บทที่ 2

วรรณคดีที่เกี่ยวข้อง

ในการศึกษาวรรณคดีที่เกี่ยวข้อง ผู้วิจัยได้ศึกษาและนำเสนอเป็น 4 ตอน ดังนี้ คือ (1) ความเป็นมาของทฤษฎีการอ้างอิงสรุป ซึ่งกล่าวถึงความเป็นมา และพัฒนาการของข้อตกลงเบื้องต้น (2) หลักพื้นฐานของทฤษฎีการอ้างอิงสรุป ซึ่งกล่าวถึงความหมายของค่าสัมประสิทธิ์การอ้างอิงสรุป ตลอดจนเทคนิควิธีในการคำนวณค่าพารามิเตอร์ (3) การประยุกต์ทฤษฎีการอ้างอิงสรุปในการวัดและประเมินผลการศึกษาในสถานการณ์ต่าง ๆ และ (4) การสอบความเรียงซึ่งกล่าวถึง ธรรมชาติของการสอบ นิยามของข้อสอบ ข้อเสนอแนะเพื่อให้การสอบมีความเชื่อถือได้มากยิ่งขึ้น และ ผลการวิจัยที่เกี่ยวข้องกับการสอบความเรียงในประเทศ ดังรายละเอียดต่อไปนี้

ความเป็นมาของทฤษฎีการอ้างอิงสรุป

เพื่อให้ผู้อ่านงานวิจัยได้เข้าใจทฤษฎีการอ้างอิงสรุปได้ดียิ่งขึ้น ผู้วิจัยขอกล่าวถึงความเป็นมาของทฤษฎี ดังนี้

Cronbach, Rajaratnam and Gleser (1963: 137-163) ได้สรุปความเป็นมาของทฤษฎีการอ้างอิงสรุปไว้ว่า ทฤษฎีการวัดแบบดั้งเดิม (Classical theory) ใช้ค่าความเที่ยงอธิบายความแม่นยำของการวัด โดยยึดข้อตกลงคุณสมบัติคู่ขนานหรือความเท่าเทียมเป็นสำคัญ ผู้ที่ได้ชื่อว่าเป็นบิดาของทฤษฎีความเที่ยงในการวัดทางจิตวิทยาคือ Spearman จากการที่รู้ว่า ค่าสหสัมพันธ์ระหว่างคุณลักษณะที่ต่างกันสองอย่างมักจะต่ำกว่าที่ควรจะเป็น ทั้งนี้เป็นผลเนื่องมาจากความคลาดเคลื่อนของการสังเกต ในปี 1904 Spearman ได้สร้างสูตรปรับแก้ โดยการหารค่าสหสัมพันธ์ที่ได้จากการสังเกต $o_{X_1X_2}$ ด้วย $o_{X_1X_2}.o_{Y_1Y_2}$ สูตรนี้มีข้อตกลงเพียงว่า X_1 และ X_2 เป็นค่าวัดที่อิสระต่อกันและวัดสิ่งเดียวกันเท่านั้น เช่นเดียวกับ Y_1 และ Y_2 Spearman ชี้ว่า $o_{X_1X_2}$ คือค่าความแม่นยำของการสังเกต ในปี 1910 Spearmanได้พัฒนาทฤษฎีความเที่ยงอย่างจริงจัง ได้กล่าวถึงข้อตกลงเกี่ยวกับแบบสอบคู่ขนานเป็นครั้งแรกว่า แบบสอบทั้งหลายที่วัดคุณลักษณะ

เดียวกัน คะแนนของแต่ละฉบับประกอบด้วยคะแนนจริงที่เท่ากันรวมกับความคลาดเคลื่อน และมีข้อตกลงเพิ่มเติมเกี่ยวกับความคลาดเคลื่อนว่า มีค่าเฉลี่ยเป็น 0 มีความแปรปรวนเท่ากัน เป็นอิสระต่อกัน และเป็นอิสระต่อคะแนนจริง ภายใต้เงื่อนไขนี้แบบสอบแต่ละฉบับจะมีค่าเฉลี่ยเท่ากัน ความแปรปรวนเท่ากัน และค่าสหสัมพันธ์ระหว่างแบบสอบคู่ขนานดังกล่าวจะมีค่าดังนี้ (1) มีค่าเท่ากัน (2) มีค่าเท่ากับอัตราส่วนระหว่างความแปรปรวนของคะแนนจริงกับความแปรปรวนของคะแนนสังเกต (3) มีค่าเท่ากับกำลังสองของค่าสหสัมพันธ์ระหว่างคะแนนสังเกตกับคะแนนจริง ในระยะเวลาใกล้เคียงกันนั้น Brown ได้พัฒนาทฤษฎีความเที่ยง โดยเริ่มด้วยการนิยามแบบสอบคู่ขนานในขณะที่ Spearman เริ่มจากคะแนนสังเกตประกอบด้วยคะแนนจริงรวมกับความคลาดเคลื่อน แต่แนวคิดเกี่ยวกับความเที่ยงของ Brown สอดคล้องกับทฤษฎีของ Spearman

จากข้อตกลงเกี่ยวกับคุณสมบัติคู่ขนานหรือความเท่าเทียมกันนี้ นำไปสู่การพัฒนาสูตรการหาความเที่ยงมากมาย รวมทั้งสูตรความสอดคล้องภายใน (internal concistency) ที่ใช้กันอยู่ทั่วไป ได้แก่ สูตร Spearman-Brown ที่ประมาณค่าความเที่ยงจากการสอบครั้งเดียว โดยคำนวณจากค่าสหสัมพันธ์ระหว่างส่วนแบ่งของแบบสอบ 2 ส่วน อาจแบ่งเป็น ข้อคู่-ข้อคี่ ครึ่งแรก-ครึ่งหลัง ค่าสหสัมพันธ์ที่ได้เมื่อปรับด้วยสูตรแล้ว จะเป็นค่าประมาณของค่าสหสัมพันธ์ระหว่างแบบสอบฉบับนั้นกับแบบสอบฉบับอื่นที่มีคุณสมบัติคู่ขนานกัน สูตรในการหาความเที่ยงโดยวิธีแบ่งครึ่งแบบสอบนี้เขียนอีกแบบหนึ่งว่า

$$\text{ความเที่ยง} = 1 - \frac{V_{A-B}}{V_T} = 2\left[1 - \frac{V_A + V_B}{V_T}\right]$$

V_A คือ ความแปรปรวนของครึ่ง A

V_B คือ ความแปรปรวนของครึ่ง B

V_T คือ ความแปรปรวนของแบบสอบทั้งฉบับ

Rulon (1939 cited by Mehrens and Ebel 1969: 104-108) ชี้ว่า V_{A-B} คือความแปรปรวนของความคลาดเคลื่อนจากจุดนี้เอง นำไปสู่สูตรความเที่ยงในรูปของอัตราส่วนระหว่างความแปรปรวนของคะแนนจริงต่อคะแนนสังเกต Rulon ตั้งข้อตกลงเบื้องต้นว่า คะแนนของครึ่ง A และ B มีความแปรปรวนร่วมกับคะแนนจริงเท่ากัน ความคลาดเคลื่อนของส่วนแบ่งแต่ละส่วน

เป็นอิสระต่อกัน และเป็นอิสระต่อคะแนนจริง

สูตรความเที่ยงชนิดความสอดคล้องภายในอีกสองสูตรคือ KR 20 และ KR 21 ซึ่งใช้ค่าสถิติจากการสอบครั้งเดียว เป็นค่าประมาณสหสัมพันธ์ระหว่างแบบสอบคู่ขนานสองฉบับ สูตร KR 20 จะให้ค่าสหสัมพันธ์ที่เหมาะสมเมื่อข้อสอบในแบบสอบฉบับนั้นวัดองค์ประกอบเดียว ถ้าความแปรปรวนของข้อสอบแต่ละข้อมีค่าเท่ากัน สูตรนี้จะเปลี่ยนเป็นสูตร KR 21 เนื่องจากสูตรทั้งสองใช้กับการให้คะแนนแบบ 0 หรือ 1 ภายใต้การให้คะแนนระบบนี้ ข้อสอบทั้งฉบับจะวัดองค์ประกอบเดียวก็ต่อเมื่อ ข้อสอบมีค่าเฉลี่ยเท่ากัน และความแปรปรวนเท่ากัน ซึ่งนั่นย่อมหมายความว่า สูตรทั้งสองต่างยึดข้อตกลงเบื้องต้นเกี่ยวกับคุณสมบัติคู่ขนานนั่นเอง (Cronbach, et al. 1963: 139-140)

กล่าวโดยสรุป ทฤษฎีความเที่ยงแบบดั้งเดิมพัฒนามาจากข้อตกลงของความเท่าเทียมหรือความเป็นคู่ขนานกันของค่าการวัด นิยามความเที่ยงว่าเป็นค่าสัมประสิทธิ์ระหว่างค่าการวัดที่มีความเท่าเทียมหรือมีความคู่ขนานกัน อาศัยนิยามนี้ การหาค่าความเที่ยงจะต้องมีการสอบวัดให้ได้ค่าการวัดอย่างน้อยสองชุด แต่ละชุดใช้เครื่องมือวัดที่มีคุณสมบัติคู่ขนานกัน คือ มีเนื้อหาอย่างเดียวกัน ค่าเฉลี่ยเท่ากัน ค่าความแปรปรวนเท่ากัน และค่าสหสัมพันธ์ระหว่างกันเท่ากัน แล้วคำนวณค่าสหสัมพันธ์ระหว่างค่าการวัดทั้งสองชุด ผลที่ได้ถือเป็นค่าประมาณความเที่ยง ในกรณีที่ต้องการประมาณค่าความเที่ยงจากการสอบครั้งเดียว ต้องแบ่งค่าการวัดออกเป็นส่วนย่อย ๆ อาจแบ่งเป็นสองส่วน เช่น ข้อคู่-ข้อคี่ ครึ่งแรก-ครึ่งหลัง หรือแบ่งย่อยตามจำนวนข้อ โดยยึดข้อตกลงว่าส่วนแบ่งย่อยแต่ละส่วนจะต้องมีความเท่าเทียมกัน

ปัญหาในทางปฏิบัติจะพบว่านอกจากแบบสอบมาตรฐานแล้ว เราไม่สามารถจะหาเครื่องมือวัดที่มีคุณสมบัติความเท่าเทียมกันอย่างสมบูรณ์ได้ ตัวอย่างเช่น เราต้องการศึกษาความเที่ยงในการตรวจคำตอบแบบความเรียง การจะหาผู้ตรวจที่มีความเท่าเทียมกันเป็นสิ่งที่ทำได้ยากมาก ไม่ว่าจะพิจารณาค่าความแปรปรวน หรือค่าสหสัมพันธ์ระหว่างผู้ตรวจ การหาความเที่ยงของการสังเกตพฤติกรรมของนักเรียนในสนามก็เช่นกัน สถานการณ์ที่ต่างกันก็ยากที่จะมีความเท่าเทียมกัน ดังนั้นการใช้สูตรที่มีข้อตกลงความเท่าเทียมหรือคู่ขนานกันจึงเป็นการปฏิบัติโดยฝืนข้อตกลง จึงทำให้เกิด

การพัฒนาทฤษฎีที่จะหาความเที่ยงโดยไม่ยึดมั่นในข้อตกลงของความเท่าเทียมกัน เช่น แนวคิดของ Kelly, Cureton และ Burt เป็นต้น (Cronbach, et al. 1963: 143-144)

เทคนิควิธีการวิเคราะห์ความแปรปรวนของ Fisher เป็นเครื่องมือสำคัญอย่างหนึ่งในการพัฒนาทฤษฎีความเที่ยงทั้งในด้านการพัฒนาข้อตกลงและวิธีการคำนวณ โดยเฉพาะการใช้อัตราส่วนของความแปรปรวนแทนค่าความเที่ยงที่เรียกว่า "สหสัมพันธ์ intraclass" ค่าสหสัมพันธ์ intraclass นี้ถูกนำมาใช้แทนค่าสัมประสิทธิ์ความเที่ยงครั้งแรกในงานของ Cureton แต่ Pearson เป็นบุคคลแรกที่อธิบายค่าสัมประสิทธิ์สหสัมพันธ์ในรูปของอัตราส่วนของความแปรปรวนตามทฤษฎีของ Fisher ต่อมา Jackson ร่วมกับ Neyman and Johnson แห่งมหาวิทยาลัยลอนดอนได้นำแนวคิดทางสถิติไปใช้หาค่าความแม่นยำของการวัด ต่อมาลูกศิษย์ของ Johnson คือ Hoyt นำไปประยุกต์กับความเที่ยงชนิดความสอดคล้องภายใน ในช่วงเวลาใกล้เคียงกันนี้ มีผลงานของนักวิจัยหลายคน ที่อธิบายความเที่ยงโดยใช้การวิเคราะห์ความแปรปรวนและเกี่ยวข้องกับค่าสหสัมพันธ์ intraclass เช่น Burt, Jackson and Ferguson, Alexander, Ebel และ Lindquist เป็นต้น (Cronbach, et al. 1963: 140-141)

แนวคิดในการนำการวิเคราะห์ความแปรปรวนไปใช้ในการประเมินความเที่ยงเป็นดังนี้

ให้ X_{pi} แทน คะแนนสังเกตของนักเรียน p ที่สอบแบบสอบ i (หรือตรวจโดยผู้ตรวจ i) สามารถเขียนในรูปตัวแบบเชิงเส้นได้ดังนี้

$$X_{pi} = M+(M_p-M)+(M_i-M)+e_{pi}$$

เมื่อให้

M_p แทน คะแนนเอกภพ (คะแนนจริง) ของนักเรียน P ซึ่งเป็นค่าเฉลี่ยของ X_{pi} เมื่อสอบด้วยแบบสอบทุกฉบับในเอกภพของแบบสอบ

M_i เป็น ค่าเฉลี่ยของแบบสอบ i คำนวณจากผู้สอบทุกคนในประชากรของผู้สอบ

M เป็น ค่าเฉลี่ยรวม (Grand Mean)

e_{pi} คือ ความคลาดเคลื่อนเชิงสุ่มมีค่าเฉลี่ยเป็น 0 ค่าความแปรปรวนเท่ากันและเป็นอิสระต่อผลอื่น ๆ ในตัวแบบ

จากตัวแบบและข้อตกลงข้างต้นสามารถประมาณค่าความแปรปรวนของแหล่งต่าง ๆ

แล้วคำนวณค่าความเที่ยงตามนิยาม โดยคำนวณจากอัตราส่วนความแปรปรวนของคะแนนจริง V_{mp} ต่อความแปรปรวนของคะแนนสังเกต V_{xi} เรียกอัตราส่วนนี้ว่า สหสัมพันธ์ intraclass

ในระยะแรกของการใช้การวิเคราะห์ความแปรปรวนในการประเมินความเที่ยงยังคงยึดข้อตกลงความเท่าเทียมเป็นหลัก เวลาต่อมามีนักวิจัยได้นำไปใช้หาความเที่ยงโดยไม่ยึดถือข้อตกลงความเท่าเทียม กลุ่มนักวิจัยที่ได้ชื่อว่าเป็นผู้พัฒนาทฤษฎีความเที่ยงที่ไม่ยึดข้อตกลงของความเท่าเทียมอย่างเป็นระบบพร้อมกับตั้งชื่อทฤษฎีนี้ว่า " GENERALIZABILITY THEORY" ได้แก่ Cronbach et al. (1972) ต่อมา Brennan (1983) พยายามเผยแพร่แนวคิดให้ง่ายขึ้นทั้งในแง่การตีความและการคำนวณ Cardinet et al. (1976,1981,1983) ได้ขยายความทฤษฎีในบางจุด ให้สามารถประยุกต์ได้กว้างขวางกว่าเดิม ดังจะได้นำเสนอในหัวข้อต่อไป

<u>หลักพื้นฐานของทฤษฎีการอ้างอิงสรุป</u>

นักวิจัยจำเป็นต้องประเมินความแม่นยำ (precision) หรือความเที่ยงของผลการวัด ทั้งนี้เนื่องจาก ต้องการอ้างอิงสรุปจากค่าสังเกตที่วัดได้ไปยังกลุ่ม (Class) ค่าสังเกตอื่น ๆ ในเอกภพของค่าสังเกตนั้น การถามถึงความเที่ยงของการให้คะแนนคำตอบความเรียงคือการถามว่าคะแนนที่ผู้ตรวจให้ไปนั้น เป็นตัวแทนของคะแนนที่อาจจะได้ เมื่อให้ผู้ตรวจคนอื่น ๆ ตรวจ คำถามเกี่ยวกับการอ้างอิงสรุปเป็นเรื่องที่มีเนื้อหาแน่ชัด ไม่ใช่เป็นเพียงวิธีการ (Cronbach, et al. 1963: 144) ในการศึกษาความเที่ยงของค่าการวัดใด ผู้ศึกษาต้องกำหนดกลุ่มค่าสังเกตรวมทั้งเทคนิควิธีการวัดให้ชัดด้วย

เนื่องจากค่าการวัดสามารถอ้างอิงสรุปไปยังเอกภพต่าง ๆ กันได้หลายเอกภพ ดังนั้นนักวิจัยต้องระบุเอกภพที่ตนเองสนใจไว้ล่วงหน้าก่อนการอ้างอิงสรุป ประเด็นนี้ทฤษฎีความเที่ยงแบบดั้งเดิมจะไม่กล่าวไว้ นั่นหมายความว่าในทฤษฎีการวัดแบบดั้งเดิม แบบสอบทุกฉบับที่มีคะแนนจริงเท่ากันจะเป็นสมาชิกของชุดแบบสอบคู่ขนานเพียงชุดเดียว และมีค่าความเที่ยงเพียงค่าเดียวเท่านั้น Guttman (1953: 125 Cited by De Gruijter and Van der Kamp 1984: 69) ชี้ให้เห็นความไม่ชัดเจนของทฤษฎีความเที่ยงดั้งเดิมไว้ดังนี้

สมมุติแบบสอบฉบับที่ 1 ประกอบด้วยข้อสอบ 1 ข้อว่า "ให้เขียนคำที่ขึ้นต้นด้วยตัว t" จากแบบสอบนี้เราสามารถสร้างแบบสอบคู่ขนานกับแบบสอบฉบับที่ 1 อย่างน้อย 2 วิธี คือ

วิธีที่ 1 เปลี่ยนตัวอักษรขึ้นต้นคำ เช่น ฉบับที่ 2 "ให้เขียนคำที่ขึ้นต้นด้วยตัว p" ฉบับ ที่ 3 อาจจะให้ขึ้นต้นด้วยตัว d โดยการปรับเวลาที่ตอบ แบบสอบทั้ง 3 ฉบับ สามารถทำให้ค่าเฉลี่ย ความแปรปรวน และค่าสหสัมพันธ์ระหว่างกันเท่ากันได้ สมมุติว่าค่าสหสัมพันธ์ระหว่างกันเป็น .70 ฉะนั้นถ้าพิจารณาตามทฤษฎีความเที่ยงดั้งเดิม แบบสอบฉบับที่ 1 จะมีค่าสัมประสิทธิ์ความเที่ยงเป็น .70

วิธีที่ 2 เปลี่ยนตำแหน่งตัวอักษร กล่าวคือแบบสอบฉบับที่ 2 อาจจะเป็น "ให้เขียนคำ ซึ่งตัวอักษรตัวที่สองเป็นตัว t" แบบสอบฉบับที่ 3 อาจจะเปลี่ยนตำแหน่งของตัว t ไปอยู่ตำแหน่งที่สาม การกระทำเช่นนี้ สามารถทำให้แบบสอบมีคุณสมบัติคู่ขนานได้ สมมุติว่าค่าสหสัมพันธ์ระหว่างแบบสอบเป็น .60

ดังนั้นแบบสอบฉบับที่ 1 จะมีค่าความเที่ยงเป็น .70 และ .60 ในเวลาเดียวกัน จากตัวอย่างนี้ชี้ให้เห็นได้ชัดว่า การตีความของความเที่ยงในทฤษฎีการวัดแบบดั้งเดิมมีความคลุมเครือ แต่ในทฤษฎีการอ้างอิงสรุปจะไม่เกิดขึ้น เพราะในการจะศึกษาเรื่องใด ต้องระบุเอกภพของเงื่อนไขการวัดที่ต้องการอ้างอิงสรุปถึง จากตัวอย่างสมมุติว่าสนใจเอกภพ "ตัวอักษรที่ใช้" และเอกภพ "ตำแหน่งของตัวอักษร" พร้อมกัน ผู้วัดผลต้องระบุเอกภพทั้งสองไว้เป็นกรอบอ้างอิงให้ชัดเจน แล้วจึงอ้างอิงสรุปไปยังเอกภพทั้งสอง แต่ถ้าสนใจเพียง ตำแหน่งของตัวอักษรเพียงเอกภพเดียว จะต้องกำหนดให้ "ตัวอักษรที่ใช้" เป็นตัวประกอบประเภทคงที่ ค่าสัมประสิทธิ์การอ้างอิงสรุปจะสามารถอ้างอิงไปยังเอกภพของตำแหน่งตัวอักษรเท่านั้น ซึ่งค่าทั้งสองอาจจะเท่ากันหรือไม่เท่ากันก็ได้ แต่ตีความแตกต่างกัน

หลักพื้นฐานของทฤษฎีการอ้างอิงสรุปมีสาระสำคัญดังนี้

1. ข้อตกลงของทฤษฎีการอ้างอิงสรุป ข้อตกลงของทฤษฎีการอ้างอิงสรุป อาศัยข้อตกลงของแบบแผนการวิเคราะห์ความแปรปรวนที่สัมพันธ์กับรูปแบบการวัดที่ต้องการศึกษา โดยทั่วไปมีข้อตกลงดังนี้

1.1 ต้องระบุเอกภพที่ต้องการอ้างอิงให้ชัดเจน จนสามารถบอกได้ว่ามีเงื่อนไข

ใดบ้างที่เป็นสมาชิกของเอกภพนั้น

1.2 เงื่อนไขการวัดเป็นอิสระต่อกัน กล่าวคือคะแนนของนักเรียนที่ทำข้อสอบข้อ i ถูกหรือผิดไม่ขึ้นกับการตอบข้ออื่น

1.3 คะแนนสังเกต (X_{pi}) เป็นค่าการวัดในมาตราช่วง (interval scale)

2. ความหมายของค่าสัมประสิทธิ์การอ้างอิงสรุป

2.1 ค่าสัมประสิทธิ์การอ้างอิงสรุป เป็นดัชนีที่สามารถอธิบายความแม่นยำของการวัดเช่นเดียวกับค่าสัมประสิทธิ์ความเที่ยงแบบดั้งเดิม สามารถใช้คำนวณช่วงความเชื่อมั่นของคะแนนเอกภพ หรือใช้ในสมการถดถอยในการประมาณค่าคะแนนเอกภพ และใช้ในการปรับแก้ค่าสหสัมพันธ์ที่ลดลงอันเนื่องจากความคลาดเคลื่อน (correcting correlations for attenuation) (Cronbach, et al. 1963: 156)

2.2 ค่าสัมประสิทธิ์การอ้างอิงสรุป เป็นค่าประมาณของค่าเฉลี่ยของสหสัมพันธ์ระหว่างค่าการวัดที่สุ่มมาจากเอกภพรายคู่ (Cronbach 1972: 157) เช่น ค่าสัมประสิทธิ์การอ้างอิงสรุปเมื่ออ้างไปยังชุดข้อสอบ (แบบสอบ) ซึ่งประกอบด้วยข้อสอบ 20 ข้อ มีค่าเป็น .83 หมายความว่า ถ้าเราสุ่มนักเรียนจากประชากรหนึ่ง สมมุติเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ของอำเภอหนึ่งมาทำการทดสอบ สุ่มแบบสอบมาทีละ ฉบับ ๆ ละ 20 ข้อที่ไม่ซ้ำกัน ค่าเฉลี่ยของค่าสะมประสิทธิ์สหสัมพันธ์ระหว่างแบบสอบที่สุ่มมาจะมีค่าเป็น .83

2.3 ค่าสัมประสิทธิ์การอ้างอิงสรุป เป็นค่ากำลังสองของค่าสหสัมพันธ์ระหว่างคะแนนเอกภพกับคะแนนสังเกต (Brennan and Kane 1979: 40) ตัวอย่าง เช่น ค่าสัมประสิทธิ์การอ้างอิงสรุป ของแบบสอบคณิตศาสตร์เรื่องสมการของชั้นประถมศึกษาปีที่ 6 จำนวน 20 ข้อ มีค่าเป็น .90 แสดงว่า ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคะแนนเอกภพเรื่องสมการกับคะแนนสังเกตของนักเรียนประถมศึกษาปีที่ 6 ยกกำลังสองมีค่าเป็น .90 ถ้าถอดรากที่ 2 จะได้ค่าสหสัมพันธ์ระหว่างคะแนนเอกภพเรื่องสมการกับคะแนนสังเกต

2.4 ค่าสัมประสิทธิ์การอ้างอิงสรุป สามารถอธิบายในรูปอัตราส่วนระหว่างความแปรปรวนของคะแนนเอกภพกับคะแนนสังเกต (Cronbach, et al.1972; Brennan 1983) เช่น ค่าสัมประสิทธิ์การอ้างอิงสรุปเป็น .90 แสดงว่าความแตกต่างที่วัดได้ ร้อยละ 90

เป็นความแตกต่างเนื่องมาจากคะแนนเอกภพ อีกเพียงร้อยละ 10 เป็นความแตกต่างเนื่องมาจากความคลาดเคลื่อน

3. ข้อแตกต่างระหว่างค่าความเที่ยงกับค่าสัมประสิทธิ์การอ้างอิงสรุป ถึงแม้ว่าค่าสัมประสิทธิ์การอ้างอิงสรุปจะมีความหมายเช่นเดียวกับความเที่ยงของทฤษฎีการวัดแบบดั้งเดิม แต่ก็มีประเด็นที่แตกต่างกันดังนี้

3.1 การวัดแต่ละครั้งมีค่าสัมประสิทธิ์การอ้างอิงสรุปได้มากกว่า 1 ค่า

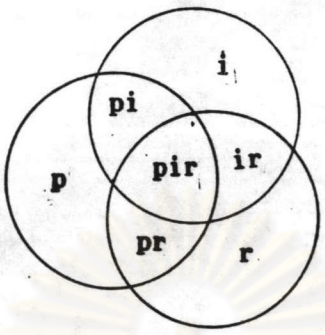
3.2 การอ้างอิงไปยังเอกภพใด จะต้องระบุและอธิบายเอกภพนั้นให้ชัดเจนและต้องสุ่มเงื่อนไขนั้นมาศึกษาด้วย

3.3 ค่าสัมประสิทธิ์การอ้างอิงสรุปสามารถบอกถึงความเป็นเอกพันธ์ของเอกภพได้ด้วย (Cronbach, et al. 1963: 159) ถ้าข้อสอบที่นำมาศึกษาเป็นตัวอย่างสุ่มจากเอกภพข้อสอบที่มีความเป็นเอกพันธ์ เราจะสามารถใช้คะแนนสังเกตแทนคะแนนเอกภพได้อย่างมั่นใจ

4. แนวคิดในการประมาณค่า parameter ของทฤษฎีการอ้างอิงสรุป Cronbach, et al. (1972) Brennan (1983) เสนอวิธีการประมาณค่า parameter เป็น 2 ขั้นตอน คือ ขั้นที่ 1 การศึกษาเพื่อการอ้างอิงสรุป (Generalizability Study ย่อว่า G Study) และ ขั้นที่ 2 การศึกษาเพื่อการตัดสินใจ (Decision Study ย่อว่า D Study) แต่ละขั้นมีรายละเอียดดังนี้

ขั้นที่ 1 G Study เป็นการประมาณค่าความแปรปรวนของแหล่งต่าง ๆ ภายใต้เงื่อนไขการวัดที่ยอมรับได้ เริ่มด้วยการกำหนดสิ่งที่ถูกวัด (object of measurement) เป็นคน (นักเรียน) หรือกลุ่มคน เช่น นักเรียนทั้งห้อง (class) หรือทั้งโรงเรียน ให้ตัวประกอบอื่น ๆ ภายใต้การสังเกตเป็นฟาเซ็ต (facets) เช่น ข้อสอบ โอกาสที่สอบ ผู้สอบ ฯลฯ กำหนดความสัมพันธ์ระหว่างฟาเซ็ตเป็นแบบ "crossed" หรือ nested แล้วใช้ ANOVA ที่สอดคล้องกับแบบแผนและความสัมพันธ์ของฟาเซ็ต ประมาณค่า mean square แล้วประมาณค่าความแปรปรวนของแหล่งต่าง ๆ เรียกค่าประมาณความแปรปรวนของสิ่งที่ถูกวัดว่า ค่าประมาณความแปรปรวนของคะแนนเอกภพ (estimated universe score variance) ซึ่งสอดคล้องกับค่าประมาณความแปรปรวนของคะแนนจริง (true score variance) ของทฤษฎีการวัดแบบดั้งเดิม เรียก

ค่าประมาณความแปรปรวนของฟาเซ็ตอื่น ๆ ว่าค่าประมาณความแปรปรวนของความคลาดเคลื่อน (estimated error score variance) การวิเคราะห์ในขั้น G Study ใช้ค่าขนาดของกลุ่มตัวอย่างที่ศึกษาจริง ๆ แต่ในการวิเคราะห์ขั้น D Study สามารถกำหนดขนาดของกลุ่มตัวอย่างแตกต่างไปจากที่วิเคราะห์ใน G Study ได้ เพื่อให้ผู้อ่านเข้าใจการวิเคราะห์ในขั้น G Study ดียิ่งขึ้น ผู้วิจัยใคร่ขอยกตัวอย่างประกอบการอธิบายตามที่ Brennan (1983: 2-6) ได้เสนอไว้ดังนี้ สมมุติว่าครูสมิธต้องการจะหาวิธีการประเมินผลความสามารถในการเขียนของนักเรียน ต้องเริ่มจากระบุข้อสอบความเรียงที่จะใช้วัด พร้อมกับกำหนดผู้ตรวจที่มีความเชี่ยวชาญ ถึงจุดนี้ สมิธยังไม่ปักใจว่าจะใช้วิธีการวัดอย่างใดอย่างหนึ่งเป็นการเฉพาะ กล่าวคือ ยังไม่กำหนดแน่นอนลงไปว่าเป็นข้อสอบข้อใด ใครจะเป็นคนตรวจ เพียงแต่บรรยายฟาเซ็ต (facet) ที่เขาเอง หรือนักวิจัยคนอื่น ๆ สนใจ ฟาเซ็ต หมายถึงชุดของเงื่อนไขการวัดที่คล้ายคลึงกัน (similar conditions of measurement) ข้อสอบแต่ละข้อถือเป็นเงื่อนไขการวัดหนึ่ง (condition) ที่ยอมรับได้ของฟาเซ็ตข้อสอบ (item facet) ผู้ตรวจแต่ละคน คือเงื่อนไขการวัดที่ยอมรับได้เงื่อนไขหนึ่งของฟาเซ็ตผู้ตรวจ (rater facet) ดังนั้น เอกภพของการสังเกตที่ยอมรับได้ (universe of admissible observation) จึงประกอบด้วยฟาเซ็ตข้อสอบและฟาเซ็ตผู้ตรวจ จากนั้นสมิธจะต้องหันมาพิจารณาความสัมพันธ์ระหว่างฟาเซ็ตผู้ตรวจและข้อสอบที่ตนเองสนใจ ถ้ายอมรับว่าในการตรวจคำตอบนั้น ผู้ตรวจ (r) แต่ละคนจะตรวจข้อสอบ (i) ทุกข้อเหมือนกัน เอกภพของการสังเกตที่ยอมรับได้ จะเป็นแบบ "crossed" ใช้สัญลักษณ์ว่า i x r อ่านว่า i crossed with r จากนั้น สมิธจะต้องเก็บรวบรวมและวิเคราะห์ข้อมูล โดยการสุ่มนักเรียนมาจำนวน n_p คน ใช้ข้อสอบจำนวน n_i ข้อ เมื่อสอบเสร็จก็สุ่มผู้ตรวจที่มีความเชี่ยวชาญมาตรวจให้คะแนน แล้วคำนวณค่าความแปรปรวนของฟาเซ็ตต่าง ๆ วิธีการดังกล่าวนี้คือ การทำ G-Study แบบของการศึกษา (design) มีชื่อว่า pxixr แบบแผนนี้มีแหล่งความแปรปรวนทั้งหมด 7 แหล่ง Cardinet, et al. (1972) จึงตั้งชื่อแบบแผนนี้ว่า แบบแผน 7 (VII Design) เขียนแทนด้วยแผนภูมิ Venn ได้ดังภาพที่ 1 ดังนี้



ภาพที่ 1 แผนภูมิ Venn แสดงองค์ประกอบความแปรปรวนของรูปแบบ pxixr เมื่อตัวประกอบที่ศึกษาทั้งหมดเป็นตัวประกอบสุ่ม

ตามรูปจะพบว่ามีผลหลัก (main effecs) อยู่ 3 ค่า คือ (1) ผลของบุคคล (p) (2) ผลของข้อสอบ (i) และ (3) ผลของผู้ตรวจ (r) ผลร่วมหรือปฏิสัมพันธ์สองระดับมี 3 ค่า คือ (1) ผลร่วมของบุคคลและข้อสอบ (pi) (2) ผลร่วมของบุคคลและผู้ตรวจ (pr) (3) ผลร่วมของข้อสอบและผู้ตรวจ (ir) และผลร่วมกันของบุคคลข้อสอบและผู้ตรวจ (pir)

ค่าความแปรปรวนของแหล่งต่าง ๆ เหล่านี้คำนวณจาก mean square และ mean square คำนวนโดยใช้ ANOVA แบบ factorial Design pxixr ในการประมาณค่าความแปรปรวนจาก mean square นั้นใช้สูตรเฉพาะอย่างที่สอดคล้องกับรูปแบบ เช่น

สูตรในการประมาณค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อฟาเซ็ตที่ศึกษาเป็นฟาเซ็ตสุ่มทั้งหมดและกำหนดรูปแบบความสัมพันธ์ระหว่างฟาเซ็ตเป็นแบบ crossed ใช้สูตรของ Cornfield & Tukey 1956 cited by Cronbach, et al. 1972: 43) ดังนี้

$$EMSp = \hat{\sigma}^2(pir,e) + n_i\hat{\sigma}^2(pr) + n_r\hat{\sigma}^2(pi) + n_in_r\hat{\sigma}^2(p)$$

$$EMSi = \hat{\sigma}^2(pir,e) + n_p\hat{\sigma}^2(ir) + n_r\hat{\sigma}^2(pi) + n_pn_i\hat{\sigma}^2(i)$$

$$EMSr = \hat{\sigma}^2(pir,e) + n_i\hat{\sigma}^2(pr) + n_p\hat{\sigma}^2(ir) + n_pn_i\hat{\sigma}^2(r)$$

$$EMSpi = \hat{\sigma}^2(pir,e) + n_r\hat{\sigma}^2(pi)$$

$$EMSpr = \hat{\sigma}^2(pir,e) + n_i\hat{\sigma}^2(pr)$$

$$EMSir = \hat{\sigma}^2(pir,e) + n_p\hat{\sigma}^2(ir)$$

$$EMSres = \hat{\sigma}^2(pir,e)$$

โครงสร้างของสมการนี้สัมพันธ์กับแผนภูมิ Venn ข้างต้น กล่าวคือ ในวงกลม p ประกอบด้วย พจน์ (term) ต่าง ๆ 4 พจน์ มี p, pi, pr และ pir,e เป็นเทอมทางขวามือของสูตรในการหาค่า EMSp ในพื้นที่ pi ซึ่งเป็นส่วนร่วมระหว่างวงกลม p และ i ประกอบด้วย 2 เทอมคือ pi และ pir,e ก็คือเทอมที่ปรากฏอยู่ทางขวามือของสูตรในการหาค่า EMSpi นั่นเอง ดังนั้นจึงสามารถสร้างสูตรหาค่า EMS ได้โดยอาศัยแผนภูมิ Venn โดยใช้ n_p เป็นตัวคูณควบถ้าองค์ประกอบนั้นไม่รวม p และคูณ n_i ถ้าองค์ประกอบนั้นไม่มี i ร่วมอยู่ด้วยและคูณ n_r ถ้าองค์ประกอบนั้นไม่มี r ร่วมอยู่ด้วย

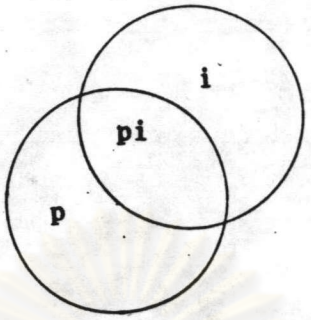
สูตรในการคำนวนเมื่อฟาเซ็ตที่ศึกษาเป็นแบบผสม คือ มีทั้งฟาเซ็ตคงที่และฟาเซ็ตสุ่มและรูปแบบความสัมพันธ์ระหว่างฟาเซ็ตเป็นแบบ crossed สมมุติให้ผู้ตรวจ (r) เป็นฟาเซ็ตคงที่สูตรในการคำนวณ รวมทั้งแหล่งความแปรปรวนที่ต้องการประมาณค่าเป็นดังนี้ (Cronbach, et al. 1972: 60)

$$EMSp = n_r\hat{\sigma}^2(pi,e|R^*) + n_in_r\hat{\sigma}^2(p|R^*)$$

$$EMSi = n_r\hat{\sigma}^2(pi,e|R^*) + n_pn_r\hat{\sigma}^2(i|R^*)$$

$$EMSpi,e = n_r\hat{\sigma}^2(pi,e|R^*)$$

เนื่องจากฟาเซ็ตผู้ตรวจ (r) เป็นฟาเซ็ตคงที่จึงหายไปจากแผนภูมิเดิม เหลือเพียงวงกลม p และ i เท่านั้นดังแผนภูมิ Venn ในภาพที่ 2 ดังนี้



ภาพที่ 2 แผนภูมิ Venn แสดงองค์ประกอบความแปรปรวนของรูปแบบ pxixr เมื่อ p และ i เป็นตัวประกอบสุ่ม r เป็นตัวประกอบคงที่

ค่าความแปรปรวนที่ได้จาก G study เป็นค่าประมาณความแปรปรวนที่แท้จริงของเอกภพหรือประชากร (parameter) ของเงื่อนไขการวัดที่มีขนาด 1 หน่วย กล่าวคือ เป็นความแปรปรวนของข้อสอบ 1 ข้อ ผู้ตรวจ 1 คน ของนักเรียน 1 คน เช่น $\hat{\sigma}^2(p)$ เป็นค่าประมาณความแปรปรวนของ $\sigma^2(p)$ หมายความว่า จากประชากรที่ทำการวัดนั้น สมมิธจะได้คะแนนของนักเรียน (p) แต่ละคนจากการสอบข้อสอบทุก ๆ ข้อ (N_i) และได้รับการตรวจจากผู้ตรวจทุก ๆ คน (N_r) ในเอกภพของการสังเกต คะแนนเฉลี่ยของนักเรียนแต่ละคนคือคะแนนเอกภพ (μ_p) ค่าความแปรปรวนของ μ_p จากการสอบวัดทุก ๆ คนในประชากรคือ $\sigma^2(U_p)$ หรือเขียนให้ง่ายว่า $\sigma^2(p)$ เป็นความแปรปรวนของคะแนนเอกภพ ซึ่งตรงกับความแปรปรวนของคะแนนจริงตามทฤษฎีการวัดแบบดั้งเดิม

ขั้นที่ 2 D Study ในขั้นนี้เน้นการใช้ค่าประมาณและการตีความองค์ประกอบความแปรปรวนสำหรับการตัดสินใจภายใต้วิธีการวัดที่เหมาะสม มีประเด็นสำคัญดังนี้

1. เอกภพของการอ้างอิง (Universes of generalization) เป้าหมายสำคัญของ D Study ได้แก่การกำหนดลักษณะเฉพาะของเอกภพของการอ้างอิงที่ผู้ตัดสินใจต้องการอ้างอิงถึง อาจจะประกอบด้วยเงื่อนไขทั้งหมดในเอกภพของการสังเกตที่ยอมรับได้ หรืออาจเป็นเซ็ตย่อย (subset) ของเอกภพการสังเกตที่ยอมรับได้ (Brennan 1983: 3)

2. ขนาดของตัวอย่าง D study (D study Sample sizes)จำนวนเงื่อนไข

ของฟาเซ็ตใน D study สามารถกำหนดแตกต่างจากจำนวนเงื่อนไขใน G study ใช้สัญลักษณ์ ($'$) แทนขนาดตัวอย่างของ D study เช่น n'_i และ n'_r แทนจำนวนข้อสอบและจำนวนผู้ตรวจ

3. โครงสร้างแบบแผน D study (D study design structure) นอกจากระบุขนาดของตัวอย่างใน D Study จะต้องระบุรูป (form) โครงสร้างของแบบแผนหรือความสัมพันธ์ของฟาเซ็ตที่ศึกษา สมมิธอาจต้องการตัดสินใจว่าในการสอบนั้น นักเรียนทุกคนต้องทำข้อสอบเหมือนกันทั้ง n'_i ข้อ และผู้ตรวจทั้ง n'_r คน ต้องตรวจทุก ๆ ข้อ แบบแผนดังกล่าวนี้เป็น p x I x R สังเกตว่าตัวอักษรตัวใหญ่แทนฟาเซ็ตใน D study ซึ่งในแบบแผน G study ใช้ตัวเล็ก ถ้าสมิธทำ D Study ตามแบบแผน p x I x R แล้ว D Study ของเขาจะตรงกับ G - study แต่ไม่จำเป็นต้องทำเช่นนั้นเสมอไป เขาสามารถตัดสินใจว่า ผู้สอบทุกคนทำข้อสอบทุกข้อ แต่ผู้ตรวจแต่ละคนตรวจคำตอบต่างข้อกัน แบบแผน D Study จะเป็น p x (I : R) ซึ่ง ":" อ่านว่า "nested Within"

4. การประมาณค่าความแปรปรวนในขั้น D Study การวิเคราะห์ข้อมูลในขั้น D Study ขึ้นอยู่กับการตัดสินใจของนักวัดผลหรือนักวิจัย ดังนั้นจึงต้องมีการประมาณค่าความแปรปรวนขึ้นมาใหม่อีกครั้ง โดยอาศัยผลจากการประมาณค่าในขั้น G Study เป็นฐาน และให้สอดคล้องกับแบบแผนและขนาดของตัวอย่างที่ต้องการตัดสินใจ เช่น ถ้าแบบแผนเป็น p x I x R ที่ข้อสอบและผู้ตรวจเป็นฟาเซ็ตสุ่ม จากตัวอย่าง สมมุติว่าสมิธต้องการอ้างอิงไปยังเอกภพข้อสอบและผู้ตรวจพร้อมกัน เมื่อการสอบของเขาต้องการใช้ $n'_i = 6$ และ $n'_r = 2$ สามารถคำนวณความแปรปรวนของ D Study ได้ดังนี้

ให้ค่าประมาณของความแปรปรวนใน G Study ของฟาเซ็ตต่าง ๆ เป็น

$\hat{\sigma}^2(p) = .30$, $\hat{\sigma}^2(i) = .25$, $\hat{\sigma}^2(r) = .10$, $\hat{\sigma}^2(pi) = .37$,

$\hat{\sigma}^2(pr) = .50$, $\hat{\sigma}^2(ir) = .25$ และ $\hat{\sigma}^2(pir) = 1.00$

ค่าความแปรปรวนใน D Study ทำได้ง่าย ๆ โดยหารค่าความแปรปรวนที่ได้จาก G Study ของผลต่าง ๆ ซึ่งเขียนในรูปสัญลักษณ์ทั่วไปว่า $\hat{\sigma}^2(\alpha)$ ด้วย $n'_i = 6$ ถ้า α เป็นผลที่มี i แต่ไม่มี r และหาร $\hat{\sigma}^2(\alpha)$ ด้วย $n'_r = 2$ ถ้า α เป็นผลที่มี r แต่ไม่มี i และหาร $\hat{\sigma}^2(\alpha)$ ด้วย $n'_i n'_r = 12$ ถ้า α มีทั้ง i และ r นั่นคือ

$\hat{\sigma}^2(p) = .30$, $\hat{\sigma}^2(I) = .04$, $\hat{\sigma}^2(R) = .05$, $\hat{\sigma}^2(pI) = .06$

$\hat{\sigma}^2(pR) = .25$, $\hat{\sigma}^2(IR) = .02$ และ $\hat{\sigma}^2(pIR) = .08$

5. ประมาณค่าความแปรปรวนของความคลาดเคลื่อน นอกจากการประมาณค่าองค์ประกอบความแปรปรวน (variance components) ตามรูปแบบและขนาดของกลุ่มตัวอย่างที่ต้องการแล้ว จะต้องประมาณค่าความแปรปรวนของความคลาดเคลื่อนซึ่งแบ่งเป็น 2 ชนิด คือ

5.1 ความแปรปรวนของความคลาดเคลื่อนแบบสัมบูรณ์ ใช้สัญลักษณ์ว่า $\sigma^2(\Delta)$ และค่าประมาณใช้ว่า $\hat{\sigma}^2(\Delta)$ เป็นความแปรปรวนของความแตกต่างระหว่างคะแนนสังเกตกับคะแนนเอกภพของผู้เข้าสอบ คำนวนจากผลบวกของค่าองค์ประกอบความแปรปรวนอื่น ๆ ทั้งหมดยกเว้นความแปรปรวนของคะแนนเอกภพ

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(I)+\hat{\sigma}^2(R)+\hat{\sigma}^2(pI)+\hat{\sigma}^2(pR)+\hat{\sigma}^2(IR)+\hat{\sigma}^2(pIR)$$

5.2 ความแปรปรวนของความคลาดเคลื่อนแบบสัมพัทธ์ (Relative error variance) ใช้สัญลักษณ์ว่า $\sigma^2(\delta)$ และค่าประมาณใช้ว่า $\hat{\sigma}^2(\delta)$ คำนวณจากผลบวกค่าขององค์ประกอบความแปรปรวนของผลร่วมระหว่างสิ่งที่ถูกวัด (p) กับฟาเซ็ตอื่น ๆ

$$\hat{\sigma}^2(\delta) = \sigma^2(pI)+\sigma^2(pR)+\sigma^2(pIR)$$

6. ประมาณค่าสัมประสิทธิ์การอ้างอิงสรุป ขั้นสุดท้ายในการทำ D Study คือการคำนวนค่าสัมประสิทธิ์การอ้างอิงสรุป (generalizability coefficient) ซึ่งเป็นดัชนีชี้บ่งถึงความเชื่อถือได้ของการวัด (dependability of measurement) (Brennan and Kane 1977: 279) ซึ่งใช้สูตรคำนวนตามนิยามความเที่ยงของทฤษฎีการวัดแบบดั้งเดิม ในรูปของ intraclass correlation ดังนี้

$$E\hat{\rho}^2 = \hat{\sigma}^2(p)/[\hat{\sigma}^2(p)+\hat{\sigma}^2(\Delta)] \text{ หรือ}$$

$$E\hat{\rho}^2 = \hat{\sigma}^2(p)/[\hat{\sigma}^2(p)+\hat{\sigma}^2(\delta)]$$

เมื่อ $E\hat{\rho}^2$ คือ ค่าสัมประสิทธิ์การอ้างอิงสรุป มีความหมายเช่นเดียวกับความเที่ยง

Cardinet, et al. (1976, 1981, 1983) ได้ขยายทฤษฎีการอ้างอิงสรุปของ Cronbach, et al. (1972) อีกทีหนึ่ง โดยเสนอแนวความคิดว่า แบบสอบทางจิตวิทยา การแนะแนว การศึกษา การบรรจุแต่งตั้ง ฯลฯ ต่างมุ่งวัดความแตกต่างของลักษณะอย่างใดอย่างหนึ่งของบุคคล เช่น ความถนัด ทักษะ ทัศนคติ ลักษณะทางบุคลิกภาพ เป็นต้น ดังนั้นความเที่ยงจึงหมายถึงความคงเส้นคงวา (stability) ของการจำแนกคะแนนของบุคคลที่ได้จากการทดสอบ

และความคลาดเคลื่อนของการวัดหมายถึง ปริมาณที่มีผลมาจากการสุ่มข้อสอบ โอกาสในการสอบ ผู้ดำเนินการสอบ ฯลฯ ในบางสถานการณ์ จุดหมายของการวัดต้องการเปรียบเทียบคุณสมบัติของข้อสอบ เช่น ความยาก หรือความแตกต่างของวิธีสอน ความเที่ยงจะหมายถึงความคงเส้นคงวาของการจำแนกข้อสอบหรือวิธีสอนนั้น การสุ่มนักเรียนหรือผู้สอบจะกลายเป็นความคลาดเคลื่อนของการวัดแทน โดยความคิดเช่นนี้ Cardinet, et al.จึงได้พัฒนาทฤษฎีการอ้างอิงสรุป เรียกแนวคิดที่พัฒนาขึ้นว่า การสมมาตรของทฤษฎีการสรุปอ้างอิง (the symmetry of generalizability theory) หลักการนี้ถือว่า ตัวประกอบ (facets) ทุกตัวที่ใช้ศึกษา ไม่ว่าจะเป็นผู้สอบ (p) ข้อสอบ (i) ผู้ตรวจ (r) ฯลฯ สามารถกำหนดเป็นสิ่งที่ถูกวัด (object of measurement) ได้ทั้งสิ้น เมื่อเราวัดสิ่งใดผลที่ได้คือ ความแตกต่าง (differentiation) ของสิ่งนั้น จึงเรียก สิ่งที่ถูกวัดว่า "ฟาเซ็ตของความแตกต่าง" (Differentiation facet) ใช้ตัวย่อว่า D facet ฟาเซ็ตที่เหลือเป็นเครื่องมือที่ใช้วัดความแตกต่าง จึงเรียกชื่อว่า "ฟาเซ็ตเครื่องมือ" (Instrument facet) ใช้ตัวย่อว่า I facet ขยายแนวคิดของ Cronbach, et al. ที่ให้สิ่งที่ถูกวัดเป็นเพียงฟาเซ็ตสุ่มเท่านั้น ให้สามารถเป็น ฟาเซ็ตคงที่ (fixed facet) ได้ด้วย

ขั้นตอนการประมาณค่าประชากรมี 4 ขั้นตอน (Cardinet, et al. 1983:19-23) ดังนี้

ขั้นที่ 1 ออกแบบการสังเกต (observation Design) ในขั้นนี้ต้องพิจารณาฟาเซ็ตที่ต้องการศึกษา พร้อมทั้งระบุความสัมพันธ์ของฟาเซ็ตว่าเป็นแบบ Crossed หรือ nested หรือ แบบ Confounded นอกจากนั้นยังต้องกำหนดจำนวนระดับหรือขนาดกลุ่มตัวอย่างของแต่ละฟาเซ็ตแล้วใช้วิธีการ ANOVA คำนวณค่า mean square ของผลต่าง ๆ

ขั้นที่ 2 ออกแบบการประมาณค่า (Estimation Design) ขั้นนี้ต้องระบุชนิดของเอกภพของแต่ละฟาเซ็ตว่า เป็นชนิดจำกัดหรือไม่จำกัด โดยพิจารณาจากการสุ่มเงื่อนไขการวัดของแต่ละฟาเซ็ต ตัวอย่างเช่นถ้า N แทนจำนวนเงื่อนไขการวัดที่สามารถยอมรับได้ในเอกภพของฟาเซ็ต (ขนาดของเอกภพ) ให้ n แทนจำนวนเงื่อนไขการวัดที่นำมาศึกษา (ขนาดของกลุ่มตัวอย่าง) วิธีการสุ่มสามารถทำได้ 3 วิธี คือ

1. สุ่มอย่างง่ายจากเอกภพที่มีขนาดไม่จำกัด ($n < N \longrightarrow \infty$)
2. สุ่มอย่างง่ายจากเอกภพที่มีขนาดจำกัด ($n < N < \infty$)

3. เลือกทุกเงื่อนไขการวัดมาจากเอกภพที่มีขนาดจำกัด ($n = N < \infty$)

วิธีการสุ่มแต่ละแบบจะเป็นตัวกำหนดชนิดของฟาเซ็ต ว่าจะเป็นแบบคงที่ เป็นแบบสุ่มแท้จริง หรือแบบสุ่มจำกัด หลังจากนั้นจะต้องประมาณค่าความแปรปรวนที่สัมพันธ์กับแบบและชนิดของฟาเซ็ตโดยใช้สูตร Cornfield and Tukey (1956) ประมาณค่าจาก mean square ในขั้นที่ 1

ขั้นที่ 3 กำหนดรูปแบบการวัด (Measurement Design) ในขั้นนี้ต้องจำแนกฟาเซ็ตแต่ละตัวว่าเป็นฟาเซ็ตของสิ่งที่ถูกวัด (D) หรือเป็นฟาเซ็ตของเครื่องมือ (I) ระบุแต่ละฟาเซ็ตเป็นแบบสุ่มหรือแบบคงที่ ในขั้นนี้ต้องการคำนวนค่าพารามิเตอร์ 3 ตัว คือ

1. ความแปรปรวนของคะแนนเอกภพซึ่งเป็นความแปรปรวนของสิ่งที่ถูกวัด

2. ความแปรปรวนของความคลาดเคลื่อนซึ่งแยกเป็นความคลาดเคลื่อนสัมบูรณ์หรือความคลาดเคลื่อนสัมพัทธ์

3. ค่าสัมประสิทธิ์การอ้างอิงสรุป

ขั้นที่ 4 การปรับรูปแบบการวัด (Optimization Design) ในขั้นตอนนี้เป็นการนำผลการคำนวนจากขั้นที่ 3 มาพิจารณาเพื่อปรับปรุงการวัดให้มีประสิทธิภาพ การปรับปรุงรูปแบบการวัดสามารถทำได้ 4 วิธี คือ

1. เพิ่มระดับของฟาเซ็ตที่ต้องการอ้างอิงสรุป

2. เปลี่ยนแปลงวิธีการสุ่มแต่ละฟาเซ็ต

3. นิยามเอกภพของฟาเซ็ตของการอ้างอิงสรุปใหม่หรือนิยามประชากรของสิ่งที่ถูกวัดใหม่

4. เปลี่ยนแปลงความสัมพันธ์ระหว่างฟาเซ็ต ในขั้นนี้ตรงกับขั้น D - Study ของ Cronbach, et al. (1972) นั่นเอง แต่ Cardinet, et al. ไม่ได้จำกัดอยู่เฉพาะด้านการตัดสินใจเท่านั้น แต่ยังมุ่งเน้นการสรุปผลการวัดในหลาย ๆ รูปแบบด้วย

ขั้นตอนการวิเคราะห์ตามที่กล่าวมาในแต่ละขั้นยังแบ่งออกเป็นขั้นตอนย่อย ๆ สามารถสรุปได้ดังตารางที่ 1 ดังนี้ (Cardinet and Allal 1983: 20-21)

ตารางที่ 1 กระบวนการวิเคราะห์การอ้างอิงสรุปตามแนวคิดของ Cardinet, et al.

ตัวแบบ	ขั้นตอนที่	สารสนเทศเกี่ยวกับฟาเซ็ต	รูปแบบการวัด	ผลการคำนวน	สัญลักษณ์	ขั้นตอนการคำนวน
การวิเคราะห์ความแปรปรวน	1	<u>การระบุสิ่งที่ศึกษา</u> ระบุลักษณะของกลุ่มข้อมูลที่สังเกต - การเลือกฟาเซ็ต - ความสัมพันธ์ของฟาเซ็ต . crossed . nested . confounded - จำนวนระดับของแต่ละฟาเซ็ต	การสังเกต (observation)	ค่าเฉลี่ยกำลังสองของผลต่างจากฟาเซ็ตที่ศึกษา (Mean - Squares)	α	1. ระบุแหล่งความแปรปรวนทั้งผลหลักและผลร่วม ตัวกำกับรวมของ (α) เขียนสัญลักษณ์ดังนี้ ตัวกำกับแรก : ตัวกำกับแผงตัวที่ 1 :...: ตัวกำกับแผงตัวที่ n เช่น ip:c:s หมายถึง ผลร่วมของข้อสอบกับนักเรียนที่แผงอยู่ในชั้นซึ่งแผงอยู่ในโรงเรียนอีกที
					MS(α)	2. หาค่าเฉลี่ยกำลังสองของผล (α) ต่าง ๆ
	2	<u>การสุ่มตัวอย่าง</u> ระบุจำนวนระดับและวิธีการสุ่มเพื่อให้รู้ว่า ฟาเซ็ตนั้นเป็นชนิด - สุ่มแท้จริง - สุ่มจำกัด - คงที่	การประมาณค่า (Estimation)	ค่าประมาณความแปรปรวนตามแบบจำลองสุ่มหรือผสมแล้วแต่กรณี	$\hat{\sigma}^2(\alpha)$ $\hat{\sigma}^2(\alpha\|M)$	3. หาค่าประมาณความแปรปรวนของแบบจำลองสุ่ม 4. หาค่าประมาณความแปรปรวนของแบบจำลองผสม

ตารางที่ 1 (ต่อ) กระบวนการวิเคราะห์การอ้างอิงสรุปตามแนวคิดของ Cardinet, et al.

ตัวแบบ	ขั้นตอนที่	สารสนเทศเกี่ยวกับฟาเซ็ต	รูปแบบการวัด	ผลการคำนวณ	สัญลักษณ์	ขั้นตอนการคำนวณ
ทฤษฎีการอ้างอิงสรุป	3	บทบาทในการวัด ระบุประชากรของสิ่งที่ถูกวัดและและเอกภพของเงื่อนไขการวัดที่ยอมรับได้ -ฟาเซ็ตของสิ่งที่ถูกวัดเป็น .ฟาเซ็ตสุ่ม (D^R) .ฟาเซ็ตคงที่ (D^F) -ฟาเซ็ตของเครื่องมือเป็น .ฟาเซ็ตคงที่หรือฟาเซ็ตสุ่มควบคุม (I^F) .ฟาเซ็ตสุ่ม (I^R)	การวัด (Measurement)	วางตำแหน่งของฟาเซ็ตในรูปแบบการวัดและระบุ Active Variance	$M(D^R/D^F/I^R/I^F)$	5.นิยามรูปแบบการวัด 1 แบบขึ้นไปเพื่อวิเคราะห์ในขั้นที่ 6 - 12 6.ควบคุมความสัมพันธ์ของรูปแบบการวัด แก้ไขไม่ให้มี D Facet แฝงอยู่ใน I Facet ซึ่งจะทำให้ความแปรปรวนของสิ่งที่ถูกวัดกับความคลาดเคลื่อนปะปนกัน 7.หาค่า Active Variance โดยขจัดองค์ประกอบที่มี I^R อยู่รวมอยู่ในตัวกำกับแรก 8.หาค่าคาดหวังของความแปรปรวนโดยการคูณ $(N_f-1)/N_f$ เข้ากับองค์ประกอบที่มีฟาเซ็ตคงที่หรือสุ่มจำกัดอยู่ในตัวประกอบแรก
				ความแปรปรวนสิ่งที่ถูกวัด	$\hat{\sigma}^2(\tau)$	9.ประมาณค่าความแปรปรวนของสิ่งถูกวัดโดยหักออกจากค่า Active Variance และผลบวกขององค์ประกอบทุกตัวที่มี D facet อยู่ในตัวกำกับเพียงตัวเดียว

ตารางที่ 1 (ต่อ) กระบวนการวิเคราะห์การอ้างอิงสรุปตามแนวคิดของ Cardinet, et al.

ตัวแบบ	ขั้นตอนที่	สารสนเทศเกี่ยวกับฟาเซ็ต	รูปแบบการวัด	ผลการคำนวณ	สัญลักษณ์	ขั้นตอนการคำนวณ
ทฤษฎีการอ้างอิงสรุป	3			ความแปรปรวนของความคลาดเคลื่อนหรือความคลาดเคลื่อนของการอ้างอิงสรุป	$\hat{\sigma}^2(\Delta)$	10.ประมาณค่าความแปรปรวนของความคลาดเคลื่อนสัมบูรณ์ โดยการหัก Active Variance ออกแล้วหาผลบวกขององค์ประกอบที่เหลือซึ่งคูณด้วย (1/ni) ทุกเทอมที่มี I facet อยู่ในตัวกำกับแรกและคูณด้วย (Ni-ni)/(Ni-1) ในเทอมที่มีฟาเซ็ตสุ่มจำกัดอยู่ในตัวกำกับแรก
					$\hat{\sigma}^2(\delta)$	11.ประมาณค่าความแปรปรวนของความคลาดเคลื่อนสัมพัทธ์ โดยการดึงผลบวกเฉพาะที่มี D facet อยู่ในตัวกำกับรวมออกจากความแปรปรวนใน ข้อ 10
				สัมประสิทธิ์การอ้างอิงสรุป	$E\hat{\rho}^2$	12.ประมาณค่าสัมประสิทธิ์การอ้างอิงสรุปโดยหารค่าความแปรปรวนของสิ่งที่ถูกวัดด้วยผลบวกของความแปรปรวนของสิ่งที่ถูกวัดและความแปรปรวนของความคลาดเคลื่อน

ตารางที่ 1 (ต่อ) กระบวนการวิเคราะห์การอ้างอิงสรุปตามแนวคิดของ Cardinet, et al.

ตำแหน่ง	ขั้นตอนที่	สารสนเทศเกี่ยวกับฟาเซ็ต	รูปแบบการวัด	ผลการคำนวณ	สัญลักษณ์	ขั้นตอนการคำนวณ
ทฤษฎีการอ้างอิงสรุป	4	การปรับปรุงรูปแบบการวัด ระบุประชากรของสิ่งที่ถูกวัดและเอกภพของการอ้างอิงสรุป โดย - ปรับความสัมพันธ์ระหว่างฟาเซ็ตจาก Crossed เป็น Nesting, Confounding - ปรับจำนวนระดับของค่าสังเกต โดยเพิ่มจำนวนระดับของ I^R facet - ปรับขนาดและวิธีการสุ่ม D facet และ I Facet ใหม่	การปรับรูปแบบการวัด (Optimization)	วางตำแหน่งของฟาเซ็ตในรูปแบบการการวัดใหม่และระบุ Active variance	$O(D/D_F^R/I/I_F^R)$	13. นิยามรูปแบบการวัดใหม่โดยการปรับในขั้นการสังเกต การประมาณค่า และการวางรูปแบบการวัดผลเพื่อลดความคลาดเคลื่อน เพิ่มความตรง หรือลดค่าใช้จ่าย โดยการคำนวณซ้ำขั้นที่ 5-8
				D-variance	$\hat{\sigma}^2(\tau')$	14. ประมาณค่าความแปรปรวนของสิ่งที่ถูกวัด โดยคำนวณซ้ำในขั้นที่ 9
				Error variance G-coefficient	$\hat{\sigma}^2(\Delta)$ $\hat{\sigma}^2(\delta')$ $E\hat{\rho}^2$	15. ประมาณค่าความแปรปรวนของความคลาดเคลื่อน โดยคำนวณซ้ำในขั้นที่ 10 และ 11 16. ประมาณค่าสัมประสิทธิ์การอ้างอิงสรุป โดยคำนวณซ้ำในขั้นที่ 12

การประยุกต์ทฤษฎีการอ้างอิงสรุปในการวัดผลการศึกษา

ในปี 1972 The US Department of Labour (cited by Shavelson and Webb 1981) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุปในการพัฒนามาตรวัด GED (General Educational Development) ที่ใช้ในการประเมินความสามารถด้านเหตุผล คณิตศาสตร์และภาษาที่จำเป็นต่องานอาชีพต่าง ๆ เพื่อประมาณเวลาที่จะใช้ในการเรียนรู้งาน การเทียบงาน และการสร้างโปรแกรมการฝึกงาน โดยใช้รูปแบบการศึกษาแบบไม่สมดุลย์ คือให้ผู้ประเมิน (r) Nested อยู่ในศูนย์ (c) ต่าง ๆ ของหน่วยงาน และ Crossed กับงานอาชีพ (j) และจำนวนครั้ง (O) ที่ประเมิน เขียนในรูปสัญลักษณ์ว่า (r:c)xjxo ความสามารถแต่ละด้าน จะได้รับการประเมินโดยผู้ประเมิน 71 คนจากศูนย์อาชีพ 11 แห่ง พบว่า ความเชื่อมั่นของการประเมิน มีค่าเพิ่มขึ้นตามการเพิ่มจำนวนผู้ประเมิน ข้อค้นพบนี้แสดงให้เห็นว่า จำนวนผู้ประเมินมีผลต่อความเชื่อมั่นในการประเมิน

Smith (1978) ได้ศึกษาถึงความคลาดเคลื่อนในการสุ่มตัวอย่างของการศึกษาการอ้างอิงสรุปชนิดหลายตัวประกอบ (multifacet study) ที่ใช้กลุ่มตัวอย่างจำนวนน้อย ๆ และเสนอวิธีลดความคลาดเคลื่อนเชิงสุ่มไว้ 3 วิธีคือ (1) สุ่มระดับของฟาเซ็ตแต่ละตัวในแบบจำลองการวิเคราะห์ความแปรปรวนให้ได้จำนวนมากที่สุดเท่าที่จะทำได้ (2) สุ่มระดับของฟาเซ็ตแต่ละตัวให้เป็นสัดส่วนกับขนาดของเอกภพของฟาเซ็ตและ (3) เปลี่ยนรูปแบบความสัมพันธ์ระหว่างฟาเซ็ต สำหรับข้อ (2) และ (3) ควรใช้เมื่อไม่สามารถสุ่มตัวอย่างให้ได้จำนวนมาก ๆ

Macready (1983: 149-157) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุปในการประเมินความยากและความเป็นเอกพันธ์ (homogeneity) ของข้อสอบอิงโดเมนที่ใช้ในการวินิจฉัย โดยมีฟาเซ็ตที่ศึกษาประกอบด้วย โดเมนการคูณจำนวนเต็ม (d) ห้องเรียน (c) จำนวนหลักของตัวคูณ (n) ข้อสอบซึ่งแฝงอยู่ในโดเมนและจำนวนหลัก (i:(d x n)) และนักเรียนซึ่งแฝงอยู่ในชั้นเรียน (s:c) จากรูปแบบการวิเคราะห์ G study แบบ (s:c)x(i:(dxn)) โดยให้ห้องเรียน นักเรียน และข้อสอบเป็นฟาเซ็ตสุ่ม ให้โดเมนและจำนวนหลักของตัวคูณเป็นฟาเซ็ตคงที่ พบว่า (s:c)x(i:(dxn)),e เป็นแหล่งความแปรปรวนที่มีค่ามากที่สุดคือ 47 % อีก 4 แหล่งที่มีค่ารอง

ลงไป ได้แก่ ห้องเรียน นักเรียนซึ่งแฝงอยู่ในห้องเรียน โดเมน และผลร่วมระหว่างนักเรียนกับโดเมน (s : c) x d โดยทั้ง 4 แหล่งมีค่ารวมกันถึง 51 % เฉพาะแหล่งความแปรปรวนสุดท้ายหมายถึง ความยากของข้อสอบในแต่ละโดเมนสำหรับนักเรียนแต่ละคนมีค่าไม่เท่ากัน แหล่งความแปรปรวนอื่น ๆ ที่เหลือมีค่าน้อยมาก โดยเฉพาะจำนวนหลักของตัวคูณ Macready สรุปว่าไม่ควรจะใช้ตัวคูณ 4 ตำแหน่ง ใช้เพียง 3 ตำแหน่งก็พอ เพราะให้ค่าความยากไม่แตกต่างกัน อีกแหล่งหนึ่งที่มีค่าน้อยคือ i:(dxn) แสดงว่าความยากของข้อสอบในทุกโดเมนที่มีตำแหน่งตัวคูณเท่ากัน มีค่าพอ ๆ กัน เมื่อตรวจดูความยากรายโดเมน พบว่าเกือบทุกโดเมนมีข้อสอบที่มีความยากเท่าเทียมกัน มีเพียงโดเมนที่ 15 ที่ค่อนข้างจะแตกต่างกัน Macready เสนอว่าควรจะแยกให้เป็นโดเมนย่อย หรือนำไปรวมกับโดเมนอื่น จากการประมาณค่าสัมประสิทธิ์การอ้างอิงสรุปของข้อสอบแต่ละโดเมน โดยใช้รูปแบบการวิเคราะห์ 2 รูปแบบ คือ sxi และ sxr เมื่อ r หมายถึงการสอบซ้ำ พบว่า ค่าสัมประสิทธิ์การอ้างอิงสรุปของข้อสอบหนึ่งข้อมีค่าอยู่ระหว่าง .338 ถึง .606

Ibrahim (1984) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุป เพื่อประมาณค่าความแปรปรวนที่มีต่อการประเมินวัตถุประสงค์ทางการศึกษา (The Rating of Evaluational Goals) โดยสุ่มตัวอย่างครู 80 คน และนักศึกษา 80 คน ในประเทศซูดาน ประเมินวัตถุประสงค์ทางการศึกษา 2 ชนิด คือ วัตถุประสงค์ที่สำคัญจริง ๆ และวัตถุประสงค์ตามที่คาดหวัง ตัวประกอบในการศึกษามี ผู้ประเมิน กลุ่มผู้ประเมิน จำนวนครั้งของการประเมิน ถิ่นที่อยู่ของผู้ประเมิน ชนิดของวัตถุประสงค์ สถานที่ทำงานของผู้ประเมิน และเพศของผู้ประเมิน พบว่า ตัวประกอบที่มีผลต่อการประเมินมากที่สุด ได้แก่ ผู้ประเมินและกลุ่มผู้ประเมิน ส่วนสถานที่ทำงานของผู้ประเมิน มีผลเล็กน้อยเท่านั้น ส่วนตัวประกอบอื่นที่เหลือไม่มีผลต่อการประเมินเลย

O'Brien (1986) ใช้ทฤษฎีการอ้างอิงสรุปในการประมาณค่าความเที่ยงของตัวแปรระดับโรงเรียน 16 ตัว ซึ่งเป็นค่าเฉลี่ยหรือร้อยละของกลุ่มตัวอย่างนักเรียน ตัวแปรแบ่งเป็น 5 กลุ่มคือ ค่าเฉลี่ยของสถานภาพทางสังคมและเศรษฐกิจของครอบครัว ค่าเฉลี่ยของผลสัมฤทธิ์ทางการเรียน ร้อยละของนักเรียนที่มีบิดาหรือมารดาอาศัยอยู่ด้วย ร้อยละของคุณภาพของห้องสมุดและการเรียนการสอนของโรงเรียน และร้อยละของนักเรียนที่ยอมรับกฎระเบียบของโรงเรียน กลุ่ม

ตัวอย่าง คือโรงเรียน 1122 โรง และนักเรียนชั้นปีที่ 2 โรงเรียนละ 36 คน ดำเนินการวิจัยโดยให้กลุ่มตัวอย่างตอบคำถามของตัวแปรแต่ละตัว จากการศึกษาพบว่า ความเที่ยงในการตอบแบบสอบถามเพิ่มขึ้นตามการเพิ่มจำนวนผู้ตอบคำถามจากแต่ละโรงเรียน จำนวนข้อคำถาม และจำนวนโรงเรียน ข้อค้นพบนี้แสดงให้เห็นว่า จำนวนผู้ประเมิน จำนวนข้อคำถามของเครื่องมือประเมินและจำนวนสถานที่ทำงานของผู้ประเมินมีผลต่อความเที่ยงของการประเมิน

Webb, Herman and Cabello (1987: 130) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุปในการวัดแบบอิงโดเมนเพื่อการสอบแบบวินิจฉัย ในการวัดด้านภาษาเรื่องสรรพนาม ได้เสนอวิธีการพัฒนาและวิเคราะห์การใช้ประโยชน์ของแบบสอบวินิจฉัยสำหรับครู โดยการเชื่อมโยงระหว่างการวัดผลแบบอิงโดเมนและทฤษฎีการอ้างอิงสรุป เพื่อหาว่าความสามารถด้านเนื้อหากลุ่มใดควรนำเสนอในสัญฐานคะแนน (profile) ของนักเรียน การศึกษาแบ่งเป็น 4 ขั้นคือ (1) กำหนดลักษณะเฉพาะของโดเมน และการสร้างแบบสอบ (2) เลือกกลุ่มเนื้อหาที่ควรเสนอในสัญฐานคะแนน (3) กำหนดจำนวนข้อสอบที่จำเป็นต่อความเที่ยงของการวัด และ (4) คำนวนค่าความแม่นยำของสัญฐานคะแนน ในขั้นที่ 2 เป็นขั้นที่เริ่มใช้ทฤษฎีการอ้างอิงสรุป รูปแบบที่ใช้วิเคราะห์ใน G study คือ s x i และใช้ความคลาดเคลื่อนแบบสัมบูรณ์คำนวณค่าสัมประสิทธิ์การอ้างอิงสรุป ทั้งแบบ univariate และ multivariate ฟาเซ็ตที่ใช้ศึกษาประกอบด้วยกลุ่มเนื้อหาย่อยของเรื่องสรรพนาม มี (1) rule ได้แก่ nominative, direct object, indirect object of preposition (2) form ได้แก่ relative, nonrelative (3) number ได้แก่ sigular, plural (3) embeddedness ได้แก่ sentence, paragraph แต่ละฟาเซ็ต "crossed" กัน แต่ข้อสอบ (i) แฝงอยู่ในฟาเซ็ตอื่น สิ่งที่ถูกวัดคือนักเรียน สุ่มมาจากนักเรียนเกรด 6 จำนวน 128 คน ให้เนื้อหาเป็นฟาเซ็ตคงที่ แต่ข้อสอบเป็นฟาเซ็ตสุ่ม ผลการวิเคราะห์ความแปรปรวนพบว่ามี 3 ฟาเซ็ตที่มีผลต่อคะแนนมากที่สุดคือ form, embeddedness และ rule มีเพียงบางฟาเซ็ตมีความสัมพันธ์กับความแตกต่างระหว่างนักเรียน ฟาเซ็ตใดที่ไม่มีปฏิสัมพันธ์กับนักเรียนแสดงว่ามีผลคงที่สำหรับนักเรียนทุก ๆ คนได้แก่ ผลหลักของ form ผลร่วมกันระหว่าง Form- embeddedness และ Rule-Form ผลร่วมหว่าง F-E แสดงว่าแต่ละ Form ของ pronoun ในแต่ละ embeddednes ข้อสอบมีความยากไม่เท่ากัน แต่ก็ยังถือว่าผลดังกล่าวมีค่าคงเส้นคงวาในนักเรียนที่ทำการทดสอบทุกคน คือ ทุกคนตอบถูกหรือผิดคล้อย

ตามกัน ดังนั้นถ้าต้องการนำเสนอสัญฐานคะแนนก็ให้อยู่ในรูปสัญฐานของกลุ่ม (group profile) ไม่จำเป็นต้องเสนอรายบุคคล (individuals'profile) ผลของฟาเซ็ตที่สัมพันธ์กับความแตกต่างระหว่างบุคคล Webb, et al. แนะว่า ควรนำเสนอในรูปสัญฐานคะแนนรายบุคคลได้แก่ rule และ Form ดังนั้นจากกลุ่มเนื้อหาทั้งหมด 32 กลุ่ม ควรจะนำเสนอในรูปสัญฐานคะแนนรายบุคคลเพียง 8 กลุ่มเท่านั้น จำนวนข้อสอบที่เหมาะสมในแต่ละกลุ่มเป็น 8 ข้อ และพบว่าค่า $E\hat{\rho}^2$ ของข้อสอบ 8 ข้อ สำหรับเนื้อหา 8 กลุ่ม มีค่าอยู่ระหว่าง .35-.75

ในประเทศไทยยังมีการประยุกต์ทฤษฎีการอ้างอิงสรุปในการวิจัยน้อยมาก แต่ก็ได้มีการเสนอบทความทางวิชาการเกี่ยวกับทฤษฎีนี้บ้าง เช่น บทความของ จักรกฤษณ์ สำราญใจ (2529: 36-47) งานวิจัยที่ประยุกต์ทฤษฎีนี้กับการประเมินผลนั้น แดง กลางท่าได่ (2531) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุปในการหาความเชื่อมั่น (reliability) ของการประเมินความตรงเชิงเนื้อหา รูปแบบการวัดในการศึกษานั้น ผู้เชี่ยวชาญทุกคนประเมินความตรง (ความเหมาะสม) ของข้อสอบทุกข้อว่าวัดตรงจุดประสงค์เพียงใด ผู้เชี่ยวชาญแฝงอยู่ในโรงเรียน รูปแบบการศึกษา G study เป็นแบบ i x (r:s) เมื่อ i คือ ฟาเซ็ตข้อสอบ r คือ ฟาเซ็ตผู้ตรวจ และ s คือ ฟาเซ็ตโรงเรียน ผลการศึกษาพบว่า แหล่งความแปรปรวนที่มีอิทธิผลต่อความเชื่อมั่นได้แก่ ความแปรปรวนของข้อสอบ ความแปรปรวนของผู้เชี่ยวชาญ ซึ่งแฝงอยู่ในโรงเรียนและผลร่วมระหว่างข้อสอบกับโรงเรียน ส่วนรูปแบบการวัดที่ให้ค่าความเชื่อมั่นในการประเมินความตรงเชิงเนื้อหาโดยผู้เชี่ยวชาญที่สูงกว่ารูปแบบการวัดอื่น ๆ ภายใต้ความสัมพันธ์ของฟาเซ็ต แบบ i x (r:s) คือ M (I/-/S/R) ส่วนขนาดของกลุ่มตัวอย่างที่จะเป็นตัวแทนของสมาชิกทั้งหมดในเอกภพของตัวประกอบ พบว่า ข้อสอบอย่างน้อย 9 ข้อ และผู้เชี่ยวชาญไม่เกิน 9 คนต่อโรงเรียน สุ่มจากโรงเรียนอย่างน้อย 7 โรงเรียน จึงจะทำให้ค่าสัมประสิทธิ์การอ้างอิงสรุปอย่างน้อย 0.80

จะเห็นว่าทฤษฎีการอ้างอิงสรุป สามารถนำไปประยุกต์เข้ากับการวัดและประเมินผลในหลาย ๆ สถานการณ์ได้ดี นับตั้งแต่การพัฒนาข้อสอบจนถึงการตีความคะแนนในรูปของ profile กล่าวได้ว่าสามารถพัฒนาการวัดผลได้ครบวงจรทีเดียว

การสอบแบบความเรียง

1. นิยามแบบสอบความเรียง Stalnaker (1951 cited by Coffman 1971: 271) ให้นิยามคำถามแบบความเรียงว่า หมายถึง ข้อคำถามซึ่งต้องการให้ผู้ตอบเขียนคำตอบขึ้นเอง โดยธรรมชาติเป็นคำตอบที่ไม่ใช่จะถูกต้องเพียงคำตอบเดียว หรือรูปแบบเดียว ความถูกต้องและคุณภาพของคำตอบต้องได้รับการตัดสิน โดยผู้มีความรู้และทักษะในเนื้อหาที่ถามเป็นอย่างดี ลักษณะเด่นของคำถามแบบความเรียง คือการให้อิสระแก่ผู้ตอบในการแสดงความคิด และคำตอบถูกไม่ใช่มีเพียงคำตอบเดียว ไม่สามารถตรวจโดยวิธีตรวจนับอย่างง่าย แม้แต่ผู้เชี่ยวชาญก็ไม่อาจลชี้ถูกหรือผิดได้อย่างเด็ดขาด เพียงแต่สามารกถจะพิจารณาถึงระดับ (degree) ของคุณภาพคำตอบได้

Coffman (1971: 271) ได้ให้นิยามการสอบความเรียง (essay examination) ว่า หมายถึงการสอบที่ใช้ข้อสอบแบบความเรียง ตั้งแต่หนึ่งข้อขึ้นไปกับกลุ่มนักเรียน ภายใต้สถานการณ์อย่างหนึ่ง เพื่อเก็บรวบรวมข้อมูลในการประเมินผล คำถามแบบความเรียงแตกต่างจากคำถามแบบตอบสั้น ๆ ตรงที่ต้องใช้ผู้เชี่ยวชาญตัดสินไม่ใช่การตรวจอย่างง่ายโดยใช้ตัวเฉลยแบบเดียวกับแบบสอบปรนัย แตกต่างจากรายงานที่ทำส่งครูในการเรียนวิชาต่าง ๆ ตรงที่การดำเนินการสอบ จะต้องดำเนินให้เป็นมาตรฐานเดียวกัน และต้องการข้อสอบที่เป็นตัวแทนผลการสอบที่ต้องการ

เยาวดี วิบูลย์ศรี (2528 : 122) ให้ความหมายของแบบสอบถามความเรียงว่าเป็นแบบสอบประเภท supply type คือ ผู้สอบต้องเรียบเรียงแนวความคิดความรู้ที่ได้เรียนมาตลอดจน เรียบเรียงภาษา ผูกเป็นรูปประโยคให้ได้ใจความชัดเจน เขียนเป็นคำตอบให้เหมาะสมกับความต้องการของคำถาม ดังนั้นข้อกระทงของแบบสอบความเรียงโดยทั่วไปจะไม่จำกัดเสรีภาพของผู้ตอบในการจัดเรียบเรียงความรู้ความคิด รวมทั้งการเรียบเรียงข้อเท็จจริงต่าง ๆ อันเป็นข้อมูลข่าวสารของคำตอบ

Tuckman (1975 อ้างถึงใน เยาวดี วิบูลย์ศรี 2528: 122) ให้ความหมายของแบบสอบความเรียงว่า เป็นแบบสอบที่ให้ผู้สอบได้แสดงความสามารถในการประยุกต์ ความรู้ วิเคราะห์ สังเคราะห์ และประเมินผลความรู้ ดังนั้นข้อกระทงของแบบสอบความเรียงต้องเป็นข้อคำถามที่ให้โอกาสผู้สอบ ได้สร้างและเรียงเรียงคำตอบในรูปสัมพันธ์ตามขอบข่ายความรู้ที่กว้าง

ทั้งนี้เพื่อให้ผู้สอบความเรียง สามารถวัดกระบานการคิดในระดับสูงตามแนวคิดของ Bloom คือ ระดับการเรียนรู้ขั้นวิเคราะห์ สังเคราะห์ และการประเมินผลเป็นส่วนใหญ่ หรือวัดในระดับการนำไปใช้แก้ปัญหาในสถานการณ์ต่าง ๆ บ้าง

โกวิท ประวาลพฤกษ์ และ สมศักดิ์ สินธุรเวชญ์ (2527: 101) ให้นิยามข้อสอบแบบความเรียงว่า เป็นข้อสอบที่ให้เด็กได้มีอิสระในการตอบเต็มที่ ให้โอกาสเด็กได้แสดงว่าเขารู้อะไร สามารถวัดสมรรถภาพด้านบูรณาการความคิด ความสามารถในการจัดรวบรวมความคิดที่สมบูรณ์ (organization ability) การสังเคราะห์ และการประเมินค่าอย่างมีประสิทธิภาพ เช่น ความรู้ในการแก้ปัญหาใหม่ ๆ และเป็นต้นคิดหรือค้นหาวิธีการใหม่ ๆ เพื่อแก้ปัญหาต่าง ๆ

จากที่กล่าวมาจะเห็นว่า ข้อสอบแบบความเรียงเป็นข้อคำถามที่ผู้ตอบจะต้องเขียนคำตอบด้วยตนเอง ผู้ตอบมีอิสระในการใช้ความสามารถด้านการเขียน เรียบเรียง ประยุกต์ความรู้ ความคิดได้กว้างขวาง คำตอบที่ถูกต้องจะมีหลายคำตอบ ผู้ที่จะตัดสินคุณภาพคำตอบจะต้องเป็นผู้เชี่ยวชาญในเรื่องที่ถามเป็นอย่างดี ดังนั้นข้อสอบแบบความเรียงจึงเหมาะที่จะใช้วัดพฤติกรรมความรู้ความคิดที่มีความซับซ้อน เช่น การสังเคราะห์และประเมินผล

2. ความเที่ยงของแบบสอบความเรียง ปัญหาหลักของการวัดแบบความเรียง คือ ความเที่ยงของแบบสอบต่ำ การที่แบบสอบมีความเที่ยงต่ำแสดงว่า ค่าที่วัดได้มีส่วนผสมที่เป็นคะแนนจริงน้อย แต่มีความคลาดเคลื่อนมาก (Ebel and Frisbie 1986; Coffman 1971: 276) งานวิจัยส่วนมากหาความเที่ยงของแบบสอบความเรียง โดยการหาค่าสหสัมพันธ์ระหว่างผู้ตรวจหลายคน หรือภายในผู้ตรวจคนเดียวกันแต่ตรวจต่างโอกาส ดังนั้นค่าสหสัมพันธ์ที่ได้จึงเป็นค่าที่แสดงความสอดคล้องกัน (agree) ระหว่างผู้ตรวจ ไม่ใช่ความเที่ยงของเครื่องมือโดยตรง ในหัวข้อนี้มีนักวัดผลได้ให้ความคิดเห็นตลอดทั้งผลงานวิจัยดังนี้

Coffman (1972: 277) กล่าวว่าปัญหาความสอดคล้องระหว่างผู้ตรวจเป็นเรื่องค่อนข้างซับซ้อนเกิดจากสาเหตุสำคัญที่ไม่สามารถหลีกเลี่ยงได้ 3 ประการ คือ

(1) ผู้ตรวจต่างคนมีแนวโน้มจะให้คะแนนความเรียงชิ้นเดียวกันแตกต่างกัน

(2) ผู้ตรวจคนเดียวกันมีแนวโน้มจะให้คะแนนความเรียงชิ้นเดียวกันแตกต่างกันตามโอกาสที่ตรวจ

(3) ความไม่สอดคล้องกันจะยิ่งมากขึ้น เมื่อข้อสอบเปิดโอกาสให้ผู้ตอบมีอิสระในการตอบมากขึ้น สาเหตุของความไม่สอดคล้องแต่ละด้านเป็นเรื่องที่ค่อนข้างซับซ้อน ความแตกต่างของคะแนนแต่ละคนอาจได้รับอิทธิพลต่อไปนี้

ก. ผู้ตรวจแต่ละคนมีความเข้มงวดแตกต่างกัน บางคนชอบให้คะแนนมาก แต่บางคนชอบให้คะแนนต่ำ ๆ

ข. การกระจายของคะแนนหรือพิสัยของคะแนนที่ได้รับจากผู้ตรวจแต่ละคนจะมีความแตกต่างกันบางคนให้คะแนนใกล้ ๆ กับค่าเฉลี่ย บางคนให้คะแนนมีพิสัยกว้าง คือ เกือบเต็มและเกือบศูนย์ เป็นต้น

ค. ผู้ตรวจต่างกันจะให้ความสำคัญกับประเด็นที่ต้องการให้คะแนนต่างกัน เช่น การตรวจความสามารถด้านการเขียน ซึ่งให้คะแนน 5 ด้านมี (1) แนวคิด (2) รูปแบบ (3) อรรถรส (4) กลไกในการเขียน และ (5) การใช้คำ ครูแต่ละคนจะให้คะแนนแต่ละด้านต่างกัน

Block (1985: 41-52) ศึกษาผลการประเมินความเรียงแบบ multiple rating ทั้งโดยวิธีใช้ผู้ตรวจหลายคน และคนเดียวตรวจหลายครั้ง เพื่อดูว่าการตรวจแต่ละครั้งหรือแต่ละคนได้ค่าคะแนนจริงตรงกันหรือไม่ ให้ผู้ตรวจ 16 คน ตรวจคำตอบความเรียง 105 ชิ้น 2 ครั้ง พบว่า ผู้ตรวจคนเดียวกันตรวจต่างกัน 2 โอกาสได้ค่าคะแนนจริงตรงกัน แต่ผู้ตรวจต่างคนจะไม่ได้คะแนนจริงตรงกัน ค่าประมาณของสหสัมพันธ์ระหว่างคะแนนจริงของผู้ตรวจหลายคน มีค่าอยู่ระหว่าง .415-.910

Finlayson (1951: 126-134) ศึกษาความเที่ยงของแบบสอบความเรียง โดยผู้ตรวจ 4 คน ตรวจความเรียง 2 ชุด พบว่ามีค่าอยู่ระหว่าง .636-.957 โดยมีค่าเฉลี่ยเป็น .810 แต่เมื่อเทียบกับการประเมินครั้งแรกโดยผู้ตรวจชุดเดียวกัน พบว่าค่าสหสัมพันธ์อยู่ระหว่าง .610-.798 ค่าเฉลี่ยเป็น .687 ค่าสหสัมพันธ์ระหว่างผู้ตรวจรายคู่อยู่ระหว่าง .591 ถึง .770 ค่าเฉลี่ยเป็น .703 จะเห็นว่าค่าความเที่ยงของความเรียงเรื่องเดียวกัน โดยผู้ตรวจชุดเดียวกัน มีค่าแตกต่างกันไปเมื่อโอกาสการตรวจเปลี่ยนไป

Coffman (1971: 278) ได้ตรวจสอบเอกสารงานวิจัยที่เกี่ยวกับแบบสอบความเรียง พบค่าความเที่ยงอยู่ระหว่าง .35-.98 โดยความเที่ยงสูงสุดเป็นการศึกษาของ Gosling (1966) Coffman ได้ให้ความเห็นเชิงวิจารณ์ไว้ว่า การตัดสินว่าความเที่ยงของแบบสอบหรือ

ข้อคำถามแบบความเรียงควรมีค่าเป็นเท่าไรเป็นเรื่องที่ตอบยาก เพราะค่าความเที่ยงจะเปลี่ยนไปตามปัจจัยที่เกี่ยวข้อง ถ้าประเมินกลุ่มนักเรียนขนาดใหญ่ที่มีพื้นฐานการเรียนการสอนแตกต่างกันมาก ความเที่ยงของการตรวจจะต่ำ แต่จะมีค่าสูงขึ้นเมื่อตรวจกลุ่มเล็กที่มีพื้นความรู้ใกล้เคียงกัน ถ้าผู้ตรวจมีความแตกต่างกันมากความเที่ยงก็จะต่ำ นอกจากนั้นเนื้อหาวิชา เป็นอีกปัจจัยหนึ่ง วิชาใดมีเนื้อหาเป็นกฎเกณฑ์ตายตัว เช่น คณิตศาสตร์ เคมี จะมีความเที่ยงสูงกว่าวิชาภาษา นอกจากผู้สอบ ผู้ตรวจ และเนื้อหาวิชาแล้ว ลักษณะของคำถามก็เป็นอีกปัจจัยหนึ่งที่มีอิทธิพลต่อความเที่ยง คำว่า "คำถาม" ในการสอบแบบความเรียง De Gruijter (1980: 245-261) ได้ให้ความหมายกินความไปถึงบริบททั้งหมดที่ปรากฏอยู่ในคำถามด้วย เช่น แนวทางที่ต้องการให้ผู้สอบตอบ รวมทั้งแนวทางที่กำหนดในการให้คะแนนของผู้ตรวจ การเปลี่ยนแปลงคำชี้แจงในคำถามจะทำให้การตอบและการตรวจเปลี่ยนไป คำถามที่ดีต้องระบุประเด็นที่มุ่งวัดอย่างเด่นชัด การปล่อยให้ผู้ตอบคิดหาแนวทางโดยอิสระมากเท่าใด ความเที่ยงก็จะยิ่งต่ำลงไปด้วย

โดยทฤษฎีแล้วการเพิ่มข้อคำถามหลายข้อในแบบสอบหนึ่ง ๆ ทำให้ความเที่ยงเพิ่มขึ้นแบบสอบความเรียงก็เช่นกัน แต่มีประเด็นที่น่าสนใจอยู่อย่างหนึ่งคือ ในคำถามข้อใดข้อหนึ่ง ข้อที่ถามยาวกว่าจะมีความเที่ยงสูงกว่า แต่การให้เวลาทำข้อสอบมากกว่าไม่ได้ทำให้ความเที่ยงสูงกว่าข้อที่ให้เวลาน้อย (Coffman 1971: 280)

นอกเหนือจากปัจจัยดังกล่าวมานั้น ความหลากหลายของเทคนิควิธีทางสถิติที่ใช้ในการคำนวณค่าความเที่ยงของแบบสอบความเรียงเป็นอีกปัจจัยหนึ่ง งานวิจัยบางเรื่องหาค่าความเที่ยง โดยวิธีหาค่าสหสัมพันธ์แบบผลคูณเพียร์สันระหว่างคะแนนที่ได้รับการตรวจ 2 ชุด วิธีนี้จะได้ค่าประมาณที่สูงกว่าความเป็นจริง เพราะว่าค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของคะแนนสองชุดนั้นจะถูกเหมาว่ามีความเท่าเทียมกัน ซึ่งจริง ๆ แล้วอาจไม่ใช่ นอกจากนั้นมีวิธีของ Ebel (1951) ซึ่งประมาณค่าความเที่ยงโดยวิธีวิเคราะห์ความแปรปรวน วิธีของ Stanley (1962) และ วิธีใช้ทฤษฎีการอ้างอิงสรุป (Cronbach, et al. 1972) ซึ่งทฤษฎีนี้ได้รับการพัฒนาต่อมาโดยนักวัดผลหลายคน เช่น Brennan (1983) Cardinet, et al. (1976, 1981, 1983) เป็นต้น

อย่างไรก็ตาม การจะเลือกแบบสอบความเรียงไปใช้ โดยพิจารณาที่ความเที่ยงนี้ขึ้นอยู่กับสถานการณ์การสอบวัดและการตัดสินใจในการใช้ผลการประเมิน ถ้าเป็นการตัดสินใจเกี่ยวกับการเรียนการสอนในชั้นเรียน ครูอาจใช้แบบสอบหรือข้อคำถามที่มีความเที่ยงต่ำได้ แต่ถ้า

เป็นการสอบคัดเลือก ควรใช้แบบสอบที่มีความเที่ยงสูง อีกประการหนึ่ง เวลาในการสอบและตรวจมีจำกัด ครูจำเป็นต้องเลือกเอาอย่างใดอย่างหนึ่ง ระหว่างคำถามหลาย ๆ ข้อเพื่อให้ได้คะแนนที่มีความเที่ยงสูงกับการถามน้อยข้อแต่สามารถวัดพฤติกรรมการเรียนรู้ขั้นสูงได้ดี (Coffman 1971: 280-281)

จากที่ผู้วิจัยกล่าวมาจะเห็นว่าแบบสอบความเรียงมีพิสัยของความเที่ยงกว้าง มีค่าจากค่อนข้างต่ำคือ .25 จนถึงสูงถึง .98 เนื่องจากธรรมชาติของการวัดแบบนี้ต้องใช้วิจารณญาณส่วนตัวของผู้ตรวจซึ่งถือว่าเป็นผู้เชี่ยวชาญตัดสิน ไม่สามารถทำเฉลยแจงนับถูกผิดได้อย่างสมบูรณ์ เช่นเดียวกับข้อสอบแบบปรนัย นอกจากนั้นยังมีปัจจัยอื่น ๆ อีกมากมายตามที่ได้กล่าวมาแล้วนั้น อย่างไรก็ตาม นักวัดผลได้พยายามหาวิธีการควบคุมแหล่งความคลาดเคลื่อนเหล่านั้น เพื่อทำให้ความเที่ยงเพิ่มขึ้นซึ่งผู้วิจัยจะได้นำเสนอพอสังเขปในหัวข้อต่อไป

3. ข้อเสนอแนะในการตรวจเพื่อให้ได้คะแนนมีความเที่ยง เนื่องจากการให้คะแนนความเรียงมีความเป็นอัตนัยสูง แต่ครูต้องสามารถหาเหตุผลอธิบายสนับสนุนผลการให้คะแนนของตนให้ได้ เมื่อถูกถามไม่ว่าจะโดยนักเรียน ครู ผู้บริหารหรือผู้ปกครองของนักเรียน หาไม่แล้วผลการประเมินก็จะไร้ความหมาย เมื่อครูสามารถให้เหตุผลได้แสดงว่า ครูมีเกณฑ์การให้คะแนน ผลของการที่มีเกณฑ์นี้ เชื่อกันว่าจะทำให้ผลการวัดมีความเที่ยงสูง นักวัดผลหลายคนให้ข้อเสนอแนะในการตรวจเพื่อให้คะแนนมีความเที่ยงไว้หลายระการ ผู้วิจัยขอยกมากล่าวในที่นี้ดังนี้

Bergman (1981: 130-131) แนะให้ปฏิบัติดังนี้

1. เตรียมคำถามให้ตอบสั้น ๆ พร้อมทั้งเตรียมโครงสร้างคำตอบหรือตัวแบบคำตอบและตรวจคำตอบในแต่ละประเด็น ว่ามีหรือไม่มี ถูกหรือไม่ถูก

2. กำหนดเค้าโครงคำตอบที่ต้องการ ช่วยให้ครูมีเกณฑ์ ในการตรวจบางครั้งครอาจจะใช้คำตอบของนักเรียน แทนการเขียนเอง

3. กำหนดประเด็นที่จะให้คะแนนในแต่ละข้อให้เด่นชัดจะทำให้การตรวจมีความยุติธรรม และง่าย

4. ตรวจแยกทีละประเด็น ตัวอย่าง การสะกด ไวยากรณ์ ลีลาการเขียน ความชัดเจนของแนวคิด สี่ประเด็นแรก เป็นประเด็นสำคัญในการประเมินความสามารถใน การ

เขียน ส่วนแนวคิดเป็นประเด็นสำคัญของการประเมินความเรียงวัดเนื้อหา

5. ตรวจทีละข้อจนครบทุกคนทำให้ครูสามารถกำหนดเกณฑ์ที่เหมาะสมในแต่ละข้อได้และสามารถขจัดความคลาดเคลื่อนแบบ halo effects

6. โดยปกติแล้ว ในการตรวจ 2-3 คนแรก ครูมักจะตั้งเกณฑ์ไว้สูง หลังจากนั้นจะค่อย ๆ ผ่อนปรนเกณฑ์ต่ำลง ดังนั้นเมื่อตรวจเสร็จแล้ว ครูควรตรวจ 5 คนแรก และ 5 คนสุดท้ายซ้ำอีกครั้ง

7. ตรวจหลายครั้งหรือใช้ผู้ตรวจหลายคนแล้วใช้คะแนนเฉลี่ย เช่น ถ้าครู 2 คน ตรวจ ก็อาจจะหาค่าเฉลี่ยของทั้งสองคน หรือให้เป็น 2 เกรด เช่น A/B,C/D

Cochran and Weideman (1937 cited by Hopkins and Stanley 1981: 223) ได้เสนอวิธีตรวจความเรียงให้มีความเที่ยง ซึ่งใช้เวลาฝึกฝนเพียง 10 นาที แต่พบว่าค่าความคงเส้นคงวาสูงถึง .80 และ .90 ดังนี้

1. ก่อนตรวจคำตอบใด ให้ศึกษาเนื้อหาในหนังสือเรียนที่เกี่ยวข้องกับคำถาม รวมทั้งสมุดโน้ตย่อที่ใช้เรียนในวิชานั้น

2. แจกแจงประเด็นสำคัญที่ควรจะตอบในแต่ละข้อ กำหนดคะแนนแต่ละประเด็นจนครบ ถือเป็นคะแนนขั้นต่ำ (minimum score) ถ้าผู้สอบคนใดตอบเพิ่มเติมนอกเหนือจากที่เตรียมไว้ ควรให้คะแนนเพิ่มเติมอีกเรียกว่า คะแนนพิเศษ (extra score) คะแนนพิเศษจะแตกต่างกันไปแต่ไม่เกินคะแนนสูงสุดที่กำหนดไว้สำหรับข้อนั้น

3. อ่านคำตอบที่สุ่มมาจำนวนหนึ่งแบบผ่าน ๆ เพื่อจะได้กำหนดคุณภาพของคำตอบที่ผู้ตรวจต้องการ

4. ตรวจแต่ละข้อจนครบทุกคนจึงตรวจข้ออื่น ทำให้เกิดผลดี 2 ประการ คือ ประการแรก ทำให้คะแนนที่ได้จากการเปรียบเทียบ มีความถูกต้องและยุติธรรม ประการที่สอง ทำให้ผู้ตรวจมีตัวแบบในการตรวจเพียงตัวแบบเดียว ทำให้ประหยัดเวลาในการตรวจ และมีความถูกต้องแม่นยำมากขึ้น

5. อ่านคำตอบให้จบไปครั้งหนึ่งก่อน แล้วจึงตรวจพิจารณารายละเอียดอีกครั้ง พยายามจดบันทึกประเด็นที่ตอบผิดที่ตรวจพบแล้วแก้ไขสั้น ๆ จดประเด็นที่ผู้สอบไม่ได้กล่าวถึง รวมทั้งค่าคะแนนในแต่ละประเด็น จะสามารถให้คะแนนตามเกณฑ์ขั้นต่ำได้ถูกต้อง ถ้ามีประเด็นพิเศษเพิ่มเติม ให้รวมกับคะแนนขั้นต่ำ ขั้นนี้อาจใช้ความเรียงตัวแบบ (essay model) เป็นเกณฑ์

เปรียบเทียบ จะทำให้การตรวจมีความเที่ยงเพิ่มขึ้น

6. ควรใช้ผู้ตรวจมากกว่าหนึ่งคน ผู้ตรวจ 2 คน แม้จะตรวจโดยให้อ่านอย่างรวดเร็วก็ยังดีกว่าผู้ตรวจคนเดียว

Linvall and Nitko (1975: 51-52) ได้เสนอแนะวิธีตรวจเพื่อให้มี ความเที่ยงไว้ดังนี้

1. ข้อสอบมีหลายข้อให้ตรวจข้อแรกของทุกคนก่อน แล้วจึงตรวจข้อที่สองในลักษณะเดิม วิธีนี้ทำให้ผู้ตรวจสามารถกำหนดเกณฑ์เดียวกันได้ และสามารถลด halo effect ของคำตอบแต่ละข้อที่จะมีต่อกันได้

2. ถ้าเป็นไปได้ควรตรวจโดยไม่ต้องดูชื่อนักเรียน เป็นการลด halo effect ซึ่งเกิดจากอิทธิพลของความประทับใจที่เกิดจากการรู้จักนักเรียน

3. ตรวจซ้ำคำตอบที่ตรวจในตอนแรก เพื่อตรวจดูว่าได้ใช้เกณฑ์เดียวกันอย่างสม่ำเสมอหรือไม่

Kubiszyn & Borich (1981: 100-101) กล่าวถึงวิธีปรับปรุงความเที่ยงของการตรวจไว้ดังนี้

1. เขียนข้อสอบให้ดี ข้อสอบที่ไม่ดีเป็นแหล่งความคลาดเคลื่อนของความเที่ยงอย่างหนึ่ง เช่น ข้อสอบที่ไม่กำหนดความยาวทำให้การตรวจขาดความเที่ยงได้ โดยทั่วไปความเรียงที่ไม่จำกัดความยามักมีความเที่ยงต่ำกว่าที่จำกัดความยาว เนื่องจากนักเรียนเมื่อยล้าและเกณฑ์การตรวจของผู้ตรวจ จะเปลี่ยนไปเมื่อตรวจหลาย ๆ คำตอบ

2. ใช้ข้อสอบแบบจำกัดคำตอบหลาย ๆ ข้อ แทนคำถามแบบขยายคำตอบเพียงข้อเดียว แต่ถ้ามีความจำเป็น ต้องใช้ข้อสอบแบบขยายคำตอบ ควรปฏิบัติตามข้อ 3.

3. ใช้คู่มือการตรวจที่กำหนดเกณฑ์ไว้ล่วงหน้า ในการประเมินทุกประเภท ปัจจัยสำคัญคือเกณฑ์ ถ้าครูไม่สามารถกำหนดเกณฑ์ที่เกี่ยวข้องได้ก่อน ความเที่ยงการตรวจจะลดลงทันที การที่ครูขาดเกณฑ์การตรวจในแต่ละข้อ อาจก่อให้เกิดผลเสียดังนี้

3.1 หลังจากตรวจไปหลาย ๆ คำตอบ เกณฑ์อาจเปลี่ยน ครูอาจให้คะแนนเข้มงวดขึ้นหรือปล่อยมากขึ้นทั้ง ๆ ที่คุณภาพของคำตอบไม่ได้เปลี่ยน

3.2 ความสามารถที่จะควบคุมเกณฑ์ให้คงที่จะเปลี่ยนไป เนื่องจากความล้าถูกรบกวน หรือ กรอบความคิดเปลี่ยน ฯลฯ

4. ใช้คู่มือการตรวจอย่างคงเส้นคงวา

5. ปิดชื่อนักเรียนก่อนตรวจให้คะแนน

6. ให้คะแนนคำถามเดียวกันจนครบทุกคนแล้วจึงเริ่มข้อใหม่

7. ปิดคะแนนข้อที่ตรวจเสร็จก่อนในขณะที่ตรวจข้อที่เหลือ

8. พยายามตรวจซ้ำอีกครั้ง ถ้าพบว่าคะแนนการตรวจซ้ำต่างจากครั้งก่อน ให้ใช้คะแนนเฉลี่ยแทน

Ebel and Frisbie (1986: 134-135) ได้เสนอแนะการตรวจเพื่อให้มีความเที่ยงไว้ดังนี้

1. ใช้วิธีวิเคราะห์ หรือวิธีประเมินรวม แต่ต้องกำหนดประเด็นที่ต้องการให้คะแนนไว้ล่วงหน้าอย่างชัดเจน ถ้าเป็นวิธีประเมินรวมจะใช้วิธี sorting เป็น 3 กอง ให้มีร้อยละดังนี้ เก่ง 25 % ปานกลาง 50 % และพวกอ่อนอีก 25 % หรือแบ่งเป็น 5 กอง ให้มีร้อยละดังนี้ อ่อน ที่สุด 5 % อ่อน 25 % ปานกลาง 40 % เก่ง 25 % และเก่งที่สุด 5 %

2. ตรวจทีละข้อคำถาม ไม่ควรตรวจให้เสร็จเป็นคน ๆ

3. ปิดชื่อผู้ ตอบไว้ก่อนจะตรวจ

4. ใช้คนตรวจมากกว่า 2 คน โดยอิสระจากกัน

เยาวดี วิบูลย์ศรี (2528) ได้เสนอแนะวิธีตรวจความเรียงเพื่อให้มีความเที่ยงระหว่างผู้ตรวจดังนี้

1. ควรตรวจแยกเป็นรายข้อ แต่ละข้อควรจะได้รับการตรวจสองครั้งจากผู้ตรวจคนเดียวหรือ ผู้ตรวจสองคน คนละครั้ง

2. ในการตรวจแต่ละครั้ง เป็นอิสระต่อกัน

3. ผู้ตรวจไม่ควรดูชื่อผู้สอบ และไม่ควรจะดูผลการตรวจที่ผ่านมาว่า แต่ละข้อหรือแต่ละคนมีการให้คะแนนอย่างไร

4. การตรวจให้คะแนน ควรแยกระบบให้เป็นคะแนนเป็นเกณฑ์ย่อย ๆ

4. ผลการวิจัยที่เกี่ยวข้องกับการสอบแบบความเรียงในประเทศ ผลงานวิจัยเกี่ยวกับแบบสอบความเรียงในประเทศ ค่อนข้างจะมีน้อยอาจเป็นเพราะว่าปัญหาเรื่องการสิ้นเปลืองแรงงานและเสียเวลาในการตรวจมาก เท่าที่มีอยู่เป็นงานวิจัยการสอบประเภทวัดความสามารถ

ทางภาษาที่เรียกว่า "เรียงความ" เป็นส่วนใหญ่ เช่น

อินทร์ ศรีคุณ (2509: 149) ได้ศึกษาองค์ประกอบที่ส่งผลต่อความสามารถในการเขียนเรียงความ ของนักเรียนที่สำเร็จชั้นประถมศึกษาปีที่4 ในโรงเรียนประถมศึกษา 6 โรง ในจังหวัดนครราชสีมา พบว่าความสามารถในการฟัง การอ่าน การเขียน สะกดคำ และการแต่งความ มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ส่วน สมบูรณ์ ชิตพงศ์ (25811: 80) ได้ศึกษาสมรรถภาพสมองที่ส่งผลต่อความสามารถในการเรียงความ พบว่าความสามารถทางภาษาส่งผลต่อเรียงความมากกว่าการหาเหตุผล และความจำ

เปี่ยมศรี นาคพัฒน์ (2500: 48) ได้ศึกษาความคิดเห็นของนักเรียนและครู ชั้นมัธยมศึกษาตอนปลาย เกี่ยวกับการเรียนการสอนเรียงความ พบว่า นักเรียนส่วนหนึ่งชอบเรียงความเพราะได้ระบายความรู้สึกนึกคิด แต่นักเรียนที่ไม่ชอบเรียงความเพราะเขียนได้ไม่ดี ไม่สามารถเขียนตามที่ต้องการได้ ขาดความรู้ในเรื่องที่เขียน รวมทั้งเขียนไม่ได้เพราะครูตั้งหัวข้อแล้วไม่ให้คำแนะนำ ส่วนครูชอบสอนเรียงความเพราะได้อ่านเรื่องราวแปลก ๆ และเห็นว่าเป็นวิชาที่วัดความรู้ด้านภาษาไทยของเด็กได้ดี แต่ครูที่ไม่ชอบสอนเรียงความเพราะ ไม่ถนัดในการแต่ง แนะนำนักเรียนไม่ได้ และไม่มีเวลาตรวจแบบฝึกหัด ซึ่งสอดคล้องกับการศึกษาของ จิตต์นิภา ภักดีชุมพล (2516: 143-148) ที่ศึกษากิจกรรมการเรียนการสอนภาษาไทยชั้นมัธยมศึกษาตอนต้นของโรงเรียนสาธิต ในกรุงเทพมหานคร พบว่าครูไม่ชอบสอนเรียงความ เพราะสอนแล้วงานมาก เสียเวลาในการตรวจแก้ไข ครูร้อยละ 28.57 มีปัญหาเรื่องข้อสอบอัตนัยตรวจยาก ส่วนนักเรียนชอบเรียนเรียงความรองจากวรรณคดีเพราะเห็นว่าเรียนสนุก มีกิจกรรมน่าสนใจและได้ความรู้

กิ่งกาญจน์ สิรสุคนธ์ (2521: 24-40) ได้ศึกษาความสัมพันธ์ระหว่างข้อสอบแบบเลือกตอบ ที่ใช้วัดความสามารถในการเขียนเรียงความ กับแบบสอบความเรียง โดยใช้แบบสอบความเรียงเป็นตัวเกณฑ์และ ใช้แบบสอบรนัยจำนวน 6 ฉบับเป็นตัวพยากรณ์ ซึ่งประกอบด้วยแบบสอบการใช้ภาษา 1 ฉบับ แบบสอบศัพท์สัมพันธ์ 1 ฉบับ และแบบสอบปรนัยที่ใช้วัดความสามารถในการเขียนเรียงความอีก 4 ฉบับ พบว่า ค่าสหสัมพันธ์พหุคูณระหว่างตัวเกณฑ์และตัวพยากรณ์ที่ดีที่สุด เป็น .29 โดยที่แบบสอบการใช้ภาษาเป็นตัวพยากรณ์ที่ดีที่สุด

จิตต์นิภา ศรีไสย์ และ อัจจิมา เทวกุล (2526: 145) ศึกษาปัจจัยที่มีอิทธิพลต่อการเขียนของนักเรียนระดับมัธยมศึกษาตอนปลาย พบว่า ตามทัศนะของครู แหล่งชุมชนมีอิทธิพลมากที่สุด คือบ้าน ห้องสมุดสาธารณะ เวลาที่เหมาะสมในการเขียนเรียงความมากที่สุดคือ เวลา

เช้า และวันหยุด สื่อมวลชนที่มีอิทธิพลต่อการเขียนของนักเรียนตามทัศนะของครูคือ หนังสือพิมพ์ รองลงมาคือ โทรทัศน์ ภาพโฆษณา และภาพยนตร์ ปัญหาที่ครูพบมากที่สุดในการเขียนเรียงความ คือ คือการสะกดคำผิด ลำดับเรื่องวกวน ตามความเห็นของนักเรียน สาเหตุที่นักเรียนเขียนเรียงความได้ดี เนื่องจากนักเรียนรักการอ่าน ชอบการเขียน รู้จักค้นคว้า และนำความรู้มาประกอบการฝึกทักษะการเขียนอย่างสม่ำเสมอ

วัลลภา เทพหัสดิน ณ อยุธยา (2526: 119) ได้ศึกษาสัมฤทธิผลทางการเขียนของนิสิต 269 คน จากมหาวิทยาลัยธรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เกษตรศาสตร์ มหิดล และ ศิลปากร พบว่า นิสิตนักศึกษาเขียนโดยยึดตนเองเป็นศูนย์กลาง ใช้สำนวนง่าย ๆ ขาดการอ้างอิง ใช้ภาษาพูดแทนภาษาเขียน บกพร่องเรื่องเครื่องหมายวรรคตอน การย่อหน้า และ การเว้นวรรค ระดับผลสัมฤทธิ์ค่อนข้างอ่อน การให้คะแนนของอาจารย์ไม่แตกต่างกันมากนักในกลุ่มที่ได้คะแนนสูง แต่ในกลุ่มต่ำ คะแนนมีความแปรปรวนมาก ผู้ที่ได้คะแนนสูงมักมีสำนวนดี มีเหตุผลอ้างอิง ลายมือสะอาด กระจ่าง น่าอ่าน

วัลลภา เทพหัสดิน ณ อยุธยา และ คณะ (2528: 254-257) ศึกษาการเขียนของนักเรียนมัธยมศึกษาตอนปลาย ตามทัศนะของอาจารย์และนักเรียน จากการศึกษา กลุ่มตัวอย่าง นักเรียนชั้นมัธยมศึกษาปีที่ 5 จำนวน 1,500 คน และ ครูที่สอนวิชาภาษาไทย จำนวน 200 คน จากทั่วประเทศ พบว่า ปัจจัยที่ส่งเสริมการเขียนของนักเรียน ได้แก่การฝึกฝนทักษะตนเองอย่างสม่ำเสมอทั้งในห้องเรียนและที่บ้าน การมีเครื่องอำนวยความสะดวกในการเขียน เช่น โต๊ะ เก้าอี้ หนังสือพิมพ์ วารสาร การเอาใจใส่ของผู้ปกครอง หลักสูตรปัจจุบันทำให้นักเรียนได้ฝึกทักษะการเขียนน้อยลง นอกจากนั้นอิทธิพลของการใช้ข้อสอบแบบเลือกตอบมีผลทำให้เด็กและครูสนใจการเรียนการสอนด้านการเขียนน้อยลง การวัดและประเมินผลยังใช้แบบดั้งเดิม คือครูเขียนวิจารณ์ไว้ตอนท้ายงานเขียนของนักเรียน ไม่มีการใช้แบบฟอร์มการประเมิน

จากงานวิจัยเกี่ยวกับแบบสอบความเรียงในประเทศที่กล่าวมา ส่วนใหญ่เป็นการศึกษาทัศนะของครูและนักเรียนต่อการวัดด้วยแบบสอบความเรียง ซึ่งพบว่าครูและนักเรียนมีทั้งชอบและไม่ชอบ ครูมักให้เหตุผล ว่าตรวจยาก เสียเวลา ขาดทักษะในการแนะนำนักเรียน นักเรียนชอบเพราะได้แสดงออกซึ่งความรู้ความคิด มีบางเรื่องที่วิจัยโดยมุ่งใช้คะแนนการสอบความเรียงเป็นเกณฑ์ในการศึกษาความตรงของแบบสอบปรนัย แต่ที่ศึกษาความเที่ยงของแบบสอบความเรียงโดย

ตรงยังไม่มี ด้วยเหตุที่แบบสอบความเรียงเป็นเครื่องมือที่มีคุณค่า และคุณสมบัติที่จำเป็นของเครื่องมือวัดทุกชนิด คือความเที่ยง หรือความสามารถในการอ้างอิงสรุป จากจำนวนเงื่อนไขสุ่มไปยังเอกภพของเงื่อนไขนั้น ดังนั้นผู้วิจัยจึงได้ศึกษา โดยนำข้อเสนอแนะจากนักวัดผลแต่ละคนที่มุ่งขจัดแหล่งความคลาดเคลื่อนที่มีอิทธิพลต่อความเที่ยงของการสอบแบบความเรียง อันมีสาเหตุมาจาก ข้อคำถาม ผู้ตรวจ วิธีตรวจ และอิทธิพลของ halo effect ข้อชี้แนะส่วนใหญ่สอดคล้องกัน แตกต่างกันบ้างก็เฉพาะประเด็นปลีกย่อยเท่านั้น ในการวิจัยครั้งนี้ ผู้วิจัยจึงได้นำข้อเสนอแนะเหล่านี้ ไปใช้ในการฝึกอบรมผู้ตรวจ เพื่อศึกษาให้แน่ชัดว่าวิธีการดังกล่าว มีผลต่อความเที่ยงของการสอบแบบความเรียงมากน้อยเพียงใดเพื่อจะได้นำผลการวิจัยไปปรับปรุงการวัดแบบความเรียงต่อไป