บทที่ ๒

วรรณคดีที่เกี่ยวข้อง

การเทียบมาตรา (Equating) เป็นคำศัพท์เฉพาะสาสตร์ทางจิตมิติ ตำราที่เกี่ยวกับการทดสอบและวัดผลทางจิตวิทยา และการศึกษาเบื้องต้นมักไม่ได้กล่าวถึง จึงเป็นเรื่องที่บุคคลทั่วไปไม่เข้าใจกัน ส่วนนักวัดผลเองโดยมากไม่ได้ให้ความสนใจเท่าที่ควร (Marco 1981) แต่อย่างไรก็ดีในช่วงเวลาประมาณ ๑๕ ปีมานี้ ได้มีนักวัดผลให้ความสำคัญกับการเทียบมาตรามากขึ้นตามลำดับ มีการศึกษาวิจัยถึงเทคนิคการเทียบมาตรารูปแบบต่าง ๆ ในสภาพความเป็นจริงหลายลักษณะ นอกจากนี้ ฮอลแลนด์ และ โรบิน (Holland and Rubin 1982) ได้เป็นบรรณาธิการจัดพิมพ์ตำรา Test Equating ขึ้นเผยแพร่ โดยมุ่งให้เกิดประเด็นอภิปราย และสร้างเสริมเทคนิควิธีการใหม่ ๆ ให้กว้างขวางและลึกซึ้งยิ่งขึ้น การเสนอวรรณคดีที่เกี่ยวข้องในบทนี้ ได้เสนอรายละเอียด ๙ หัวข้อ ดังนี้ คือ (๑) พัฒนาการทางแนวคิดและรูปแบบการเทียบมาตรา (๒) การออกแบบรวบรวมข้อมูล (๓) วิธีเทียบมาตรารูปแบบอีควิเปอร์เซนไทล์ (๔) วิธีเทียบมาตรารูปแบบเชิงเส้นตรง (๕) วิธีเทียบมาตรารูปแบบอิงทฤษฎีการตอบข้อสอบ (๖) ความยาวแบบสอบร่วม (๗) ความคลาดเคลื่อนของการเทียบมาตรา (๘) การประเมินความเพียงพอของการเทียบมาตรา และ (๙) งานวิจัยที่เกี่ยวข้องกับการเปรียบเทียบรูปแบบการเทียบมาตรา

พัฒนาการทางแนวคิดและรูปแบบการเทียบมาตรา

๑. แนวคิดเชิงทฤษฎี

กัลลิคเสน (Gulliksen 1950: 298-304) ให้ความหมายของการเทียบมาตราว่า เป็นวิธีการทำคะแนนที่ได้จากแบบสอบสองชุดของวิชาเดียวกัน ให้เป็นคะแนนสมมูลที่เปรียบเทียบกันได้โดยตรง วิธีการเทียบมาตราที่ใช้กับการสอบทั้งสองชุดของแบบสอบกับกลุ่มคนเพียงกลุ่มเดียวเป็นวิธีง่าย ๆ โดยแปลงคะแนนแต่ละชุดให้เป็นคะแนนมาตรฐาน แล้วนำคะแนนแปลงมาเทียบกันโดยตรง แต่ก่อนอื่นต้องทำการตรวจสอบให้แน่ใจเสียก่อนว่า คะแนนทั้งสองชุดนั้นมีสัมประสิทธิ์ความ

โค้ง และความเบ้ของการแจกแจงคล้ายคลึงกัน หรือถ้าคะแนนสองชุดนั้นได้ทำให้เป็นมาตรฐานในเทอมของเกณฑ์ใดเกณฑ์หนึ่ง ก็ต้องตรวจสอบดูว่า แบบสอบทั้งสองชุดมีค่าสัมประสิทธิ์สหสัมพันธ์กับเกณฑ์เท่ากันโดยประมาณ

ฟลานาแกน (Flanagan 1951: 747-748) ได้กล่าวถึงการเทียบมาตราว่าเป็นวิธีการที่ทำคะแนนจากแบบสอบต่างชุดให้มีคุณลักษณะที่เปรียบเทียบกันได้ คำว่า "ความสามารถในการเปรียบเทียบกันได้" มีความหมายเฉพาะที่ว่า เมื่อกำหนดประชากรให้ ถ้าการแจกแจงของคะแนนจริงจากแบบสอบทั้งสองจากกลุ่มตัวอย่างที่เลือกมาขนาดใหญ่ใด ๆ มีลักษณะเหมือนกันแล้ว คะแนนดิบของแบบสอบทั้งสองชุดจึงจะสามารถเปรียบเทียบกันได้ หรือถ้าความเที่ยงของแบบสอบสองชุดนั้นเท่ากันในประชากรนั้นแล้ว ก็สามารถเปรียบเทียบการแจกแจงของค่าที่ได้เช่นกัน จากความหมายดังกล่าวเป็นนิยามเชิงทฤษฎี ในทางปฏิบัติได้นิยามไว้ว่า ในประชากรที่กำหนด ถ้าคะแนนจากแบบสอบสองชุดมีค่าเฉลี่ยเท่ากัน หรือเกือบเท่ากันในทุก ๆ กลุ่มตัวอย่างขนาดใหญ่ใด ๆ การเปรียบเทียบจะทำได้ อย่างไรก็ดีเมื่อวิเคราะห์ถึงความเป็นจริงที่จะให้เกิดลักษณะตามที่นิยามเป็นเรื่องยากมาก ปัจจัยสำคัญที่ต้องคำนึง คือ ความเป็นคู่ขนานของแบบสอบ ฟลานาแกนได้แนะนำว่า ควรเริ่มตั้งแต่การสร้างแบบสอบให้มีคุณสมบัติของความเป็นคู่ขนานกัน แล้วเลือกวิธีการเทียบมาตราที่เหมาะสม ซึ่งเสนอไว้ ๔ วิธี คือ (๑) เทียบค่าเฉลี่ย โดยคำนวณค่าเฉลี่ยของการแจกแจงคะแนนทั้งสองชุด ถ้าความแตกต่างของค่าเฉลี่ยอยู่ภายในขอบเขตความแปรผันเชิงสุ่มแล้ว ถือว่าคะแนนสองชุดนั้นเปรียบเทียบกันได้ แต่ถ้าความแตกต่างมีนัยสำคัญ ให้ใช้วิธีบวกเข้าหรือลบออกเท่าจำนวนที่แตกต่างจากคะแนนชุดหนึ่ง เพื่อให้เกิดคะแนนสมมูลกับอีกชุดหนึ่ง (๒) ใช้เทคนิคของสมการถดถอย ใช้ค่าประมาณดีที่สุด ("best estimate") ของคะแนนจากแบบสอบชุดหนึ่งซึ่งรู้ค่าของอีกชุดหนึ่ง (๓) ใช้คะแนนมาตรฐาน เป็นวิธีปรับคะแนนอย่างคงที่ตลอดการแจกแจง และ (๔) ใช้อีควิเปอร์เซนไทล์ เป็นวิธีหาค่าคะแนนโดยการเปรียบเทียบตามสัดส่วนของการแจกแจงคะแนน ซึ่งเป็นวิธีที่เหมาะสมที่สุด

จากมโนทัศน์ดังกล่าว มีผลทำให้การเทียบมาตราในสมัยทศวรรษ ๑๙๕๐ มุ่งเน้นเทคนิคการสร้างแบบสอบชุดต่าง ๆ ให้มีคุณสมบัติของความเป็นคู่ขนานมากที่สุด เมื่อแบบสอบเป็นคู่ขนานแล้ว คะแนนที่ได้จากทั้งสองชุด ย่อมเป็นคะแนนสมมูลกัน (equivalent scores) หรือถ้าหากมีความแตกต่างเกิดขึ้น เช่น ค่าเฉลี่ยหรือส่วนเบี่ยงเบนมาตรฐานต่างกันในขอบเขตที่ยอมรับ

จะใช้การปรับค่าความแตกต่างตามที่กัลลิคเสนไว้แนะนำดังกล่าวมาแล้ว ก็สามารถทำให้คะแนนทั้งสองเป็นคะแนนสมมูลได้ เทคนิคการสร้างแบบสอบคู่ขนานที่กัลลิคเสนแนะนำมีดังนี้ (Gulliksen 1950: 207-210)

(๑) วิเคราะห์ความยากและค่าอำนาจจำแนกรายข้อ แล้วทำจุดลงบนกระดาษกราฟ

(๒) พิจารณาข้อสอบที่เกาะกลุ่มกันอยู่ว่า แต่ละกลุ่มวัดความสามารถอะไรบ้าง

(๓) แยกข้อสอบที่เกาะกลุ่ม และวัดสิ่งเดียวกันให้อยู่คนละชุดโดยสุ่ม

วิธีนี้ช่วยให้ผู้พัฒนาแบบสอบสามารถควบคุมความหลากหลายของคำถามจากแบบสอบชุดหนึ่งไปสู่อีกชุดหนึ่งได้ แต่วิธีนี้ ฟลานาแกน (Flanagan 1951: 749-50) พบว่า มีจุดอ่อนอยู่ ๕ ประการ คือ

(๑) ความยากที่วิเคราะห์จากฉบับเริ่มแรกที่มีจำนวนข้อสอบมาก ๆ นั้นมีค่าแตกต่างจากความยากของข้อสอบข้อเดียวกัน เมื่อกระจายเข้าไปอยู่ในฉบับสมมูล ทั้งนี้เนื่องจากตำแหน่งของการจัดเรียงข้อสอบ และบริบทของแบบสอบมีผลกระทบต่อค่าดังกล่าว

(๒) โดยปกติการแบ่งข้อสอบจากฉบับเริ่มแรกไปยังฉบับสมมูล โดยไม่ทำให้สูญเสียลักษณะการสุ่มเนื้อหาในแต่ละชุดนั้นเป็นเรื่องที่ทำไม่ได้

(๓) การปรับปรุงแบบสอบชุดหลัง ๆ ทำได้ยาก เพราะจะต้องคงสภาพความยาก คุณสมบัติทั่วไป และจำนวนข้อในแต่ละตอนให้เหมือนฉบับเดิม

(๔) การดำเนินการสอบที่จัดขึ้นต่างเวลากัน มีความหมายถึงการเปลี่ยนแปลงสถานการณ์ของการทดสอบ ซึ่งมีผลกระทบต่อการแปรเปลี่ยนของการแจกแจงคะแนนดิบ จะนำมาเปรียบเทียบโดยตรงไม่ได้

(๕) ความสัมพันธ์ของข้อสอบข้อเดียวกันในแบบสอบต่างชุด ย่อมแตกต่างกัน ซึ่งทำให้ค่าความคงที่ภายในเปลี่ยนไป ผลที่ตามมาจากการทดสอบจะได้การแจกแจงที่ไม่เหมือนกันด้วย

ดังนั้น การเทียบมาตราของแบบสอบด้วยการจัดกระทำในขั้นตอนการสร้างแบบสอบคู่ขนาน จึงมีข้อจำกัดมาก นอกจากนี้การทดสอบในสภาพการณ์จริงบางกรณี โดยเฉพาะที่เป็นโปรแกรมการสอบระดับชาติ จำเป็นต้องยึดนโยบาย และข้อกำหนดของระเบียบเป็นสำคัญในการปฏิบัติ เทคนิคการเทียบมาตราจึงหันมาเน้นกระบวนการปรับโดยอาศัยวิธีการทางสถิติ ด้วยเหตุนี้การศึกษาค้นคว้า

หาวิธีการเทียบมาตรา จึงมุ่งหาวิธีการทางสถิติที่เหมาะสมเพื่อใช้กับแบบสอบที่มีคุณสมบัติน้อยกว่าที่กำหนดตามนิยามของแบบสอบคู่ขนาน กล่าวโดยสรุปได้ว่า การเทียบมาตราเป็นกระบวนการที่รวมเอาเทคนิคการสร้างแบบสอบอย่างพิถีพิถัน ตลอดจนวิธีการใช้สถิติที่เหมาะสมในการปรับคะแนนที่ได้จากต่างชุดของแบบสอบ เพื่อชดเชยความแตกต่างของความยากที่ต่างกัน (Holland and Rubin 1982: 2) ดังนั้น การศึกษาถึงเทคนิคการเทียบมาตราจะเป็นวิธีการทางสถิติมาก โดยถือว่าแบบสอบสองชุดที่ต้องการเทียบนั้นได้รับการพัฒนาถึงระดับที่วัดความสามารถในสิ่งเดียวกัน และอยู่ภายใต้ข้อจำกัดของสภาพความเป็นจริงที่หลีกเลี่ยงไม่ได้

๒. <u>รูปแบบการเทียบมาตรา</u> (Equating models)

๒.๑ <u>รูปแบบการเทียบมาตราตามประเพณีนิยม</u> (Traditional models)

แองกอฟ (Angoff 1984: 85-93) แบ่งรูปแบบการเทียบมาตราออกเป็น ๒ รูปแบบ คือ Equating model และ Calibration model ซึ่งมีความแตกต่างในเชิงแนวคิด ดังนี้

Equating model เป็นรูปแบบการแปลงคะแนนที่ยึดนิยามของคะแนนสมมูลเป็นเป้าหมาย (equivalent scores) นิยามของคะแนนสมมูลกล่าวไว้ว่า คะแนนสองจำนวน จำนวนหนึ่งมาจากแบบสอบชุด X ส่วนอีกจำนวนหนึ่งมาจากแบบสอบชุด Y โดยที่ทั้ง X และ Y วัดพังชั่นเชิงจิตวิทยาเดียวกันด้วยแบบสอบที่มีความเที่ยงระดับเดียวกัน คะแนนสองจำนวนนี้นับว่าเป็นคะแนนสมมูล ถ้าต่างก็สมนัยกับตำแหน่งเปอร์เซนไทล์ของประชากรที่กำหนดให้ที่ตำแหน่งเดียวกัน ตามนิยามนี้ ถ้าแบบสอบ ๒ ชุด มีความยากต่างกัน ซึ่งหมายถึง การแจกแจงของคะแนนดิบของแบบสอบสองชุดนั้นย่อมต่างกันด้วยนั้น คะแนนสมมูลที่เกิดจากการใช้วิธีการเทียบมาตรานี้ย่อมเกิดจากการยืดหรือรวบมาตราของคะแนนชุดหนึ่ง เพื่อทำให้การแจกแจงคะแนนเหมือนกับการแจกแจงของอีกชุดหนึ่งนั่นเอง ดังนั้น การเทียบมาตราด้วยรูปแบบ Equating นี้มีผลทำให้ผู้สอบแต่ละคนได้คะแนนแปลงเหมือนกันไม่ว่าผู้นั้นจะรับการทดสอบด้วยแบบสอบชุดใด ในกรณีที่แบบสอบ ๒ ชุด มีลักษณะสำคัญ ๆ ใกล้เคียงกันมากตามทฤษฎีซึ่งเป็นเหตุผลที่จะกำหนดข้อสมมุติฐานที่ว่า การแจกแจงคะแนนดิบของแบบสอบสองชุดมีรูปร่างเหมือนกัน และการแปลงคะแนนจากชุดหนึ่งไปยังอีกชุดหนึ่งสามารถทำได้

โดยเปลี่ยนจุดเริ่มต้นของมาตราและหน่วยการวัดเท่านั้น หรือกล่าวอีกนัยหนึ่งว่า ทำการปรับโมเมนต์ที่หนึ่งและสองนั่นเอง วิธีการเช่นนี้ คือ การแปลงเชิงเส้นตรง จึงอาจกล่าวสรุปได้ว่า การเทียบมาตราด้วยการแปลงคะแนนเชิงเส้นตรงให้ค่าใกล้เคียงอย่างมากกับวิธีการอีควิเปอร์เซนไทล์ เมื่อรูปร่างการแจกแจงคะแนนดิบของแบบสอบทั้งสองคล้ายคลึงกัน ดังนั้น ถ้าผู้วิเคราะห์สามารถกำหนดข้อตกลงความคล้ายคลึงของการแจกแจงได้ ก็สามารถใช้วิธีการเชิงเส้นตรง ซึ่งให้ผลน่าพอใจกว่าอีควิเปอร์เซนไทล์ตรงที่สามารถหลีกเลี่ยงความลำเอียงที่อาจเกิดขึ้นในขณะทำการปรับเส้นโค้ง แต่อย่างไรก็ตามการแจกแจงของคะแนนจากแบบสอบสองชุด โดยธรรมชาติมีลักษณะการแจกแจงแตกต่างกัน การเทียบมาตราด้วยวิธีอีควิเปอร์เซนไทล์จึงดูเหมือนให้ความถูกต้องและเหมาะสมกว่า (Angoff 1984: 87)

Calibration model เป็นรูปแบบการเทียบมาตราที่ต้องการให้ผลของคะแนนแปลงสะท้อนลักษณะความยากหรือง่ายกว่าของแบบสอบชุดใหม่ ถ้าพิจารณาย้อนกลับไปที่รูปแบบ Equating วิธีการของอีควิเปอร์เซนไทล์จะกำหนดให้คะแนนชุดใหม่อยู่ในกรอบของชุดเก่า คะแนนสูงสุดของชุดใหม่ซึ่งวัดระดับความสามารถสูงกว่าชุดเก่าจะถูกกำหนดให้เสมอกัน สภาพเช่นนี้สารสนเทศจากการแปลงคะแนนจะถูกจำกัด แองกอฟ (Angoff 1984: 89) ได้เปรียบเทียบการเทียบมาตราตามรูปแบบ Calibration ว่า เสมือนหนึ่งเทอร์โมมิเตอร์สองชนิดที่สร้างขึ้นเพื่อวัดอุณหภูมิที่มีขีดจำกัดต่างกัน เช่น วัดอุณหภูมิระหว่าง ๔๐ องศาฟาร์เรนไฮท์ ถึง ๑๐๐ องศาฟาร์เรนไฮท์ กับอีกอันหนึ่งวัดอุณหภูมิระหว่าง ๕๔ องศาฟาร์เรนไฮท์ กับ ๑๐๔ องศาฟาร์เรนไฮท์ ทั้งสองอันนี้ต้องการเทียบมาตราเพื่ออ่านอุณหภูมิในหน่วยเดียวกัน รูปแบบการ Calibration จะช่วยให้เทอร์โมมิเตอร์ทั้งสองอันให้ค่าจากการวัดอุณหภูมิที่สะท้อนให้เห็นความแตกต่างอุณหภูมิที่วัดได้อย่างถูกต้อง ครอนแบค (Cronbach 1970: 111) ได้อธิบายวิธีการ Calibration ว่า เหมือนกับวิธีการทำเครื่องหมายหน่วยวัดบนเครื่องวัดความดันอากาศที่ใช้สารอื่นเพื่อให้มีความถูกต้องเช่นเดียวกับเครื่องวัดที่ใช้ปรอท วิธีที่ใช้เทียบมาตราในรูปแบบ Calibration ที่เหมาะสม คือ การเทียบเชิงเส้นตรง

แองกอฟ (๑๙๘๔: ๙๓) ได้สรุปไว้ว่า การปรับมาตราที่ได้คำนึงถึงความแตกต่างของความยาก และพิสัยความสามารถที่วัดมีอยู่ ๒ รูปแบบ คือ Calibration เป็นการปรับเชิงเส้นตรง หรือใช้คะแนนมาตรฐาน (Z score) ซึ่งเป็นวิธีที่ต้องการให้คะแนนแปลงสะท้อนให้เห็นความสามารถในการจำแนกตามที่แบบสอบต้องการ อีกรูปแบบหนึ่ง คือ Equating เป็นการปรับที่ไม่เป็นเส้นตรง (curvilinear) หรืออีควิเปอร์เซนไทล์ ซึ่งบางครั้งวิธีเชิงเส้นตรงใช้ประมาณค่าได้ดี และโดยความจริงที่ว่าวิธีเชิงเส้นตรงสมมูลกับอีควิเปอร์เซนไทล์ เมื่อรูปร่างการแจกแจงของคะแนนดิบเหมือนกัน ซึ่งหมายความว่า นอกจากสองโมเมนต์แรกแล้ว โมเมนต์มาตรฐานที่เหลือของการแจกแจงคะแนนดิบของแบบสอบสองชุดของกลุ่มผู้สอบที่กำหนดให้เหมือนกันหมด ความแตกต่างของการเทียบมาตราสองรูปแบบนี้ จะเห็นได้ชัดเมื่อใช้เทียบคะแนนจากแบบสอบที่มีความยากไม่เท่ากัน ๒ ชุด ไปสู่มาตราใหม่ที่เป็นคนละมาตรากับมาตราคะแนนดิบเดิม จะเห็นความแตกต่างที่สำคัญชัดเจน วิธีการเชิงเส้นตรงจะแสดงลักษณะให้ปรากฏในคะแนนที่มาจากวิธีการแปลง และถ้าแบบสอบมีความยากต่างกัน อาจเกิดเหตุการณ์ที่ว่า คะแนนของชุดหนึ่งอาจสูงเกินกว่าที่คนสอบอีกชุดหนึ่งจะทำได้ถึง ส่วนในกรณีของอีควิเปอร์เซนไทล์จะปรับความยากเหล่านี้ เพื่อให้คะแนนของแต่ละคนมีคงเดิมโดยไม่สนใจว่า ผู้สอบนั้นจะได้รับการทดสอบด้วยแบบสอบชุดใด

จากการเสนอรูปแบบการเทียบมาตราที่กล่าวมา พบว่า ต่างมีเป้าหมายร่วมกัน คือ ต้องการแปลงหน่วยในระบบมาตราของคะแนนชุดหนึ่งไปสู่หน่วยของระบบมาตราคะแนนอีกชุดหนึ่ง โดยให้ตัวเลขภายหลังกระบวนการแปลงสามารถนำมาเปรียบเทียบในเชิงปริมาณได้ สำหรับบทความ และผลงานวิจัยที่ปรากฏในเวลาต่อมา ไม่พบว่ามีผู้ใดเน้นความแตกต่างของรูปแบบทั้งสองอีกเลย และการเทียบมาตราที่ปรากฏในระยะต่อมานี้ ได้ใช้คำว่า Equating เป็นคำทั่วไป ส่วนผลการเทียบจะสะท้อนให้เห็นความยากที่ต่างกันหรือไม่นั้นได้เป็นส่วนหนึ่งที่นำมาพิจารณาถึงสภาพของแบบสอบ และสถานการณ์ของการทดสอบในการวิเคราะห์ เพื่อสร้างคะแนนสมมูลภายใต้การออกแบบในการทดลอง ดังนั้น การเทียบมาตราตามประเพณีนิยมมาแต่เดิมอาจสรุปได้ว่ามีสองรูปแบบ คือ รูปแบบอีควิเปอร์เซนไทล์ และรูปแบบเชิงเส้นตรง (Kolen and Whitney 1982)

๒.๒ <u>รูปแบบอิงทฤษฎีการตอบข้อสอบ</u> (Item response theory: IRT models)

เป็นรูปแบบการเทียบมาตราที่พัฒนาต่อมาจากทฤษฎีการสอบที่ว่าด้วยการวัดความสามารถจริงที่แฝงอยู่ในแต่ละบุคคล (Latent trait theory) ทฤษฎีว่าด้วยการตอบข้อสอบ เป็นทฤษฎีที่จำลองคำตอบ (response) ของผู้สอบจากการตอบคำถามข้อสอบรายข้อในแบบทดสอบ ทฤษฎีนี้มีประโยชน์ต่อการออกแบบของแบบสอบ ต่อการบรรยายและประเมินข้อสอบและแบบสอบ ต่อการให้คะแนนคำตอบของผู้สอบรายบุคคลได้อย่างเหมาะสม ต่อการทำนายคะแนนของผู้สอบรายคนและกลุ่มผู้สอบ และต่อการจัดกระทำและแปลความหมายของคะแนนสอบ ซึ่งรวมถึงการเทียบมาตราของแบบสอบต่างชุดด้วย (Lord 1982 (b): 141) ทฤษฎีนี้ได้พัฒนาขึ้นและนำสู่การประยุกต์กับการวัดผลมาไม่น้อยกว่า ๓๐ ปี ในทศวรรษที่ผ่านมามีงานวิจัยมุ่งเน้นการประยุกต์ทฤษฎีนี้อย่างจริงจัง จนอาจถือว่าเป็น "ยุคของการปฏิวัติของสาขาการวัดผลทางการศึกษา" (Marco 1977) รูปแบบการเทียบมาตราที่พัฒนาขึ้นภายใต้ทฤษฎีนี้ที่รู้จักกันดี ได้แก่ รูปแบบหนึ่งพารามิเตอร์ (The one parameter model หรือ Rasch model) และรูปแบบโลจิสสามพารามิเตอร์ (The three-parameter logistic model) ซึ่งลอร์ดเป็นผู้นำในการพัฒนาทฤษฎีนี้และนำไปสู่การประยุกต์กับปัญหาการทดสอบมากมาย รวมถึงการเทียบมาตราคะแนนมาเป็นเวลาเกือบ ๒๐ ปีแล้ว (Marco 1977) สำหรับงานวิจัยนี้ได้เลือกรูปแบบโลจิสสามพารามิเตอร์เพียงอย่างเดียว การเสนอวรรณคดีที่เกี่ยวข้อง จึงจำกัดเฉพาะเรื่องราวของโลจิสสามพารามิเตอร์เท่านั้น

นิยามการเทียบมาตรารูปแบบอิงทฤษฎีการตอบข้อสอบ (เรียกย่อ ๆ ว่า IRT) เป็นนิยามที่มุ่งเน้นความสำคัญของความเสมอภาค หรือความเป็นธรรมที่บุคคลจะได้รับจากการทดสอบอย่างเท่าเทียมกัน ถึงแม้การสอบนั้นจะใช้แบบสอบต่างชุดกัน (Lord 1980: 195) ความจริงมโนทัศน์ของความเสมอภาคนี้ แองกอฟ (Angoff 1984: 85) ได้ให้ไว้ในทำนองเดียวกันว่า ผู้ที่รับการทดสอบต้องไม่ได้รับประโยชน์ หรือสูญเสียประโยชน์เป็นพิเศษอันเนื่องมาจากการได้รับการทดสอบด้วยแบบสอบชุดใดชุดหนึ่งที่มีความยากกว่า หรือง่ายกว่ากัน นิยามเชิงทฤษฎีในทำนองนี้เทคนิคการเทียบมาตราที่แล้วมามีข้อจำกัดในทางปฏิบัติอยู่มาก ลอร์ดจึงให้นิยามในเชิงปฏิบัติว่า ความเสมอภาค หมายถึง การที่ผู้สอบทุกคนที่ทุกระดับความสามารถ θ มีการแจกแจงความถี่ตาม

เงื่อนไขความสามารถของคะแนน x ($\xi_{x|\theta}$) เหมือนกับการแจกแจงความถี่ตามเงื่อนไขความสามารถของคะแนนแปลง x(y) ซึ่งเขียนเป็นสัญญลักษณ์ได้ว่า

$$f_{x|\theta} \equiv f_{x(y)|\theta}$$

เมื่อ x(y) เป็นฟังชั่นหนึ่งต่อหนึ่งของ y และกำหนดให้ความแปรปรวนของการแจกแจงตามเงื่อนไขเท่ากันด้วย (Lord 1980: 195-6) เมื่อวิเคราะห์จากนิยาม พบว่า การเทียบมาตราดังกล่าวจะทำได้เมื่อข้อมูลคะแนนมีความสมบูรณ์ในการระบุความสามารถที่แท้จริงของบุคคลได้ ดังนั้นเมื่อมีการพัฒนาทฤษฎีการวัดความสามารถแฝง และจำเพาะลงมาถึงการตอบข้อสอบ (Item response theory) ซึ่งสามารถหาค่าประมาณความสามารถจริงของบุคคลได้จากการตอบข้อสอบจึงเป็นขั้นการพัฒนาทฤษฎีของการเทียบมาตราอีกระดับหนึ่ง การเทียบมาตราในระดับนี้ เรียกว่าการเทียบมาตราคะแนนจริง (True-score equating) โดยทฤษฎีการเทียบคะแนนจริงมีข้อกำหนด ๓ ประการ (Lord 1980: 199) คือ

(๑) ความเสมอภาค (equity) อธิบายไว้ว่า ถ้าหากพิจารณาที่ระดับความสามารถ (θ) ใด ๆ การแจกแจงความถี่อย่างมีเงื่อนไขของคะแนนที่แปลงด้วยเทคนิคการเทียบมาตรา IRT จะต้องเหมือนกับการแจกแจงความถี่อย่างมีเงื่อนไขของคะแนนจากแบบสอบที่ต้องการเทียบ เขียนในรูปของสมการได้ดังนี้

$$f_{x|\theta} \equiv f_{x(y)|\theta}$$

เมื่อ x(y) เป็นฟังชั่นหนึ่งต่อหนึ่งของ y

(๒) ความไม่แปรเปลี่ยนเมื่อเปลี่ยนกลุ่ม (Invariance across groups) คะแนนแปลง x(y) จะเหมือนกันโดยไม่ขึ้นกับตัวแปรด้านอื่น ๆ ของประชากรที่นำมาสร้างสมการสำหรับการเทียบมาตรา

(๓) ความสมมาตร (Symmetry) หมายถึง การเทียบมาตราคะแนนจะต้องเหมือนกัน ไม่ว่าการเทียบนั้นจะเทียบจากแบบสอบ X ไปหาแบบสอบ Y หรือแบบสอบ Y ไปหาแบบสอบ X

ข้อกำหนดที่สำคัญนี้ โดยทั่วไปจะหาไม่พบในการแปลงคะแนนด้วยคะแนนสอบที่มีความคลาดเคลื่อนรวมอยู่ (fallible scores) นอกจากจะต้องทำการเทียบมาตราคะแนนด้วยคะแนนจริง (true scores) เท่านั้น

๑. <u>สรุปนิยามการเทียบมาตรา</u> (Marco 1981)

นิยามที่ ๑ คะแนนจากแบบสอบ X และแบบสอบ Y นำมาเทียบมาตรากันได้ ถ้า $M_{y'} = M_x$ และ

$SD_{y'} = SD_x$ สำหรับประชากร P

เมื่อ M และ SD คือ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน

y' คือ คะแนนแปลงที่ได้จากฟังชั่นการเทียบมาตรา

$e_X(Y)$

ในกรณีเช่นนี้ รูปแบบการเทียบมาตราเชิงเส้นตรงสามารถแปลงคะแนน Y ไปสู่มาตราของ X อย่างเพียงพอ ด้วยสมการ

$$y' = (SD_x/SD_y)\, y + M_x - (SD_x/SD_y)\, M_y$$

โดยทั่วไป M_y ไม่เท่ากับ M_x และ SD_y ไม่เท่ากับ SD_x เพราะว่า มีความหมายแตกต่างกันในความยากของคำถามในแบบสอบต่างชุดกัน การแปลงเชิงเส้นตรงจะปรับ M_y ให้เท่ากับ M_x และ SD_y เท่ากับ SD_x และเป็นการอ้างถึงการจัดการให้คะแนนเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานให้เหากัน นิยามนี้ไม่มีข้อกำหนดว่า แต่ละรายบุคคลจะต้องมี y เท่ากับ x เพียงแต่กำหนดให้คะแนนเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานจะต้องเท่ากันในประชากรที่ศึกษา และโดยนิยามนี้ไม่มีข้อกำหนดว่า แบบสอบ X และ Y จะต้องวัดในสิ่งเดียวกัน เช่น ความรู้หรือความสามารถ หรือทักษะต่าง ๆ

นิยามที่ ๒ คะแนนจากแบบสอบ X และแบบสอบ Y นำมาเทียบมาตรากันได้ ถ้า y และ x มีตำแหน่งเปอร์เซนไทล์เดียวกันในการแจกแจงของคะแนน X และ Y ในประชากร P

นิยามนี้ตรงกับนิยามรูปแบบอีควิเปอร์เซนไทล์ ฉะนั้นใช้วิธีคำนวณตำแหน่งเปอร์เซนไทล์ของคะแนนการแจกแจงใน X และ Y แล้วหาตำแหน่งคะแนนที่สมนัยกัน ในทางปฏิบัติการแจกแจงของ X และ Y จะถือเสมือนหนึ่งเป็นการแจกแจงต่อเนื่องมากกว่าเป็นคะแนนขาดตอน และตำแหน่งเปอร์เซนไทล์ใช้การคำนวณด้วย linear interpolation การแจกแจงน่าจะปรับเสียก่อนการคำนวณ นิยามนี้ไม่ได้มีข้อกำหนดว่า X และ Y ต้องวัดในสิ่งเดียวกัน

นิยามที่ ๓ คะแนนจากแบบสอบ X และแบบสอบ Y จะนำมาเทียบมาตรากันได้ เมื่อ $M_{y'} = M_x$ สำหรับประชากรกลุ่มย่อยที่มีความสามารถ ความรู้ และทักษะที่ทำการวัดอยู่ในระดับเดียวกัน นั่นคือ

$$M_{y'|a} = M_{x|a}$$

a คือ ระดับความสามารถที่กล่าวถึง

ฟังชั่นการเทียบมาตราที่เป็นไปตามนิยาม ต้องเป็นฟังชั่นที่ไม่ใช่เส้นตรง (curvilinear) การใช้รูปแบบเชิงเส้นตรงอาจให้การประมาณที่เป็นประโยชน์ได้ ส่วนรูปแบบอีควิเปอร์เซนไทล์อาจให้ผลลัพธ์ที่คงเส้นคงวาตามนิยาม หรืออาจไม่เป็นเช่นนั้นก็ได้ แต่ที่ชัดเจน คือ ต้องใช้วิธีการที่สลับซับซ้อนยิ่งขึ้น จึงเพียงพอที่จะให้เกิดการแปลงที่ให้คะแนนสมมูลได้ เช่น วิธีที่อาศัยรูปแบบอิงทฤษฎีการตอบข้อสอบ (Lord 1980, chapter 13) ซึ่งจะให้ค่าประมาณของระดับความสามารถที่อยู่ภายในกลุ่ม นิยามที่ ๓ นี้ ได้กำหนดให้ X และ Y วัดในสิ่งเดียวกัน วิธีการทางสถิติที่จะให้ผลลัพธ์อย่างน่าพอใจตามนิยามก็ต่อเมื่อได้เป็นไปตามข้อตกลงเบื้องต้นของนิยาม ไม่มีข้อกำหนดใด ๆ ที่เกี่ยวกับความคมชัดของแบบสอบ ฉะนั้นจึงสามารถใช้เทียบมาตรากับแบบสอบที่มีความเที่ยงต่างกันได้

นิยามที่ ๔ คะแนนจากแบบสอบ X และแบบสอบ Y นำมาเทียบมาตรากันได้ เมื่อ $M_{y'} = M_x$ ในทุก ๆ กลุ่มย่อยที่มีความสามารถระดับเดียวกันในประชากร P นั่นคือ

$$M_{y'|a} = M_{x|a}$$

และคะแนนความคลาดเคลื่อนมาตรฐานของการวัด y' และ x เท่ากันในประชากร P

นิยามนี้เพิ่มข้อกำหนดจากนิยามที่ ๓ อีกหนึ่งข้อ คือ แบบสอบ X และ Y ต้องมีความแม่นยำในการวัดเท่ากัน ซึ่งหมายถึงมีค่าความเที่ยงเท่ากัน ในกรณีนี้แบบสอบ X และ Y ควรประกอบด้วยข้อสอบจำนวนที่เท่ากัน

นิยามที่ ๕ คะแนนจากแบบสอบ X และแบบสอบ Y นำมาเทียบมาตรากันได้ เมื่อ $M_{y'} = M_x$ และ $SD_{y'} = SD_x$ สำหรับทุก ๆ กลุ่มย่อยที่มีระดับความสามารถเดียวกันในประชากร P นั่นคือ

$$M_{y'|a} = M_{x|a} \quad \text{และ}$$
$$SD_{y'|a} = SD_{x|a} \quad \text{สำหรับทุก ๆ } a$$

นิยามนี้ยากแก่การจัดการให้มีข้อกำหนดตามที่ระบุไว้ เพราะไม่ใช่เฉพาะแต่ความคลาดเคลื่อนมาตรฐานโดยส่วนรวมจะต้องเท่ากัน แต่หมายถึงความคลาดเคลื่อนมาตรฐานของทุก ๆ กลุ่มย่อยต้องเท่ากันหมดด้วย โดยทางปฏิบัตินิยามนี้ไม่มีทางทำได้เลย นอกจากต้องอาศัยการดำเนินการสอบด้วยคอมพิวเตอร์ และแต่ละคนจะต้องได้รับจำนวนคำถามที่มากพอที่จะให้การประมาณที่แม่นยำเพื่อให้เป็นไปตามนิยาม

นิยามที่ ๖ คะแนนจากแบบสอบ X และแบบสอบ Y จะเทียบมาตรากันได้ ถ้าการแจกแจงความถี่ของ y' และ x เหมือนกันในแต่ละกลุ่มตัวอย่างย่อยที่มีความสามารถอยู่ในระดับเดียวกัน

นิยามที่ ๖ นี้ มีข้อกำหนดมากขึ้นกว่านิยามที่ ๕ ในเรื่องของรูปร่างของการแจกแจงของคะแนน y อย่างมีเงื่อนไขที่ระดับความสามารถใด ๆ ทั้งนี้ย่อมหมายถึงว่า $\mu_{y|a}$ และ $SD_{y|a}$ จะต้องเข้าชุดกับรูปร่างของการแจกแจงอย่างมีเงื่อนไขของ X ลอร์ด (Lord 1980) ได้แสดงให้เห็นว่า นิยามนี้เป็นไปได้ต่อเมื่อคำถามในแบบสอบ Y มีความเท่าเทียมหรือสมมูลในเชิงหน้าที่กับคำถามในแบบสอบ X เท่านั้น หรือว่าแบบสอบทั้งสองฉบับสามารถให้ผลการวัดเป็นคะแนนสมบูรณ์ (perfectly reliable scores) ในกรณีเช่นนี้ $y = x$ นิยามนี้ดูเหมือนจะไม่มีประโยชน์ แต่เป็นนิยามในเชิงทฤษฎีที่ต้องการให้มีความหมายของความไม่แตกต่างของการสอบ X หรือ Y ของบุคคลใด ๆ อย่างแท้จริง

<u>การออกแบบการรวบรวมข้อมูลคะแนน</u>

การเทียบมาตราเป็นวิธีการเชิงประจักษ์ เพื่อกำหนดการแปลงคะแนนที่ได้จากแบบสอบชุดหนึ่งไปสู่แบบสอบอีกชุดหนึ่งในความหมายที่เท่าเทียมกัน ลักษณะของกระบวนการเชิงประจักษ์นี้จึงเกี่ยวข้องกับการออกแบบเพื่อรวบรวมข้อมูล และจัดกระทำในทางสถิติเพื่อระบุการแปลงคะแนนในที่สุด (Marco 1981) แองกอฟ (Angoff 1984: 94-123) ได้บรรยายทั้งสองประเด็นไว้อย่างครอบคลุมในแบบการเทียบมาตรา ๖ แบบ (designs) มาร์โคได้นำมาบรรยายซ้ำ แต่ได้แยกประเด็นของการออกแบบเพื่อรวบรวมข้อมูลกับการจัดกระทำทางสถิติออกจากกัน ทั้งนี้เพราะวิธีการทางสถิติบางวิธีใช้ได้กับแบบแผนการรวบรวมข้อมูลมากกว่าหนึ่งแบบ ซึ่งสรุปสาระสำคัญดังนี้ คือ

แบบแผนที่ ๑ กลุ่มเดียว ทำแบบสอบทั้งสองชุด คือ ชุดใหม่และชุดเก่า การทดสอบอาจจัดทำในวันเดียว หรือ ๒ วันติดต่อกัน แต่หลักการที่สำคัญ คือ ช่วงระยะห่างของการทำแบบสอบ ๒ ชุด ต้องเป็นเวลาสั้น ๆ เพื่อมิให้เกิดประสบการณ์แทรกซ้อนที่มีผลกระทบต่อคะแนน ลำดับการทดสอบชุดใดก่อนไม่ใช่ปัญหา เพราะถือว่าองค์ประกอบต่าง ๆ อันได้แก่ การเรียนรู้ การฝึกฝน ความล้า มีผลต่อคะแนนน้อยมาก

แบบแผนที่ ๒ กลุ่มสุ่มสองกลุ่ม แต่ละกลุ่มได้รับการทดสอบทั้งสองชุดในลักษณะการจัดลำดับการสอบก่อนหลังสลับให้เกิดความสมดุลย์กัน แบบแผนที่ ๒ ดัดแปลงมาจากแบบแผนที่ ๑ เป็นแบบแผนที่จะรับประกันอิทธิพลขององค์ประกอบการเรียนรู้ การฝึกฝน และความล้า เป็นการขจัดความลำเอียงที่จะเกิดขึ้นกับผลการสอบชุดใดชุดหนึ่ง

แบบแผนที่ ๓ กลุ่มสุ่มสองกลุ่ม ให้แต่ละกลุ่มทำการสอบเพียงชุดเดียว สิ่งสำคัญของแบบแผนที่ ๓ คือ ต้องมีกลุ่มสุ่ม ๒ กลุ่ม ที่มีความคล้ายคลึงกันทั้งทางด้านความรู้ ความสามารถ หรือทักษะที่ต้องการวัดเพื่อให้คะแนนสอบที่ได้มาไม่ใช่เป็นผลของความแตกต่างของความสามารถของกลุ่ม แบบแผนนี้ได้เคยนำไปใช้กับการทดสอบ the Graduate Management Adminission Test, the Law School Admission Test และ the GRE Aptitude Test

แบบแผนที่ ๔ กลุ่มสุ่มสองกลุ่ม แต่ละกลุ่มทำแบบสอบเพียงชุดเดียว และทำแบบสอบร่วมเหมือนกันอีกส่วนหนึ่ง แบบแผนนี้มีข้อได้เปรียบกว่าแบบแผนที่ ๓ ในประเด็นที่ว่า ได้ข้อสนเทศจากแบบสอบร่วมจากกลุ่มสุ่มทั้งสอง แบบสอบร่วมอาจเป็นฉบับแยกจากแบบสอบทั้งสองชุด (external) หรือเป็นส่วนที่ผนวกเข้าในแบบสอบทั้งสองชุด (internal) ก็ได้ ส่วนที่เป็นแบบสอบร่วมควรประกอบด้วยคำถามที่คล้ายคลึงกับคำถามในแบบสอบที่ต้องการเทียบมาตราทั้งสองชุด ประโยชน์จากแบบสอบร่วมที่มีต่อการเทียบมาตราของแบบสอบสองชุด จะมีมากหรือน้อยขึ้นอยู่กับความสัมพันธ์ของแบบสอบร่วมกับแบบสอบสองชุดนั้น เหตุผลที่แบบแผนที่ ๔ มีข้อได้เปรียบ เพราะคะแนนชุดเก่าและใหม่สามารถรับการปรับเพื่อสะท้อนให้เห็นความแตกต่างที่ปรากฏอยู่ในกลุ่มสุ่มทั้งสองตามผลของแบบสอบร่วมได้

แบบแผนที่ ๕ กลุ่มไม่ได้จากการสุ่มสองกลุ่ม แต่ละกลุ่มทำแบบสอบเพียงชุดเดียว และทำแบบสอบร่วมเหมือนกันอีกส่วนหนึ่ง เช่นเดียวกับแบบแผนที่ ๔ กลุ่มผู้สอบที่ไม่ได้มาจากการสุ่มมักจะเกิดขึ้นกับการสอบตามธรรมชาติ ปัจจัยสำคัญของแบบแผนนี้อยู่ที่แบบสอบร่วม ซึ่งต้องให้มีความคล้ายคลึงมากที่สุดกับแบบสอบสองชุดที่ต้องการทำการเทียบมาตรา โดยปกติแล้วไม่มีวิธีการทางสถิติใดที่จะทำการปรับคะแนนในกรณีที่กลุ่มที่ไม่ได้มาจากการสุ่ม แต่จะทำได้ก็ต่อเมื่อมีแบบสอบร่วมที่มีความเป็นคู่ขนานกับแบบสอบชุดใหม่และเก่า หน้าที่ของแบบสอบร่วม คือ กำหนดจุดอ้างอิงให้เกิดขึ้นกับทั้งสองชุด

ในทำนองกับการกำหนดจุดเดือด หรือจุดเยือกแข็งของน้ำ เพื่อใช้กำหนดมาตราบนเทอร์โมมิเตอร์ ฟาร์เรนไฮท์ให้เท่ากับองศาบนเทอร์โมมิเตอร์แบบเซนติเกรดได้ แบบแผนที่ ๕ นี้ได้นำไปใช้กับการเทียบมาตราในแบบสอบ SAT มาแล้ว

ประเด็นของวิธีการทางสถิติที่ใช้ในการเทียบมาตรา โดยหลักทั่วไป แบบแผนที่สามารถได้ข้อมูลคะแนนสอบทั้งสองชุดจากผู้สอบคนเดียวกันแล้ว การหาคะแนนสมมูลทำได้ด้วยรูปแบบของการเทียบที่ตำแหน่งเปอร์เซนไทล์ และรูปแบบการเทียบเชิงเส้นตรงแบบง่าย คือ เทียบด้วยคะแนนมาตรฐาน เช่น แบบแผนที่ ๑ ๒ และ ๓ ส่วนในกรณีที่ผู้สอบคนเดียวทำเพียงชุดเดียว จะมีปัญหาในการประมาณค่าคะแนนสอบของคนทั้งหมดในการทำแบบสอบทั้งสองชุด ปัญหานี้เมื่อใช้แบบแผนที่ ๔ และ ๕ จะช่วยขจัดได้ วิธีทางสถิติที่ใช้ คือ การวิเคราะห์การถดถอย (regression analysis) และการประมาณด้วยความเป็นไปได้สูงสุด (Maximum likelihood) หลังจากได้ค่าประมาณผลสอบของคนทั้งหมดในแบบสอบทั้งสองชุดแล้ว ก็ทำการหาคะแนนสมมูลได้ (Marco 1981)

จากแบบแผนการเก็บรวบรวมข้อมูล และการใช้สถิติในการเทียบมาตรานี้ พบว่า แบบแผนที่สำคัญที่ก่อให้เกิดการประยุกต์ใช้ได้อย่างมาก คือ แบบแผนที่มีการใช้แบบสอบร่วม ซึ่งยังคงใช้อยู่ในโครงการทดสอบของ SAT และสอดคล้องกับเงื่อนไขในสภาพแวดล้อมปัจจุบันด้วย

<u>วิธีเทียบมาตรารูปแบบอีควิเปอร์เซนไทล์</u>

๑. <u>วิธีการทั่วไป</u>

มโนทัศน์ของการเทียบมาตรารูปแบบอีควิเปอร์เซนไทล์ เริ่มจากการแจกแจงของคะแนนจากแบบสอบชุด X และชุด Y ที่มีลักษณะคล้ายกันหรือถ้ามีความแตกต่างเกิดขึ้นบ้าง ก็มีเพียงเล็กน้อย เช่น ต่างกันที่ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน การเทียบหาคะแนนสมมูลทำได้โดยใช้คะแนน ณ ตำแหน่งเปอร์เซนไทล์เดียวกันของคะแนน ๒ ชุดนั้น ผลการเทียบมาตราแสดงด้วยกราฟ ขั้นตอนการแปลงคะแนนมีดังนี้ คือ เลือกกลุ่มผู้สอบที่มีความสามารถกระจาย มีทั้งเก่ง ปานกลาง และอ่อน แล้วแบ่งเป็นกลุ่มย่อยโดยสุ่ม ๒ กลุ่ม ให้กลุ่มหนึ่งทำแบบสอบ X และอีกกลุ่มทำแบบสอบ Y

ผลการสอบนำมาทำการแจกแจงคะแนน X และ Y คำนวณหาจุดกลางเปอร์เซนไทล์ของแต่ละการแจกแจง อ่านและทำเครื่องหมายสำหรับค่า X และ Y จากการแจกแจงคะแนนที่สมนัยกันบนกระดาษกราฟ ทำจุดบนกระดาษประมาณ ๓๐ จุด แล้วลากเส้นเชื่อมจุดจะเกิดเป็นเส้นกราฟ ทำการปรับเส้นให้เรียบ เส้นกราฟนี้จะใช้อ่านค่า x ที่สมนัยกับ y หรืออ่านค่า y ที่สมนัยกับ x ก็ได้ จากนั้นทำตารางสำเร็จเพื่ออ่านค่าคะแนนแปลง โดยปกติการเทียบมาตรารูปแบบอีควิเปอร์เซนไทล์ ให้ภาพของการแปลงที่สะท้อนถึงระดับความยากง่ายของแบบสอบสองชุด ถ้าแบบสอบสองชุดมีความยากใกล้เคียงกัน เส้นกราฟจะมีลักษณะใกล้เคียงเส้นตรง แต่ถ้าแบบสอบมีความยากต่างกัน เส้นกราฟจะเป็นเส้นโค้ง (curvilinear) คะแนนสมบูลที่เกิดขึ้นจะถูกยึดหรือหย่อนคะแนนคืน เพื่อให้คงรักษามาตราให้เหมือนชุดก่อนตามต้องการ (Angoff 1984: 97-101)

๒. <u>การเทียบมาตราโดยใช้แบบสอบร่วม</u> (CIE)

บรูน และ ฮอลแลนด์ (Braun and Holland 1982: 19-22) ได้แสดงความสัมพันธ์ของแบบสอบร่วมในกระบวนการเทียบมาตรารูปแบบอีควิเปอร์เซนไทล์ไว้ดังนี้ ให้ T เป็นประชากรสังเคราะห์ ได้มาจากประชากร P และประชากร Q เขียนเป็นสมการแทนได้ว่า

$$T = fP + (1-f)Q$$

โดยที่ P และ Q รวมกันด้วยสัดส่วนน้ำหนัก f และ (1-f) ตามลำดับ การคำนวณการแจกแจงของ T ทำได้โดยคำนวณการแจกแจงตามเงื่อนไข (condition distributions) ของ P และ Q ก่อนแล้วหาค่าเฉลี่ยที่ถ่วงน้ำหนักด้วย f และ (1-f) สมการปกติ คือ

$$f = N_1 / (N_1 + N_2)$$

เมื่อ N_1 และ N_2 เป็นขนาดกลุ่มตัวอย่างของ P และ Q ตามลำดับ ค่าของ f นี้ เป็นค่าเฉลี่ยของ P และ Q

ในรูปแบบการเทียบมาตราอีควิเปอร์เซนไตล์ ใช้ข้อตกลง (assumption) น้อยที่สุด วัตถุประสงค์ คือ หาค่าประมาณ $F_T(X)$ และ $G_T(Y)$ ซึ่งเป็นการแจกแจงของแบบสอบ X และ Y ในประชากรสังเคราะห์ T นั่นเอง ข้อมูลปริมาณที่ต้องนำมาวิเคราะห์ให้นิยามไว้ดังนี้

$F_P(x|v)$ เป็นฟังชั่นการแจกแจงตามเงื่อนไขของ x เมื่อกำหนด v ; (v in P)

$F_Q(x|v)$ " " x v ; (v in Q)

$G_P(y|v)$ " " y v ; (v in P)

$G_Q(y|v)$ " " y v ; (v in Q)

$K_P(v)$ เป็นฟังชั่นการแจกแจงของ v ในประชากรย่อย P

$K_Q(v)$ " " v " " Q

$K_T(v)$ " " v ในประชากรสังเคราะห์ T

ดังนั้น จากนิยาม $K_T(v) = fK_P(v) + (1-f)K_Q(v)$ แต่ $F_Q(x|v)$ และ $G_P(y|v)$ ไม่สามารถประมาณได้จากข้อมูลจริงเลย จึงจำเป็นต้องระบุเป็นข้อตกลงที่ไม่มีการทดสอบ ดังนี้

$$F_T(x) = \int F_P(x|v)dK_P(v)f + \int F_Q(x|v)dK_Q(v)(1-f)$$

และ

$$G_T(y) = \int G_P(y|v)dK_P(v)f + \int G_Q(y|v)dK_Q(v)(1-f)$$

การอินทีเกรตของสมการข้างต้นนี้ ความจริงแล้วเป็นผลรวม เพราะทุก ๆ การแจกแจงเป็นจำนวนเลขไม่ต่อเนื่อง แต่การเขียนอย่างนี้รัดกุมมากกว่า และเพื่อให้เกิดความง่ายขึ้น จึงกำหนดข้อตกลงว่า การแจกแจงของ X เมื่อกำหนด $V = v$ เหมือนกันในประชากรย่อย P และ Q และการแจกแจงของ Y เมื่อกำหนด $V = v$ เหมือนกันทั้งใน P แล้ว Q ฉะนั้นรูปสมการที่ง่ายขึ้น คือ

ถ้า $F_P(x|v) = F_Q(x|v)$

และถ้า $G_P(y|v) = G_Q(y|v)$ แล้ว

$$F_T(x) = \int F_P(x|v)dK_T(v)$$

และ

$$G_T(y) = \int G_Q(y|v)dK_T(v)$$

ซึ่งหมายความว่า $F_T(X)$ และ $G_T(Y)$ หาได้ในเทอมของปริมาณโดยประมาณจากข้อมูลที่รวบรวมได้ ฟังชั่นเทียบมาตราจะเป็นดังนี้

$$\hat{e}_X(Y) = \hat{F}_T^{-1}(\hat{G}_T(Y))$$

๓. <u>ประเด็นอภิปรายในการใช้รูปแบบอีควิเปอร์เซนไทล์</u>

การใช้รูปแบบอีควิเปอร์เซนไทล์ มีประเด็นที่เป็นข้ออภิปรายเสมอ คือ เทคนิคของการเกลาหรือปรับเส้นให้เรียบ แองกอฟ (Angoff 1984: 11-12) ได้ให้รายละเอียดการเกลาเส้นด้วยมือ โดยให้เขียนจุดสร้างกราฟลงบนกระดาษกราฟชนิดที่เป็นความน่าจะเป็นปกติ (normal probability paper) ซึ่งจะทำให้การแจกแจงที่เป็นโค้งปกติมีภาพเป็นเส้นตรง และง่ายต่อการเกลาด้วยมือ อุปกรณ์ที่ช่วยในการลากเส้น ได้แก่ กระดูกงู หลักการเกลาโดยทั่วไป คือ ให้ลากเส้นผ่านจุดต่าง ๆ ที่ไม่อยู่ในลู่เดียวกันในลักษณะให้เฉลี่ยความใกล้เคียงกันทั้งสองข้าง ความชำนาญของผู้ปฏิบัติจะสามารถมองเห็นความผิดปกติของเส้นได้ นอกจากการเกลาด้วยมือแล้ว มีเทคนิคการเกลาอีกกลุ่มหนึ่ง คือ เทคนิควิธีเชิงวิเคราะห์ซึ่งจะทำการวิเคราะห์ก่อนเขียนกราฟ วิธีเชิงวิเคราะห์เป็นวิธีเกลาเส้นอย่างเป็นปรนัย ไม่เกิดความลำเอียงดังเช่นวิธีใช้มือ วิธีเชิงวิเคราะห์มีหลายวิธี เช่น วิธีของเคอร์ทัน และ ทูกี ใช้วิธีการทำให้จุดต่าง ๆ ที่อยู่กันเป็นชุดเฉลี่ยให้มีแนวโน้มเป็นพาราโบลิคหรือคิวบิค (Cureton and Tukey 1951 cited by Angoff 1984: 12) ถ้าการแจกแจงเชิงทฤษฎีกับการแจกแจงเชิงประจักษ์สอดคล้องกัน แองกอฟได้แนะนำให้ใช้วิธีการ (the negative hypergeometric (Angoff 1984: 13)) นอกจากนี้มีวิธีการปรับเกลาเทียบคะแนนเชิงเส้นตรง (smoothing linearly interpolated equated score) ซึ่งลินเช และ ปริชาด

(Limdsay and Prichard 1971 cited by Kolen 1984) เสนอไว้ แต่วิธีนี้มีข้อจำกัดสองประการ คือ ผลการเทียบให้คะแนนแปลงที่ขาดคุณสมบัติของความสมมาตร และไม่สามารถระบุความคลาดเคลื่อนมาตรฐานของคะแนนเทียบเมื่อเปลี่ยนจุด โคเลน (Kolen 1984) จึงได้ทำการศึกษาแก้ไขจุดอ่อนโดยใช้เทคนิค Cubic splines ซึ่งเป็นวิธีทางคณิตศาสตร์ ตามปกติ Cubic splines เป็นเครื่องมือที่ใช้ปรับเส้นโค้งธรรมดาที่ใช้ในการเขียนแบบ ปัจจุบันได้พัฒนาไปใช้กับสถานการณ์ต่าง ๆ ที่ต้องการทำเส้นโค้งให้พิดกับข้อมูล การประเมินประสิทธิภาพของการทดลองใช้วิธีนี้ตัดสินโดยเปรียบเทียบผลลัพธ์จากการปรับอีควิเปอร์เซนไทล์กับผลจากการเทียบมาตราด้วยวิธีอื่น ๆ จากการวิเคราะห์กลุ่มสอบทานผล ข้อมูลได้จากกลุ่มตัวอย่างสุ่มสมมูลประมาณ ๓๐๐๐ คนต่อการทำแบบสอบ ๑ ชุด จากแบบสอบทั้งหมด ๔ ชุด ผลการศึกษา พบว่า การเทียบมาตราโดยทำการปรับเส้นด้วยวิธีเชิงวิเคราะห์ตามเทคนิคของ cubic splines เป็นวิธีที่มีความเพียงพอสูงสุดในการเทียบมาตราของแบบสอบ AAP โดยเฉพาะอย่างยิ่งกับชุดแบบสอบที่มีความคล้ายคลึงน้อยที่สุด

ลอร์ด (Lord 1982a) ยอมรับว่า การปรับหรือเกลาเส้นโค้งของการเทียบมาตราในรูปแบบอีควิเปอร์เซนไทล์ช่วยลดความคลาดเคลื่อนเชิงสุ่มได้ส่วนหนึ่ง แต่มีโอกาสทำให้เกิดความลำเอียงได้เช่นกัน และมีความเห็นว่า การระบุความคลาดเคลื่อนของคะแนนแปลงที่ปรับแล้วเป็นงานหนักเกินความจำเป็น

รูปแบบอีควิเปอร์เซนไทล์ มีข้อจำกัดที่ต้องคำนึงอยู่หลายประการ คือ

(๑) ในการเขียนเส้นกราฟที่แสดงคะแนนสมมูลที่ยังไม่ได้ปรับหรือเกลาให้เรียบจะมีความคลาดเคลื่อนมาก แต่เทคนิคการปรับเส้นเท่าที่ทำกันมานั้นยังไม่มีวิธีใดที่ให้หลักประกันผลได้อย่างน่าเชื่อถือ ลักษณะเช่นนี้ยังก่อให้เกิดความลำเอียงที่ยังต้องหาทางปรับปรุงอีก (Angoff 1984: 97; Potthoff 1982: 209)

(๒) รูปแบบอีควิเปอร์เซนไทล์ มีความไวต่อความแปรปรวนเชิงสุ่มมาก โดยเฉพาะขนาดตัวอย่างที่มีขนาดเล็ก (Angoff 1984: 97; Potthoff 1982: 210)

(๓) ในกรณีที่แบบสอบสองชุดมีความเที่ยงต่างกันมาก ผลของการเทียบมาตราขาดความคงที่ (Potthoff 1982: 210)

(๔) การเทียบคะแนนสมมูลของแบบสอบสองชุดทำได้เฉพาะในช่วงพิสัยของคะแนนที่มีความถี่ของคะแนนสังเกตเท่านั้น ส่วนที่อยู่นอกพิสัยดังกล่าวจะมีความคลาดเคลื่อนของการเทียบมาตราสูงมาก (Angoff 1984: 97)

<u>วิธีเทียบมาตราแบบเชิงเส้นตรง</u>

รูปแบบการเทียบมาตราเชิงเส้นตรง เป็นรูปแบบที่ได้รับความนิยมมากมาเป็นเวลานาน ได้ใช้ในโครงการทดสอบระดับชาติหลายโครงการ เช่น การทดสอบความถนัดเชิงวิชาการ (SAT) (Donlon and Angoff 1971: 32) เริ่มด้วยการเทียบมาตราคะแนนชุดต่าง ๆ ที่ใช้เวลาต่อ ๆ มากับแบบสอบฉบับต้นแบบที่ใช้เมื่อมิถุนายน ๑๙๔๑ ตัวอย่างการเทียบมาตราที่จัดกระทำเมื่อปี ค.ศ. ๑๙๔๑ ได้ดำเนินการดังนี้ คือ ออกแบบเพื่อการเทียบมาตราด้วยการใช้ข้อสอบจำนวนหนึ่งที่เป็นของชุดที่ใช้เมื่อเมษายน ๑๙๔๑ เข้าไปรวมอยู่ในชุดใหม่ที่ใช้เมื่อมิถุนายน ๑๙๔๑ คะแนนเฉลี่ยและคะแนนเบี่ยงเบนมาตรฐานของกลุ่มข้อสอบจำนวนนี้ที่ร่วมกัน (the carried over items หรือ common items) ได้ปรากฏว่า แตกต่างกันในกลุ่มผู้สอบในเดือนเมษายนและมิถุนายน ความแตกต่างของค่าสถิติทั้งสองนี้ได้นำไปใช้ปรับค่าสถิติอื่น ๆ ของกลุ่มผู้สอบเหล่านั้นที่รับการทดสอบจากแบบสอบต่างชุดกัน หลังจากการปรับได้ทำให้เกิดคะแนนเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของคะแนนดิบที่ปรับแล้วของชุดใหม่ที่เท่ากับคะแนนเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของคะแนนมาตราของชุดเก่า ด้วยวิธีการเช่นนี้ทำให้เกิดการแปลงคะแนนเชิงเส้นตรง (a linear conversion) จากคะแนนดิบไปสู่คะแนนมาตราสำหรับแบบสอบชุดใหม่ได้ ในการสอบเมื่อเดือนเมษายน ค.ศ. ๑๙๔๑ ได้ดำเนินการเทียบด้วยวิธีการเดียวกัน

๑. <u>วิธีเทียบโดยใช้แบบสอบร่วม</u> (CIE)

ผู้ที่คิดการเทียบมาตราคะแนนโดยใช้แบบสอบร่วมในรูปแบบเชิงเส้นตรง คือ เลดยาร์ด ทักเกอร์ (Ledyard Tucker) และได้ตั้งชื่อวิธีการเทียบมาตรานี้ว่า Tucker equating เพื่อเป็นเกียรติแก่ผู้ริเริ่ม (Angoff 1961 cited by Donlon and Angoff 1971: 38; Braun and Holland 1982: 23) วิธีนี้เริ่มต้นด้วยการตั้งข้อตกลงเบื้องต้นว่า การเทียบมาตราคะแนนมีพังชั่นเป็นเส้นตรงดังนี้

$$e_X(Y) = \mu_X + \frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) \quad \text{----------(๔)}$$

เมื่อ μ_X σ_X μ_Y และ σ_Y เป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของประชากร T ทุกคนทำแบบสอบชุด X และ Y ฟังชั่นค่าเฉลี่ยของประชากรในแบบสอบชุด X คือ

$$\mu_X = \int E_P(X|v)dK(v)f + \int E_Q(X|v)dK_Q(v)(1-f) \quad \text{-------(๕)}$$

และมีค่าความแปรปรวนของ X คือ

$$\sigma_X^2 = \int Var_P(X|v)dK_p(v)f + \int Var_Q(x|v)dK_Q(v)(1-f)$$
$$+ \int E_P^2(X|v)dK_p(v)f + \int E_Q^2(X|v)dK_Q(v)(1-f) - \mu_X^2$$
$$\text{----------(๓)}$$

สำหรับฟังชั่นค่าเฉลี่ยและความแปรปรวนของประชากรในแบบสอบชุด Y เป็นในทำนองเดียวกัน ซึ่งสัญญลักษณ์ต่าง ๆ มีความหมายแทนสิ่งต่อไปนี้ คือ

P คือ ประชากร P ที่ทำแบบสอบชุด X และแบบสอบร่วม V

Q คือ ประชากร Q ที่ทำแบบสอบชุด Y และแบบสอบร่วม V

T คือ ประชากรสังเคราะห์มาจากการรวม P และ Q

f เป็นสัดส่วนน้ำหนักของประชากร P ใน T

1-f เป็นสัดส่วนน้ำหนักของประชากร Q ใน T

E และ Var คือ ค่าคาดหมายและความแปรปรวนของประชากร

$K_P(v)$ และ $K_Q(v)$ เป็นฟังชั่นการแจกแจงของแบบสอบร่วม V ใน P และ Q ตามลำดับ

เทอมต่าง ๆ ในสมการ (๒) และ (๓) ที่สามารถหาค่าประมาณจากข้อมูล คือ

$$E_P(X|v) = \int x dF_P(x|v)$$

$$Var_P(X|v) = \int [x - E_P(X|v)]^2 \, dF_P(x|v)$$

และเป็นทำนองเดียวกันในเทอม $E_Q(X|v)$ และ $Var_Q(X|v)$

จากสมการข้างต้นจะพบว่า ค่า μ_X และ σ_X^2 หาได้ก็ต่อเมื่อได้กำหนดเงื่อนไขเป็นข้อตกลงเบื้องต้น ๒ ประการก่อน คือ

(๑) ค่าประมาณเฉลี่ยและค่าความแปรปรวนอย่างมีเงื่อนไขของ X เมื่อกำหนด v ในประชากรย่อย P เหมือนกับในประชากร Q และในกรณี Y เมื่อกำหนด v ในประชากรย่อย Q ย่อมเหมือนกับในประชากร P เขียนในรูปสูตรดังนี้

ถ้า $E_P(X|v) = E_Q(X|v)$ และ

ถ้า $Var_P(X|v) = Var_Q(X|v)$ แล้ว

$$\mu_X = \int E_P(X|v) dK_T(v) \quad \text{และ} \quad \text{----------} \quad (๔)$$

$$\sigma_X^2 = \int Var_P(X|v) dK_T(v) + \int E_P^2(X|v) dK_T(v) - \mu_X^2 \quad -- \quad (๕)$$

สำหรับค่า μ_Y และ σ_Y^2 หาได้ในทำนองเดียวกัน (๔) และ (๕)

(๒) เป็นข้อตกลงเบื้องต้นตามรูปแบบ (model assumption) เพราะเป็นการยอมรับการทำ $E_P(X|v)$ $E_Q(X|v)$ $Var_P(X|v)$ และ $Var_Q(X|v)$ ให้มีรูปที่ง่ายขึ้น ข้อตกลงเบื้องต้นเหล่านี้เป็นแบบฉบับที่กำหนดในการวิเคราะห์การถดถอยว่า ค่าคาดหมายอย่างมีเงื่อนไขเป็นเส้นตรง และความแปรปรวนอย่างมีเงื่อนไขเป็นค่าคงที่ ดังเช่น

$$E_P(X|v) = (a_{X|V.P})v + b_{X|V.P}$$

$$E_Q(Y|v) = (a_{Y|V.Q})v + b_{Y|V.Q}$$

$$Var_P(X|v) = \sigma^2_{X|V.P}$$

$$Var_Q(Y|v) = \sigma^2_{Y|V.P}$$

ผลจากการกำหนดข้อตกลงเบื้องต้น ๒ ประการที่กล่าวมานี้ สรุปเป็นหลักของการเทียบมาตราตามวิธีการของทักเกอร์ (Tucker equating) ได้ดังนี้

ถ้า $E_P(X|v) = E_Q(X|v) = (a_{X|V.P})\, v + b_{X|V.P}$

และ $Var_P(X|v) = Var_Q(X|v) = \sigma^2_{X|V.P}$ แล้วจะได้

$$\mu_X = (a_{X|V.P})\, \mu_{V.T} + b_{X|V.P} \text{ ---------- (๖)}$$

และ

$$\sigma^2_X = \sigma^2_{X|V.P} + (a_{X|V.P})^2\, \sigma^2_{V.T} \text{ -------- (๗)}$$

ในทำนองเดียวกัน สามารถหาค่า μ_Y และ σ^2_Y จาก (๖) และ (๗) ดังนี้

$$\mu_Y = (a_{Y|V.Q})\, \mu_{V.T} + b_{Y|V.Q} \text{ ---------- (๘)}$$

และ

$$\sigma^2_Y = \sigma^2_{Y|V.Q} + (a_{Y|V.Q})^2\, \sigma^2_{V.T} \text{ ------- (๙)}$$

ในการคำนวณหาค่าต่าง ๆ ตามข้อตกลงเบื้องต้นนี้ ให้ใช้ค่าจากกลุ่มตัวอย่างเพื่อหาค่าประมาณกำลังสองน้อยที่สุด (least squares estimates) ของ $a_{X|V.P}$ $b_{X|V.P}$ $\sigma^2_{X|V.P}$ $a_{Y|V.Q}$ $b_{Y|V.Q}$ และ $\sigma^2_{Y|V.Q}$ ข้อมูลทั้งหมดเมื่อนำมารวมเข้าด้วยกันจะประมาณค่า $\mu_{V.T}$ และ $\sigma^2_{V.T}$ จากนั้นนำค่าประมาณเหล่านี้แทนในสมการ (๑) จะได้ฟังชั่นของการเทียบมาตราเชิงเส้นตรงของ Y ไปสู่ X (Braun and Holland

1982: 23-24)

การหาค่าประมาณค่าเฉลี่ยและความแปรปรวนของประชากรของคะแนนจากชุด X และชุด Y ด้วยวิธีการของทักเกอร์ต้องประมาณการตอบของกลุ่มตัวอย่าง Q ในการทำแบบสอบ X และพิจารณาผู้สอบ ๒ กลุ่ม ที่มีความแตกต่างกันตามธรรมชาติ ต่อมาลอร์ด (Lord 1955) ได้คิดวิธีการหาค่าประมาณค่าเฉลี่ยและความแปรปรวนของประชากร โดยพิจารณากลุ่มผู้สอบเมื่อได้รับการสุ่มมาจากประชากรเดียวกัน จะประมาณค่าการตอบหรือคะแนนของทั้งประชากรในแบบสอบทั้งสองชุด แทนที่จะประมาณเพียงกลุ่ม Q ในการตอบชุด X เท่านั้น วิธีการของลอร์ดที่นำมาใช้หาค่าประมาณ คือ วิธีการความเป็นไปได้สูงสุด (the maximum likelihood method) ได้สมการที่ใช้หาค่าประมาณดังนี้

$$\hat{\mu}_X = M_{X\alpha} + b_{XV\alpha}(\hat{\mu}_V - M_{V\alpha}) \quad -- (๑๐)$$

$$\hat{\mu}_Y = M_{Y\beta} + b_{YV\beta}(\hat{\mu}_V - M_{V\beta}) \quad -- (๑๑)$$

$$\hat{\sigma}^2_X = S^2_{X\alpha} + b^2_{XV\alpha}(\hat{\sigma}^2_V - S^2_{V\alpha}) \quad -- (๑๒)$$

$$\hat{\sigma}^2_Y = S^2_{Y\beta} + b^2_{YV\beta}(\hat{\sigma}^2_V - S^2_{V\beta}) \quad -- (๑๓)$$

เมื่อ

$$\hat{\mu}_V = M_{Vt}$$

$$\hat{\sigma}^2_V = S^2_{Vt}$$

และ

$$t = \alpha + \beta$$

ค่าประมาณเหล่านี้นำไปแทนในสมการที่ (๑) จะเป็นสมการเทียบมาตราเชิงเส้นตรง

ในการศึกษาครั้งนี้ ผู้วิจัยได้ใช้วิธีการสุ่มกลุ่ม ๒ กลุ่ม ให้ทำแบบสอบ X หรือ Y กลุ่มละแบบสอบเดียว และทั้ง ๒ กลุ่ม ได้ทำแบบสอบร่วม V ด้วย การเลือกวิธีการทางสถิติมาใช้หาค่าประมาณประชากรต่าง ๆ จึงได้พิจารณาเลือกใช้วิธีการความเป็นไปได้สูงสุด (the maximum likelihood) ตามที่ลอร์ดได้เสนอไว้ (Angoff 1984: 105; Lord 1955: 198)

๒. ประเด็นอภิปรายในการใช้รูปแบบเชิงเส้นตรง

โดยทั่วไปการเทียบมาตราเชิงเส้นตรง เป็นรูปแบบที่ให้ทั้งความสะดวก และความเป็นปรนัย แต่ก็มีข้อจำกัดในการใช้อยู่มาก โดยเฉพาะแบบสอบที่มีความยากต่างกัน พอททอฟ (Potthoff) ได้แนะนำทางเลือกอื่นเมื่อพบว่า รูปแบบเชิงเส้นตรงยังไม่เพียงพอ ดังนี้ (Potthoff 1982: 209-210)

(๑) ปรับปรุงวิธีการให้คะแนน เพื่อให้คะแนนไม่เกาะกลุ่มมากเกินไป ซึ่งเป็นสาเหตุหนึ่งที่ทำให้การเทียบมาตราเชิงเส้นตรงไม่ได้ผลน่าพอใจ

(๒) ควรใช้วิธีการตรวจสอบที่เป็นปรนัย เพื่อช่วยในการตัดสินใจว่า ควรใช้รูปแบบเชิงเส้นตรงหรือไม่ ในกรณีที่ใช้วิธีการเทียบที่มีข้อตกลงเบื้องต้นของความเที่ยงเท่ากัน อาจทดสอบความแตกต่างของการแจกแจงของคะแนนภายหลังการเทียบดูว่า ยังคงเหมือนกันหรือไม่ ถ้าพบว่าไม่สามารถปฏิเสธสมมุติฐานศูนย์ ก็แสดงว่ารูปแบบเชิงเส้นตรงใช้ได้แล้ว ไม่จำเป็นต้องหาวิธีการอื่น ๆ ที่ไม่ใช่เชิงเส้นตรง

(๓) ถ้าพบว่ารูปแบบเชิงเส้นตรงไม่เหมาะในการเทียบมาตราแล้ว ให้พิจารณารูปแบบอื่น แต่ถ้าพบว่ารูปแบบอื่นก็ยังให้ผลไม่ดีไปกว่าเชิงเส้นตรง อาจตัดสินใจใช้เชิงเส้นตรง แต่ต้องมีรายงานประกอบเพื่อช่วยให้ผู้นำผลของคะแนนแปลงไปใช้ได้ เห็นข้อจำกัดที่ปรากฏอยู่ สิ่งที่ควรมีในรายงาน คือ

ก. ระบุลักษณะเฉพาะของแบบสอบ คะแนนแปลง และให้ข้อสังเกตเพื่อเตือนถึงผลกระทบของคะแนนต่อการเทียบในช่วงพิสัยต่าง ๆ

ข. น่าจะเสนอพิสัยคะแนนแปลงในช่วงความเชื่อมั่นที่เหมาะสม โดยเฉพาะในกรณีที่แบบสอบสองชุดมีความเที่ยงไม่เท่ากัน ช่วงความเชื่อมั่นย่อมต่างกัน

(๔) ในบางกรณีที่พบว่า รูปแบบเชิงเส้นตรงให้ประสิทธิภาพดีมาก ยกเว้นคะแนนที่กระจายอยู่ที่ปลายสุดทั้งสองข้าง อาจดำเนินการเทียบมาตราช่วยเสริมในส่วนนี้ ปัญหานี้เกิดขึ้นจากแบบสอบ ๒ ชุด อาจมีความยาก (หรือง่าย) ไม่เพียงพอที่จะจำแนกได้ดีในส่วนปลายที่สูงสุดหรือต่ำสุด การแก้ไข คือ นำแบบสอบ ๒ ชุดนี้ไปทำการสอบกับคนที่มีระดับความสามารถในระดับความยากนั้น ๆ เพิ่มเติม

<u>วิธีการเทียบมาตราด้วยรูปแบบอิงทฤษฎีการตอบข้อสอบ</u>

ลอร์ด (Lord 1977) ได้แนะนำการพิจารณาเลือกวิธีเทียบมาตราคะแนนของแบบสอบสองชุดไว้ว่า ถ้าไม่รู้แน่ว่าแบบสอบ ๒ ฉบับ ที่กำลังจะทำการเทียบมาตรา มีความเป็นคู่ขนานกันหรือไม่ วิธีการที่ปลอดภัยที่สุด คือ ให้จัดกระทำกับแบบสอบนั้นในฐานะไม่เป็นคู่ขนานกัน และใช้วิธีการประเภทที่ไม่ใช่โมเดลเส้นตรง การออกแบบเพื่อศึกษาการเทียบมาตราคะแนนโดยใช้ทฤษฎีการตอบข้อสอบ สามารถนำมาใช้กับคน ๒ กลุ่ม ที่ต่างกัน และใช้กับแบบสอบ ๒ ชุด ที่ไม่เป็นคู่ขนานกัน ซึ่งวิธีการเทียบมาตราคะแนนแบบดั้งเดิมไม่สามารถทำได้เหมาะสมนัก ในกรณีเช่นนี้การเทียบมาตราใช้แบบสอบร่วมเข้ามาช่วยพิจารณา และปรับความแตกต่างของความสามารถ ๒ กลุ่ม แต่ละกลุ่มของผู้สอบจะสอบเพียงฉบับเดียวกับแบบสอบร่วม ตัวอย่างการเทียบมาตราคะแนนด้วยแบบแผนนี้ เช่น มาร์โค (Marco 1977 cited by Lord 1977) ได้วางแผนศึกษาการเทียบมาตราคะแนนสอบวัดผลสัมฤทธิ์วิชาแคลคูลัส ฉบับเพื่อการจัดชั้นเรียนในระดับวิทยาลัย ซึ่งมี ๘๕ ข้อ (College Board Advanced placement program: AP) และฉบับโปรแกรมการสอบระดับวิทยาลัย (The College Level Examination Program: CLEP) ซึ่งง่ายกว่าฉบับแรกเล็กน้อย มี ๕๐ ข้อ และมีแบบสอบร่วม (Anchor test) ๑๗ ข้อ แทรกอยู่ในแต่ละฉบับ กลุ่มตัวอย่างผู้สอบชุด AP จำนวน ๑๒๒๐ คน เมื่อพฤษภาคม ๑๙๗๒ ส่วนกลุ่มตัวอย่างผู้สอบชุด CLEP จำนวน ๑๒๐๙ คน เมื่อพฤษภาคม ๑๙๗๔ กระบวนการเทียบมาตราคะแนนเริ่มด้วยการใช้โปรแกรมคอมพิวเตอร์ประมาณค่าประจำกรข้อสอบ และความสามารถจากการทำงานของเครื่อง ๑ ครั้ง ได้ค่าประมาณพร้อม ๆ กันทั้งหมด ทั้งที่เป็นค่าประมาณความสามารถ มี ๒๒๒๖ ตัว (ทั้ง ๒ กลุ่มรวมกัน และหักออก ๒๙๓ คน ที่ตอบไม่สมบูรณ์) และค่าประมาณข้อสอบ ๗๙ ข้อ (45+50-17) เป็นค่า a b และ c สำหรับทุกข้อ รายละเอียดการใช้แบบสอบร่วมในรูปแบบนี้ จะแยกกล่าวเป็น ๒ ส่วน

คือ การเทียบคะแนนจริง และคะแนนสังเกต

๑. <u>การเทียบมาตราคะแนนจริง</u>

พิจารณาจากความสัมพันธ์เชิงคณิตศาสตร์ระหว่างความสามารถ และจำนวนคะแนนนับตามข้อที่ทำถูก คะแนนจริง ξ จากแบบสอบ X และ η จากแบบสอบ Y เป็นดังต่อไปนี้

$$\xi = \xi(\theta) = \sum_{i=1}^{n_X} P_i(\theta) \quad ------ (๑๔)$$

$$\eta = \eta(\theta) = \sum_{j=1}^{n_Y} P_j(\theta) \quad ------ (๑๕)$$

เมื่อ n_X และ n_Y เป็นจำนวนข้อในแบบสอบชุด X และชุด Y ตามลำดับ สมการทั้งสองเป็นสมการคณิตศาสตร์ ดังนั้น ถ้าแทนค่าความสามารถ θ ลงในสมการทั้งสองก็จะได้คะแนนจริงที่สมมูลกันบนมาตราเดียวกัน หรืออีกวิธีหนึ่งทำสมการทั้งสองให้เท่ากัน โดยกำจัดค่า θ ให้หายไป ก็จะสามารถหาคะแนนสมมูลของ X และ Y ได้ การหาค่าที่เทียบบนมาตราเดียวกันนี้ ใช้คำนวณจากค่า $P_i(\theta)$ และ $P_j(\theta)$ ซึ่งใช้ค่าประมาณแทน คือ $\hat{P}_i(\theta)$ และ $\hat{P}_j(\theta)$ ที่ได้จากการแทนค่า $\hat{a}$, $\hat{b}$ และ $\hat{c}$ (Lord 1977)

การใช้โปรแกรม LOGIST จะวิเคราะห์ข้อมูลการตอบของทุกคน และข้อสอบทุกข้อพร้อม ๆ กัน โดยถือว่าข้อสอบชุดหนึ่งที่ผู้สอบไม่ได้ทำนั้นเป็นส่วนที่ผู้สอบทำไม่ถึง (not reached) วิธีนี้ทำให้การหาค่าประมาณประชากรปรากฏออกมาบนมาตราเดียวกันหมด นับว่าเป็นวิธีที่มีประสิทธิภาพ โดยมีสมมุติฐานเบื้องต้นว่า ความสามารถ ณ θ ใด ๆ ของผู้สอบไม่มีความแตกต่างกันจากแบบสอบหนึ่งไปยังอีกแบบสอบหนึ่ง (Lord 1980: 20)

การใช้แบบสอบร่วมทำได้ ๒ กรณี คือ ให้แบบสอบร่วมผนวกเข้าเป็นชุดเดียวกับแบบสอบที่ต้องการเทียบ ซึ่งเรียกว่า แบบสอบร่วมภายใน (internal anchor test) เช่น ตัวอย่างแบบสอบร่วม ๑๗ ข้อ ในวิชาแคลคูลัสที่กล่าวถึงข้างต้น ส่วนในกรณีที่จัดแยกเป็นชุดแบบสอบร่วม

ต่างหากจากแบบสอบที่ต้องการเทียบ เรียกว่า แบบสอบร่วมภายนอก (external anchor test) จุดมุ่งหมายของแบบสอบร่วม คือ เชื่อมข้อมูลคำตอบเข้าด้วยกัน เพื่อทำให้ค่าพารามิเตอร์ที่วิเคราะห์ออกมาอยู่บนมาตราเดียวกัน ถ้าขาดข้อสอบร่วมจำนวนนี้แล้ว จะไม่สามารถเทียบกันได้ นอกเสียจากว่ากลุ่มผู้สอบ X และ Y จะมีการแจกแจงของความสามารถเหมือนกัน (Lord 1980: 202)

๒. <u>การเทียบมาตราคะแนนสังเกต</u> (Observed score equating)

จากการเทียบมาตราด้วยคะแนนจริงดังที่กล่าวมาแล้ว ทำให้เห็นปัญหาในทางปฏิบัติ คือ การประมาณคะแนนของผู้สอบจากคำตอบ $\hat{\xi} = \Sigma_i \hat{P}_i(\hat{\theta})$ แท้ที่จริง คือ คะแนนที่มีความคลาดเคลื่อนอีกชุดหนึ่งนั่นเอง ไม่มีคุณสมบัติของคะแนนจริง ซึ่งลอร์ด (Lord 1980: 196-198, 202) ได้พิสูจน์ให้เห็นแล้วว่า การเทียบมาตราจะทำให้ได้ผลอย่างเข้มงวดไม่ได้ นอกเสียจากว่าแบบสอบ ๒ ฉบับนี้ต้องคู่ขนานกันจริง ลอร์ดได้เสนอการเทียบด้วยคะแนนสังเกต ซึ่งเป็นวิธีการเทียบมาตราโดยประมาณที่ทำได้ในทางปฏิบัติที่สถานการณ์การสอบไม่สามารถจัดให้ผู้สอบทำแบบสอบได้ทั้งสองชุด กระบวนการเริ่มด้วยการประมาณการแจกแจงความสามารถของคนในกลุ่มผู้สอบ $\hat{r}(\theta)$ (หมายถึง a, หรือ b หรือ c) โดยใช้การแจกแจงจริงของ $\hat{\theta}_a$ ในกลุ่มเป็นตัวประมาณ $r(\theta)$ การแจกแจงของคะแนนคิบหาได้จากสมการต่อไปนี้

$$\hat{\phi}_x(x) = \frac{1}{N} \sum_{a=1}^{N} \phi_x(x \mid \hat{\theta}_a) \quad \text{เมื่อ} \quad a = 1, 2, \ldots, N$$

ถ้า $\hat{r}(\theta)$ เป็นค่าต่อเนื่อง จะประมาณค่า $\hat{\phi}(x)$ จากสมการ

$$\hat{\phi}_x(x) = \int_{-\infty}^{\infty} \phi_x(x \mid \theta)\, \hat{r}(\theta)\, d\theta$$

ฟังชั่น $\phi_x(x \mid \hat{\theta}_a)$ ใช้ประมาณค่าประจากรข้อสอบในฉบับ X ในทำนองเดียวกัน ประมาณค่าประจากรของแบบสอบ Y และเนื่องจากแบบสอบ X และ Y ต่างมีการแจกแจงเป็นอิสระกัน เมื่อกำหนดให้ θ คงที่ ก็สามารถประมาณส่วนที่เป็นการกระจายร่วมกันของคะแนน X และ Y สำหรับกลุ่มที่ถูกประมาณโดย

$$\hat{\phi}(x,y) = \frac{1}{N} \sum_{a=1}^{N} \hat{\phi}_x(x|\hat{\theta}_a)\,\hat{\phi}_y(y|\hat{\theta}_a)$$

หรือโดย

$$\phi(x,y) = \int_{-\infty}^{\infty} \phi_x(x|\theta)\;\phi_y(y|\theta)\;r(\theta)\,d\theta \text{----} \quad (๑b)$$

จากที่กล่าวมานี้ จะเห็นบทบาทของแบบสอบร่วมยัดเยน ทำให้สามารถประมาณการแจกแจงร่วม (joint distribution) ของ X และ Y ถึงแม้จะไม่มีผู้ใดสอบทั้ง ๒ ฉบับ

สมการ ๑b เป็นการแจกแจงร่วมกันของตัวแปรสามตัว คือ θ, x และ y เนื่องจากว่า θ เป็นตัวระบุคะแนนจริง ξ และ η การแจกแจงนี้จึงนับว่าเป็นการแจกแจงร่วมกันของตัวแปร ๔ ตัว คือ ξ η x และ y สมการนี้ไม่มีทางอื่นใดจะประมาณได้ นอกจากระบุความสัมพันธ์เชิงอีควิเปอร์เซนไทล์ระหว่าง x และ y จากสมการสุดท้ายข้างต้น ๒ สมการ วิธีการเป็นการเทียบมาตราคะแนนของแบบสอบโดยประมาณ

๓. <u>ประเด็นอภิปรายเกี่ยวกับการเทียบมาตราแบบ IRT</u>

จากทฤษฎีการเทียบคะแนนคิบของแบบสอบสองชุดด้วยวิธีการเทียบคะแนนจริง และวิธีเทียบคะแนนสังเกต หรือการเทียบอีควิเปอร์เซนไทล์ใน IRT ดังที่กล่าวมาแล้วนั้น มีทั้งข้อได้เปรียบและเสียเปรียบ วิธีเทียบคะแนนจริง ξ และ η ให้ความหมายของคะแนนสมบูรณ์เฉพาะคะแนนที่อยู่เหนือค่าเฉลี่ยของการเดา ส่วนวิธีการใช้การประมาณคะแนนสังเกต ก็มีจุดอ่อนที่เป็นเพียงการเทียบมาตราอย่างประมาณเท่านั้น คอยข้อคำถามว่า วิธีใดจะให้ผลการเทียบดีกว่ากันนั้น ลอร์คได้กล่าวว่า ทั้งสองวิธีมีความสอดคล้องกันมาก แต่การสรุปเปรียบเทียบอ้างอิงทั่วไปต้องทำอย่างพิถีพิถัน (Lord 1980: 203)

โคเลน (Kolen 1981) ได้ทดลองใช้วิธีอีควิเปอร์เซนไทล์ IRT พบว่าจากตัวอย่างแบบสอบและสถานการณ์สอบเฉพาะในการวิจัย วิธีนี้เป็นวิธีที่ดี โดยใช้เกณฑ์ประเมินความคงเส้นคงวาในการวิเคราะห์กลุ่มตัวอย่างสอบหาผล

ปี ๑๙๘๔ ลอร์ด และ วินเกอสกี (Lord and Wingersky 1984) ได้ศึกษาเปรียบเทียบผลการเทียบมาตราตามรูปแบบ IRT ด้วยวิธีคะแนนจริงและคะแนนสังเกต พบว่า ทั้งสองวิธีแทบจะไม่ได้มีข้อแตกต่างกันเลย และได้วิจารณ์ไว้ว่า วิธีคะแนนสังเกตมีความซับซ้อนในการคำนวณและลงทุนด้านอื่น ๆ แพงกว่าวิธีคะแนนจริง

ความยาวของแบบสอบร่วม

จากการเสนอแบบแผนการจัดการเทียบมาตราที่ใช้แบบสอบร่วมอย่างละเอียดของแองกอฟ (Angoff 1984: 106-107) ได้แนะนำให้ใช้แบบสอบร่วมที่มีความยาวเพียงพอที่จะก่อให้เกิดข้อสนเทศเพื่อนำไปปรับความแตกต่างระหว่างกลุ่มผู้สอบอย่างมีประสิทธิภาพ จำนวนข้อที่เหมาะสม คือ ต้องไม่น้อยกว่า ๒๐ ข้อ หรือไม่น้อยกว่าร้อยละ ๒๐ ของจำนวนข้อในแบบสอบที่ต้องการเทียบมาตรา แล้วแต่จำนวนไหนจะมากกว่ากัน ขณะเดียวกันไรท์ (Wright 1979: 98) ได้กล่าวว่า ควรเป็นข้อสอบที่วัดเรื่องเดียวกันกับแบบสอบที่ต้องการศึกษา และมีเพียง ๑๐ ข้อ ก็เพียงพอแล้ว ที่กล่าวมานี้ยังขาดผลเชิงประจักษ์ในการสนับสนุน ต่อมาปี ค.ศ. ๑๙๘๕ บูเดส (Budescu 1985: 13-20) ได้ศึกษาวิเคราะห์เชิงทฤษฎี ส่วนคลีน และ โคเลน (Klein and Kolen 1985) ได้ศึกษาเชิงประจักษ์ ซึ่งจะเสนอต่อไปนี้

๑. การวิเคราะห์เชิงทฤษฎีของบูเดส (๑๙๘๕, ๑๓ – ๒๐) พบว่า ตัวแปรสำคัญที่ส่งผลถึงประสิทธิภาพของการเทียบมาตรา คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างแบบสอบร่วมกับแบบสอบสองชุดที่ต้องการเทียบมาตรา และในเชิงทฤษฎีวัดผลดั้งเดิม (classical test theory) ตัวแปรความยาวของแบบสอบร่วมมีความสัมพันธ์ต่อความเที่ยง อธิบายได้ดังต่อไปนี้

กลุ่ม α ทำแบบสอบชุด X กลุ่ม β ทำแบบสอบชุด Y และทั้งสองกลุ่มทำแบบสอบร่วม V เหมือนกัน ค่าเฉลี่ยกับส่วนเบี่ยงเบนมาตรฐานของแบบสอบทั้งสองชุด และของแบบสอบร่วมประมาณได้ตามลำดับกลุ่มดังนี้ ให้ $M_{y\alpha}$ $S_{X\alpha}$ $M_{V\alpha}$ $S_{V\alpha}$ เป็นค่าสถิติที่ได้จากกลุ่ม α และ $M_{Y\beta}$ $S_{Y\beta}$ $M_{V\beta}$ $S_{V\beta}$ เป็นค่าสถิติที่ได้จากกลุ่ม β ในลักษณะเดียวกัน ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของแบบสอบร่วม V ของกลุ่มตัวอย่างรวม T ($\alpha + \beta$) ได้โดยตรงจากผลการสอบ ในกรณีที่กลุ่มผู้สอบไม่ได้มาจากการสุ่ม วิธีการประมาณค่าจึงใช้วิธีของทักเกอร์

(Tucker equating method) แต่ถ้าการจัดดำเนินการสอบได้ใช้กลุ่มผู้สอบอย่างสุ่ม การประมาณค่าจะใช้วิธีเป็นไปได้สูงสุด (Maximum likelihood) สูตรที่ใช้มีความแตกต่างกันเล็กน้อย สาระสำคัญของวิธีการใช้แบบสอบร่วมอยู่ที่การประมาณค่าเฉลี่ยและความแปรปรวนของชุด X และ Y ในกลุ่มตัวอย่างรวม ซึ่งหาได้ด้วยการใช้ค่าต่าง ๆ ที่รู้แล้ว คือ ค่าสหสัมพันธ์ระหว่างคะแนนของแบบสอบร่วม V กับคะแนนจากแบบสอบ X และ Y และค่าความแตกต่างระหว่างคะแนนในแบบสอบร่วม V ของกลุ่มตัวอย่างทั้ง ๒ ดังสูตร

$$\hat{M}_{Xt} = M_{X\alpha} + \frac{r_{XV\alpha} S_{X\alpha}}{S_{V\alpha}} (M_{Vt} - M_{V\alpha}) \quad \text{---------} \quad (๑๗)$$

และ $$\hat{S}^2_{Xt} = S^2_{X\alpha} + \frac{r^2_{XV\alpha} S^2_{X\alpha}}{S^2_{V\alpha}} (S^2_{Vt} - S^2_{V\alpha}) \quad \text{---------} \quad (๑๘)$$

ทำนองเดียวกัน สูตรที่ใช้สำหรับแบบสอบชุด Y มีดังนี้

$$\hat{M}_{Yt} = M_{Y\beta} + \frac{r_{YV\beta} S_{Y\beta}}{S_{V\beta}} (M_{Vt} - M_{V\beta}) \quad \text{---------} \quad (๑๙)$$

$$\hat{S}^2_{Yt} = S^2_{Y} + \frac{r^2_{YV\beta} S^2_{Y\beta}}{S^2_{V\beta}} (S^2_{VT} - S^2_{V\beta}) \quad \text{---------} \quad (๒๐)$$

เมื่อแทนค่าประมาณต่าง ๆ ลงในสมการเส้นตรง และจัดเทอมใหม่ จะได้สมการเพื่อการเทียบมาตราดังนี้

$$Y = AX + B = \frac{\hat{S}_{Yt}}{\hat{S}_{Xt}} X + (\hat{M}_{Yt} - \frac{\hat{S}_{Yt}}{\hat{S}_{Xt}} \hat{M}_{Yt}) \quad \text{---------} \quad (๒๑)$$

จากสมการข้างต้นเห็นได้ชัดเจนว่า ถ้าสัมประสิทธิ์สหสัมพันธ์ระหว่างแบบสอบร่วมกับชุด X หรือ Y เป็นศูนย์ แบบสอบร่วมที่นำมาใช้ในงานนี้ไม่มีความหมายเลย ตรงกันข้ามถ้าสัมประสิทธิ์สหสัมพันธ์ดังกล่าวมีค่าเพิ่มสูงขึ้น ๆ จะมีผลต่อการเพิ่มคุณภาพของตัวประมาณประชากรรวมอย่างสูงขึ้นตามกัน (monotonic increasing) การที่จะให้ได้มาซึ่งความสัมพันธ์สูง ๆ จะต้องดำเนินการสร้างและจัดชุดแบบสอบร่วมอย่างพิถีพิถันที่ให้มีลักษณะคู่ขนานกับแบบสอบทั้งสองชุด และที่ต้องการเน้นในงานของบูเดส (๑๙๘๕: ๑๕) คือ ความยาวของแบบสอบร่วม ซึ่งการวิเคราะห์ต่อไปนี้ จะแสดงให้เห็นความสำคัญของความยาวของแบบสอบร่วม ดังนี้

กลุ่ม α ทำชุด X และ V ให้ได้คะแนนรวม Z

ดังนั้น Z ของแต่ละคน คือ $Z = X + V$ ---------- (๒๒)

ให้ความเที่ยงของแบบสอบรวม (X + V) เป็น r_{ZZ} แสดงในเทอมของพารามิเตอร์ของ X และ V ดังนี้ (Horst 1968: 281 cited by Budescu 1985: 16)

$$r_{ZZ} = \frac{(S_x^2 r_{xx} + S_v^2 r_{vv} + 2 r_{xv} S_v S_x)}{(S_x^2 + S_v^2 + 2 r_{xv} S_x S_v)} \quad \text{---- (๒๓)}$$

เมื่อ r_{xx} และ r_{vv} เป็นค่าประมาณความเที่ยงของแบบสอบชุด X และแบบสอบร่วม V

ให้ p ($0 < p < 1$) แทนสัดส่วนของข้อสอบชุด X ที่อยู่ใน Z และ $q = (1 - p)$ เป็นสัดส่วนของแบบสอบร่วม V ในแบบสอบรวม Z ถ้า ๒ แบบสอบ X และ V ต่างเป็นคู่ขนานกัน ความเที่ยงเขียนในรูปของฟังชั่นของ r_{ZZ} โดยอาศัยสูตร Spearman - Brown ดังนี้

$$r_{xx} = \frac{p\, r_{ZZ}}{(1 - q\, r_{ZZ})} \quad \text{และ}$$

$$r_{vv} = \frac{q\, r_{vv}}{(1 - p\, r_{ZZ})} \qquad \text{---------- (๒๔)}$$

ในทำนองเดียวกัน ถ้า s_Z^2 เป็นความแปรปรวนของแบบสอบฉบับ z ความแปรปรวนของส่วนประกอบสามารถเขียนในรูปของฟังชั่นของความยาวสัมพัทธ์ได้ดังนี้

$$s_x^2 = s_Z^2\, p\, (1 - q\, r_{ZZ}) \quad \text{และ}$$

$$s_v^2 = s_Z^2\, q\, (1 - p\, r_{ZZ}) \qquad \text{---------- (๒๕)}$$

แทนสมการ (๒๔) และ (๒๕) ใน (๒๓) จะพบว่า r_{xv} เป็นฟังชั่นของความเที่ยงของแบบสอบฉบับรวม z และความยาวของส่วนประกอบ x และ v ดังนี้

$$r_{xv} = r_{ZZ} \left[\frac{p\, q}{(1 - p\, r_{ZZ})\,(1 - q\, r_{ZZ})} \right]^{\frac{1}{2}} \qquad \text{------ (๒๖)}$$

นิพจน์ในสูตร (๒๖) เป็นอิสระจากค่าเฉลี่ยและความแปรปรวนรวมของแบบสอบที่เกี่ยวข้องทั้ง ๒ ดังนั้น สามารถเอาไปใช้กับกลุ่มผู้สอบอีกกลุ่ม (β) ได้เหมือน ๆ กัน สรุปจากการวิเคราะห์ได้ว่า โดยธรรมชาติแล้วสหสัมพันธ์ระหว่างแบบสอบ x กับแบบสอบรวมเป็นฟังชั่นเพิ่มขึ้นตามความเที่ยงของแบบสอบฉบับรวม (r_{ZZ}) เมื่อความเที่ยงของแบบสอบรวม (r_{ZZ}) คงที่ ค่าสหสัมพันธ์ระหว่างแบบสอบ x กับแบบสอบรวม (r_{xv}) มีค่าสูงสุดที่ความยาวแบบสอบทั้งสองเท่ากับ ($p = q = .5$) ค่า r_{xv} มีลักษณะสมมาตรและเพิ่มขึ้นตามผลคูณของ p และ q จากจุดสำคัญนี้ทำให้ได้ค่าอ้างอิง คือ ประสิทธิภาพสูงสุดของการเทียบมาตราเมื่อจัดสัดส่วนของแบบสอบที่ต้องการเทียบกับแบบสอบรวมให้มีสัดส่วนเท่ากัน ในกรณีเช่นนี้จะได้สหสัมพันธ์ระหว่างแบบสอบที่ต้องการเทียบกับแบบสอบรวม (r_{xv}) มีค่าเท่ากับ $r_{ZZ}/(2 - r_{ZZ})$ (๑๙๘๕: ๑๖ – ๑๗)

จากการวิเคราะห์ที่กล่าวมาข้างต้น บูเดส (๑๙๘๕: ๑๗) ได้นำมาสร้างเป็นดัชนีของประสิทธิภาพสัมพัทธ์ของการเทียบมาตรา เรียกว่า ความพร่องสัมพัทธ์ (relative deficiency) ตามนิยามของสมการ (๒๗) ดังนี้

$$R.D(q) = 1 - \frac{r_{xv}(q)}{r_{xv}(0.5)} = 1 - \left[\frac{(2 - r_{ZZ})^2 pq}{(1 - p\, r_{ZZ})(1 - q\, r_{ZZ})} \right]^{\frac{1}{2}} \quad -- (๒๗)$$

มาตรการนี้จะแสดงให้เห็นถึงการเพิ่มประสิทธิภาพในเชิงเปรียบเทียบกับค่าอ้างอิงสูงสุด เมื่อ $p = q = .5$ ดังนั้น กระบวนการเทียบมาตราใด ๆ ที่มีค่าความพร่องสัมพัทธ์ต่ำกว่าหรือเท่ากับ ๐.๑๐ จัดว่าเป็นการเทียบมาตราที่สามารถให้ผลที่น่าพอใจ (๑๙๘๕: ๑๗)

ในกรณีสูตรสำเร็จของแองกอฟที่ใช้ ๒๐% ถ้าจะให้มีประสิทธิภาพสัมพัทธ์ในระดับที่น่าพอใจแล้ว จะต้องเป็นแบบสอบที่มีค่าความเที่ยงสูงกว่า ๐.๘๖๖ (๑๙๘๕: ๑๗)

กล่าวโดยสรุป การวิเคราะห์ของบูเดสเป็นการวิเคราะห์เชิงทฤษฎีที่ชี้ให้เห็นความสัมพันธ์ของความยาวแบบสอบร่วมกับประสิทธิภาพของการเทียบมาตรา ในเชิงปฏิบัติไม่สามารถให้ข้อกำหนดทั่วไปที่จะถือปฏิบัติกัน ทั้งนี้เพราะการปฏิบัติจะต้องคำนึงถึงตัวแปรแวดล้อมอื่น ๆ อีก เช่น เวลา การลงทุน และข้อจำกัดในทางปฏิบัติอื่น ๆ อีก แต่สิ่งที่ควรพิจารณา คือ ให้ประสิทธิภาพของการเทียบมาตรานั้น ๆ อยู่ในระดับของการยอมรับเฉพาะกรณี (๑๙๘๕: ๑๙)

๒. <u>การศึกษาเชิงประจักษ์ของคลีน และ โคเลน</u> (๑๙๘๕) คลีน และ โคเลน ได้เลือกศึกษากรณีการเทียบมาตราที่กลุ่มผู้สอบไม่ได้มาจากการสุ่ม ซึ่งวิธีการใช้แบบสอบร่วมเป็นสิ่งจำเป็นอย่างยิ่งที่ให้พื้นฐานเพื่อการประมาณความแตกต่างของกลุ่มผู้สอบสองกลุ่ม ประเด็นของแบบสอบร่วมที่เขาทำการศึกษา คือ เรื่องของความยาวของแบบสอบร่วม โดยได้ทำการตรวจสอบผลกระทบของความยาวต่อผลการเทียบมาตรา ดังนี้

ข้อมูลเป็นผลการสอบเพื่อขอรับประกาศนียบัตรวิชาชีพ แบบสอบเป็นชนิดเลือกตอบ มี ๒๕๐ ข้อ ได้ทำการสอบเมื่อปี ค.ศ. ๑๙๘๔ ออกแบบการวิจัยด้วยการแบ่งแบบสอบเป็น ๒ ชุด และแบบสอบร่วม ความยาวต่าง ๆ กัน ๕ ขนาด ในแต่ละขนาดประกอบด้วยข้อสอบที่มีความสมดุลย์ทั้งใน

ด้านเนื้อหา ความยากเฉลี่ย และอำนาจจำแนกเฉลี่ยของแบบสอบทั้งฉบับ แบบสอบร่วมที่ใช้จึงมีขนาด ๒๐ ข้อ ๔๐ ข้อ ๖๐ ข้อ ๘๐ ข้อ และ ๑๐๐ ข้อ แต่ละแบบสอบร่วมจึงเป็นซับเซทของฉบับที่ยาวกว่า และทุกฉบับต่างก็เป็นซับเซทของจำนวน ๒๕๐ ข้อ

การจัดเตรียมกลุ่มตัวอย่างผู้สอบ ผู้วิจัยมีจุดประสงค์จะศึกษาผลการเทียบมาตรา แบบทักเกอร์ (Tucker equating) ซึ่งเป็นรูปแบบเชิงเส้นตรง (Angoff 1971: 579-583 cited by Klein and Kolen 1985: 4) กับกลุ่มผู้สอบที่มีความสามารถคล้ายคลึงกัน และกับกลุ่มผู้สอบที่มีความสามารถแตกต่างกัน ผู้วิจัยจึงได้แบ่งคนสอบจำนวน ๑,๗๐๕ คน ออกเป็น ๒ กลุ่ม ไม่ซ้ำซ้อนกันตามเกณฑ์ของความสามารถคล้ายคลึงและแตกต่างที่มีข้อยฐาน (meadin)

ผลการเทียบมาตราที่เกิดกับการใช้แบบสอบร่วมขนาดความยาว ๕ ขนาด กับกลุ่มผู้สอบ ๒ ลักษณะ รวมมีผลการเทียบมาตรา ๑๐ ชุด จากนั้นได้คำนวณค่าความคลาดเคลื่อน (root-mean-squared error: RMSE) ของแต่ละกรณีด้วยสูตร

$$\widehat{RMSE} = \left[\Sigma_i N_i (X'_i - X_i)^2 \,/\, \Sigma_i X_i \right]^{\frac{1}{2}} \quad \text{---} \quad (๒๘)$$

เมื่อ X_i เป็นคะแนนแถวที่ i ของแบบสอบทั้งหมด

N_i เป็นจำนวนผู้สอบที่ได้คะแนน i จากแบบสอบทั้งหมด

X'_i เป็นคะแนนสมมูลของ X_i ที่ประมาณด้วยวิธีการของทักเกอร์ในรูปแบบเชิงเส้นตรง

นอกจากนี้ยังได้ประมาณค่าเฉลี่ยของความคลาดเคลื่อนของการเทียบมาตรา ซึ่งเป็นตัวที่ทำให้เกิด $\widehat{RMSE}$ โดยใช้สมการ

$$\widehat{BIAS} = \bar{X}' - \bar{X} \quad \text{----------} \quad (๒๙)$$

เมื่อ $\bar{X}'$ เป็นค่าเฉลี่ยของค่าประมาณของคะแนนดิบสมมูลย์ และ

$\bar{X}$ เป็นค่าเฉลี่ยของคะแนนดิบ

เปรียบเทียบค่าดัชนีความคลาดเคลื่อน ($\widehat{RMSE}$) กับส่วนเบี่ยงเบนมาตรฐานของแบบสอบ ซึ่งเท่ากับ ๒๐.๗๒ พบว่า ความคลาดเคลื่อนของการเทียบมาตราสำหรับแบบสอบร่วมที่มีขนาด ๒๐ ข้อ และ ๔๐ ข้อ ของกลุ่มความสามารถแตกต่างกันมีค่าสูง

คลีนและโคเลนอภิปรายไว้ว่า กรณีแบบสอบที่นำมาศึกษาได้จัดกระทำให้เป็นแบบสอบเทียบมาตรากับแบบสอบร่วม แล้วทำการเทียบมาตรากับแบบสอบเดิม เมื่อแบบสอบร่วมมีความสมนัยกับแบบสอบทั้งในด้านความครอบคลุมของเนื้อหา ความยากเฉลี่ย และค่าอำนาจจำแนกเฉลี่ย ผลการวิจัยไม่สามารถสรุปว่า ความยาวของแบบสอบร่วมมีความสัมพันธ์อย่างมีระบบกับค่าประมาณความคลาดเคลื่อนมาตรฐานของการเทียบมาตรา ความแม่นยำของการเทียบมาตราที่ใช้แบบสอบร่วมขนาด ๒๐ ๔๐ ๖๐ ๘๐ และ ๑๐๐ ในกลุ่มผู้สอบที่คล้ายกัน สำหรับกลุ่มผู้สอบที่มีความสามารถแตกต่างกัน พบว่า ค่าประมาณความคลาดเคลื่อนพุ่งสูงขึ้นอย่างมากเมื่อจำนวนแบบสอบร่วมลดลงจาก ๖๐ ข้อ ลงเหลือ ๔๐ ข้อ และจาก ๔๐ ข้อ ลดลงเหลือ ๒๐ ข้อ

กล่าวโดยสรุป เมื่อแบบสอบที่นำมาเทียบมาตรามีความคล้ายคลึงกันมากเช่นในกรณีที่ทำการศึกษา คือ เป็นข้อสอบชุดเดียวกัน และกลุ่มผู้สอบมีระดับความสามารถคล้ายคลึงกัน การเพิ่มความยาวให้กับแบบสอบร่วมไม่ได้เพิ่มความแม่นยำของการเทียบมาตราอย่างชัดเจน แต่สำหรับกลุ่มผู้สอบที่มีความสามารถแตกต่างกันแล้ว การเพิ่มความยาวของแบบสอบร่วมมีผลต่อการเพิ่มความแม่นยำของการเทียบมาตราอย่างชัดเจน จากกรณีการศึกษาครั้งนี้ พบว่า เมื่อแบบสอบร่วมมีความยาวน้อยกว่าร้อยละ ๒๕ ของแบบสอบทั้งฉบับแล้ว ผลของความคลาดเคลื่อนของการเทียบมาตราจะอยู่ในระดับที่มากจนไม่เป็นที่ยอมรับได้ และที่ต้องย้ำอีกครั้ง คือ เป็นกรณีที่แบบสอบร่วมมีความสมนัยกับแบบสอบทั้งฉบับอย่างมากในด้านความครอบคลุมของเนื้อหา ความยาว และอำนาจจำแนก การศึกษานี้ไม่มีหลักฐานที่จะชี้ว่าความยาวของแบบสอบร่วมเป็นตัวแปรสำคัญเพียงตัวแปรเดียว

จากผลการศึกษาของบูเดส (Budescue) คลีน และ โคเลน (Klein and Kolen) ที่เกี่ยวกับความยาวของแบบสอบร่วมนี้ พบว่า ตัวแปรที่มีความสัมพันธ์และร่วมส่งผลต่อประสิทธิภาพของการเทียบมาตราที่สำคัญ คือ คุณภาพของแบบสอบร่วม และจากผลการศึกษาของ

คลิน และ โคเลน ยังพบประเด็นของความไม่คงที่ของการเทียบมาตรา เมื่อเปลี่ยนลักษณะของกลุ่มผู้สอบอีก

สำหรับการศึกษาที่กำลังดำเนินอยู่ในครั้งนี้ ผู้วิจัยได้จัดกระทำกับตัวแปรความยาวของแบบสอบร่วมให้มีต่างกัน ๓ ระดับ คือ คิดเป็นอัตราร้อยละเทียบกับแบบสอบที่ทำการเทียบมาตรา เป็น ๒๐ ๔๐ และ ๖๐ โดยมีกลุ่มผู้สอบที่เป็นไปตามธรรมชาติของการสอบจริง ๒ สถานการณ์ คือ สถานการณ์การสอบแข่งขันเพื่อคัดเลือกเข้าทำงาน และสถานการณ์การสอบเพื่อสัมฤทธิผลรายวิชาในระดับปริญญาตรี โดยผู้วิจัยได้ตั้งสมมุติฐานจากหลักของแองกอฟ บทวิเคราะห์ของบูเคส และผลการวิจัยส่วนหนึ่งของคลิน และ โคเลนว่า แบบสอบร่วมที่ยาวกว่าย่อมได้ผลการเทียบมาตราที่ดีกว่า

<u>ความคลาดเคลื่อนของการเทียบมาตรา</u>

การประเมินความคลาดเคลื่อนของการเทียบมาตราที่สมบูรณ์ คือ การวิเคราะห์แหล่งความลำเอียง (bias) และความแปรผัน (variability) แล้วนำมาเสนอให้เห็นเป็นภาพเดียวกัน เพื่อช่วยในการตัดสินใจว่า การเทียบมาตราด้วยวิธีนั้น ๆ ควรทำหรือไม่อย่างไร (Braun and Holland 1982: 32) รายละเอียดที่จะเสนอต่อไปนี้ ได้แบ่งเป็น ๒ ตอน คือ เรื่องของความแปรปรวน ซึ่งมีทฤษฎีค่อนข้างหนักแน่นแล้ว และเรื่องของความลำเอียง ซึ่งเป็นบทวิเคราะห์ที่ยังต้องการวิจัยสนับสนุนอีกมาก

๑. <u>ความแปรผันเชิงสุ่ม และความคลาดเคลื่อนมาตรฐานของการเทียบมาตรา</u>

วิธีการเทียบมาตราทุกวิธี ไม่ว่าจะอยู่ในรูปแบบการเทียบมาตราใด เมื่อกลุ่มตัวอย่างผู้สอบเป็นกลุ่มสุ่มจากประชากรเดียว หรือหลายประชากร ย่อมมีความแปรผันเชิงสุ่มเกิดขึ้น จึงนิยมใช้เทคนิคการประมาณค่าความคลาดเคลื่อนเชิงสุ่ม (estimated sampling errors) กับการเทียบมาตราต่าง ๆ ความคลาดเคลื่อนเชิงสุ่มนี้จึงมีสมมุติฐานเบื้องต้นว่า ตัวอย่างมาจากการสุ่ม และใช้ความคลาดเคลื่อนมาตรฐานของการเทียบมาตรา (Standard error of equating ย่อว่า SEE) เป็นมาตรการวัดความแปรผันประเภทนี้

๑.๑ ความคลาดเคลื่อนมาตรฐานของการเทียบมาตราในรูปแบบเชิงเส้นตรง

ลอร์ด (Lord 1950 cited by Angoff 1984: 96) ได้สร้างสูตรสำหรับคำนวณหาความคลาดเคลื่อนมาตรฐานของรูปแบบเชิงเส้นตรงหลายสูตร ตามแบบแผนการรวบรวมข้อมูล ซึ่งได้มีการนำไปใช้อย่างกว้างขวางมานาน เช่น การเทียบมาตรากับแบบสอบ SAT ตั้งแต่ทศวรรษที่ ๑๙๕๐ (Donlon and Angoff 1971: 33) บรูน และ ฮอลแลนท์ (Braun and Holland 1982: 33-35) ได้นำมาเสนอให้อยู่ในรูปแบบทั่วไป ดังนี้

ให้ n_X คือ จำนวนคนสอบที่ได้จากการสุ่มจากประชากร P และทำแบบสอบ X

n_Y คือ จำนวนคนสอบที่ได้จากการสุ่มจากประชากร P เช่นกัน แต่ทำแบบสอบ Y

n_X และ n_Y ไม่จำเป็นต้องเท่ากัน

X_i คือ คะแนนของคนที่ i ในกลุ่มที่ทำแบบสอบ X

Y_j คือ คะแนนของคนที่ j ในกลุ่มที่ทำแบบสอบ Y

$\{X_i\}$ และ $\{Y_i\}$ ต่างมีการแจกแจงอย่างอิสระด้วยฟังชั่น F(x) และ G(y) ตามลำดับ ตัวอย่าง X และ Y เป็นตัวอย่างอิสระจากกัน สัญญลักษณ์โมเมนต์ที่ ๑ และ ๒ ของการแจกแจงใน X และ Y มีดังนี้

$$\mu_X = E(X) \qquad \mu_Y = E(Y) \quad ----- (๓๐)$$

$$\sigma_X^2 = Var(X) \qquad \sigma_Y^2 = Var(Y) \quad --- (๓๑)$$

ฟังชั่นเทียบมาตราของ Y ไปยัง X ในรูปเชิงเส้นตรง คือ

$$x^*_{(y)} = \mu_X + \frac{\sigma X}{\sigma Y} (y - \mu_Y)$$

ซึ่ง y จะแปรผันไปตลอดคะแนนดิบของ Y เขียนให้อยู่ในรูปของค่าประมาณประชากร ดังนี้

$$x^*_{(y)} = \bar{X} + \frac{s_x}{s_y}(y - \bar{Y})$$

เมื่อ $\bar{X}$ และ s_X^2 เป็นค่าเฉลี่ยและความแปรปรวนของ $\{X_i\}$

ความแปรปรวนแอสซิมโทต (Asymptotic variance) ของ $\hat{x}^*_{(y)}$ ได้มาจากทฤษฎีบทดังต่อไปนี้ —

$$\mathrm{Var}(\hat{X}^*_{(y)}) = \sigma^2_X \left\{ \frac{1}{n_X} + \frac{1}{n_Y} + \left(\frac{\mathrm{Skew}(X)}{n_X} + \frac{\mathrm{Skew}(Y)}{n_Y} \right) Z(y) + \left(\frac{2 + \mathrm{kurt}(X)}{4n_X} + \frac{2 + \mathrm{kurt}(Y)}{4n_Y} \right) Z^2(y) \right\}$$

$$+ \text{(higher order terms)} \qquad \text{---------- (๓๒)}$$

เมื่อ $\mathrm{Skew}(X) = \mu_{3X} / \sigma^3_X$

$\mathrm{kurt}(X) = (\mu_{4X} / \sigma^2_X)^{-3}$

$\mu_{3X} = E(X - \mu_X)^3$

$\mu_{4X} = E(X - \mu_X)^4$

สำหรับเทอมต่าง ๆ ของ Y หาได้ในทำนองเดียวกัน และ

$Z(y) = (y - \mu_Y) / \sigma_Y$

สำหรับเทอมต่าง ๆ ที่มีกำลังสูงขึ้นเป็นกำลังที่เป็น + ของ n_X^{-1} และ n_Y^{-1} จากทฤษฎีบทนี้ ได้อาศัยบทแทรก ๒ บท ช่วยทำให้สูตรชัดเจนขึ้น

บทแทรกที่ ๑ ถ้า X และ Y มีการแจกแจงปกติแล้ว

Skew(X) = 0 Skew(Y) = 0 และ

kurt(X) = 0 kurt(Y) = 0

ดังนั้น $Var(x^*_{(y)}) = (\sigma^2_X / n_h)(2 + Z^2(y)) + (\text{higher order terms})$

เมื่อ $n_h = [\frac{1}{2}(n_X^{-1} + n_Y^{-1})]^{-1}$ คือ ตัวกลางฮาโมนิคของ n_X และ n_Y

บทแทรกที่ ๒ ถ้า X และ Y มีการแจกแจงเป็นปกติ และ

$n_X = n_Y = N/2$ แล้ว

$$Var(x^*_{(y)}) = \frac{2\sigma^2_X}{N}(2 + Z^2(y)) + (\text{higher order terms}) \quad -- (๓๓)$$

โดยทั่วไปสภาพข้อมูลมักไม่เป็นไปตามบทแทรกที่ ๑ คือ การแจกแจงของคะแนนย่อมมีลักษณะเบ้ แต่เหมาะกับกรณีที่ n_X ไม่เท่ากับ n_Y ความคลาดเคลื่อนมาตรฐานของการเทียบมาตรา คือ รากที่ ๒ ของ $Var(X^*_{(y)})$ หรือ $SEE^2_{x^*}$

สูตรสำหรับหาความคลาดเคลื่อนมาตรฐาน เมื่อข้อมูลแตกต่างจากที่กล่าวมานี้จะแตกต่างไปตามเงื่อนไข ซึ่งสรุปได้ดังนี้

(๑) เมื่อกลุ่มสุ่มสองกลุ่มทำแบบสอบกลุ่มละชุด (เหมือนที่กล่าวมาข้างต้น) หาค่าประมาณความคลาดเคลื่อนจากรากที่สองของสูตรต่อไปนี้ (Lord 1950 cited by Angoff 1984: 97)

$$SEE^2_{x^*} = \frac{2S^2_X}{N_t}(Z^2_y + 2) \quad ----- (๓๔)$$

เมื่อ $N_t = n_X + n_Y$

$Z_y = (X - M_Y)/S_Y$

(๒) เมื่อกลุ่มสุ่มสองกลุ่ม แต่ละกลุ่มทำแบบสอบทั้งสองชุด แต่จัดทำเป็นการทดสอบให้ผู้สอบรับการทดสอบ X หรือ Y ก่อนหลังในลักษณะสลับ สูตรที่ใช้ คือ (Lord 1950 cited by Angoff 1984: 103)

$$SEE^2_{X^*} = s^2_X (1 - r_{XY}) \frac{z^2_y (1 + r_{XY})}{N_t} + 2 \quad -- (๓๔)$$

เมื่อ r_{XY} คือ สัมประสิทธิ์สหสัมพันธ์ของคะแนนชุด X และ Y พิจารณาจากสูตรนี้จะเห็นว่า ถ้า X และ Y มีความสัมพันธ์สูง SEE จะมีค่าต่ำลง เช่น ถ้า r_{XY} มีค่า .๘๐ ค่า $SEE^2_{X^*}$ ที่ $z_y = 0$ จะมีค่าเป็น ๑/๑๐ ของ $SEE^2_{X^*}$ ใน (๑) ซึ่งอธิบายได้ในแง่ของจำนวนผู้สอบว่า กรณีที่เทียบมาตรฐานด้วย (๑) จะต้องใช้จำนวนผู้สอบมากกว่า (๒) เป็น ๑๐ จึงจะได้ผลที่มีความคงชัดเท่ากัน

(๓) เมื่อกลุ่มสุ่มสองกลุ่มทำแบบสอบกลุ่มละชุด (X หรือ Y) และทำแบบสอบร่วมเหมือนกัน (Angoff 1984: 104) ค่าประมาณความคลาดเคลื่อนหาได้จากสูตร (Lord 1950 cited by Angoff 1984: 106) ดังนี้

$$SEE^2_{X^*} = 2\sigma^2_X (1 - \hat{r}^2)\frac{(1 + \hat{r}^2) z^2_y + 2}{N_t} \quad -- (๓๒)$$

โดยมีข้อตกลงว่า

$$\hat{r} = \frac{b_{xv\alpha}\ \hat{\sigma}_v}{\hat{\sigma}_X} = \frac{b_{yv\beta}\ \hat{\sigma}_v}{\hat{\sigma}_Y}$$

เมื่อ $b_{xv\alpha}$ และ $b_{yv\beta}$ คือ สัมประสิทธิ์ถดถอย

และ $\hat{\sigma}_v$ $\hat{\sigma}_X$ และ $\hat{\sigma}_Y$ คือ ค่าประมาณส่วนเบี่ยงเบนมาตรฐานของ V X และ Y ตามลำดับ

พิจารณาจากสูตร ถ้า $\hat{r} = 0$ ค่าประมาณความแปรปรวนคลาดเคลื่อน จะมีขนาดเท่ากับหาได้จาก (๑) ถ้า $\hat{r} = .7$ ค่าที่ได้เป็นเพียงครึ่งหนึ่งที่ $z_y = 0$

๑.๖ ความคลาดเคลื่อนมาตรฐานของการเทียบมาตรารูปแบบอีควิเปอร์เซนไทล์

การศึกษาความคลาดเคลื่อนมาตรฐานของการเทียบมาตรารูปแบบอีควิเปอร์-เซนไทล์ เป็นผลงานที่ค่อนข้างใหม่ของลอร์ด (Lord 1982) ในการพัฒนาสูตรได้กำหนดให้กลุ่มตัวอย่าง ๒ กลุ่ม ทำแบบสอบกลุ่มละชุด และสร้างข้อตกลงให้คะแนนสอบเป็นตัวแปรต่อเนื่อง การหาความแปรปรวนแอสซิมโทติคมีดังนี้

ให้ $F(X)$ และ $G(Y)$ แทนการแจกแจงความถี่สะสมของคะแนน X และ Y ในประชากร P ให้กลุ่มตัวอย่าง N_2 ทำแบบสอบชุด Y แล้วหาสัดส่วน q ที่มีคะแนนต่ำกว่าค่าคงที่ y_o ที่เลือกมาตามความสนใจ ให้กลุ่มตัวอย่าง N_1 ทำแบบสอบชุด X ให้ N_{1gth} เป็นสถิติอันดับในกลุ่มตัวอย่างนี้ โดยที่ x^* สมมูลกับ y_o ดังนั้น แต่ละค่าของ y_o ที่สนใจ จะพบความแปรปรวนเชิงสุ่มของ x^* สำหรับแต่ละค่าของ y_o ถ้าให้ q คงที่ x^* จะมีการแจกแจงอย่างปกติด้วยค่าเฉลี่ย $\mu_{x^*|q}$ มีความสัมพันธ์เป็น $F(\mu_{x^*|q}) = q$ -------- (๓๗)

และความแปรปรวน $\sigma^2_{x^*|q} = pq / N_1 [g(\mu_{x^*|q})]^2$ --------- (๓๘)

เมื่อ $p \equiv 1 - q$

และ $f(x)$ คือ ความน่าจะเป็นของความหนาแน่นที่ X

แต่อย่างไรก็ตาม q เป็นสัดส่วนตัวอย่าง จึงต้องหา $\text{Var } x^*$ เมื่อ q ไม่คงที่จาก

$$\text{Var } x^* = \text{Var}(\mu_{x^*|q}) + E(\sigma^2_{x^*|q}) \quad \text{-------} \quad (๓๙)$$

เมื่อ Var และ E แทนความแปรปรวนและค่าคาดหมายของ x^* สำหรับทุกค่าของ q ที่ y_o คงที่

$$\frac{d}{dq}\mu_{x^*|q} = \frac{1}{g(\mu_{x^*|q})}$$

เนื่องจาก $\text{Var } q = PQ / N_2$

เพราะฉะนั้น $\text{Var}(\mu_{x^*|q}) = PQ / N_2 g_o^2$ ---------- (๔๐)

เมื่อ Q เป็นค่าประจากรของ q ซึ่งนิยามไว้ว่า

$$Q \equiv G(y_o) \qquad \text{---------- (๔๑)}$$

$$P \equiv 1 - Q \qquad \text{และ}$$

$$g_o \equiv g(x_o)$$

$$G(x_o) \equiv Q \qquad \text{---------- (๔๒)}$$

สมการที่ (๔๐)ขึ้นอยู่กับ N_2 ไม่ใช่ N_1 เพราะว่า ความผันแปรใน $\mu_{x^*|q}$ เกิดขึ้นจากความผันแปรของ q ซึ่งเป็นสถิติตัวหนึ่งที่มีความแปรปรวน PQ / N_2 เพื่อประเมิน $E\sigma^2_{x^*|q}$ ให้เขียน (๓๔)โดยขยายอนุกรมและไม่ต้องสนใจเทอมที่มีกำลังสูง ๆ ดังนี้

$$E\sigma^2_{x^*|q} = E \frac{pq}{N_1 g_o^2 \left(1 - \frac{g - g_o}{g_o}\right)^2}$$

$$= E \frac{pq}{N_1 g_o^2} \left(1 - 2\frac{g - g_o}{g_o} + \ldots\right)$$

$$= E \frac{P - P^2}{N_1 g_o^2} = \frac{1}{N_1 g_o^2} \left(P - \text{Var } p - P^2\right)$$

$$\simeq \frac{PQ}{N_1 g_o^2} \qquad \text{---------- (๔๓)}$$

เมื่อ $g \equiv g(\mu_{x^*|q})$ แทน (๔๐) และ (๔๓) ลงใน (๓๔) ได้สูตรสุดท้าย

$$\text{Var } X^* \cong \frac{PQ}{g_0^2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right) \text{ ----------(๔๔)}$$

$$\text{SEE}^2 \cong \frac{pq}{g_0^2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right) \text{ ----------(๔๕)}$$

๑.๓ ความคลาดเคลื่อนมาตรฐานอิงทฤษฎีการตอบข้อสอบ (IRT)

ในทฤษฎีการตอบข้อสอบ (IRT) ค่าประมาณคะแนนจำนวนข้อที่ตอบถูก (number-right scores) ξ ของผู้สอบที่ทำแบบสอบ X และ η ของผู้สอบที่ทำแบบสอบ Y มีค่าเท่ากับฟังชั่นคุณลักษณะที่ประเมินที่ระดับความสามารถ (θ) เดียวกัน ให้ความหมายของความเท่ากันในเชิงคะแนนสมบูรณ์ ในทางปฏิบัติจึงนำความสัมพันธ์เชิงฟังชั่นของ ξ และ η มาใช้เทียบมาตรคะแนน X และ Y จากคะแนนที่สอบได้จริงของคนทั้งสองกลุ่ม การประมาณเพื่อให้เทียบ η ไปยัง ξ ตามฟังชั่นจะต้องใช้ค่าประมาณของประชากรข้อสอบ ค่าประมาณเหล่านี้ คือ ที่มาของความคลาดเคลื่อนเชิงสุ่ม (sampling error) ในการเทียบมาตรา (Lord 1981: 2)

ลอร์ด (Lord 1981) ได้วิเคราะห์สูตรสำหรับหาความคลาดเคลื่อนมาตรฐาน (the asymptotic standard error) ของการเทียบมาตรคะแนนจริงในรูปแบบอิงทฤษฎีการตอบข้อสอบแบบสามพารามิเตอร์ของโลจิส ออกแบบการเทียบมาตราโดยให้กลุ่มผู้สอบ ๒ กลุ่ม ทำการสอบจากแบบสอบคนละฉบับ (X หรือ Y) และทำแบบสอบร่วมชนิดภายนอก (V) เหมือนกัน การเทียบมาตรคะแนน η ไปยัง ξ ใช้วิธีการทำให้ค่าความสามารถอยู่คงที่ในระดับเดียวกัน ในรูปแบบดังกล่าวมีงาน ๒ ขั้นตอน คือ ขั้นแรกเทียบ ω ไปยัง ξ ในผู้สอบกลุ่มแรก ขั้นที่สองเทียบ η ไปหา ω ในผู้สอบกลุ่มที่สอง การทำงาน ๒ ขั้นตอนนี้ จะทำให้เทียบ η ไปยัง ξ ได้ การหาค่าประมาณคะแนนจริงจากฟังชั่นใช้แทนค่าประมาณประชากรข้อสอบที่ได้จากการประมาณด้วยวิธีการความน่าจะเป็นสูงสุด (Maximum likelihood estimates) ซึ่งโดยอาศัย

โปรแกรมสำเร็จรูป LOGIST ช่วยในด้านคำนวณ สมการคะแนนจริงมีดังนี้

$$\hat{\xi} = \sum_{g} \hat{P}_{g1}(\theta_1) \qquad \text{---------- (๔๖)}$$

$$\hat{\omega} = \sum_{g} \hat{P}_{g2}(\theta_1) \qquad \text{---------- (๔๗)}$$

$$\hat{\omega} = \sum_{g} \hat{P}_{g3}(\theta_2) \qquad \text{---------- (๔๘)}$$

$$\hat{\eta} = \sum_{g} \hat{P}_{g4}(\theta_2) \qquad \text{---------- (๔๙)}$$

สมการทั้งสี่นี้ แสดงให้เห็นว่า $\hat{\eta}$ เป็นฟังชั่นของค่าประมาณประชากรข้อสอบทุกค่ารวมกันตามการกำหนดค่า ξ

จากการหาอนุพันธ์ของข้อสอบในแต่ละพารามิเตอร์ a b c ของข้อสอบแต่ละข้อ เมื่อกำหนดให้ ξ คงที่ ทำให้หาค่าประมาณคะแนนจริงของ $\hat{\eta}$ และได้สมการความแปรปรวน $\hat{\eta}$ ดังนี้

$$\text{Var}\ \hat{\eta} \doteq \sum_{p=1}^{4} \sum_{g=1}^{n_p} \sum_{r=1}^{3} \sum_{s=1}^{3} \eta'_{rgp}\, \eta'_{sgp}\, \text{COV}(\hat{t}_{rgp}, \hat{t}_{sgp}) \qquad \text{-----(๕๐)}$$

ค่าที่ได้เป็นค่าประมาณขั้นต่ำ สัญญลักษณ์ที่ใช้มีดังนี้

p คือ จำนวนฟังชั่นของการประมาณคะแนนจริง ซึ่งมี ๔ ฟังชั่น

คือ ๑ หมายถึง กลุ่มที่หนึ่งทำแบบสอบ X

๒ หมายถึง กลุ่มที่หนึ่งทำแบบสอบ ω

๓ หมายถึง กลุ่มที่สองทำแบบสอบ ω

๔ หมายถึง กลุ่มที่สองทำแบบสอบ Y

r และ s หมายถึง พารามิเตอร์ข้อสอบ มี ๓ ตัว คือ

๑ หมายถึง a ๒ หมายถึง b และ ๓ หมายถึง c

n = 1 ถึง n_p หมายถึง ลำดับข้อสอบในแบบสอบแต่ละฉบับ

$\eta'_{rgp} = \partial\eta/\partial t_{rgp}$ คือ ค่าอนุพันธ์ของคะแนนจริง η ที่เทียบกับพารามิเตอร์ของข้อสอบ

$Cov(\hat{t}_{rgq}, \hat{t}_{sgp})$ หมายถึง ความแปรปรวนร่วมของค่าประมาณสูงสุดของพารามิเตอร์ เมื่อกำหนดค่า θ คงที่

จากการศึกษาความคลาดเคลื่อนมาตรฐานของการเทียบมาตรา IRT ลอร์ด (Lord 1981) ได้ใช้ตัวอย่างคะแนนจากการสอบจริงในโครงการทดสอบ SAT ชุด VSA4 มีข้อสอบ ๔๐ ข้อ เทียบมาตราไปสู่แบบสอบชุด XSA2 ซึ่งมี ๔๕ ข้อ คนสอบทุกคนได้ทำแบบสอบร่วม ๔๐ ข้อ เหมือนกัน ผลการประมาณค่าพารามิเตอร์ข้อสอบ และความสามารถได้จากผู้สอบ ๒๒๖๕ คน เมื่อปี ๑๙๗๓ และ ๒๒๘๖ คน เมื่อปี ๑๙๗๕ การวิเคราะห์ความคลาดเคลื่อนได้ผลดังนี้ การเทียบมาตรา IRT ไม่เป็นเส้นตรงเมื่อคะแนนอยู่นอกพิสัยจาก ๓๕๐ ถึง ๖๕๐ คะแนน ค่าความคลาดเคลื่อนเพิ่มขึ้นขณะที่คะแนนจริงน้อยลงจนต่ำสุดที่คะแนนจริง – ๕.๕ แต่ด้านตรงกันข้าม ค่าความคลาดเคลื่อนน้อยลงเมื่อคะแนนมาตราสูงสุดของพิสัย เหตุผลที่อธิบาย คือ คะแนนในส่วนนี้เข้าใกล้คะแนนเต็ม (perfect score) ความคลาดเคลื่อนมีแนวโน้มใกล้ศูนย์ สำหรับคะแนนในด้านต่ำที่ต่ำกว่าระดับการเดาเฉลี่ย ใน IRT ไม่มีคำอธิบายในส่วนนี้ การหาความคลาดเคลื่อนจึงไม่คำนวณเพราะไม่มีความหมาย

๒. <u>ความลำเอียงของการเทียบมาตรา</u> (Bias error)

ความลำเอียง เป็นอีกองค์ประกอบหนึ่งของความคลาดเคลื่อนของการเทียบมาตรา ซึ่งจะต้องพิจารณากันอย่างรอบคอบ เมื่อต้องการใช้ผลของการเทียบมาตรา เมื่อมีการสุ่มตัวอย่างอย่างง่าย สัดส่วนตัวอย่างย่อมเป็นตัวประมาณที่ไม่เอนเอียงของสัดส่วนประชากรที่สมนัยกัน ในกรณีเช่นนี้เมื่อไปใช้กับการเทียบมาตราของแบบสอบคู่ขนานสองชุด ด้วยการใช้รูปแบบการเทียบที่ตำแหน่งเปอร์เซนไทล์ (equipercentile method) ย่อมจะไม่มีความลำเอียง แต่ถ้าในกรณีที่การแจกแจงของแบบสอบสองชุดที่นำมาเทียบนั้นมีการแจกแจงคนละรูปร่างแตกต่างกันตั้งแต่โมเมนต์ที่สองขึ้นไป การใช้ฟังชั่นเทียบมาตราเชิงเส้นตรงย่อมจะทำให้เกิดความลำเอียงได้ (Jaeger 1981: 25)

บรูน และ ฮอลแลนด์ ได้เสนอบทวิเคราะห์ไว้ดังนี้ (๑๙๘๒: ๓๖ – ๓๙) ความลำเอียงเชิงสถิติ มีความหมาย ๒ ประการ คือ

๑. ความแตกต่างของค่าเฉลี่ยของค่าตัวประมาณตลอดการทำการสุ่มซ้ำ ๆ จากประชากรเดียวกัน กับค่าของประชากรที่ถูกประมาณ

๒. ความแตกต่างระหว่างฟังชั่นของการเทียบมาตราประมาณ (estimated equating function) และค่าที่แท้จริงของฟังชั่นการเทียบมาตรา (equating function)

แหล่งของความลำเอียงของการเทียบมาตราที่สำคัญมี ๒ แหล่ง คือ ความผันแปรของประชากร (population variability) และความคลาดเคลื่อนของรูปแบบ (model errors) ซึ่งมีคำอธิบาย ดังนี้

๒.๑ <u>ความแปรเปลี่ยนของประชากร</u> (population variability)

เนื่องจากการเทียบมาตราไม่ได้จัดกระทำกับประชากรเพียงกลุ่มเดียวโดยตลอด ฟังชั่นของการเทียบมาตราที่สร้างขึ้นเพื่อใช้กับประชากรหนึ่ง อาจนำไปใช้กับอีกประชากรหนึ่งไม่ได้ เช่น ถ้า $x_p^*(y)$ เทียบมาตรา Y ไปยังมาตรา X ในประชากร P และ $Y_Q^*(Z)$ เทียบ Z ไปยัง Y ในประชากร Q แล้ว จะรีบด่วนสรุปว่า ฟังชั่นเชิงซ้อน $X_P^*(Y_Q(Z))$ เทียบมาตราจาก Z ไปยัง X ในประชากร P หรือ Q ย่อมไม่ได้ ในแบบแผนการเทียบมาตราที่ใช้แบบสอบร่วมได้แก้ไขปัญหานี้ โดยการสร้างประชากรสังเคราะห์ (synthesis population) ซึ่งเป็นผลรวมของประชากร P และ Q ตามสัดส่วน แต่ก็ยังไม่รับประกันเลยว่า จะไม่มีความลำเอียงในทุก ๆ ประชากรที่แปรเปลี่ยนไป

๒.๒ <u>ความคลาดเคลื่อนรูปแบบ</u> (model errors)

ความคลาดเคลื่อนนี้ เกิดขึ้นจากการจัดกระทำข้อสมมุติฐานที่ว่าด้วยรูปการแจกแจงให้ง่ายขึ้นอย่างผิดพลาด ซึ่งแบ่งได้เป็น ๒ พวก คือ

(๑) ความคลาดเคลื่อนรูปแบบที่ทดสอบได้ (testable model errors) เช่น ความเป็นเส้นตรงของฟังชั่นการถดถอย สามารถทดสอบได้จากข้อมูลจำนวนมาก ซึ่งจะเห็นผลว่า มีความเพียงพอหรือไม่

(๒) ความคลาดเคลื่อนรูปแบบที่ทดสอบไม่ได้ (Nontestable model errors) เป็นกรณีที่ไม่สามารถหาข้อมูลมาทดสอบเพื่อพิสูจน์ว่าสมมุติฐานที่ว่าไว้เพียงพอหรือไม่ ตัวอย่างที่พบ เช่น ความคลาดเคลื่อนที่เกิดจากการประมาณความถี่ในวิธีการเทียบมาตราของทักเกอร์ ซึ่งกล่าวว่า การแจกแจงความถี่ของ X อย่างมีเงื่อนไขเมื่อกำหนด v ใน P เหมือนกับการแจกแจงความถี่ของ X อย่างมีเงื่อนไขเมื่อกำหนด v ใน Q ความคลาดเคลื่อนของการกำหนดเป็นข้อสมมุติฐานนี้ หากเกิดขึ้นจะมีผลก่อให้เกิดความลำเอียงในการประมาณของฟังชั่นการเทียบมาตราได้อย่างมาก

สำหรับเรื่องของการวิเคราะห์แหล่งความคลาดเคลื่อนทั้งที่เป็นความแปรปรวนของการสุ่มและความลำเอียง ยังต้องมีการวิจัยอีกมากเพื่อสร้างวิธีปฏิบัติที่เหมาะสมกับปัญหาในแง่มุมต่าง ๆ (Braun and Holland 1982: 32) ดังนั้น การวิจัยครั้งนี้ผู้วิจัยจะประเมินรูปแบบการเทียบมาตราต่าง ๆ ด้วยการใช้มาตรการของความคลาดเคลื่อนมาตรฐานของการเทียบมาตราที่แนะนำโดยออร์คเท่านั้น จะไม่พิจารณาในประเด็นของความลำเอียงของการเทียบมาตรา

<u>การประเมินความเพียงพอของการเทียบมาตรา</u>

วิธีการเทียบมาตราละแนนแต่ละวิธี ประกอบขึ้นด้วยรูปแบบของการเทียบมาตรา (models) ซึ่งมีข้อกำหนดที่ว่าด้วยสมมุติฐานเบื้องต้น (assumptions) ของแต่ละรูปแบบ และประกอบด้วยการออกแบบ (designs) เพื่อจัดเก็บข้อมูลให้เป็นไปตามข้อตกลงต่าง ๆ ถ้าหากทุกอย่างเป็นไปตามเงื่อนไข เชื่อได้ว่าผลของการเทียบมาตราจะมีความถูกต้อง (accurate) และความคงรัด (precise) ตามทฤษฎี แต่ความเป็นจริงของการทดสอบมักไม่ได้เป็นไปตามอุดมการ เพราะมีหลายสิ่งหลายอย่างที่อยู่นอกเหนือการควบคุม เช่น โปรแกรม

การทดสอบระดับชาติ ซึ่งจะมีทั้งนโยบาย และกฎหมายที่อยู่นอกเหนือข้อตกลงเชิงทฤษฎี ตัวแบบสอบเองก็จำเป็นต้องมีการเปลี่ยนชุดไป ดังนั้น ข้อมูลที่นักจิตมิตินำมาใช้ จึงไม่สามารถระบุว่าเป็นตัวอย่างประชากรได้อย่างชัดเจน และโดยความเป็นจริงเป็นการจัดกระทำกับค่าประชากรมากกว่า (population quantities) ไม่ใช่ค่าประมาณ (sample estimates) (Braun and Holland 1982: 10)

ด้วยสภาพความเป็นจริงดังกล่าว ทำให้การเทียบมาตราที่ต้องจัดกระทำอยู่นั้นอยู่ในภาวะที่มีเงื่อนไขน้อยกว่าความพอดีตามข้อตกลงในแต่ละรูปแบบ ดังนั้น การเทียบมาตราที่ได้พัฒนาขึ้น จึงจำเป็นต้องมีการตรวจสอบความเพียงพอของรูปแบบ (the adequacy of equating models) (Petersen, Marco and Stewart 1982: 71) วิธีการประเมินความเพียงพอมีผู้เสนอแนวความคิดและวิถีปฏิบัติไว้ดังนี้

• <u>ดัชนีตรวจสอบความเพียงพอของเจเกอร์</u> (Jaeger)

เจเกอร์ (Jaeger 1981: 26) ได้เสนอดัชนี ๕ ตัว เพื่อตรวจสอบความเพียงพอของการใช้รูปแบบเชิงเส้นตรงว่า เทคนิควิธีที่ได้นำมาใช้ในการเทียบเพียงพอกับการปรับความแตกต่างระหว่างการแจกแจงของคะแนนจากแบบสอบหรือไม่ หรือจำเป็นต้องมีรูปแบบอื่นที่มีความสลับซับซ้อนมากขึ้น ดัชนี ๕ ตัว มีดังนี้ คือ

๑.๑ ดัชนีความคลาดคลึงของการแจกแจงคะแนนสะสมของแบบสอบเก่าและชุดใหม่ โดยการปรับความแตกต่างระหว่างค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน การทดสอบความเหมือนของการแจกแจงใช้การทดสอบด้วย the Kolmogorov-Smirnov two-sample test (Smirnov 1948 cited by Jaeger 1981: 27)

๑.๒ รูปแบบของการแจกแจงคะแนนดิบกับคะแนนแปลง (Shape of the raw score tranformation) เหตุผลของการใช้ดัชนีตัวนี้ คือ ถ้าการเทียบมาตราด้วยรูปแบบเชิงเส้นตรงสามารถอธิบายความแตกต่างในการแจกแจงของคะแนนดิบทั้งสองชุดอย่างเพียงพอ ก็เป็นเหตุผลอย่างเพียงพอเช่นกันที่จะยอมรับว่าการแปลงคะแนนดิบจากแบบสอบชุดใหม่ไปยังแบบสอบชุดเก่าเป็นเส้นตรงอย่างแน่นอน

๑.๓ ความคงเส้นคงวาของผลลัพธ์ของการเทียบมาตราตามรูปแบบเชิงเส้นตรงกับการเทียบที่ตำแหน่งเปอร์เซนไทล์ (Consistency of linear and equipercentile equating results) การวิเคราะห์นี้อาศัยข้อตกลงที่เป็นสมมุติฐานเบื้องต้นที่ว่า ถ้ารูปแบบเชิงเส้นตรงมีความเพียงพอแล้ว ฟังชั่นของรูปแบบการเทียบที่ตำแหน่งเปอร์เซนไทล์จะแปรผันไปโดยสุ่มรอบ ๆ ฟังชั่นของรูปแบบเชิงเส้นตรงที่สมนัยกัน

๑.๔ ความคล้ายคลึงของการแจกแจงความยากของข้อสอบ (Similarity of item difficulty distributions) โดยอาศัยหลักการที่ว่า การใช้รูปแบบเชิงเส้นตรงมีความเพียงพออย่างแท้จริงกับแบบสอบที่มีคุณสมบัติเป็นคู่ขนานกัน ถ้ามีการเบี่ยงเบนจากความเป็นคู่ขนานมากเท่าใด แสดงว่าต้องการรูปแบบการเทียบมาตราที่ซับซ้อนขึ้น เพราะการแจกแจงของแบบสอบที่ไม่ใช่คู่ขนานจะมีความแตกต่างเกิดขึ้นในระดับโมเมนท์ที่สูงขึ้น

๑.๕ ความคล้ายคลึงของค่าอำนาจจำแนกของข้อสอบ (Similarity of item discrimination distributions) มีเหตุผลทำนองเดียวกับดัชนีตัวที่ ๔

๒ <u>ดัชนีความแตกต่าง</u> (Discrepancy Indecies)

ปีเตอร์สัน มาร์โค และ สตีเวอร์ท (Petersen, Marco and Stewart 1982: 91) ได้เสนอวิธีประเมินความเพียงพอของรูปแบบการเทียบมาตราโดยการหาความแตกต่างระหว่างคะแนนแปลง (an estimated criterion score, t') ซึ่งเป็นผลจากการเทียบมาตรากับคะแนนเกณฑ์ (criterion score, t) ที่สมนัยกัน ถ้าความแตกต่างมีค่าน้อย มีความหมายว่า ความคลาดเคลื่อนที่เกิดขึ้นจากการใช้รูปแบบการเทียบมาตรานั้น ๆ น้อยด้วย รูปแบบการเทียบมาตราดังกล่าวย่อมมีความเหมาะสมกับสถานการณ์ที่ใช้

ดัชนีความแตกต่างที่นำมาเปรียบเทียบระหว่างรูปแบบและสถานการณ์ต่าง ๆ กัน คือ กำลังสองของค่าเฉลี่ยของความแตกต่างที่ถ่วงน้ำหนักด้วยความแปรปรวนของคะแนนเกณฑ์ (the weigthed mean-square difference) เป็นค่ามาตรฐาน ค่าที่คำนวณออกมานี้เรียกว่า ความคลาดเคลื่อนรวม (the total error) ซึ่งมีสูตรดังนี้

$$\text{total error} = \sum_j f_j d_j^2 / n s_t^2 \quad \text{--------} \quad (๔๐)$$

เมื่อ d_j = $(t - t')$

n = จำนวนคะแนนที่ใช้

s_t^2 = ความแปรปรวนของคะแนน t

๓ ดัชนีเปรียบเทียบเปอร์เซนไทล์ (the percentile comparision index)

เป็นมาตราการวัดความไม่สอดคล้องระหว่างการแจกแจงของคะแนนในแบบสอบชุด X กับแบบสอบชุด Y ที่ได้แปลงไปอยู่ในมาตราคะแนนของ X แล้วตามวิธีที่ได้พัฒนาขึ้น ดัชนีเปรียบเทียบเปอร์เซนไทล์ คือ ค่าเฉลี่ยกำลังสองของความแตกต่าง (the mean-squared difference) ที่ได้จากการแจกแจงของคะแนนต่าง ๆ ของเกณฑ์ X กับคะแนนแปลง X^* ที่แปลงมาจาก Y ด้วย วิธีการเทียบมาตราที่ระบุไว้ ๓ ที่ตำแหน่งเปอร์เซนไทล์เดียวกัน ดัชนีตัวนี้ได้เสนอโดย โคเลน (Kolen 1982) ได้แนะนำให้ใช้ข้อมูลจากกลุ่มตัวอย่างสอบทานผลซึ่งได้สุ่มมาจากประชากรเดียวกันกับกลุ่มตัวอย่างที่ใช้พัฒนาตารางคะแนนแปลง สูตรการคำนวณมีดังนี้ คือ

$$C = \sum_i (x_i - x_i^*)^2 / nk \quad \text{--------} \quad (๔๒)$$

เมื่อ n คือ จำนวนของคะแนนดิบของกลุ่มสอบทานผล

k คือ จำนวนข้อสอบในแบบสอบร่วมที่ใช้

ค่า C ที่ได้ ถ้ามีค่าน้อยจะให้ความหมายว่า รูปแบบการเทียบมาตราที่นำมาสร้างตารางคะแนนแปลงนั้นมีความเหมาะสมและเพียงพอที่จะให้ผลของการแปลงคะแนนอย่างคงเส้นคงวา

วิธีการประเมินความเพียงพอที่ได้กล่าวมานี้ อาจจำแนกเป็นสองพวก พวกแรก เป็นการประเมินก่อนดำเนินการ เช่น ดัชนีความคล้ายคลึงของการแจกแจงคะแนนสะสมของแบบสอบสองชุด

ดัชนีความคล้ายคลึงของการแจกแจงของค่าความยากของข้อสอบ เป็นต้น พวกหลัง เป็นการประเมินผลของการเทียบมาตรา ซึ่งอาศัยคะแนนเกณฑ์ที่เลือกแล้วเป็นหลักในการเทียบหาความแตกต่าง สำหรับดัชนีที่เสนอโดย ปิเตอร์สันและคณะนั้น คะแนนเกณฑ์ที่ใช้ คือ ผลการแปลงคะแนนด้วยรูปแบบอิงพฤษฎีการตอบข้อสอบ (๑๙๘๒: ๔๒) ค่าดัชนีที่คำนวณมีลักษณะเป็นหน่วยมาตรฐานแล้ว สามารถนำค่าเหล่านี้ที่ได้จากการใช้รูปแบบที่ต่างกันตลอดจนสถานการณ์ได้ข้อมูลที่ต่างกันมาเปรียบเทียบกันโดยตรงได้ ส่วนดัชนีที่เสนอโดย โคเลน (Kolen 1982) ได้ใช้ข้อมูลคะแนนจากผู้สอบเองเป็นเกณฑ์ในการหาความแตกต่าง ข้อมูลเหล่านี้ได้มาจากการออกแบบด้วยการใช้กลุ่มตัวอย่างสอบหานผล ซึ่งผู้สอบในกลุ่มตัวอย่างนี้ได้รับการทดสอบด้วยแบบสอบทั้งสองชุด ดังนั้น การใช้คะแนนของคนเองเป็นเกณฑ์จึงมีความอิสระไม่ต้องขึ้นกับกระบวนการแปลงคะแนนอื่น ๆ เช่นวิธีที่เสนอโดย ปิเตอร์สันและคณะด้วยเหตุผลดังกล่าว ผู้วิจัยจึงได้เลือกวิธีการประเมินความเพียงพอจากการวิเคราะห์ผลในกลุ่มตัวอย่างสอบหานผล แต่การหาค่าดัชนีได้ดัดแปลงจากสูตรของโคเลน ไปใช้ตามแนวความคิดของปิเตอร์สันและคณะ คือ ใช้ค่าความแปรปรวนเป็นตัวถ่วงน้ำหนักเพื่อให้ค่าที่ได้มีหน่วยเป็นมาตรฐาน

สูตรที่ดัดแปลงและนำมาใช้ในการศึกษาครั้งนี้ คือ

$$C = \sum_{i} (x_i - x_i^*)^2 \ / n\, s_x^2 \qquad \text{----------} \ (๔๓)$$

เมื่อ x_i คือ คะแนนเกณฑ์หรือคะแนนจากการสอบชุด X ของคนที่ i

x_i^* คือ คะแนนที่ได้เทียบด้วยตารางแปลงคะแนนที่สมนัยกันของคนที่ i

n คือ จำนวนคนในกลุ่มตัวอย่างสอบหานผลที่นำมาวิเคราะห์

s_X^2 คือ ค่าความแปรปรวนของคะแนน X

<u>งานวิจัยที่เกี่ยวข้องกับการเปรียบเทียบรูปแบบการเทียบมาตรา</u>

งานวิจัยที่มุ่งตรวจสอบผลจากวิธีการเทียบมาตราแบบต่าง ๆ ในสถานการณ์ที่ได้ออกแบบจัดกระทำข้อมูลต่าง ๆ กันในปัจจุบันนี้ นับว่ามีมากพอที่จะให้วิเคราะห์ในเชิงวรรณคดี แต่การลงข้อสรุปที่ว่า วิธีการใดให้ผลที่แม่นยำที่สุดนั้นยังไม่อาจทำได้ ในการวิเคราะห์จากวรรณคดีพบประเด็นที่เป็น

เงื่อนไขต่อผลการสรุปที่ต้องพิจารณาต่าง ๆ กัน ซึ่งจะกล่าวในรายละเอียดต่อไป

สไลด์ และ ลิน (Slinde and Linn 1977, 1978, 1979) ได้ทำการตรวจสอบถึงปัญหาของการเทียบมาตราในแนวดิ่ง (vertical equating) ของแบบสอบสองชุดที่สร้างขึ้นเพื่อใช้กับประชากรที่มีระดับความสามารถต่างกันด้วยวิธีออม จากการศึกษาได้ให้ข้อเสนอแนะว่า การใช้รูปแบบการเทียบมาตราสามรูปแบบ คือ เชิงเส้นตรง อีควิเปอร์เซนไทล์ และ IRT ชนิดหนึ่งพารามิเตอร์รูปแบบโลจิสอาจมีข้อจำกัดในกระบวนการเทียบมาตราในแนวดิ่ง โดยเฉพาะอย่างยิ่งเมื่อทำการเทียบมาตราโดยใช้แบบสอบสองชุดที่มีความแตกต่างกันมาก และกลุ่มตัวอย่างสองกลุ่มที่มีระดับความสามารถต่างกันมาก จากการศึกษานี้ให้ข้อสังเกตว่า ถ้าใช้รูปแบบ IRT ชนิดสามพารามิเตอร์ของโลจิสอาจให้ผลการเทียบมาตราที่ดีกว่าในสถานการณ์เช่นนั้น

ปีเตอร์สัน มาร์โค และ สตีเวอร์ท (Petersen, Marco and Stewart 1982; 71-135) เป็นกลุ่มนักวิจัยที่ทำการศึกษาเชิงประจักษ์ถึงเทคนิคการเทียบมาตราต่าง ๆ อย่างกว้างขวางที่สุด การเทียบมาตราได้จัดกระทำโดยเทียบแบบสอบชุด SAT (Scholastic Aptitude Test) ที่จัดสอบเมื่อเดือนเมษายน ๑๙๗๕ กับแบบสอบชุด TSWE (Test of Standard Written English) ที่จัดสอบเมื่อพฤศจิกายน ๑๙๗๕ ข้อมูลที่นำมาใช้เป็นบันทึกผลการสอบรายข้อของผู้เข้าสอบทั้งหมด กลุ่มตัวอย่างได้สร้างขึ้นโดยการกำหนดลักษณะเฉพาะ และเลือกคะแนนตามจุดมุ่งหมาย มีทั้งคะแนนสอบที่ต้องการเทียบ และแบบสอบร่วม คำนวณค่าสถิติของข้อสอบรายข้อ และข้อมูลอื่น ๆ ที่ต้องการใช้ มีกลุ่มตัวอย่างพื้นฐาน ๓ กลุ่ม จากการสอบเมื่อเดือนเมษายน ๑๙๗๕ และอีก ๓ กลุ่ม จากการสอบเมื่อเดือนพฤศจิกายน ๑๙๗๕ กลุ่มตัวอย่างไม่มีคนสอบซ้ำประกอบด้วยกลุ่มละ ๔๗๓๐ คน แล้วจัดเป็นกลุ่มย่อยโดยออกแบบการสุ่มให้ลักษณะแตกต่างกัน ๓ กลุ่มย่อย คือ กลุ่มสุ่มอย่างง่าย กลุ่มความสามารถคล้ายคลึงกัน (ใช้ค่าเฉลี่ยความสามารถทางภาษาเป็นเกณฑ์) และกลุ่มไม่คล้ายคลึงกัน การออกแบบเทียบมาตราได้ใช้แบบสอบ SAT ส่วนที่เป็นภาษามาแยกเป็นแบบสอบ ๒ ชุด แล้วนำมาเทียบมาตรากันโดยใช้แบบสอบร่วมลักษณะต่าง ๆ กัน คือ ต่างกันในประเด็นของตำแหน่งแบบสอบร่วม (ภายนอกและภายใน) ประเด็นของเนื้อหา (คล้ายคลึงกันและไม่คล้ายคลึงกัน) ประเด็นของความยาก (คล้ายคลึงกันและไม่คล้ายคลึงกัน) ส่วนในเรื่องของกลุ่มตัวอย่างที่นำเข้ามามีทั้ง ๓ ลักษณะดังกล่าวแล้ว นอกจากนี้ได้ทำการเทียบมาตราแบบสอบ SAT กับแบบสอบที่สร้าง

ขึ้นจากแบบสอบ TSWE โดยใช้แบบสอบร่วมชนิดภายใน ลักษณะของแบบสอบร่วมที่ใช้กับแบบสอบที่เทียบตัวเอง (SAT) ได้ศึกษาทั้งที่มีเนื้อหาคล้าย และไม่คล้ายกับฉบับที่ต้องการเทียบ การประเมินความเพียงพอของวิธีการเทียบมาตราในสถานการณ์ต่าง ๆ ที่ทำการศึกษาอาศัยดัชนีความแตกต่าง (Discrepancy index) ดัชนีนี้ คือ ค่าที่ถ่วงน้ำหนักแล้วของค่าเฉลี่ยกำลังสองของความแตกต่างระหว่างค่าประมาณ หรือคะแนนแปลงกับคะแนนเกณฑ์ (The weighted mean-square difference) คะแนนเกณฑ์ที่ใช้ คือ คะแนนแปลงจากคะแนนคิบชุดเดียวกัน ด้วยรูปแบบการเทียบมาตรา IRT ค่านี้ทำให้เป็นมาตรฐานด้วยการทำให้เป็นสัดส่วนของความเบี่ยงเบนมาตรฐานของคะแนนเกณฑ์ ซึ่งนำไปใช้เปรียบเทียบผลที่ได้จากการวิเคราะห์ต่างสถานการณ์และต่างรูปแบบที่ใช้ ค่าดัชนีให้ความหมายตามหลักการ ดังนี้ ค่าดัชนีความแตกต่าง หรือที่เรียกว่า ค่าความคลาดเคลื่อนรวม (TE) ซึ่งคำนวณจากสูตรที่ (๕๑) เมื่อส่วนเบี่ยงเบนมาตรฐานของคะแนนเกณฑ์ที่ใช้มีค่าเท่ากับ ๑.๐๐ ปริมาณของ TE มีความหมายในเชิงคุณภาพของวิธีการเทียบมาตรา ดังนี้

	TE	$\leq$	๒๕	คือ น่าพอใจอย่างมาก	(Zone A)
๒๕	$<$ TE	$\leq$	๑.๐๐	คือ น่าพอใจ	(Zone B)
๑.๐๐	$<$ TE	$\leq$	๑๒๕	คือ ปานกลาง	(Zone C)
๑๒๕	$<$ TE	$\leq$	๔.๐๐	คือ ไม่น่าพอใจ	(Zone D)
๔.๐๐	$<$ TE			คือ ไม่น่าพอใจอย่างมาก	(Zone E)

ระดับต่าง ๆ ที่กำหนดนี้มาจากหลักเหตุผลที่ว่า ถ้าความคลาดเคลื่อนรวม (TE) เกิดขึ้นเนื่องจากความลำเอียงเพียงสาเหตุเดียวแล้ว ย่อมเป็นเหตุผลที่ทำให้นักจิตมิติยอมรับได้ว่า ผลการเทียบมาตราอยู่ในระดับน่าพอใจอย่างมาก เมื่อค่าเฉลี่ยความคลาดเคลื่อนต่างจากความจริง (truth) ไม่เกินไปกว่าร้อยละ ๕ ของส่วนเบี่ยงเบนมาตรฐานของคะแนนเกณฑ์ และจะไม่เป็นที่ยอมรับอย่างแน่นอนถ้าผลการเทียบมีค่าเฉลี่ยต่างจากความจริงถึงร้อยละ ๒๐ ของส่วนเบี่ยงเบนมาตรฐานของคะแนนเกณฑ์ จุดตัดที่กำหนดขึ้นนี้ถึงแม้จะเป็นการกำหนดโดยคณะผู้วิจัยเอง แต่เป็นวิธีการที่มีเหตุผลและสามารถทำให้ศึกษาเปรียบเทียบผลการเทียบมาตราที่เกิดจากการใช้วิธีการ และสถานการณ์ที่ต่างกันได้ ผลการวิจัยสรุปไว้ ดังนี้

(๑) รูปแบบการเทียบเชิงเส้นตรงโดยใช้เทคนิคของพอททอฟ (Potthoff) ให้ผลไม่น่าพอใจเกือบทุกสถานการณ์ที่ศึกษา

(๒) ในกรณีที่ใช้กลุ่มตัวอย่างขนาดเท่ากัน รูปแบบอีควิเปอร์เซนไทล์ให้ค่าความคลาดเคลื่อนมากกว่าการเทียบด้วยวิธีการเชิงเส้นตรง

(๓) รูปแบบเชิงเส้นตรงโดยเทคนิคของทักเกอร์ ๑ ให้ค่าความคลาดเคลื่อนขนาดใหญ่กว่าเชิงเส้นตรงในกรณีที่ใช้กลุ่มตัวอย่างแตกต่างกัน

(๔) การใช้รูปแบบเชิงเส้นตรงให้ผลไม่เหมาะสม เมื่อเกิดความสัมพันธ์เชิงเส้นโค้ง ซึ่งเนื่องจากความยากของแบบสอบสองชุดต่างกัน

(๕) รูปแบบอีควิเปอร์เซนไทล์ให้ผลการเทียบดีกว่ารูปแบบเชิงเส้นตรง เมื่อแบบสอบสองชุดมีความสัมพันธ์เชิงเส้นโค้ง ซึ่งเกิดจากความยากที่ต่างกันของแบบสอบสองชุด

(๖) ถ้าความยากของแบบสอบร่วมต่างกับแบบสอบฉบับเทียบตัวเองมาก ความคลาดเคลื่อนจะเกิดมากขึ้นด้วย

(๗) แบบสอบที่ทำการเทียบมาตราสองชุด ถ้ามีความยากต่างกันมาก จะเกิดความคลาดเคลื่อนมากด้วย

(๘) ถ้าแบบสอบร่วมได้สร้างให้มีลักษณะคล้ายกับฉบับที่เทียบมาตรามาก หรืออาจเรียกว่าเป็นฉบับย่อส่วนแล้ว การเทียบมาตราจะให้ผลดีที่สุด

ผู้วิจัยได้ให้ข้อสังเกตว่า การสรุปผลการวิจัยเป็นการทั่วไปนั้นต้องระมัดระวังมาก เพราะการศึกษานี้ใช้เฉพาะแบบสอบส่วนที่เป็นภาษาเท่านั้น การศึกษายังไม่ได้แยกผลของความแตกต่างในส่วนที่เกี่ยวกับเนื้อหา และความยากอย่างอิสระ และสุดท้าย คือ เกณฑ์การเทียบอาจให้ความลำเอียงในสถานการณ์บางอย่าง อย่างไรก็ดีผลการวิจัยอาจให้ประโยชน์คือการนำไปใช้โดยที่ผู้ใช้ต้องพิจารณาความเหมาะสมของรูปแบบกับสถานการณ์ที่กำหนด และลักษณะเฉพาะของกลุ่มตัวอย่าง นอกจากนี้ผู้วิจัยมีความเห็นว่า ควรมีการพัฒนาในส่วนที่เป็นกฎการตัดสินใจ เพื่อชี้ให้เห็นว่า ควรเลือกใช้รูปแบบใดกับลักษณะแบบสอบและกลุ่มตัวอย่างคนที่แปรเปลี่ยนไป

ปีเตอร์สัน คุก และ สตอกกิง (Petersen, Cook and Stocking 1981) ทำการศึกษาเปรียบเทียบความคงเส้นคงวาของมาตราที่เกิดจากการเทียบด้วยรูปแบบ IRT กับรูปแบบดั้งเดิมสองรูปแบบ คือ อีควิเปอร์เซนไทล์ และรูปแบบเชิงเส้นตรง รูปแบบเชิงเส้นตรงที่นำมาศึกษามี ๓ วิธี คือ Tucker model, Levine Equally Reliable model และ Levine unequal Reliable model ส่วน IRT ใช้ ๓ วิธี คือ วิธีปัจจุบัน วิธีกึ่งวิเคราะห์ค่าโดยกำหนดค่าความยากคงที่ และวิธีกึ่งวิเคราะห์ค่าโดยกำหนดค่าความยากที่เทียบแล้ว การศึกษาครั้งนี้ต้องการประเมินขนาดของการเบี่ยงเบนของมาตราในกระบวนการเทียบมาตราที่เป็นลูกโซ่ การเบี่ยงเบนของมาตรานี้ คือ ความแตกต่างของคะแนนแปลงในแบบสอบ (โดยได้ผ่านการแปลงหลายครั้งเหมือนลูกโซ่) กับคะแนนเดิมก่อนแปลง เช่น คะแนน v4 เทียบกับ v4 ที่ได้จากการใช้วิธีการเทียบมา ๔ ครั้ง กลุ่มตัวอย่างสุ่มจากการทดสอบ ๔ ครั้ง ในรอบ ๖ ปี แยกเป็นกลุ่มของฉบับภาษา และกลุ่มของฉบับคณิตศาสตร์ แต่ละฉบับมีกลุ่มตัวอย่างย่อย ๑๒ กลุ่ม กลุ่มละ ๒๖๗๐ คน โดยประมาณ แบบสอบร่วมมีเนื้อหากำหนดแน่นอนไม่มีการเปลี่ยนแปลง การเปรียบเทียบความคลาดเคลื่อนใช้การพิจารณาจากภาพกราฟ และดัชนีความแตกต่าง (a discrepancy index) ซึ่งเป็นค่าถ่วงน้ำหนักแล้วของค่าเฉลี่ยความแตกต่างกำลังสอง (The weighted mean square difference) ความแตกต่างคำนวณได้จากผลต่างของคะแนนมาตราเกณฑ์, t (คะแนนดั้งเดิมของ v4) กับคะแนนมาตราที่แปลงด้วยวิธีการต่าง ๆ ที่ศึกษา, t′ ดัชนีนี้เป็นดัชนีสรุปช่วยในการประเมินประสิทธิภาพของรูปแบบการเทียบมาตราต่าง ๆ ผลการวิจัยพบว่า ในการใช้รูปแบบเชิงเส้นตรงกับการเทียบมาตราสายแบบสอบภาษา และสายแบบสอบคณิตศาสตร์ของ SAT มีผลที่คล้ายคลึงกัน

Tucker model ให้ผลความคลาดเคลื่อนมากที่สุด ส่วน Levine equally Reliable model ให้ความคลาดเคลื่อนน้อยที่สุด รูปแบบอีควิเปอร์เซนไทล์ให้ผลโดยทั่วไปไม่น่าพอใจเมื่อเทียบกับรูปแบบเชิงเส้นตรง เหตุผลที่น่าจะอธิบายได้ คือ รูปแบบอีควิเปอร์เซนไทล์ไม่สามารถให้สารสนเทศได้ตลอดพิสัยของคะแนนของแบบสอบ ถ้ากลุ่มตัวอย่างที่ใช้ไม่มีการแจกแจงโดยตลอด รูปแบบ IRT ซึ่งทำการศึกษา ๓ วิธีนั้นพบว่า ลักษณะผลการเทียบโดยทั่วไปที่ทำกับแบบสอบภาษาและแบบสอบคณิตศาสตร์ มีความแตกต่างกัน วิธีที่ใช้ค่า b ที่เทียบแล้วให้ผลน่าพอใจที่สุดกับแบบสอบภาษา แต่ให้ผลตรงกันข้ามกับแบบสอบคณิตศาสตร์ สำหรับวิธี IRT ปัจจุบัน ให้ผลด้อยกว่าวิธีที่ใช้ค่า b เทียบแล้วเพียงเล็กน้อยในแบบสอบภาษา แต่เมื่อเปลี่ยนเป็นแบบสอบคณิตศาสตร์ วิธี IRT ปัจจุบันให้ผลดีกว่ามาก เปรียบเทียบผลของรูปแบบ IRT กับรูปแบบดั้งเดิม พบว่า ความสัมพันธ์ที่เกิดขึ้นกับแบบสอบภาษาต่างกับแบบสอบคณิตศาสตร์ สำหรับการเทียบในสายของภาษารูปแบบ IRT ดีกว่า แต่ในสายคณิตศาสตร์รูปแบบ Levine และ IRT ปัจจุบันให้ผลคล้ายคลึงกันมาก ผลการวิจัยได้สรุปเป็นข้อเสนอว่า ถ้าหากจำเป็นต้องเลือกรูปแบบการเทียบเพียงรูปแบบเดียวในการเทียบกับวิชาทางภาษาและคณิตศาสตร์แล้ว รูปแบบ IRT ปัจจุบันเป็นวิธีที่เหมาะสมที่สุดใน ๗ วิธีที่ได้ศึกษามา แต่ไม่ได้หมายความว่า รูปแบบ IRT ปัจจุบันจะให้ผลดีกว่าเสมอไปเมื่อข้อมูลที่ใช้ต่างไปจาก SAT

โคเลน (Kolen 1981) เปรียบเทียบผลการเทียบมาตราระหว่างรูปแบบดั้งเดิมสองวิธี คือ รูปแบบเชิงเส้นตรง และรูปแบบอีควิเปอร์เซนไทล์ กับรูปแบบ IRT ๓ วิธี โดยใช้ข้อมูลการทดสอบนักเรียนในรัฐไอโอวา ปี ๑๙๗๕ ในโครงการ the Iowa Tests of Educational Development (ITED) ข้อสอบที่ทำการศึกษา คือ แบบสอบวัดสัมฤทธิผล ที่ใช้กับระดับที่ ๑ คือ ระดับการศึกษาเกรด ๙ และ ๑๐ ระดับที่ ๒ คือ ที่ใช้กับเกรด ๑๑ และ ๑๒ ในแต่ละระดับแบบสอบได้พิมพ์ใช้ในปีต่าง ๆ สำหรับฉบับที่นำไปศึกษา คือ ฉบับที่พิมพ์ครั้งที่ ๗ เทียบไปหาครั้งที่ ๖ โดยให้ X7 และ Y7 ใด ๆ เป็นแบบสอบสองชุดที่ต้องการเทียบกัน เป็นแบบสอบที่ไม่เป็นคู่ขนานกัน การเทียบมาตรานั้น ต่างเทียบ X7 และ Y7 ไปสู่มาตราร่วมที่กำหนดขึ้น คือ X6 โคเลน ได้ออกแบบการรวบรวมข้อมูลโดยสุ่มโรงเรียน ๓๔ แห่ง กำหนดให้นักเรียนระดับ ๑ และ ๒ ได้รับการทดสอบในทำนองเดียวกัน คือ ใช้วิธีกำหนดให้นักเรียนได้รับการทดสอบเพียงชุดเดียว X6 X7 หรือ Y7 โดยให้สอบชุดละ ๑ ใน ๓ ของคนสอบทั้งหมดโดยการสุ่ม สำหรับข้อมูลในกลุ่มสอบทานผล คือ นักเรียน

ทุกคนที่ ๓ ของแต่ละชุดแต่ละระดับของแบบสอบที่ใช้ ขั้นตอนการดำเนินการเทียบรูปแบบอีควิเปอร์-เซนไทล์ และรูปแบบเชิงเส้นตรงได้จัดกระทำตามวิธีของแองกอฟ (Angoff 1984) ส่วนรูปแบบ IRT ที่ศึกษามี ๗ วิธี คือ ใช้ ๑, ๒ หรือ ๓ พารามิเตอร์ แต่ละวิธีเทียบด้วยค่าประมาณคะแนนจริง (Estimated true score equating) และเทียบด้วยการประมาณค่าสังเกต (Estimated observed score equating) วิธีสุดท้าย คือ Rasch model เกณฑ์ที่ใช้ในการประเมินผลการเทียบมาตราวิธีต่าง ๆ มีหลายวิธี สำหรับโคเลน ได้เลือกวิธีการตรวจสอบผลสุดท้ายของการเทียบมาตราในแต่ละวิธีจากกลุ่มตัวอย่างสอบทานผล ซึ่งเป็นกลุ่มตัวอย่างอิสระ ดัชนีที่ใช้เปรียบเทียบ คือ ค่าเฉลี่ยกำลังสองของความแตกต่าง (mean squared difference) ระหว่างคะแนนจากแบบสอบชุดเก่ากับชุดใหม่ที่แปลงไปสู่มาตราชุดเก่าด้วยวิธีต่าง ๆ กันในแต่ละตำแหน่งเปอร์เซนไทล์เดียวกัน ค่าสถิติจากการวิเคราะห์กลุ่มสอบทานผลนำมาทดสอบความแตกต่างอย่างมีนัยสำคัญทางสถิติด้วยการทดสอบฟรายด์แมน (Friedman test) ผลการวิจัยพบว่า ผลการเทียบมาตรา ๗ วิธี แตกต่างอย่างมีนัยสำคัญที่ระดับ .๐๑ และ .๐๕ เมื่อใช้กับแบบทดสอบระดับที่ ๑ และ ๒ ตามลำดับ เมื่อแยกพิจารณาแบบสอบระดับที่ ๑ ซึ่งมีความยากต่างกันระหว่างแบบสอบชุดเก่าและชุดใหม่ พบว่า วิธี IRT สามพารามิเตอร์ประมาณค่าการสังเกตให้ผลได้อย่างแน่นอนมากที่สุด ส่วนวิธีที่ให้ผลไม่แน่นอนที่สุด คือ วิธีรูปแบบเชิงเส้นตรง และถัดขึ้นมา คือ วิธี IRT พารา-มิเตอร์เดียว สำหรับแบบสอบของระดับที่ ๒ วิธีที่ให้ผลแน่นอนที่สุด คือ วิธี IRT สามพารามิเตอร์ประมาณคะแนนจริง ส่วนวิธีที่ให้ผลแน่นอนน้อยที่สุด คือ IRT พารามิเตอร์เดียว และถัดขึ้นมา คือ วิธีรูปแบบเชิงเส้นตรง โคเลน ได้เสนอการอภิปรายสรุปโดยส่วนรวมว่า วิธีที่ให้ผลการเทียบมาตราที่ไม่เพียงพอที่สุด คือ วิธี IRT พารามิเตอร์เดียว ทั้งนี้มีข้อน่าสังเกตปัจจัยที่อาจเป็นสาเหตุ คือ การเดาซึ่งโดยรูปแบบการประมาณค่าพารามิเตอร์ของข้อสอบไม่ได้ระบุ แต่ถือว่าเป็นส่วนหนึ่งของความสามารถ ดังนั้น ลักษณะข้อมูลที่มาจากข้อสอบที่มีความยากต่างกัน การเทียบมาตราอาจได้รับผลกระทบ และทำให้ประสิทธิภาพต่ำลง ข้อสังเกตนี้เป็นทำนองเดียวกับผลการวิจัยของ สไลน์ และ ลินท์ (Slinde and Linn 1979) ในกรณีการเทียบโดยใช้วิธี IRT สามพารามิเตอร์ประมาณคะแนนจริง ซึ่งให้ผลการเทียบที่น่าพอใจที่สุดใน ๗ วิธีที่ศึกษา ยังมีปัญหาที่ต้องศึกษา คือ ผลของการเทียบมาตราที่ต่ำกว่าระดับการเดา ซึ่งต้องอาศัยวิธีการประมาณอื่น เช่น วิธีประมาณเชิงเส้นตรงที่นำมาใช้ในครั้งนี้ วิธีอีควิเปอร์เซนไทล์ให้ผลการเทียบมาตราอยู่ในระดับเพียงพอปานกลางสำหรับการ

ศึกษาในครั้งนี้ เป็นวิธีที่ผู้นำไปใช้ไม่สามารถคาดหวังที่จะได้ผลสมบูรณ์ แต่เป็นวิธีสำคัญในการเทียบแบบสอบที่มีความยากต่างกันในขอบเขตที่กลายการศึกษานี้ สำหรับรูปแบบการเทียบเชิงเส้นตรงให้ผลการเทียบที่ไม่น่าพอใจอย่างชัดเจนกับแบบสอบที่มีความยากไม่เท่ากัน

คุก ดันบาร์ และ ไอเนอร์ (Cook, Dunbar and Eegnor 1981) ศึกษาวิธีการเทียบมาตราด้วยรูปแบบ IRT ที่จะนำไปสู่การแก้ปัญหาการทดสอบในการปฏิบัติ โดยการเปรียบเทียบผลการเทียบที่ใช้รูปแบบดั้งเดิม วิธีทั้งหมดที่ทำการศึกษามี ๔ วิธี คือ วิธีเชิงเส้นตรง วิธีอีควิเปอร์-เซนไทล์ธรรมดา วิธีอีควิเปอร์เซนไทล์ที่ใช้ประมาณความถี่ และวิธี IRT ประมาณค่าคะแนนจริง เกณฑ์การเปรียบเทียบความแตกต่างของการเทียบมาตราได้อาศัยการพิจารณา ๒ ทาง ทางแรก คือ ความสอดคล้องสัมพัทธ์โดยพิจารณาจากภาพการแจกแจงของคะแนนแปลง ที่ใช้วิธีเทียบแบบดั้งเดิมวิธีหนึ่ง กับการแจกแจงของคะแนนแปลงเมื่อใช้วิธีเทียบด้วยรูปแบบ IRT เกณฑ์การเปรียบเทียบอีกทางหนึ่ง คือ การเปรียบเทียบเชิงตัวเลข ซึ่งเป็นค่าดัชนีความแตกต่าง ซึ่งมาร์โค เคยใช้มาแล้ว (cited in Marco et al 1979) เป็นค่าที่ถ่วงน้ำหนักแล้วของค่าเฉลี่ยกำลังสองของความแตกต่างระหว่างค่าประมาณจากการเทียบมาตราวิธีที่กำหนดกับคะแนนเกณฑ์ เกณฑ์ที่ใช้ในการศึกษา คือ คะแนนแปลงที่ผ่านกระบวนการเทียบมาตราตามรูปแบบ IRT ข้อมูลที่ใช้ในการศึกษาเป็นผลการสอบรายบุคคลจากการสอบสองครั้งที่ได้จัดดำเนินการไปแล้ว แบบสอบที่ใช้มี ๒ ชุด ซึ่งสร้างขึ้นโดย the College Board Admissions Testing Program เป็นแบบสอบหลายตัวเลือก มีความยาวและความยากต่างกัน รูปแบบการเทียบมาตรา IRT ที่เลือกไว้ศึกษา คือ การเทียบโดยใช้การประมาณคะแนนจริงแบบสามพารามิเตอร์โลจิส การประมาณค่าพารามิเตอร์ข้อสอบ และความสามารถผู้สอบดำเนินการเช่นเดียวกับการวิจัยของผู้อื่น คือ ใช้โปรแกรมคอมพิวเตอร์ LOGIST ของวูด และ ลอร์ด ความน่าจะเป็นของการตอบข้อสอบข้อใด ๆ ได้ถูกต้องในรูปแบบ IRT มีฟังชั่น ดังนี้

$$P_i(\theta) = c_i + (1 - C_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}} \quad ---(๕๔)$$

เมื่อ a_i b_i และ c_i เป็นพารามิเตอร์ของข้อสอบที่บรรยายข้อสอบข้อที่ i และ θ คือ ระดับความสามารถของผู้สอบ คะแนนจริงของผู้สอบในแต่ละระดับความสามารถประมาณจากสมการ

$$\hat{\xi} = \sum_{i=1}^{n} \hat{P}_i(\theta) - \left[\sum_{i=1}^{n} \hat{Q}_i(\theta) \right] / A - 1 \quad -(๕๕)$$

$$\hat{\eta} = \sum_{j=1}^{m} \hat{P}_j(\theta) - \left[\sum_{j=1}^{m} \hat{Q}_j(\theta) \right] / A - 1 \quad -(๕๖)$$

เมื่อ A คือ จำนวนตัวเลือกในแต่ละข้อ $\hat{\xi}$ และ $\hat{\eta}$ เป็นคะแนนจริงจากการประมาณแบบสอบคนละชุด สำหรับคะแนนจริงที่อยู่ต่ำกว่าระดับการเคาใช้วิธีการ linear interpolation ผลการวิจัยพบวิธีการเทียบมาตราแบบดั้งเดิม ๓ วิธี ให้ผลสอดคล้องอย่างมากกับวิธี IRT ตรงช่วงคะแนนส่วนใหญ่ของมาตรา ความแตกต่างที่เบี่ยงเบนออกจากผลของ IRT เกิดขึ้นเพราะช่วงปลายของการแจกแจง ผลที่ปรากฏนี้จะสรุปว่าเป็นเช่นนั้นเลย ยังอาจมีข้อโต้แย้งได้ เพราะเกณฑ์ที่ใช้ประเมินนั้นเป็นเกณฑ์ที่ยึดการเทียบมาตรา IRT เป็นหลัก ความเหมาะสมของเกณฑ์ถึงแม้จะได้รับการสนับสนุนจากผลการวิจัยก่อนก็ตาม ดังนั้น การอภิปรายจึงไม่สามารถสรุปได้หนักแน่น แต่ให้ความจริงจากการค้นพบในการศึกษาครั้งนี้ได้ว่า สภาพการทดสอบที่นำมาศึกษาประกอบด้วยแบบสอบที่มีความยากต่างกัน กลุ่มตัวอย่างไม่ได้เป็นกลุ่มสมมูลโดยสุ่ม แต่อย่างไรก็ตามแบบสอบ ๒ ชุดนี้ได้ทำให้เกิดการแจกแจงคะแนนในแต่ละชุดของกลุ่มตัวอย่างมีลักษณะเหมือนกัน จึงน่าจะเป็นเหตุผลที่ทำให้รูปแบบเชิงเส้นตรงและเส้นโค้งมีผลใกล้เคียงกันมาก

ความแตกต่างที่ปรากฏตรงปลายของการแจกแจงคะแนนเมื่อใช้รูปแบบทั้งที่เป็นเส้นตรงและเส้นโค้ง ส่วนมากเกิดขึ้นเพราะข้อมูลในส่วนนี้มีน้อย หายากมาก ในทางทฤษฎีจึงแนะนำให้ใช้รูปแบบการเทียบ IRT แบบสามพารามิเตอร์ เพราะไม่มีผลกระทบจากการขาดแคลนข้อมูล แต่อย่างไรก็ดีในส่วนที่อยู่ทางปลายด้านคะแนนต่ำ IRT เองไม่สามารถประมาณความสัมพันธ์ได้ ต้องอาศัยวิธีอื่น ดังนั้น ในประเด็นของการเปรียบเทียบผลการเทียบคะแนนที่อยู่ส่วนปลายจึงยังคงไม่สามารถให้ข้อสรุปได้

โคเลน และ วิทนี (Kolen and Whitney 1982) ทำการศึกษาเปรียบเทียบวิธีการเพื่อเทียบมาตรา ๔ วิธี คือ อีควิเปอร์เซนไทล์ เชิงเส้นตรง IRT พารามิเตอร์เดียว และ IRT สามพารามิเตอร์ ว่าวิธีใดให้ความเพียงพอในการเทียบมาตราแบบสอบ General Educational Development (GED) มากกว่ากัน แบบสอบ GED เป็นแบบสอบวัดผลสัมฤทธิ์เพื่อใช้ตัดสินให้ประกาศนียบัตรแก่ผู้ที่ต้องการเทียบความรู้ระดับเตรียมอุดมศึกษาทั่วประเทศ แบบสอบที่พัฒนาขึ้นใช้มี ๑๒ ชุด แต่ละชุดประกอบด้วย ๕ ฉบับย่อย การออกแบบรวบรวมข้อมูลใช้กลุ่มตัวอย่างร่วม คือ กำหนดให้ผู้สอบแต่ละคนสอบ ๒ ชุด ในจำนวน ๑๑ ชุด ส่วนชุดที่ ๑๒ เป็นแบบสอบร่วมที่ทุกคนในกลุ่มตัวอย่างต้องทำเหมือนกันหมด การจัดให้ผู้สอบได้รับการทดสอบ ๒ ชุดใด ๆ ใช้การสุ่มและให้เกิดความสมดุลย์ในการจัดการ กลุ่มตัวอย่างผู้สอบมาจากการสอบในปี ๑๙๘๐ จำนวนมากกว่า ๘๐๐,๐๐๐ คน ทำการสุ่มอย่างแบ่งชั้นภูมิหลายขั้นจากประเภทโรงเรียนเขตทางภูมิศาสตร์ สถานภาพทางสังคม สุดท้ายได้ตัวอย่างผู้สอบรายคนในแต่ละโรงเรียน ๆ ละ ๒๐ คน จำนวนคนในแต่ละชุดของแบบสอบที่ทำการเทียบมาตราเท่ากับ ๒๐๐ คนโดยประมาณ ซึ่งเป็นขนาดกลุ่มตัวอย่างที่ค่อนข้างเล็กในการศึกษาเกี่ยวกับการเทียบมาตรา แต่ผู้วิจัยได้ให้เหตุผลว่า ถึงอย่างไรก็เป็นขนาดที่เป็นปกติของการทดสอบที่พบโดยทั่วไป นอกจากนี้การออกแบบโดยใช้กลุ่มตัวอย่างร่วมเป็นวิธีที่ทำให้เกิดความแม่นยำได้ตามการสนับสนุนของแองกอฟ (๑๙๘๔) และลอรด์ (๑๙๘๐) และมีข้อสังเกตเพิ่มเติมว่า หากข้อค้นพบใดที่เป็นข้อที่ดีจากการวิจัยนี้ ย่อมจะต้องดีกว่าเมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น การดำเนินการเทียบมาตราแต่ละชุดทั้ง ๑๑ ชุด ได้เทียบไปสู่มาตราคะแนนสังเกตของแบบสอบร่วม รูปแบบเชิงเส้นตรง และรูปแบบอีควิเปอร์เซนไทล์ ใช้วิธีที่แองกอฟได้บรรยาย คือ IA - 1 และ IA - 2 ตามลำดับ ส่วนรูปแบบ IRT สามพารามิเตอร์ และพารามิเตอร์เดียว เริ่มด้วยการประมาณค่า

พารามิเตอร์ของข้อสอบ และความสามารถของผู้สอบด้วยโปรแกรมสำเร็จรูป LOGIST ของ Wood, Wingersky และ Lord เฉพาะแบบสอบร่วมก่อน จากนั้นใช้ค่าพารามิเตอร์ความสามารถของผู้สอบกำหนดให้เป็นค่าคงที่ในการวิเคราะห์แบบสอบชุดอื่น ๆ ๑๐ ชุด เพื่อประมาณค่าพารามิเตอร์ของข้อสอบอีก ๑๐ ชุด ให้ค่าประมาณต่าง ๆ มีมาตราเดียวกัน ขั้นต่อไปสร้างคะแนนสมมูลระหว่างแบบสอบร่วม กับแบบสอบอีก ๑๐ ชุดนั้น โดยใช้การประมาณคะแนนจริงของผู้สอบซึ่งคำนวณจากผลบวกของค่าความน่าจะเป็นของการตอบถูกต้องตลอดทุกข้อในแบบสอบ สำหรับรูปแบบสามพารามิเตอร์ใช้วิธีการ linear interpolation หาค่าคะแนนสมมูลของส่วนที่อยู่ต่ำกว่าคะแนนเดา ผลการเทียบมาตราทั้ง ๔ วิธี ได้นำไปประเมินความเพียงพอโดยการวิเคราะห์จากกลุ่มตัวอย่างสอบทานผล ซึ่งเป็นกลุ่มอิสระมีขนาดร้อยละ ๑๐ ของกลุ่มตัวอย่างเทียบมาตรา ทุก ๆ คนในกลุ่มนี้ได้รับการทดสอบแบบสอบทั้ง ๒ ชุด ดัชนีที่ใช้ คือ percentile comparison index เป็นมาตรการที่ใช้วัดความแตกต่างระหว่างการแจกแจงของคะแนนในแบบสอบร่วม และคะแนนแปลงของอีกชุดหนึ่ง ดัชนีนี้ คือ ค่าเฉลี่ยกำลังสองของความแตกต่างของผู้สอบทุกคนในกลุ่มตัวอย่างสอบทานผลระหว่างตัวเลขคะแนนในแบบสอบร่วมกับคะแนนแปลงจากแบบสอบอีกชุดหนึ่ง ณ ที่ตำแหน่งเปอร์เซนไทล์เดียวกัน ดัชนีนี้ โคเลน (Kolen 1981) ได้พัฒนาขึ้นใช้ ผลการวิจัยได้ข้อค้นพบดังนี้ จากการประเมินผลสุดท้ายของการเทียบมาตราวิธีต่าง ๆ กับกลุ่มสอบทานผล ทำการทดสอบความมีนัยสำคัญของค่าอันดับจากการแปลงตามปริมาณของค่าดัชนีด้วยการทดสอบนันพาราเมตริก Friedman Test สรุปได้ว่า รูปแบบเชิงเส้นตรงให้ผลการเทียบที่มีความเพียงพอมากกว่ารูปแบบอีควิเปอร์เซนไทล์ และรูปแบบ IRT สามพารามิเตอร์ ขณะเดียวกันรูปแบบ IRT พารามิเตอร์เดียว ให้ผลการเทียบที่เพียงพอกว่ารูปแบบ IRT สามพารามิเตอร์ ข้ออภิปรายเกี่ยวกับผลการวิจัยครั้งนี้ ผู้วิจัยได้ตั้งข้อสังเกตไว้หลายประการดังนี้ คือ การประมาณค่าพารามิเตอร์ตามทฤษฎีของสามพารามิเตอร์ ทำให้มีค่าที่ประมาณสูงมาก ซึ่งเป็นธรรมชาติของทฤษฎีที่สามารถควบคุมความยากของแบบสอบ เมื่อมาใช้ในกรณีของกลุ่มตัวอย่างขนาดเล็กเกินไป ทำให้เกิดความผันแปรเชิงสุ่มจนทำให้ผลการเทียบมาตราขาดความแน่นอนไป แต่จุดอ่อนนี้สามารถแก้ไขได้โดยเพิ่มตัวอย่างให้มีขนาดใหญ่ขึ้นตามคำแนะนำของลอร์ด (Lord 1980: 209-210) รูปแบบอีควิเปอร์เซนไทล์ให้ผลการเทียบมาตราด้อยกว่ารูปแบบเชิงเส้นตรง เนื่องจากตัวอย่างมีขนาดเล็กเกินไป ทำให้คะแนนแปลงมีลักษณะผิดปกติมาก ซึ่งสันนิษฐานว่ามาจากความคลาดเคลื่อนเชิงสุ่ม การเกลาเส้นให้เรียบในส่วนที่มีลักษณะผิดปกติได้ใช้วิธีเกลาด้วยมือ ทำให้ความคลาด

เคลื่อนที่ปรากฏในกลุ่มสอบทานผลน้อยลง แต่ก็ยังไม่สามารถปรับปรุงได้มากกว่านั้น จึงได้ลงความเห็นว่า ตามสถานการณ์ที่ศึกษาในกรณีของแบบสอบ GED นี้ เมื่อออกแบบด้วยการใช้กลุ่มตัวอย่างร่วม หรืออาจจะเป็นกลุ่มตัวอย่างสุ่มสมมูลแล้ว รูปแบบที่เหมาะสมที่สุด คือ วิธีการเชิงเส้นตรง เป็นวิธีที่ง่าย และให้ผลการเทียบที่แน่นอนมากกว่า

สรุปการวิเคราะห์วรรณคดีที่เกี่ยวข้องกับการเปรียบเทียบรูปแบบการเทียบมาตรา แยกได้เป็น ๒ ประเด็น คือ ประเด็นของผลการศึกษา และประเด็นของวิธีพิจารณาตัดสินการเปรียบเทียบ ประเด็นของผลการศึกษา จากการวิจัยที่เริ่มมาตั้งแต่ปี ๑๙๗๗ จนถึง ๑๙๘๒ สรุปได้ว่า ไม่มีรูปแบบการเทียบมาตราใดที่ให้ผลในการประยุกต์ได้ดีที่สุดกับทุกสถานการณ์ที่กำหนด แต่จากผลการวิจัยที่ดำเนินการมานั้นได้ให้ข้อความจริงจากการค้นพบในเชิงประจักษ์ที่พอจะเป็นแนวทางเพื่อเลือกใช้ให้ใกล้เคียงเหมาะสมกับสภาพที่ต้องการของแต่ละกรณี การพิจารณาทางเลือกที่สำคัญก่อนตัดสินใจเลือกรูปแบบที่เป็นหลักในการกำหนดคะแนนสมมูล คือ ลักษณะของแบบสอบ ๒ ชุดที่ต้องการเทียบ หรือมากกว่า ๒ ชุด หรือการเทียบอยู่ในลักษณะหลายชุดเทียบกันเป็นสายโซ่ ลักษณะของกลุ่มตัวอย่างผู้สอบ ซึ่งเป็นเรื่องของการออกแบบรวบรวมข้อมูลคะแนนเพื่อสร้างคะแนนสมมูล นอกจากนี้ยังมีตัวแปรปลีกย่อยซึ่งเป็นข้อจำกัดเฉพาะกรณี ข้อสอบอย่างกว้าง ๆ มีดังนี้ (๑) เมื่อแบบสอบที่นำมาเทียบมีความยากแตกต่างกัน กลุ่มตัวอย่างไม่ได้เป็นกลุ่มสมมูล ซึ่งส่วนมากเป็นการเทียบมาตราในแนวดิ่ง (vertical equating) รูปแบบ IRT สามพารามิเตอร์ให้ผลน่าพอใจกว่ารูปแบบอื่น (Slinde and Linn 1979; Petersen, Cook and Stocking 1981; Kolen 1981; Cook, Dunbar, Eignor 1981) (๒) เมื่อแบบสอบมีความคล้ายคลึงในระดับความยาก และกลุ่มตัวอย่างเป็นกลุ่มตัวอย่างร่วม หรือกลุ่มสุ่มสมมูล ควรใช้วิธีการของรูปแบบเชิงเส้นตรง ซึ่งสะดวกและให้ความแน่นอนกว่า (Petersen, Cook and Stocking 1981; Kolen, Whitney 1982) (๓) เมื่อแบบสอบมีความยากต่างกัน แต่กลุ่มตัวอย่างเป็นกลุ่มสุ่มสมมูลแล้ว วิธีการของรูปแบบอีควิเปอร์เซนไทล์ ซึ่งให้ความสัมพันธ์ของการแปลงที่ไม่เป็นเส้นตรงให้ความเหมาะสมต่อการแปลงได้ แต่ถ้าไม่เป็นกลุ่มสมมูลบางครั้งอาจเกิดการแจกแจงคะแนนที่แตกต่างมาก จนทำให้เกิดความคลาดเคลื่อนที่ทำให้ผลการเทียบขาดความเพียงพอของการสร้างคะแนนสมมูล (Cook, Dunbar, Eignor 1981; Kolen 1982) (๔) ในสภาพข้อมูลที่ค่อนข้างน้อย และแบบสอบโดยส่วนใหญ่

ทำหน้าที่วัดในเรื่องเดียวกัน รูปแบบเชิงเส้นตรงและ IRT พารามิเตอร์เดียว จะให้ผลการเทียบที่เพียงพอกว่ารูปแบบอีควิเปอร์เซนไทล์กับ IRT สามพารามิเตอร์ ประเด็นของวิธีการพิจารณาตัดสินการเปรียบเทียบที่ใช้ในการศึกษาทั้งหมดเป็นเกณฑ์ภายนอก (external criteria) โดยใช้พิจารณาจากความคลาดเคลื่อน (discrepancy) ระหว่างคะแนนแปลงที่เกิดจากการใช้ผลการเทียบมาตราวิธีที่ต้องการศึกษากับคะแนนเกณฑ์ ค่าความคลาดเคลื่อนใช้เป็นดัชนีชี้ความแตกต่างเชิงปริมาณ การตัดสินความเพียงพอของวิธีเทียบมาตราใช้หลักของการเปรียบเทียบค่าดัชนีความคลาดเคลื่อนเป็นการตัดสินด้วยความเพียงพอสัมพัทธ์ (relative adequacy) รายละเอียดของเทคนิคการวิเคราะห์แตกต่างกันในแต่ละการวิจัยดังนี้ (๑) ดัชนีความคลาดเคลื่อน (discrepancy index) เป็นค่าที่ถ่วงน้ำหนักแล้วของค่าเฉลี่ยกำลังสองของความแตกต่างระหว่างคะแนนที่ประมาณกับคะแนนเกณฑ์ ค่านี้ใช้การถ่วงน้ำหนักให้เป็นมาตรฐานด้วยความเบี่ยงเบนมาตรฐานของการแจกแจงในคะแนนเกณฑ์ ซึ่งจะทำให้ค่าดัชนีที่ได้จากการเทียบมาตราต่างกลุ่มผู้สอบ หรือต่างชุดของแบบสอบสามารถเปรียบเทียบได้โดยตรง ค่าตัวเลขของดัชนีที่น้อยให้ความหมายว่า วิธีการเทียบมาตราที่ใช้กับสภาพการทดสอบนั้น ๆ มีความแน่นอนในการแปลงคะแนน (stability) สำหรับการวิจัยที่ใช้วิธีการเปรียบเทียบนี้ ได้แก่ ปีเตอร์สัน มาร์โค และ สตีเวอร์ท (Petersen Marco and Stewart 1982) ซึ่งใช้คะแนนเกณฑ์ที่เป็นคะแนนแปลงโดยรูปแบบ IRT กับข้อมูลคะแนนเดียวกันกับการแปลงด้วยวิธีอื่น ๆ ที่ต้องการศึกษา การใช้ผลการเทียบด้วย IRT เป็นเกณฑ์นี้ ถึงแม้ผู้วิจัยจะได้อ้างอิงผลการวิจัยต่าง ๆ ให้เป็นเหตุผลที่หนักแน่นขึ้นก็ตาม แต่ผู้วิจัยเองก็ไม่กล้าให้คำยืนยันว่าเมื่อเปลี่ยนสถานการณ์ทดสอบผิดไปจากที่ศึกษาแล้ว ความลำเอียงจะไม่เกิดขึ้น ปีเตอร์สัน คุก และสตอกคิง (Petersen, Cook and Stocking 1981) ได้เปลี่ยนเกณฑ์เป็นคะแนนดิบของแบบสอบชุดแรกในกระบวนการเทียบมาตราแบบสายโซ่ แบบสอบชุดหลังแต่ละครั้งจะเทียบไปยังชุดก่อนหน้านั้น หลังจากผ่านกระบวนการเทียบไป ๒ ถึง ๓ ครั้ง จะนำคะแนนแปลงไปสู่มาตราของชุดแรกมาเทียบกับคะแนนดั้งเดิมของชุดแรกซึ่งใช้เป็นคะแนนเกณฑ์ ความแตกต่างที่เกิดขึ้นจะให้เป็นค่าดัชนีความแตกต่าง ค่านี้จะให้ความหมายของความคงเส้นคงวาของการแปลง (consistency) ในกระบวนการที่เป็นสายโซ่ ส่วนงานวิจัยของโคเลน (Kolen 1981) ได้ใช้เทคนิคการวิเคราะห์กลุ่มสอบทานผลเพื่อสร้างเกณฑ์ที่ปราศจากความลำเอียง การวิเคราะห์กลุ่มตัวอย่างสอบทานผล เป็นวิธีตรวจสอบความเพียงพอของรูปแบบการเทียบมาตราตอนปลายทาง โดยใช้กลุ่มตัวอย่าง

อิสระจากประชากรเกี่ยวกับกลุ่มเทียบมาตรา พิจารณาจากความแตกต่างระหว่างการแจกแจงของคะแนนในแบบสอบชุดแรกกับคะแนนแปลงของแบบสอบชุดหลังในมาตราของชุดแรก ความแตกต่างในแนวคิดนี้ คือ ค่าเฉลี่ยกำลังสองของความแตกต่างของผู้สอบทุกคนในกลุ่มตัวอย่างสอบทานผลระหว่างตัวเลขคะแนนของแบบสอบชุดแรกกับตัวเลขของคะแนนแปลงที่ตำแหน่งเปอร์เซนไทล์เดียวกัน โคเลน (๑๙๘๑) เรียกดัชนีนี้ว่า percentile comparison index โคเลน และ วิทนี (Kolen and Whitney 1982) ได้ปรับปรุงดัชนีนี้เพื่อให้ใช้เปรียบเทียบกับการเทียบมาตราที่เกิดจากการใช้แบบสอบที่มีความยาวของแบบสอบต่างกัน โดยการถ่วงน้ำหนักด้วยจำนวนข้อสอบ (k) ซึ่งช่วยให้ค่าดัชนีสามารถนำมาเปรียบเทียบเมื่อเปลี่ยนชุดของแบบสอบ

จากการวิเคราะห์วรรณคดีที่เกี่ยวกับการวิจัยเปรียบเทียบวิธีการเทียบมาตราดังกล่าวมาแล้วนี้ ผู้วิจัยได้พัฒนาวิธีการที่จะใช้ประเมินเพื่อตัดสินวิธีการเทียบมาตรา ๓ รูปแบบ จากแนวคิดของการใช้วิเคราะห์กลุ่มสอบทานผล ดัชนีความแตกต่างที่ใช้ คือ ดัชนีความแตกต่างมาตรฐาน (index of standard discripancy) ซึ่งได้จากค่าเฉลี่ยกำลังสองความแตกต่างของคะแนนแปลงกับคะแนนเกณฑ์ของบุคคลเดียวกันทุกคนในกลุ่มตัวอย่างสอบทานผล และได้ถ่วงน้ำหนักให้เป็นมาตรฐานด้วยส่วนเบี่ยงเบนมาตรฐานของการแจกแจงของคะแนนเกณฑ์ สำหรับคะแนนเกณฑ์ที่ใช้ คือ คะแนนการสังเกตของแบบสอบชุด X ที่กลุ่มตัวอย่างสอบทานผลทำได้ ค่านี้ให้ความหมายของการเปรียบเทียบเชิงปริมาณไม่ว่าจะเป็นผลจากการทดสอบด้วยแบบสอบใด หรือรูปแบบการเทียบใด ค่าที่น้อยกว่าหมายถึง รูปแบบการทดสอบนั้น ๆ มีความเพียงพอในการแปลงคะแนนแบบสอบชุดหลังให้อยู่ในคะแนนมาตราของชุดแรก การลงข้อสรุปใช้หลักการเปรียบเทียบ และทดสอบว่าการทดสอบนั้นพาราเมตริกเพื่อทดสอบความมีนัยสำคัญทางสถิติของดัชนีความแตกต่างด้วยนันพาราเมตริก