimension) ด้วยการวิเคราะห์ตัวประกอบหลัก (principal components analysis) พบว่า ตัวประกอบแรกที่ได้ โดยไม่มีการหมุนแกน และใช้ค่าสหสัมพันธ์แบบเตตราคอริคแสดงถึงความเป็นเอกมิติได้พอเพียงในการใช้แบบจำลองทฤษฎีการตอบข้อสอบ สำหรับกลุ่มผิวขาวโดยอธิบายความแปรปรวนได้ร้อยละ 25 ในขณะที่ค่อนข้างน้อยสำหรับผิวดำ คือ อธิบายได้ร้อยละ 16

ใช้ทฤษฎีการตอบข้อสอบแบบ 2 พารามิเตอร์และแบบ 3 พารามิเตอร์ชนิดบังคับให้ $c_R = c_F = c = .23$

ใช้วิธีการวัดพื้นที่ 4 วิธี จากพารามิเตอร์ข้อสอบที่เทียบมาตรแล้ว คือ ESA ซึ่งได้แก่ พื้นที่แบบปิดคิดเครื่องหมาย (Exact Signed Area) ได้จากการคำนวณพื้นที่ตลอดช่วงความสามารถโดยไม่จำกัดขอบเขต และ EUA คือพื้นที่แบบปิดไม่คิดเครื่องหมาย (Exact Unsigned Area) กรณีที่เป็นพื้นที่ที่คิดเครื่องหมาย กลุ่มอ้างอิง (R) คือ กลุ่มผิวดำดังนั้นค่าที่เป็นบวกของ CSA หรือ ESA จะแสดงว่าข้อสอบข้อนั้นลำเอียงเข้าข้างกลุ่มผิวดำ

ข้อกระทงที่สร้างเพื่อให้เกิดความลำเอียงเข้าข้างผิวดำ 10 ข้อ ให้ค่าเป็น 1 อีก 40 ข้อ ที่เหลือเป็น 0 แล้วหาค่าสหสัมพันธ์ แบบ point biserial ระหว่างค่าดังกล่าวกับการวัด 4 วิธี เพื่อดูผลการตรวจค้นแต่ละวิธี และหาค่าสหสัมพันธ์ระหว่างวิธีต่าง ๆ เพื่อดูความคล้ายคลึงกัน

การประมาณค่าอัตราความคลาดเคลื่อนของการตรวจค้นความลำเอียง กระทำโดยการมีข้อตกลงว่า การวัดพื้นที่แต่ละวิธีมีการกระจายเป็นปกติ ค่าเฉลี่ยและค่าความเบี่ยงเบนมาตรฐานของการวัดแต่ละวิธี ประมาณจากข้อสอบ 40 ข้อ กำหนดค่าวิกฤตแบบทางเดียวเป็น 2 ค่า คือ .05 และ .01 ข้อที่มีพื้นที่ใหญ่กว่าค่าวิกฤตจะเป็นข้อสอบที่ลำเอียงการจัดประเภทข้อสอบเช่นนี้ทำให้เกิดความคลาดเคลื่อน 2 ชนิด ได้แก่ ความคลาดเคลื่อนชนิดระบุว่าไม่ลำเอียง ทั้ง ๆ ที่ลำเอียง (FN = False Negatives) และระบุว่าลำเอียง ทั้ง ๆ ที่ไม่ลำเอียง (FP = False Positives) และเมื่อใช้ p = .05 พบว่ามี FP มากกว่าที่ .01

ผลจากการวิเคราะห์ข้อมูลพบว่า เกือบไม่มีความแตกต่างระหว่างการคำนวณพื้นที่แบบช่วงเปิด และแบบช่วงปิดในการตรวจค้นความลำเอียงของข้อสอบ ส่วนการใช้พื้นที่แบบไม่คิดเครื่องหมายไม่ว่าจะเป็นแบบช่วงเปิดหรือปิดมีความไวต่อการตรวจค้นความลำเอียงของข้อสอบมากกว่าแบบคิดเครื่องหมาย นอกจากนั้นพบว่าการคำนวณพื้นที่ช่วงเปิดนั้นพื้นที่ชนิดมีเครื่องหมาย และไม่คิดเครื่องหมาย มีความสัมพันธ์กันมากกว่าการคำนวณพื้นที่ช่วงปิด

ในด้านของความคลาดเคลื่อนในการตรวจค้นความลำเอียง พบว่าวิธี 3 พารามิเตอร์ โดยใช้พื้นที่แบบคิดเครื่องหมาย มีอัตราความคลาดเคลื่อนต่ำกว่าวิธีอื่น ๆ ไม่มีความคลาดเคลื่อนเลยในระดับ .01 และ มีความคลาดเคลื่อนในการระบุข้อสอบว่าลำเอียงโดยไม่ลำเอียงจริง จำนวน 2 ข้อ ที่ระดับ .05 ส่วนวิธี 3 พารามิเตอร์ แบบไม่คิดเครื่องหมายมีอัตราความคลาดเคลื่อนสูงกว่าวิธีอื่น ๆ เล็กน้อย

Ryan (1991) ได้ตรวจค้นความลำเอียงของข้อสอบด้วยวิธีการที่เสนอโดย Mantel และ Haenszel (1959) โดยใช้ตัวอย่างกลุ่มอ้างอิง 500 คน และกลุ่มเปรียบเทียบ 100 คน เพื่อตรวจสอบความคงที่ของค่าประมาณความลำเอียงด้วย MH ข้ามกลุ่มตัวอย่างและขนาดตัวอย่างต่าง ๆ กัน และเพื่อศึกษาว่า MH มีความแกร่งต่อบริบทของข้อสอบเพียงไร โดยมีข้อสอบร่วมจำนวน 40 ข้อ และข้อสอบที่ใช้สลับกันอีกจำนวน 35 ข้อ กลุ่มตัวอย่างเป็นชาวผิวขาวและผิวดำจำนวน 5015 คน และ 670 คน ตามลำดับโดยมีเกณฑ์การแยกประเภทข้อสอบจากผลการวิเคราะห์เป็น 3 ประเภท

ประเภท A ข้อสอบที่ทำหน้าที่ได้ตรง มีค่า $\alpha_{MH}$ < 1.00 และมีค่าไม่ต่างจาก 0 อย่างมีนัยสำคัญ

ประเภท B ข้อสอบที่พอจะนำมาใช้ได้ แต่ควรเลือกข้อที่มีค่า $\alpha_{MH}$ ต่ำคือ $1 = \alpha_{MH} < 1.5$ และไม่ต่างจาก 1 อย่างมีนัยสำคัญ

ประเภท C เลือกมาใช้ถ้าตรงตามวัตถุประสงค์เฉพาะที่ระบุไว้อย่างมีนัยสำคัญว่าต้องการข้อสอบลักษณะนี้ได้แก่ข้อที่มีค่า $\alpha_{MH} > 1.5$ และมากกว่า 1 อย่างมีนัยสำคัญ

ขั้นตอนในการวิเคราะห์ขั้นแรกมีการวิเคราะห์ 5 แบบ ในการสร้างฐานเรียกว่า ชุด 1 ทำโดยการสุ่มตัวอย่างผิวขาว 670 คน เป็นกลุ่มเป้าหมาย ที่เหลือเป็นกลุ่มอ้างอิงแล้วสุ่มกลุ่มผิวขาวมาอีก 4 กลุ่ม เป็นกลุ่มอ้างอิง ได้แก่ $W_1$ $W_2$ $W_3$ $W_4$ และวิเคราะห์ข้อสอบร่วม 40 ข้อด้วย MH จับคู่การวิเคราะห์เป็น W/W W/$W_1$ W/$W_2$ W/$W_3$ W/$W_4$ คะแนนที่ใช้แบ่งระดับความสามารถ คือ คะแนนรวมจากข้อสอบ 40 ข้อ

การวิเคราะห์ชุด 2 ทำในลักษณะเดียวกับ ชุด 1 แต่ใช้กลุ่มเป้าหมายเป็นผิวดำวิเคราะห์ด้วย MH 5 คู่ ได้แก่ BW $BW_1$ $BW_2$ $BW_3$ $BW_4$ เป็นการวิเคราะห์เพื่อดูความคงที่ของค่า $\alpha_{MH}$ ข้ามกลุ่มและขนาดตัวอย่าง

ชุด 3 เป็นการวิเคราะห์ผลของบริบทของข้อสอบ $BW_1$ $BW_2$ $BW_3$ และ BW4 โดยใช้ข้อสอบ 75 ข้อจากฟอร์ม 1 ฟอร์ม 2 ฟอร์ม 3 ฟอร์ม 4 ใช้คะแนนรวมเป็นเกณฑ์แบ่งช่วงความสามารถ (แต่ละฟอร์มมีข้อสอบร่วม 40 ข้อ)

และชุด 4 เป็นการวิเคราะห์เช่นเดียวกับชุด 3 แต่ใช้ข้อสอบ 4 ฟอร์มที่ไม่มีข้อสอบร่วม (มีข้อสอบ 35 ข้อ) และใช้คะแนนรวมจาก 35 ข้อ เป็นเกณฑ์แบ่งระดับความสามารถ

ผลการวิเคราะห์พบว่า ความลำเอียงสำหรับกลุ่มผิวขาวด้วยกัน 5 คู่นั้นมีความสัมพันธ์กันต่ำมาก ไม่มีความแตกต่างอันเนื่องมาจากความลำเอียง แต่มีความแตกต่างจากความคลาดเคลื่อนในการสุ่มตัวอย่างที่มีความซ้ำซ้อนกันมากกว่า จึงอาจเป็นการลำเอียงแบบไม่แท้ก็ได้

เกณฑ์การแบ่งช่วงระดับความสามารถ ซึ่งเป็นบริบทที่ศึกษาไม่มีผลต่อค่า $\alpha_{MH}$ ไม่ว่าจะใช้เกณฑ์คะแนนรวม 35 ข้อหรือ 75 ข้อ ก็ได้ค่าสหสัมพันธ์ของการเปรียบเทียบแต่ละคู่เท่า ๆ กัน

ผลจากการจำแนกประเภทข้อสอบมีเพียง 1 ข้อ ที่มีความสอดคล้องกับทุกการวิเคราะห์ คือ เป็นประเภท B ทุกครั้ง นอกนั้นมีความแปรปรวนไปตามการวิเคราะห์

Harris และ Carlton (1993) ได้ศึกษาลักษณะกว้าง ๆ ของข้อสอบเพื่อชี้ให้เห็นจุดเด่นและจุดด้อย สำหรับผู้สอบหญิงและชายโดยแบ่งลักษณะที่ศึกษาเป็นประเด็นที่วัด รูปแบบและเนื้อหาของข้อสอบ

ข้อสอบที่นำมาศึกษาเป็นข้อมูลการตอบข้อสอบในแบบสอบ SAT-M จำนวน 6 ฟอร์ม ซึ่งเป็นตอนที่ทดสอบเฉพาะคณิตศาสตร์ ประกอบด้วยข้อสอบ 60 ข้อ เป็นการแก้ปัญหาทางคณิตศาสตร์ทั่วไป 40 ข้อ และการเปรียบเทียบปริมาณตัวเลข 20 ข้อ ผู้สอบเป็น

นักเรียนชั้นมัธยมทั้งต้นและปลาย เป็นชาย 181,228 คน หญิง 198,668 คน ทุกคนรายงานด้วยตนเองว่า ภาษาอังกฤษเป็นภาษาที่แต่ละคนใช้ได้ดีที่สุด จำนวนกลุ่มตัวอย่างที่ใช้ใน SAT-M ฟอร์ม 1-6 เป็นชาย 6,329-74,283 คน หญิง 6,712-83,945 คน

ผู้วิจัย ได้ระบุลักษณะเดิมของข้อสอบโดยใช้ข้อมูลจากการศึกษาในครั้งก่อน ๆ ของผู้อื่น เช่น Bleistein, 1986; Donlon, Ekstrom และ Lockheed, 1979; Doolittle และ Cleary, 1987; McPeek และ Wild, 1987; Schmitt และ Dorans, 1987; Wendler และ Carlton, 1987 เป็นต้น สิ่งที่ระบุได้แก่ ประเด็นที่วัด (นั่นคือ เนื้อหาหลักได้แก่ เลขคณิต พีชคณิต เรขาคณิต หรือทั่ว ๆ ไป) รูปแบบข้อสอบ (มีการใช้ตัวเลือกที่ว่า "ไม่สามารถตัดสินได้หรือไม่") และเนื้อหาของข้อสอบ (มีการอ้างอิงด้านเพศหรือไม่)

ตรวจสอบด้วย วิธี MH และพิจารณาความแตกต่างของค่าเดลต้า ซึ่งใช้แสดงถึงความยากของข้อสอบใน SAT อยู่แล้ว โดยที่ค่าเดลต้า = 1.00 แสดงว่าข้อสอบมีความยาก สำหรับกลุ่มหนึ่งมากกว่าอีกกลุ่มเท่ากับ 1 เดลต้า หรือเท่ากับความแตกต่างร้อยละ 10 โดยประมาณ

ค่าเดลต้าที่เป็นลบแสดงว่ายากสำหรับหญิงมากกว่าชาย ค่าบวกแสดงว่ายากสำหรับชายมากกว่าหญิง

ผลการวิเคราะห์โดยจับคู่ชายและหญิงในด้านความสามารถ ซึ่งได้แก่ คะแนนรวมจากแบบสอบแล้ว ได้ค่าเดลต้ามีค่า -1.86 ถึง 1.27 ค่าเฉลี่ย -.01 และค่าความเบี่ยงเบนมาตรฐาน =.51 และใช้การวิเคราะห์ความแปรปรวนทดสอบนัยสำคัญของความแตกต่างเฉลี่ยระหว่างข้อที่มีและไม่มีลักษณะดังที่ระบุไว้ พบว่า โดยทั่ว ๆ ไป ถ้าเนื้อหาหลักเป็นเรขาคณิตหรือเลขคณิตแล้วชายจะทำได้ดีกว่าหญิง แต่กรณีเป็นเนื้อหาทั่ว ๆ ไป (เช่น โครงสร้างของระบบจำนวน เซท ฯลฯ) หญิงจะทำได้ดีกว่าชาย ถ้าเป็นเลขคณิตและพีชคณิต หญิงจะทำได้ดีกว่าชาย แต่ถ้าเป็นเลขคณิตและเรขาคณิตแล้ว ชายทำได้ดีกว่าหญิง การพบว่าหญิงทำได้ดีกว่าในกรณีของพีชคณิตและชายทำได้ดีกว่าในเรขาคณิต (โดยที่หญิงและชายมีความสามารถรวมเท่ากัน) นั้นสอดคล้องกับการค้นพบของ Doolitte และCleary (1987)

นอกจากนั้นพบว่า ชายจะทำได้ดีกว่าในข้อสอบที่ต้องการกระบวนการทางสมองในระดับสูงกว่า ในข้อที่เป็นการประยุกต์นำมาใช้ในชีวิตประจำวัน ส่วนหญิงทำได้ดีกว่ากรณีที่มีตัวแปร (เช่น x, a, b) ในปัญหาและในตัวเลือก กรณีเป็นปัญหาตรงไปตรงมาจากหลักสูตรหรือหนังสือเรียนโดยไม่มีการนำมาประยุกต์ และในกรณีมีการอ้างอิงถึงบุคคลโดยไม่มีการแสดงถึงเพศหรือสถานภาพทางเชื้อชาติ

Raju และคณะ (1993) ได้ศึกษาเปรียบเทียบผลการทดสอบค่า Z ของ Raju ทั้งที่ได้จากพื้นที่ชนิดมีและไม่คิดเครื่องหมาย การทดสอบค่า $\chi^2$ ของ Lord (ใช้โมเดล 2 พารามิเตอร์) และการทดสอบ MH-$\chi^2$ โดยใช้เกณฑ์ตัดสิน ดังนี้

$\chi^2$ ของ Lord = 13.80

Z มีค่าระหว่าง $\pm$3.30 และ

MH-$\chi^2$ มีค่า = 10.83

ข้อมูลที่ใช้ในการตรวจค้นความลำเอียงของข้อสอบ คือ ผลการตอบแบบสอบ Gater-Mac Ginitie Reading Tests (GMRT) เฉพาะแบบสอบย่อยที่วัดคำศัพท์ 45 ข้อ เป็นแบบเลือกตอบ 5 ตัวเลือก กับนักเรียนระดับ 10 และ 12 ในปี 1987 จำนวน 839 คน ในจำนวนนี้มีนักเรียนผิวดำ 245 คน ผิวขาว 436 เป็นหญิง 440 และชาย 399 คน

ผลจากการศึกษา ในกรณีการตรวจสอบความลำเอียงระหว่างนักเรียนผิวดำและผิวขาว พบว่า มีข้อสอบลำเอียง 1 ข้อ สำหรับการทดสอบด้วย $\chi^2$ และ Z ทั้งพื้นที่ชนิดคิดเครื่องหมายและไม่คิดเครื่องหมาย ได้แก่ ข้อ 41 ส่วน MH พบข้อสอบลำเอียง 2 ข้อ ได้แก่ ข้อ 14 และ 27 ทั้ง 3 ข้อ เป็นข้อที่มีความยากต่างกันมากสำหรับผู้สอบผิวดำและผิวขาว

ส่วนผลการตรวจสอบความลำเอียงระหว่างชายและหญิง 3 วิธีแรก พบข้อสอบลำเอียงซ้ำกัน 4 ข้อ ส่วน MH พบ 5 ข้อ ซ้ำกับ 3 วิธีแรก 4 ข้อ (ร้อยละ 80) ได้แก่ ข้อ 2, 18, 23, 33 และข้อที่ไม่ซ้ำกับ 3 วิธีแรกคือ ข้อ 41 ทุกข้อมีค่าความยากสูงกว่าข้อที่ไม่ลำเอียง

การพบว่า MH ค้นพบข้อสอบลำเอียงได้มากกว่าวิธีที่ใช้ทฤษฎีการตอบข้อสอบเช่นนี้ค้านกับการค้นพบของ สแวมมินนาธาน และโรเจอรส์ (1990) ที่พบว่าวิธีที่ใช้ทฤษฎีการตอบข้อสอบค้นพบข้อสอบลำเอียงได้มากกว่า MH

ข้อควรสังเกต คือ การทดสอบ Z ของราชูอาจให้ผลถูกต้องกว่าการกำหนดเกณฑ์ตัดสินตามอำเภอใจ (arbitary) สำหรับพื้นที่ความแตกต่างระหว่างโค้งเพราะค่าความคลาดเคลื่อนมาตรฐานในการวัดเป็นตัวหนึ่งที่เข้ามาเกี่ยวข้องทำให้ข้อสอบข้อที่มีความแตกต่างของพื้นที่น้อยกว่าเป็นข้อสอบลำเอียง และข้อสอบที่มีค่าความแตกต่างของพื้นที่มากกว่าเป็นข้อสอบไม่ลำเอียงได้

Roussos และ Stout (1993) ได้ศึกษาผลของการใช้จำนวนตัวอย่างขนาดเล็ก ความแตกต่างกันของค่าพารามิเตอร์ข้อสอบรวมทั้งความแตกต่างในค่าเฉลี่ยของความสามารถของกลุ่มอ้างอิง และกลุ่มเปรียบเทียบเพื่อศึกษาถึงผลที่มีต่อความคลาดเคลื่อนชนิดที่ 1 (Type I error) ซึ่งเกิดจากการปฏิเสธสมมุติฐานศูนย์ที่เป็นจริงหรือในกรณีนี้ คือ การยอมรับว่าข้อสอบลำเอียงทั้ง ๆ ที่ข้อสอบไม่ลำเอียง

การศึกษาที่ 1 ใช้กลุ่มตัวอย่าง และกลุ่มเปรียบเทียบกลุ่มละ 100 200 500 และ 1000 คน เลือกข้อสอบ 1 ข้อ จากแบบสอบ ASVAB เลือกข้อที่มีค่าพารามิเตอร์ของข้อสอบคือ a b และ c ใกล้เคียงกับค่าเฉลี่ยของแบบสอบมากที่สุด คือ 1.32 0.03 และ 0.25 (ค่าเฉลี่ยคือ 1.22 0.09 และ 0.02) และเป็นข้อสอบที่ไม่ลำเอียง ความสามารถ ( $\theta$ ) ของกลุ่มตัวอย่างทั้ง 2 กลุ่มถูกจำลองให้มีการกระจายเป็นปกติมีค่าความแปรปรวนเป็น 1.00 ความแตกต่างระหว่างค่าเฉลี่ยของการแจกแจงของ 2 กลุ่ม ถูกจำลองให้มีค่า 0.0 0.5 และ 1.0 แล้วคำนวณค่าความสามารถเฉลี่ยของแต่ละกลุ่มให้จุดกลางของค่าเฉลี่ยอยู่บนระดับความยากเฉลี่ยของแบบสอบ จำลองการตอบ 400 ครั้ง และวิเคราะห์ 400 ครั้ง ในแต่ละสถานการณ์

ผลการใช้ SIBTEST และ MH ตรวจสอบความลำเอียงของข้อสอบที่เลือกขึ้นมาดังกล่าว พบว่านอกจากกรณีกลุ่มตัวอย่างขนาด 100 คน และค่าความแตกต่างของความสามารถ ($d_T$) = 0.0 แล้ว อัตราการปฏิเสธสมมุติฐานศูนย์ หรืออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีทั้งสองแตกต่างกันอย่างไม่มีนัยสำคัญ ทั้งยังมีความแตกต่างจากค่าระดับนัยสำคัญที่กำหนด คือ 0.05 อย่างไม่มีนัยสำคัญเช่นกัน สำหรับความแตกต่างที่พบในจำนวนตัวอย่าง 100 คน และ $d_T$ = 0.0 นั้น อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้งสองวิธีต่ำกว่า 0.05 อย่างไรก็ตาม SIBTEST มีอัตราความเคลื่อนประเภทที่ 1 สูงกว่า MH เล็กน้อย สำหรับการวิเคราะห์ความลำเอียงโดยใช้กลุ่มตัวอย่างไม่เกิน 1000 คน และมีแนวโน้มว่าอัตราความคลาดเคลื่อนจะสูงขึ้นเมื่อกลุ่มตัวอย่างมากขึ้นทั้ง 2 วิธี

การศึกษาที่ 2 ใช้แบบสอบเดิม แต่เลือกค่า a b และ c สำหรับข้อสอบที่นำมาวิเคราะห์ให้แตกต่างกันโดย

a = 0.4 1.0 และ 2.5

b = -1.5 -0.5 0.0 0.5 และ 1.5

c = 0.20

จำลองการกระจายของความสามารถเป้าหมายให้เป็นปกติ $d_T$ มีค่า 0.0 และ 1.0 นอกจากนั้นการจำลองทุกอย่างเป็นเช่นเดียวกับการศึกษาที่ 1 แต่ใช้จำนวนผู้สอบขนาดใหญ่ คือ 500 1,000 และ 3,000 คน และแต่ละสถานการณ์จำลองการตอบ 100 ครั้ง และตรวจสอบโดยใช้ SIBTEST และ MH พบว่าเมื่อค่า a สูงขึ้นและ b ต่ำ อัตราการปฏิเสธสมมุติฐานจะสูงกว่าระดับที่กำหนดไว้คือ 0.05 ทั้ง 2 วิธี และ MH จะมีอัตราการปฏิเสธสูงกว่า SIBTEST และยิ่งจำนวนกลุ่มตัวอย่างใหญ่ขึ้น อัตราการปฏิเสธจะยิ่งสูงขึ้น ในขนาดกลุ่มตัวอย่าง 3,000 คน ที่ค่า a = 2.5 และ b = -1.5 อัตราการปฏิเสธของ MH สูงถึง 1.00 และ SIBTEST = 0.91 ในกรณีที่ $d_T$ = 1.0 และ c = 0.20 ในภาพรวมแล้ว SIBTEST มีอัตรา

การปฏิเสธต่ำกว่า MH แต่เมื่อใช้ค่า $d_T$ = 0.0 ไม่มีค่า c ผลปรากฏว่า อัตราการปฏิเสธของทั้ง 2 วิธีอยู่ในระดับใกล้เคียงกับ 0.05 แม้ว่า SIBTEST จะสูงกว่า MH แต่ไม่มีนัยสำคัญ

สรุปว่าเมื่อใช้จำนวนตัวอย่างน้อย ไม่เกิน 1,000 คน MH และ SIBTEST ให้ผลในด้านความคลาดเคลื่อนประเภทที่ 1 ใกล้เคียงกัน และอยู่ภายในระดับ 0.05 ในขณะที่พบว่า SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่า MH ในกรณีที่จำนวนตัวอย่างมีขนาดใหญ่ (มากกว่า 1,000 คน) และค่าเฉลี่ยของการแจกแจงของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีความแตกต่างกัน รวมทั้งข้อสอบมีความยากต่ำ และค่าอำนาจจำแนกสูง นอกจากผลทางด้านอัตราการปฏิเสธแล้ว ยังพบว่าตัวประมาณความลำเอียงคือ $\beta$ และ $\Delta$ ที่มีค่าต่ำกว่าเกณฑ์หรือมีค่าแสดงความลำเอียงเพียงเล็กน้อยก็อาจมีนัยสำคัญได้ ($\Delta$ = 1.5 และ 1.0 และ $\beta$ = 0.10 หรือ 0.05)

การศึกษาเกี่ยวกับการตรวจค้นความลำเอียงของข้อสอบในต่างประเทศเท่าที่ผ่านมาจะเห็นว่า มีการใช้ทั้งข้อมูลจริงและข้อมูลจำลอง ข้อมูลจำลองที่นำมาใช้ในการศึกษาเป็นข้อมูลที่จำลองขึ้นตามทฤษฎีการตอบข้อสอบทั้งสิ้น ซึ่งเป็นจุดที่น่าวิจารณ์ว่าเป็นผลให้พบว่า การตรวจสอบข้อสอบลำเอียงด้วย IRT มีความถูกต้องกว่าวิธีอื่น แม้จะยังมีความคลาดเคลื่อนในการตรวจสอบทั้งชนิด FN และ FP ก็ตาม

ข้อค้นพบที่สอดคล้องกับอีกประการ คือ ส่วนมากพบความสัมพันธ์กันในระดับสูงระหว่างวิธี IRT ชนิด 2 และ 3 พารามิเตอร์ และ MH แม้จะมีการพบข้อขัดแย้งกันอยู่เช่นกันเมื่อใช้ข้อมูลจริงในด้านจำนวนข้อสอบลำเอียงที่ตรวจพบ ซึ่ง Swaminathan และRogers (1990) พบว่า IRT ตรวจพบข้อสอบลำเอียงได้มากกว่า MH ในขณะที่ Raju และคณะ (1993) พบว่า MH ตรวจพบได้มากกว่า IRT รวมทั้งการพบความไม่คงที่ของการตรวจพบข้อสอบลำเอียงเมื่อใช้กลุ่มตัวอย่างต่างกันอันเนื่องมาจากความไวของค่าสถิติทดสอบต่อจำนวนกลุ่มตัวอย่าง และปัญหาการใช้กณฑ์เพื่อใช้ตัดสินว่าข้อสอบลำเอียง ซึ่งส่วนมากกำหนดเกณฑ์แตกต่างกันตามอำเภอใจ (arbitary) ( Hambleton 1986 Thissen Steinberg และ Wainer 1988 Linacre 1988 Perlman และคณะ 1988 Hambleton และRogers 1989 Baghi และ Ferrara 1990 Swaminathan และRogers 1990 และRaju และคณะ 1993

จากข้อค้นพบดังกล่าว ทำให้ผู้วิจัยสนใจในการพัฒนาเกณฑ์ที่จะใช้ตัดสินข้อสอบลำเอียงขึ้นจากข้อมูลเชิงประจักษ์ โดยเลือกวิธีการที่ผู้วิจัยเห็นว่ามีแนวคิด ในเรื่องของความลำเอียงเหมาะสม เป็นแนวเดียวกันและมีผลการศึกษาที่ใกล้เคียงกัน 3 วิธี คือ วิธี IRT-2 พารามิเตอร์ วิธี MH และวิธี SIBTEST โดยนำดัชนีที่ได้จาก 3 วิธีมาใช้ในการพัฒนาเกณฑ์ 4 ตัว ได้แก่ SA UA $\alpha_{MH}$ และ $\beta_{SIB}$ ซึ่งจากเอกสารรายงานที่เกี่ยวข้องดังกล่าว สรุปการใช้เกณฑ์ที่ผ่านมาของวิธีตรวจค้นความลำเอียงดังตารางที่ 3

จากข้อค้นพบดังกล่าว ทำให้ผู้วิจัยสนใจในการพัฒนาเกณฑ์ที่จะใช้ตัดสินข้อสอบลำเอียงขึ้นจากข้อมูลเชิงประจักษ์ โดยเลือกวิธีการที่ผู้วิจัยเห็นว่ามีแนวคิด ในเรื่องของความลำเอียงเหมาะสม เป็นแนวเดียวกันและมีผลการศึกษาที่ใกล้เคียงกัน 3 วิธี คือ วิธี IRT-2 พารามิเตอร์ วิธี MH และวิธี SIBTEST โดยนำดัชนีที่ได้จาก 3 วิธีนี้มาใช้ในการพัฒนาเกณฑ์ 4 ตัว ได้แก่ SA UA $\alpha_{MH}$ และ $\beta_{SIB}$

ตารางที่ 3 สรุปเกณฑ์ที่พบว่ามีการนำมาใช้ตัดสินข้อสอบลำเอียง

| วิธี | เกณฑ์การตัดสิน |
|---|---|
| IRT | 1. การทดสอบนัยสำคัญของค่า $\chi^2$ ของ Lord<br>2. $\chi^2$ > 13.80 UA (หรือ $\Phi$) > .20 UA > .40<br>3. SA > .468<br>4. การทดสอบค่า Z (Raju) |
| MH | 1. ทดสอบนัยสำคัญของ MH-$\chi^2$<br>2. MH-$\chi^2$ > 6.64<br>3. $\alpha_{MH}$ > 1.00<br>4. MH-$\chi^2$ > 10.83 |
| SIB-TEST | ทดสอบนัยสำคัญของค่า Z |