การประยุกต์ใช้ทฤษฎีการอ้างอิงสรุป เพื่อให้คะแนนการสอบเรียงความ และการพัฒนาโปรแกรมคอมพิวเตอร์ใช้งาน

โดย

รศ. ดร. สุพัตเก์ สุกมลสันต์

สถาบันภาษา
จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2540

การประยุกต์ใช้ทฤษฎีการอ้างอิงสรุป เพื่อให้คะแนนการสอบเรียงความ และการพัฒนาโปรแกรมคอมพิวเตอร์ใช้งาน

โดย

รศ. ดร. สุพัฒน์ สุกมลสันต์

สถาบันภาษา
จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2540

คำนำและกิตติกรรมประกาศ

งานวิจัยเรื่องนี้เกิดขึ้นเพราะผู้วิจัยมีความสนใจที่จะติดตามพัฒนาการใหม่ ๆ ทางด้านการทดสอบ วัดและประเมินผลทางการศึกษาและการเรียนการสอนภาษาอังกฤษ เนื่องจากงานวิจัยเกี่ยวกับการประยุกต์ใช้ทฤษฎีการอ้างอิงสรุป (Generalizability Theory) เป็นเรื่องที่ใหม่มากสำหรับวงการวิจัยในประเทศไทย จึงเป็นการยากที่จะหาเอกสารและวรรณกรรมที่เกี่ยวข้องในประเทศ และเท่าที่สืบค้นเอกสารต่าง ๆ ขณะนี้ ผู้วิจัยยังไม่พบว่ามีงานวิจัยในประเทศที่ใช้ทฤษฎีนี้ในวงการเรียนการสอนภาษาอังกฤษ ด้วยเหตุนี้จึงไม่มีผลการวิจัยที่เกี่ยวข้องโดยตรง

อนึ่ง สำหรับท่านผู้อ่านที่อาจจะไม่คุ้นเคยกับทฤษฎีการอ้างอิงสรุป ผู้วิจัยขอแนะนำให้ทำความเข้าใจกับทฤษฎีนี้ก่อนในบทที่ 2 ก่อนที่จะอ่านบทอื่น ๆ ของงานวิจัยนี้

ในการทำการวิจัยเรื่องนี้ ผู้วิจัยได้รับความช่วยเหลืออย่างดียิ่งหลายอย่างจากบุคคลหลายฝ่ายที่จะต้องขอขอบคุณเป็นอย่างยิ่งไว้ ณ ที่นี้ด้วย เช่น

ขอขอบคุณคณะกรรมการสร้างและพัฒนาแบบทดสอบ CU-TEP ทุกท่านเป็นอย่างมากที่มีส่วนช่วยให้แบบทดสอบดังกล่าวมีคุณภาพเป็นที่เชื่อถือได้ทั่วไปในขณะนี้ โดยเฉพาะอย่างยิ่ง ผศ. ศิริพร พงษ์สุรพิพัฒน์ และ รศ. ดร. อัจฉรา วงศ์โสธรที่มีส่วนพัฒนางานดังกล่าวมาก และผู้วิจัยได้นำข้อทดสอบส่วนหนึ่งมาใช้ในงานวิจัยนี้

ขอขอบคุณ รศ. อัญชิการ์ โรงสะอาด ผศ. พัชรี ชินธรรมมิตร ผศ. ลลิตา หมอกพริ้ง อาจารย์ระรองแพ่งสภา อาจารย์นิรดา สีมากุล และอาจารย์สุกัญญา เป็นอย่างยิ่งที่ช่วยอนุเคราะห์ตรวจข้อทดสอบให้สำหรับงานวิจัยนี้โดยเฉพาะ

ขอขอบคุณคณะกรรมการวิจัยของสถาบันภาษาทุกท่านเป็นอย่างมากที่ช่วยกรุณาให้คำแนะนำหลายอย่าง เพื่อให้งานวิจัยนี้มีความถูกต้องและสมบูรณ์ยิ่งขึ้น

ขอขอบคุณคุณจรัสศรี โพธิวัตฉุธรรมมากที่ช่วยพิมพ์ต้นฉบับงานวิจัยนี้ให้ได้อย่างถูกใจ และท้ายสุดขอขอบคุณคุณกิ่งกมน ทวิชชาติวิทยากุลมากที่ช่วยประสานงานหลายอย่างจนทำให้งานวิจัยนี้สำเร็จได้ในที่สุด

สุพัฒน์ สุกมลสันต์
11 มีนาคม 2540

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ที่สำคัญเพื่อประยุกต์แนวความคิดของทฤษฎีการอ้างอิงสรุปมาใช้ในการตวรจให้คะแนนการสอบเรียงความในหลาย ๆเงื่อนไขและเพื่อสร้างโปรแกรมคอมพิวเตอร์สำหรับใช้ในการคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุปในเงื่อนไขต่าง ๆของการทดสอบ ปัจจัยประกอบ (facets) ของการศึกษาครั้งนี้ได้แก่ ผู้สอบจำนวน 200 คนที่เป็นแบบสุ่ม ข้อทดสอบ 3 ข้อที่เป็นแบบสุ่ม และผู้ตรวจข้อสอบ 2 ประเภทที่เป็นแบบคงที่ คือผู้ที่มีประสบการณ์ตรวจข้อทดสอบที่ใช้ในการวิจัยโดยตรงและผู้ที่มีประสบการณ์ตรวจทางอ้อม จึงทำให้เอกภพของสิ่งสังเกตที่ยอมรับได้ (universe of admissible observations) มีขนาด 200 x 3 x 2 = 1,200 เงื่อนไข และเอกภพของการอ้างอิงสรุป (universe of generalizability) ครั้งนี้มี 6 เงื่อนไข

ผลของการศึกษาสามารถสรุปได้ดังนี้

ก. ด้านการประยุกต์ใช้แนวคิดทฤษฎีการอ้างอิงสรุปเพื่อการตรวจข้อทดสอบเรียงความ

1. ในกรณีที่กำหนดให้ผู้ตรวจข้อทดสอบทุกประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน หรือ [P×I×R]

1.1 เมื่อ R = fixed และ P,I = random ปรากฏว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง มีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญ กล่าวคือ $\Phi_1(\lambda_0) = 0.9696 > \Phi_2(\lambda_0) = 0.9596$ และ $\Sigma\rho_1^2 = 0.84055 > \Sigma\rho_2^2 = 0.72610$ ($p = 0.05$)

1.2 เมื่อ I,P,R = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญ เช่นเดียวกันกับข้อ 1.1 กล่าวคือ $\Phi_1(\lambda_0) = 0.9211 > \Phi_2(\lambda_0) = 0.9189$ และ $\Sigma\rho_1^2 = 0.66086 > \Sigma\rho_2^2 = 0.52018$ ($p = 0.05$)

2. ในกรณีที่กำหนดให้ผู้ตรวจข้อทดสอบบางประเภทตรวจข้อทดสอบบางข้อของผู้สอบทุกคน หรือ [Px(R:I)]

2.1 เมื่อ R = fixed และ P,I = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง มีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญ กล่าวคือ $\Phi_1(\lambda_0) = 0.9696 > \Phi_2(\lambda_0) = 0.9596$ และ $\Sigma\rho_1^2 = 0.84055 > \Sigma\rho_2^2 = 0.72610$ ($p = 0.05$)

2.2 เมื่อ I,P,R = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง มีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญเช่นเดียวกับข้อ 2.1 กล่าวคือ $\Phi_1(\lambda_0) = 0.9211 > \Phi_2(\lambda_0) = 0.9192$ และ $\Sigma\rho_1^2 = 0.66086 > \Sigma\rho_2^2 = 0.52018$ ($p = 0.05$)

3. เมื่อกำหนดให้ผู้ตรวจบางประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบบางคน หรือ [Ix(P:R)]

3.1 เมื่อ R = fixed และ P,I = random ปรากฏว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยเท่ากับค่าดังกล่าวของผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อมและมีสัมประสิทธิ์การอ้างอิงสรุปต่ำกว่าอย่างมีนัยสำคัญ กล่าวคือ $\Phi_1(\lambda_0)=0.9999 = \Phi_2(\lambda_0)=0.9999$ และ $\Sigma\rho_1^2=0.99971 < \Sigma\rho_2^2=0.99991$ $(p=0.05)$

3.2 เมื่อ I,P,R = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยต่ำกว่าค่าดังกล่าวของผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ทางอ้อมแต่มีสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าอย่างมีนัยสำคัญ กล่าวคือ $\Phi_1(\lambda_0)=0.8153 < \Phi_2(\lambda_0)=0.8207$ และ $\Sigma\rho_1^2=0.44340 > \Sigma\rho_2^2=0.31235$ $(p=0.05)$

ดังนั้น จะเห็นได้ว่าเงื่อนไขที่ดีที่สุดสำหรับการตรวจข้อสอบแบบเรียงความ คือเงื่อนไขที่ 3.1

ข. ด้านการสร้างโปรแกรมคอมพิวเตอร์เพื่อใช้งาน

ปรากฏว่า โปรแกรมที่สร้างขึ้นสามารถคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุป ($\Sigma\rho^2$) และค่าดัชนีความเชื่อถือ [$\Phi(\lambda_0)$] สำหรับเงื่อนไขต่างๆ ได้อย่างถูกต้องเมื่อเปรียบเทียบกับค่าต่างๆ ดังกล่าว จากข้อมูลและผลลัพธ์ของการคำนวณจากหนังสือที่ใช้ในการอ้างอิง

ABSTRACT

The main purposes of this study were to apply the concepts of the Generalizability Theory for marking an essay test and to write a computer program for calculating the generalizability coefficients of a test when it was administered under various conditions. The facets of the test were 200 testees (random), 3 essay items (random) and 2 types of markers (fixed). They were those who had direct experience in marking the test for this study and those who had indirect experience. Therefore, the universe of admissible observations was 1,200 conditions. As for the universe of generalizability, there were 6 conditions. The data of this study were analyzed by the computer program written by the author.

The findings can be summarized as follows:

A. The Applications of Generalizability Theory

1. If all the markers mark all the items of all the testees or [P x I x R]:

1.1. When R = fixed and P, I = random, it was found that those markers who had direct experience in marking the test could mark the test, on average, with a significant higher index of dependability and a higher generalizability coefficient than those who had indirect experience in marking the test. Statistically, $\Phi_1(\lambda_0) = 0.9696 > \Phi_2(\lambda_0) = 0.9596$ and $\Sigma\rho_1^2 = 0.84055 > \Sigma\rho_2^2 = 0.72610$ ($p = 0.05$)

1.2. When R, P, I = random, it was found that those markers who had direct experience in marking the test could mark the test, on average, with a significant higher index of dependability and a higher generalizability coefficient than those who had indirect experience in marking the test. Statistically, $\Phi_1(\lambda_0) = 0.9211 > \Phi_2(\lambda_0) = 0.9189$ and $\Sigma\rho_1^2 = 0.66086 > \Sigma\rho_2^2 = 0.52018$ ($p = 0.05$)

2. If some types of the markers mark some items of the test of all the testees or [P x (R:I)]:

2.1 When R = fixed and P, I = random, it was found that those markers who had direct experience in marking the test could mark the test, on average, with a significant higher index of dependability and a higher generalizability coefficient than those who had indirect experience in marking the test. Statistically, $\Phi_1(\lambda_0) = 0.9696 > \Phi_2(\lambda_0) = 0.9596$ and $\Sigma\rho_1^2 = 0.84055 > \Sigma\rho_2^2 = 0.72610$ ($p = 0.05$)

2.2. When R, P, I = random, it was found that those markers who had direct experience in marking the test could mark the test, on average, with a significant higher index of dependability and a higher generalizability coefficient than those who had indirect experience in marking the test. Statistically, $\Phi_1(\lambda_0) = 0.9211 > \Phi_2(\lambda_0) = 0.9189$ and $\Sigma\rho_1^2 = 0.66086 > \Sigma\rho_2^2 = 0.52018$ $(p = 0.05)$

3. If some types of the markers mark all items of the test of some of the testees or [I x (P:R)]:

3.1 When R = fixed and P, I = random, it was found that those markers who had direct experience in marking the test, marked the test, on average, with the same index of dependability as those who had indirect experience and with a lower generalizability coefficient than the other group of markers. Statistically, $\Phi_1(\lambda_0) = 0.9999 = \Phi_2(\lambda_0) = 0.9999$ and $\Sigma\rho_1^2 = 0.99971 < \Sigma\rho_2^2 = 0.99991$ $(p = 0.05)$

3.2. When R, P, I = random, it was found that those markers who had direct experience in marking the test marked the test, on average, with a significant lower index of dependability but a higher generalizability coefficient than those who had indirect experience in marking the test. Statistically, $\Phi_1(\lambda_0) = 0.8153 < \Phi_2(\lambda_0) = 0.8207$ and $\Sigma\rho_1^2 = 0.44340 > \Sigma\rho_2^2 = 0.31235$ $(p = 0.05)$

Therefore, the most appropriate model for marking an essay test was no. 3.1.

B. Computer Program Verification

When compared with the results from referenced textbooks, the new program could yield the same results in various conditions in producing the generalizability coefficients and indices of dependability.

สารบัญเรื่อง

บทที่ 1
บทนำ

ความสำคัญและที่มาของปัญหาที่ทำการวิจัย

แบบทดสอบเรียงความเป็นเครื่องมือที่มีความสำคัญอย่างยิ่งอย่างหนึ่งต่อการวัดและประเมินผลการศึกษาเพราะเป็นแบบทดสอบที่ให้อิสระในการแสดงความคิด ส่งเสริมการจัดระเบียบความคิด การประยุกต์ความรู้กับสภาพการณ์ใหม่ และความคิดสร้างสรรค์ รวมทั้งสามารถวัดผลการเรียนรู้ตามจุดประสงค์ด้านพุทธิปริเขต (cognitive domain) ในระดับสูงได้อีกด้วย ได้แก่ การวิเคราะห์ การสังเคราะห์ และการประเมินค่า (Thorndike and Hargen, 1977 : 263) นอกจากคุณภาพในแง่ของการวัดผลการเรียนที่ซับซ้อนได้เหมาะสมแล้ว แบบทดสอบเรียงความยังส่งเสริมการเรียนการสอนให้มีประสิทธิภาพมากยิ่งขึ้น สามารถสร้างและดำเนินการสอบได้ง่าย ไม่สิ้นเปลืองวัสดุอุปกรณ์ในการจัดพิมพ์ และสามารถทดสอบได้ทุกเวลาที่ทำการสอน เช่นเพียงแต่เขียนข้อคำถามบนกระดานดำนักเรียนทั้งห้องก็สามารถใช้ร่วมกันได้ทันที (Coffman, 1971 : 104) และยังเป็นเครื่องมือที่มีความตรงเชิงประจักษ์ (face validity) ในด้านการวัดความสามารถในการใช้ภาษาด้านการเขียนได้เป็นอย่างดี (Hopkins and Stanley 1981 : 225)

จากผลการวิจัยเปรียบเทียบกับแบบทดสอบปรนัยพบว่า ผู้เรียนที่เรียนโดยใช้แบบทดสอบเรียงความเป็นเครื่องมือในการวัดผลมีผลสัมฤทธิ์ทางการเรียนสูงกว่าผู้เรียนโดยใช้การวัดผลแบบปรนัย (Bergman 1981 : 127) Coffman (1971 : 286) กล่าวว่าครูควรใช้แบบทดสอบเรียงความในการประเมินผลระหว่างการเรียนการสอนให้มาก เพราะนอกจากจะเปิดโอกาสให้นักเรียนได้ฝึกฝนทักษะการเขียนและบูรณาการความรู้ที่ได้เรียนมาแล้ว ถ้าหากมีผู้สอบไม่มาก การตรวจเรียงความอาจใช้เวลาน้อยกว่าการสร้างข้อทดสอบปรนัยที่ดี

แต่อย่างไรก็ตาม แม้ว่าแบบทดสอบเรียงความจะมีจุดเด่นหลายประการดังได้กล่าวมาแล้ว แต่แบบทดสอบเรียงความก็มีจุดอ่อนหรือปัญหาในด้านการวัดผลที่สำคัญอยู่ 2 ประการ คือ ปัญหาความเที่ยง (reliability) และการสิ้นเปลืองเวลาและแรงงานในการตรวจ กล่าวคือแบบทดสอบเรียงความมักจะมีความเที่ยงต่ำและเสียเวลาในการตรวจมาก จากการสำรวจงานวิจัย Coffman (1971 : 277) พบว่า งานวิจัยบางเรื่องแบบทดสอบเรียงความมีค่าความเที่ยง 0.35 เท่านั้น Godshalk, Swineford and Coffman (1966 : 39-40) พบว่า ค่าความเที่ยงของข้อทดสอบเรียงความ 1 ข้อ เมื่อใช้ผู้ตรวจ 5 คนเป็น 0.40 แต่ถ้าใช้ 5 ข้อ และตรวจคนเดียวจะมีค่าเป็น 0.25 การที่แบบทดสอบเรียงความมีค่าความเที่ยงต่ำ Ebel and Frisbie (1986 : 129) ชี้ว่าเกิดจากสาเหตุสำคัญ 3 ประการ คือ

1) คำถามมีน้อยไม่ครอบคลุมเนื้อหา

2) ความไม่จำเพาะเจาะจงของคำถาม และ

3) ความเป็นอัตนัยของการตรวจที่เกิดจากจากตัวแปรต่าง ๆ ที่ไม่เกี่ยวข้องกับสิ่งที่มุ่งทดสอบ

จากผลการวิจัยพบว่า คำตอบเรียงความประเภทวัดความสามารถในการเขียนหรือเรียงความที่มีคุณภาพปานกลาง ถ้าตรวจตามหลังเรียงความที่มีคุณภาพต่างกันจะได้คะแนนต่างกัน กล่าวคือ เมื่อตรวจตามหลังเรียงความที่มีคุณภาพต่ำมักจะได้คะแนนสูงกว่าเมื่อตรวจตามหลังชุดที่มีคุณภาพสูง (Hughes, Keeling & Tuck 1984 : 277-281) นอกจากนี้ ยังมีงานวิจัยที่ชี้ให้เห็นถึงอิทธิพลของตัวแปรอื่น ๆ ที่มีผลต่อค่าความเที่ยงของแบบทดสอบเรียงความอีก เช่น การที่ผู้ตรวจรู้จักประวัติการเรียนของผู้ตอบ ลายมือ ความเรียบร้อย การสะกดการันต์ และความบกพร่องในด้านไวยากรณ์ เป็นต้น (Chase 1983 : 293-297)

ปัญหาความเที่ยงของแบบทดสอบเรียงความที่ต่ำนี้ มีนักวัดผลการศึกษาได้ให้ข้อเสนอแนะเพื่อให้ผลการวัดมีความแม่นยำไว้หลายประการ เช่น De Gruijter (1980 : 221) กล่าวว่า การควบคุมแหล่งความคลาดเคลื่อนไว้ล่วงหน้าจะทำให้ผลการวัดมีความเที่ยงมากขึ้น ในการวัดผลแบบเรียงความ นั้น ปัจจัยประกอบ (facet) ที่สำคัญจำเป็นต้องควบคุม คือ ข้อทดสอบ และผู้ตรวจ ปัจจัยประกอบทั้งสองนี้เกี่ยวข้องซึ่งกันและกัน ข้อทดสอบเรียงความที่ไม่ดีไม่สามารถยกระดับให้ดีขึ้นได้ด้วยวิธีการตรวจใด ๆ ไม่ว่าวิธีตรวจนั้นจะกำหนดเกณฑ์การตรวจไว้เด่นชัดเพียงใดก็ตาม ในทางตรงกันข้าม ถ้าข้อทดสอบที่สร้างอย่างดีแต่ใช้วิธีตรวจไม่เหมาะสม ข้อทดสอบนั้นก็จะไร้คุณค่าเช่นกัน นอกจากนี้ Mehrens and Lehmann (1972 : 228) ได้เสนอแนะว่า เพื่อให้การตรวจมีความเที่ยงเพิ่มขึ้นผู้ตรวจควรจะต้องปฏิบัติตามหลักการต่อไปนี้อย่างเคร่งครัด คือ

1) ใช้วิธีการตรวจที่เหมาะสมที่สามารถขจัดอคติได้

2) การตรวจต้องมุ่งเฉพาะประเด็นที่เกี่ยวข้องกับผลการเรียนที่มุ่งทดสอบเท่านั้น

3) ระมัดระวังอย่าให้นิสัยส่วนตัวของผู้ตรวจมีอิทธิพลต่อคะแนน

4) ต้องใช้เกณฑ์การตรวจกับนักเรียนทุกๆ คนอย่างคงเส้นคงวา

นอกจากหลักการดังกล่าวแล้ว นักวัดผลคนอื่น ๆ ได้ให้ข้อเสนอแนะคล้ายคลึงกันว่าในการตรวจควรตรวจทีละข้อจนครบทุก ๆ คน ปกปิดชื่อของนักเรียนเอาไว้อย่าให้รู้ว่ากำลังตรวจคำตอบของใครและถ้าทำได้ควรตรวจหลายครั้ง หรือตรวจหลายคน (multiple rating) คือให้ผู้ตรวจคนเดียวตรวจหลายครั้งหรือในการตรวจครั้งหนึ่งใช้ผู้ตรวจหลายคน แล้วใช้คะแนนเฉลี่ยหรือคะแนนรวมแทนคะแนนความสามารถของผู้สอบ วิธีการต่าง ๆ ดังกล่าวแล้วจะให้ผลการวัดที่มีความเที่ยงสูงขึ้นได้ (Bergman,1981 : 42-70)

อนึ่ง วิธีตรวจข้อทดสอบเรียงความที่นักวัดผลกล่าวถึงสามารถจำแนกได้เป็น 2 วิธี (Mehrens and Lehmann 1984 : 113-116; Glover, Bruningand Filbeck 1982 : 420) คือ

1. การตรวจแบบวิเคราะห์ (Analytic method) หรือการตรวจครั้งละประเด็น (Point method) คือการตรวจแยกทีละประเด็น ให้คะแนนสูงสุดในแต่ละประเด็นเท่ากับค่าน้ำหนักคะแนนที่กำหนดไว้

2. การตรวจแบบสังเคราะห์ (holistic method) หรือการตรวจประเมินรวม (Global method) คือการตรวจที่นำเอาประเด็นต่าง ๆ ที่เกี่ยวข้องกับสิ่งที่มุ่งทดสอบมารวมกันแล้วให้คะแนนโดยอาศัยความประทับใจจากการอ่านอย่างรวดเร็ว

การตรวจแต่ละวิธีเหมาะสมกับสถานการณ์การประเมินและเนื้อหาวิชาแตกต่างกัน เช่น วิธีแรก เหมาะกับการตรวจเรียงความประเภทวัดเนื้อหา และในสถานการณ์การประเมินผลการเรียนซึ่งมีผู้สอบไม่มากนัก ข้อเสียของวิธีนี้คือเสียเวลาตรวจมากเพราะต้องตรวจทีละประเด็น และผลการตรวจประเด็นต้น ๆ มักก่อให้เกิดความประทับใจแก่ผู้ตรวจและมีผลกระทบต่อการให้คะแนนของประเด็นหลัง ๆ เช่น ประเด็นแรกได้คะแนนสูงมักทำให้ประเด็นหลัง ๆ ได้คะแนนสูงด้วย ความคลาดเคลื่อนนี้เรียกว่า"ผลกระทบทรงกรด" หรือ halo effect (Coffman 1971 : 292-293) นอกจากนั้น การแยกประเด็นย่อยมากๆ อาจวัดคุณลักษณะที่ไม่ตรงกับจุดมุ่งหมายการเรียนการสอนที่ต้องการได้ (Wiseman 1949 อ้างถึงใน Coffman 1971 : 293) การตรวจวิธีที่ 2 สามารถกระทำได้รวดเร็วจึงเหมาะกับการตรวจเรียงความประเภทวัดความสามารถในการเขียน และในสถานการณ์ที่มีผู้สอบมากๆ ปกติแล้วการตรวจวิธีประเมินรวมมักให้ค่าความเที่ยงต่ำกว่าการตรวจวิธีวิเคราะห์ (De Gruijter 1980 : 247) ดังนั้น การจัดกลุ่มคุณภาพของคำตอบ หรือใช้คำตอบต้นแบบ (essay model) เป็นตัวอย่างจะช่วยทำให้ความเที่ยงของการตรวจวิธีประเมินรวมมีค่าสูงขึ้นได้ (Ebel and Frishbie 1986 : 134)

อนึ่ง วิธีหนึ่งที่จะทำให้ผลการตรวจแบบทดสอบเรียงความมีความเที่ยงมากขึ้นอาจทำได้โดยการเพิ่มจำนวนข้อทดสอบให้มีมากขึ้น หรือเพิ่มจำนวนผู้ตรวจมากขึ้น แต่การเพิ่มข้อทดสอบเรียงความหรือการเพิ่มผู้ตรวจ เป็นอีกปัญหาหนึ่งที่จะต้องพิจารณาให้รอบคอบทั้งในแง่ทฤษฎีและการปฏิบัติ เพราะการสอบแบบเรียงความผู้ตอบต้องใช้เวลาในการเขียนมากและครูก็ต้องใช้เวลาในการตรวจมากเช่นกัน ในการเพิ่มข้อทดสอบหรือผู้ตรวจเพื่อเพิ่มค่าความเที่ยงนั้นทฤษฎีการทดสอบแบบประเพณีนิยม (Classical Test Theory) อธิบายด้วยสูตรของ Spearman-Brown ซึ่งเหมาะกับการวัดที่คำนึงแต่ข้อทดสอบเป็นสำคัญเพียงด้านเดียว แต่การสอบแบบเรียงความมีปัจจัยประกอบ (facets) ที่สำคัญหลายอย่าง เช่น ผู้ตรวจ และผู้สอบ การเพิ่มข้อทดสอบหรือผู้ตรวจจึงไม่สามารถจะใช้สูตร Spearman-Brown คำนวณหาค่าความเที่ยงได้ (Cronbach and Others 1972 : 172) ในทางปฏิบัติการเพิ่มข้อทดสอบเรียงความแม้เพียง 1 ข้อ ครูต้องสิ้นเปลืองเวลาและแรงงานในการตรวจมาก ทำให้กิจกรรมการวัดและประเมินผลเป็นเรื่องน่าเบื่อหน่าย และในบางครั้งจำนวนครูมีน้อยไม่เพียงพอกับระดับความแม่นยำที่ต้องการ

ในปัจจุบัน นี้มีทฤษฎีใหม่ด้านการทดสอบและประเมินผลชื่อว่าทฤษฎีการอ้างอิงสรุป (Generalizability Theory) ซึ่งกลุ่มนักวัดผลนำโดย Cronbach (Cronbach and Others 1972) ได้พัฒนาขึ้น และต่อมามีนักวัดผลอีกหลายท่านได้ช่วยกันพัฒนาแนวคิดของทฤษฎีนี้ให้สมบูรณ์มากยิ่งขึ้น ทฤษฎีนี้มุ่งขยายความคิดของทฤษฎีความเที่ยงแบบประเพณีนิยมให้ชัดเจนยิ่งขึ้น Vann der Kamp (1976 : 173-174) กล่าวว่า ทฤษฎีการอ้างอิงสรุปสามารถใช้อธิบายความเที่ยงได้ทุกสถานการณ์ โดยเฉพาะอย่างยิ่งการสอบแบบเรียงความ ทั้งนี้เนื่องจากทฤษฎีนี้ไม่ได้กำหนดข้อตกลงเบื้องต้นเกี่ยวกับคุณสมบัติความเท่าเทียมของข้อทดสอบ ผู้ตรวจแต่ละคนไม่จำเป็นต้องมีคุณสมบัติความเท่าเทียมกัน และยอมรับว่าความแปรปรวนของผลการวัดเกิดขึ้นได้จากปัจจัยประกอบต่าง ๆ หลายแหล่งด้วยกัน จึงพยายามประมาณค่าของความแปรปรวนจากแหล่งต่าง ๆ ที่เกี่ยวข้องโดยอาศัยการวิเคราะห์ความแปรปรวนจากปัจจัยประกอบหลายแหล่ง (multifacet analysis of variance) การทราบค่าความแปรปรวนจากแหล่งต่าง ๆ นี้ จะช่วยให้นักวัดผลสามารถตรวจสอบถึงแหล่งความแปรปรวนที่ไม่พึงประสงค์ที่มีผลต่อความเที่ยงของการทดสอบได้ ทฤษฎีนี้สามารถกำหนดมโนทัศน์ (concept) เกี่ยวกับเอกภพหรือปริเขต (universe or domain) เป็นกรอบอ้างอิงได้อย่างชัดเจน เช่น ผู้ตรวจทั้งหมดที่ผู้วัดต้องการจะอ้างอิงถึงหมายถึงใคร มีจำนวนเท่าใด และข้อทดสอบที่ต้องการนำมาสอบมีขอบเขตเนื้อหากว้างเพียงใด การที่ถามว่าการตรวจเรียงความมีค่าความเที่ยงหรือความคล้องกันระหว่างผู้ตรวจ (rater agreement) เพียงใด หมายถึงถามว่าผู้วัดหรือผู้ประเมินผลสามารถจะอ้างอิงสรุป (generalize) ผลการวัดจากคะแนนที่ได้ชุดหนึ่งไปยังคะแนนจากการตรวจชุดอื่น ๆ ได้ดีเพียงใดนั่นเอง และค่าความเที่ยงจากการวัดของคะแนนแต่ละชุดถือว่าเป็นตัวอย่างหนึ่งของเอกภพของค่าการวัดที่เป็นไปได้ทั้งหมด ความเที่ยงตามทฤษฎีนี้ หมายถึงความแม่นยำในการอ้างอิงสรุปไปยังกลุ่ม (class) ทฤษฎีนี้เชื่อว่า คะแนนจริง (true score) คือคะแนนเอกภพ (universe score) ซึ่งได้แก่ค่าเฉลี่ยของคะแนนสังเกต (observed score) ที่เป็นไปได้ทั้งหมดภายในเอกภพที่ผู้วัดผลสนใจ และคะแนนเอกภพของผู้สอบแต่ละคนมีได้หลายค่าเช่นเดียวกับค่าสัมประสิทธิ์การอ้างอิงสรุป (generalizability coefficient) และค่าดัชนีความเชื่อถือ (index of dependability) ทั้งนี้ขึ้นอยู่กับเอกภพที่ผู้วัดหรือประเมินผลสนใจในการอ้างอิงสรุป รวมทั้งระดับของจุดตัดของคะแนน (cutting score) ที่ใช้ในการอ้างอิงความเชื่อมั่นด้วย

นอกจากนี้ การวิเคราะห์ความแปรปรวนจากปัจจัยประกอบหลายแหล่ง ยังมีประโยชน์อีกหลายอย่าง เช่น

1) ช่วยขจัดความกำกวมที่มีและซ่อนเร้นอยู่ในสูตรการคำนวณหาค่าความเที่ยงตามทฤษฎีการทดสอบแบบประเพณีนิยมได้ ในกรณีที่มีปัจจัยประกอบเกี่ยวข้องหลายอย่างและมีการวัดในหลายๆ สถานการณ์หรือเงื่อนไข (condition)

2) สามารถตรวจสอบผลของอันตรกริยา (interaction) ของปัจจัยประกอบหลายอย่างในสถานการณ์ต่าง ๆ ได้ จึงช่วยให้นักวัดผลเข้าใจเรื่องของการวัดมากยิ่งขึ้น และวิธีการนี้ไม่สามารถทำได้โดยใช้ทฤษฎีการทดสอบแบบประเพณีนิยม

3) สามารถตอบคำถามที่เกี่ยวข้องกับการสอบได้หลายอย่างจากการวิเคราะห์เพียงครั้งเดียว ซึ่งถ้าใช้แนวคิดตามทฤษฎีการทดสอบแบบประเพณีนิยมจะต้องใช้ข้อมูลหลาย ๆ ชุด

4) ช่วยให้มีการวางแผนเก็บรวบรวมข้อมูลอย่างมีประสิทธิภาพ

ดังนั้น จากประโยชน์ของทฤษฎีการอ้างอิงสรุปดังได้กล่าวมาแล้ว ผู้วิจัยจึงมีความสนใจที่จะทำการประยุกต์แนวความคิดของทฤษฎีดังกล่าวมาใช้กับการตรวจให้คะแนนการสอบเรียงความ รวมทั้งพัฒนาโปรแกรมคอมพิวเตอร์เพื่อการวิเคราะห์ความแปรปรวนจากปัจจัยประกอบหลายแหล่งด้วย

วัตถุประสงค์ของการวิจัย

1. เพื่อประยุกต์แนวความคิดของทฤษฎีการอ้างอิงสรุปมาใช้ในการตรวจให้คะแนนการสอบแบบเรียงความ ในเงื่อนไขต่อไปนี้

1). เมื่อกำหนดให้ผู้ตรวจทุกประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน

2). เมื่อกำหนดให้ผู้ตรวจบางประเภท ตรวจข้อทดสอบบางข้อของผู้สอบทุกคน

3). เมื่อกำหนดให้ผู้ตรวจบางประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบบางคน

2. เพื่อสร้างโปรแกรมคอมพิวเตอร์สำหรับใช้กับเครื่องไมโครคอมพิวเตอร์เพื่อคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุป (generalizabiligy coefficient) สำหรับเงื่อนไขต่าง ๆ ในการสอบ

ขอบเขตของการวิจัย

การวิจัยครั้งนี้มีขอบเขตจำกัดดังนี้

1. ปัจจัยประกอบมีเพียง 3 อย่างคือ ผู้สอบ ข้อทดสอบ และผู้ตรวจ

2. เงื่อนไขที่เป็นไปได้ในทางปฏิบัติไม่เกิน 6 อย่างตามที่จะระบุไว้ในสมมุติฐานการวิจัย

3. จุดตัดของคะแนนอยู่ระหว่าง 0.0% - 100.0% เมื่อมีพิสัย = 0.10%

4. การตรวจข้อทดสอบเป็นแบบสังเคราะห์ โดยอาศัยแถบระดับความสามารถ (ability band) เป็นเกณฑ์ให้คะแนนผลการสอบ

ตัวแปรในการวิจัย

ก. ตัวแปรต้น (independent variables) ได้แก่ ปัจจัยประกอบ 3 อย่างที่มีรายละเอียดดังต่อไปนี้

1). ข้อทดสอบ (i) เป็นแบบเรียงความของแบบทดสอบวัดสมิทธิภาพภาษาอังกฤษของจุฬาลงกรณ์มหาวิทยาลัย (Chulalongkorn University Test of English Proficiency : CU-TEP) จำนวน 3 ข้อจากข้อทดสอบจำนวนอนันต์

ดังนั้น ข้อทดสอบ 3 ข้อ เป็นตัวอย่างแบบจำลองสุ่ม (random model) จากเอกภพของข้อทดสอบจำนวนไม่จำกัด ($n=3<N \propto$)

2). ผู้ตรวจ (r) มี 2 ประเภท คือ ผู้ตรวจที่มีประสบการณ์การตรวจโดยตรง ซึ่งได้แก่ผู้ที่เป็นกรรมการของแบบทดสอบ CU-TEP และมีประสบการณ์ในการตรวจให้คะแนนแบบทดสอบเรียงความของข้อทดสอบ CU-TEP มาแล้ว 6 คน และผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม ซึ่งได้แก่ผู้ที่เป็นอาจารย์สอนภาษาอังกฤษแต่ไม่เคยมีประสบการณ์ในการตรวจข้อทดสอบ CU-TEP มาก่อนอีก 6 คน

ดังนั้น ผู้ตรวจ 2 ประเภท เป็นตัวอย่างแบบจำลองคงที่ (fixed model) จากเอกภพจำนวนจำกัด ($n=N=2$)

3). ผู้สอบ (p) ได้แก่ผู้ที่สมัครสอบแบบทดสอบ CU-TEP จำนวน 200 คน ที่ได้มาจากคนจำนวนอนันต์

ดังนั้น ผู้สอบ 200 คน เป็นตัวอย่างแบบจำลองสุ่มจากเอกภพของคนจำนวนไม่จำกัด ($n=200<N \propto$)

ข. ตัวแปรตาม (dependent variables) ได้แก่

1). ค่าสัมประสิทธิ์การอ้างอิงสรุป ($\Sigma\rho^2$) ของการตรวจให้คะแนนแต่ละเงื่อนไข หรือสถานการณ์

2). ดัชนีความเชื่อถือ [$\Phi(\lambda_0)$] ของผลการสอบแต่ละเงื่อนไข หรือสถานการณ์ เมื่อจุดตัดของคะแนนแตกต่างกัน

<u>สมมุติฐานในการวิจัย</u>

ในการวิจัยครั้งนี้ผู้วิจัยมีสมมุติฐานในการวิจัยดังต่อไปนี้

<u>สมมุติฐานที่ 1</u>

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงจำนวน 2 คน เมื่อตรวจ**ข้อทดสอบทุกข้อของผู้สอบทุกคน** จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ตรง และ fixed >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed

<u>สมมุติฐานที่ 2</u>

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงจำนวน 2 คน เมื่อตรวจ**ข้อทดสอบบางข้อของผู้สอบทุกคน**จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ตรง และ fixed >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed

<u>สมมุติฐานที่ 3</u>

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงจำนวน 2 คน เมื่อตรวจ**ข้อทดสอบทุกข้อของผู้สอบบางคน** จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ตรง และ fixed >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed

สมมุติฐานที่ 4

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ตรง และ random >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ทางอ้อม และ random

สมมุติฐานที่ 5

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบบางข้อของผู้สอบทุกคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ตรง และ random >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ random

สมมุติฐานที่ 6

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบบางคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ตรง และ random >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ random

ข้อตกลงเบื้องต้น

การวิจัยครั้งนี้มีข้อตกลงเบื้องต้นดังนี้

1. ผู้สอบทุกคนตอบข้อทดสอบทั้ง 3 ข้อ อย่างเต็มความสามารถเนื่องจากการสอบ CU-TEP ผู้สอบทุกคนต้องสมัครมาสอบและต้องเสียค่าลงทะเบียนสอบเพื่อให้ได้ผลการสอบดีที่สุด เพราะผลการสอบมีประโยชน์โดยตรงต่อผู้สอบ เนื่องจากเป็นการสอบแข่งขันเพื่อความก้าวหน้าของผู้สอบเอง

2. ผู้ตรวจทุกคนตรวจให้คะแนนผลการสอบทุกข้ออย่างเต็มความสามารถเพราะผู้ตรวจครึ่งหนึ่งเป็นกรรมการของการทดสอบ CU-TEP จึงถือว่าเป็นหน้าที่ความรับผิดชอบโดยตรง และผู้ตรวจอีกครึ่งหนึ่งเป็นผู้ที่มีความเต็มใจที่จะช่วยเหลือโครงการวิจัยนี้

3. ข้อทดสอบเรียงความที่ใช้ในการวิจัยเป็นข้อทดสอบที่ดี เพราะได้ผ่านการพัฒนาเป็นอย่างดีแล้วจากผู้ที่มีความรู้ความสามารถทางการทดสอบทางภาษาอังกฤษเป็นอย่างดี

ความจำกัดของการวิจัย

งานวิจัยนี้มีข้อจำกัดบางอย่างต่อไปนี้

1. การตรวจข้อทดสอบของผู้ตรวจแต่ละประเภท ผู้ตรวจแต่ละคนไม่ได้ตรวจข้อทดสอบของผู้สอบทั้งหมด หากแต่แบ่งกันตรวจคนละประมาณ 70 ฉบับ (คน) การกระทำดังกล่าวนี้อาจมีผลต่อการอ้างอิงสรุป

2. ผู้วิจัยประสงค์ที่จะพัฒนาโปรแกรมคอมพิวเตอร์เพื่อการคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุปโดยใช้เครื่องไมโครคอมพิวเตอร์ แต่เนื่องจากโปรแกรมที่พัฒนาขึ้นเป็นโปรแกรมที่ยาวมากเพราะต้องการครอบ

คลุมเงื่อนไขหลาย ๆ อย่าง และโปรแกรมเขียนด้วยภาษา FORTRAN เมื่อทำการเปลี่ยนโปรแกรมเป็นภาษาเครื่อง (compile) ปรากฏว่าเกินขีดจำกัดของโปรแกรม (ความจุมากกว่า 640Kb) จึงต้องพัฒนาโปรแกรมขึ้นใช้กับเครื่องคอมพิวเตอร์ขนาดใหญ่ (mainframe) จึงทำให้โปรแกรมมีข้อจำกัดในการใช้ (ใช้ไม่สะดวกเท่าที่ควร)

<u>คำนิยามศัพท์ที่ใช้ในการวิจัย</u>

ในการวิจัยครั้งนี้มีคำศัพท์หลายคำที่มีความหมายจำเพาะดังต่อไปนี้

1. แบบทดสอบสมิทธิภาพทางภาษาอังกฤษของจุฬาลงกรณ์มหาวิทยาลัย (Chulalongkorn University Test of English Proficiency) หรือ **แบบทดสอบ CU-TEP** หมายถึง แบบทดสอบมาตราฐานเพื่อทดสอบสมิทธิภาพทั่วไปทางภาษาอังกฤษของนิสิต นักศึกษา และบุคคลที่สนใจทั่วไปในระดับอุดมศึกษา แบบทดสอบนี้มุ่งวัดความรู้ความสามารถ 3 ด้าน คือ ทักษะการฟังเข้าใจความ การอ่านเข้าใจความ และการเขียน ซึ่งมี 2 ส่วน คือ ส่วนที่ทดสอบด้วยข้อทดสอบปรนัย และส่วนที่ทดสอบด้วยข้อทดสอบเรียงความ ซึ่งมีอยู่ 3 ข้อ แบบทดสอบ CU-TEP นี้พัฒนาขึ้นโดยสถาบันภาษาและฝ่ายวิชาการของจุฬาลงกรณ์มหาวิทยาลัย

2. **ผู้ตรวจที่มีประสบการณ์การตรวจโดยตรง** หมายถึง ผู้ตรวจข้อทดสอบเรียงความที่เป็นกรรมการของแบบทดสอบ CU-TEP และมีประสบการณ์ในการตรวจให้คะแนนแบบทดสอบเรียงความของข้อทดสอบ CU-TEP มาแล้ว

3. **ผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม** หมายถึง ผู้ตรวจข้อทดสอบเรียงความที่เป็นอาจารย์สอนภาษาอังกฤษและเคยตรวจข้อทดสอบเรียงความมาก่อน แต่ไม่เคยตรวจข้อทดสอบ CU-TEP มาก่อนเลย

4. **เงื่อนไขหรือสถานการณ์** (condition) หมายถึงผลรวมผสม (combination) ระหว่าง จำนวนข้อทดสอบเรียงความ จำนวนประเภทของผู้ตรวจ และจำนวนผู้สอบข้อทดสอบเรียงความที่ใช้ในการวิจัยนี้ทั้งหมด

5. **ปัจจัยประกอบ** (facet) หมายถึง แหล่งของความแปรปรวนที่เกี่ยวข้องกับการทดสอบครั้งนี้ ซึ่งได้แก่ ข้อทดสอบเรียงความ ผู้ตรวจ และผู้สอบ

6. **เอกภพ** (universe) หมายถึง ประชากรของทั้งหมดของปัจจัยประกอบแต่ละอย่างซึ่งสามารถแบ่งออกได้ 2 อย่าง คือ

1.) **เอกภพของสิ่งสังเกตที่ยอมรับได้** (universe of admissible observations) หมายถึง สถานการณ์หรือเงื่อนไขผสมที่เป็นไปได้ระหว่างปัจจัยประกอบ ด้านข้อทดสอบ ผู้ตรวจ และผู้สอบตามจำนวนเงื่อนไขที่มีอยู่จริง เช่น ผู้ตรวจประเภทที่ 1 (มีประสบการณ์โดยตรง) ตรวจข้อทดสอบข้อที่ 1 ของผู้สอบคนที่ 1 เป็นต้น

2.) **เอกภพของการอ้างอิงสรุป** (universe of generalizability) หมายถึง สถานการณ์หรือเงื่อนไขผสมที่อาจเป็นไปได้สำหรับประชากรของแต่ละปัจจัยประกอบที่ผู้วิจัยต้องการอ้างอิงสรุป (generalize) ผลการศึกษา เช่น ผู้ตรวจประเภทที่ 2 (มีประสบการณ์ทางอ้อม) ใด ๆ ตรวจข้อทดสอบข้อใด ๆ ของผู้สอบคนใดก็ได้ เป็นต้น

7. **คะแนนเอกภพ** (universe score) หมายถึง คะแนนที่บ่งบอกความสามารถที่แท้จริง (true score) ของผู้สอบแต่ละคนในเงื่อนไขหรือสถานการณ์ที่ระบุจำเพาะ ซึ่งในการวิจัยครั้งนี้ได้แก่คะแนนเฉลี่ยของคะแนนสังเกต (observed score) ในเอกภพของการอ้างอิง และเนื่องจากเอกภพของการอ้างอิงมีได้หลายเอกภพ ดังนั้นคะแนนเอกภพจึงอาจมีได้หลายค่า

8. **คะแนนสังเกต** (observed score) หมายถึง คะแนนที่ได้จากการตรวจข้อทดสอบเรียงความแต่ละข้อหรือคะแนนรวม

9. **ค่าสัมประสิทธิ์การอ้างอิงสรุป** (generalizability coefficient) หมายถึง ค่าความเที่ยงของการตรวจวัดให้คะแนนข้อทดสอบในเงื่อนไขหรือสถานการณ์ที่กำหนด ค่าสัมประสิทธิ์การอ้างอิงสรุปนี้เป็นค่าประมาณของค่าเฉลี่ยของค่าสหสัมพันธ์ระหว่างค่าการวัดซึ่งสุ่มมาจากเอกภพรายคู่ เช่น ค่าสัมประสิทธิ์อ้างอิงสรุปเมื่ออ้างอิงไปยังชุดของข้อทดสอบจำนวน 3 ข้อ มีค่าเท่ากับ 0.80 หมายความว่า ถ้าสุ่มผู้สอบมาจากประชากรหนึ่งในจำนวนหนึ่ง เพื่อสอบข้อทดสอบ 3 ข้อที่ไม่ซ้ำกัน ค่าเฉลี่ยของสัมประสิทธิ์การอ้างอิงสรุป ระหว่างการทดสอบแต่ละครั้งจะมีค่าเท่ากับ 0.80 เป็นต้น ค่านี้มีค่าเท่ากับ KR_{20} (Kuder-Richardson Formular 20) หรือค่า Cronbach α (Brennan and Kane 1977:280)

นอกจากนี้ค่าสัมประสิทธิ์การอ้างอิงสรุป ยังหมายถึงอัตราส่วนระหว่างความแปรปรวนของคะแนนเอกภพกับคะแนนสังเกต เช่น ค่าสัมประสิทธิ์การอ้างอิงสรุปเท่ากับ 0.80 หมายความว่า ความแตกต่างที่วัดได้ร้อยละ 80 เป็นความแตกต่างเนื่องมาจากคะแนนเอกภพ อีกร้อยละ 20 เป็นความแตกต่างเนื่องมาจากความคลาดเคลื่อน เป็นต้น

10. **ดัชนีความเชื่อถือ** (index of dependability) หมายถึงความคงที่ของคะแนนการสอบของผู้สอบแต่ละบุคคลว่าแตกต่างจากคะแนนจุดตัดมากน้อยเพียงใด ดัชนีนี้คือค่าความเที่ยงของแบบทดสอบอิงเกณฑ์

<u>ประโยชน์ที่คาดว่าจะได้รับ</u>

1.ทำให้ครู-อาจารย์ผู้สอนทราบถึงขนาดของความแปรปรวนของแหล่งต่าง ๆ ที่มีอิทธิพลต่อความเที่ยงของการวัดผลแบบเรียงความ และสามารถเลือกใช้แบบจำลองการวัดที่เหมาะสมได้เพื่อเป็นการประหยัดเวลา และแรงงาน แต่ได้ผลถูกต้องที่สุด

2. ทำให้ครู-อาจารย์ผู้สอนสามารถอ้างอิงสรุป (generalize) ผลการเรียนการสอนและการทดสอบของผู้เรียนในสถานการณ์ต่าง ๆ ไปยังกลุ่มประชากรได้อย่างมั่นใจยิ่งขึ้น ทำให้ทราบระดับความรู้และความสามารถของผู้เรียนได้ดียิ่งขึ้น

3. ทำให้ครู-อาจารย์ตระหนักถึงความสำคัญของแหล่งความแปรปรวนต่าง ๆ ที่เกี่ยวข้องกับการทดสอบเพื่อนำมาพิจารณาร่วมในการประเมินผลการเรียนการสอนได้ถูกต้อง และยุติธรรมยิ่งขึ้น

4. เป็นแนวทางให้นักทดสอบนำความรู้เกี่ยวกับทฤษฎีการอ้างอิงสรุปไปใช้ในการทดสอบและการเรียนในสถานการณ์หรือเงื่อนไขต่าง ๆ มากขึ้นได้

5. ทำให้ครู-อาจารย์ผู้สอนได้โปรแกรมคอมพิวเตอร์ที่สามารถใช้คำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุปได้อย่างสะดวกและถูกต้องสำหรับเงื่อนไขต่าง ๆ ในการสอบเรียงความ ซึ่งอาจเป็นการ กระตุ้นให้ครู-อาจารย์นำการสอบแบบเรียงความมาใช้มากยิ่งขึ้นได้

บทที่ 2
เอกสารและวรรณกรรมที่เกี่ยวข้อง

ในบทนี้ผู้วิจัยจะขอกล่าวถึงเอกสารและวรรณกรรมที่เกี่ยวข้องกับเรื่องที่วิจัยเป็น 4 หัวข้อที่สำคัญ คือ

1. ทฤษฎีการอ้างอิงสรุป
2. การประยุกต์ทฤษฎีอ้างอิงสรุปมาใช้ในการวัดผลการศึกษา
3. การทดสอบแบบเรียงความ และ
4. ผลการวิจัยที่เกี่ยวข้องกับการเรียนการสอนและการสอบแบบเรียงความในประเทศ

ต่อไปนี้เป็นรายละเอียดของหัวข้อหลักที่สำคัญทั้ง 4 ดังกล่าวแล้ว

1. ทฤษฎีการอ้างอิงสรุป

ก. ประวัติความเป็นมาโดยสังเขป

แนวความคิดพื้นฐานเชิงทฤษฎีของทฤษฎีการอ้างอิงสรุป (Generalizability Theory) เป็นแนวคิดที่ปรากฏในบทความทางวิชาการของ Cronbach, Rajaratnam and Gleser ในปี 1963 และบทความของ Gleser, Cronbach and Rajaratnam ในปี 1965 และต่อมา Cronbach, Gleser, Nanda and Rajaratnam ได้ช่วยกันเขียนหนังสือขึ้นเผยแพร่ความคิดเรื่องนี้อย่างละเอียดในปี 1972 ชื่อ The Dependability of Behavioral Measurement และในระยะต่อมาได้มีนักทดสอบและวัดผลอีกหลายท่านได้เขียนหนังสือเผยแพร่แนวคิดของทฤษฎีนี้อย่างแพร่หลาย เช่น Brennan, Kane, Gillmore, Van der Kamp, Shavelson และ Webb เป็นต้น (Brennan 1983:1) Cronbach, Rajaratnam and Gleser ได้สรุปความเป็นมาของทฤษฎีอ้างอิงสรุปไว้ว่า ทฤษฎีทดสอบแบบประเพณีนิยม (Classical Test Theory) ใช้ค่าความเที่ยงอธิบายความแม่นยำของการวัด โดยยึดข้อตกลงเบื้องต้นว่าข้อทดสอบที่ใช้สอบเป็นข้อทดสอบคู่ขนาน หรือมีความเท่าเทียมเป็นสำคัญ ตามทฤษฎีเดิมนี้นิยามความเที่ยงว่าเป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าการวัดที่มีความเท่าเทียมหรือมีความคู่ขนานกัน จากนิยามนี้ การหาค่าความเที่ยงจะต้องมีการสอบวัดให้ได้ค่าการวัดอย่างน้อยสองชุด และแต่ละชุดใช้เครื่องมือวัดที่มีคุณสมบัติคู่ขนานกัน คือ มีเนื้อหาอย่างเดียวกัน ค่าเฉลี่ยเท่ากัน ค่าความแปรปรวนเท่ากัน และค่าสหสัมพันธ์ระหว่างกันเท่ากัน แล้วคำนวณหาค่าสหสัมพันธ์ระหว่างค่าการวัดทั้ง 2 ชุด ผลที่ได้ถือเป็นค่าความเที่ยง ในกรณีที่ต้องการประมาณค่าความเที่ยงจากการสอบครั้งเดียว ต้องแบ่งค่าการวัดออกเป็นส่วนย่อย ๆ เช่น ข้อคู่ - ข้อคี่ ครึ่งแรก -ครึ่งหลัง หรือแบ่งย่อยตามจำนวนข้อ โดยยึดข้อตกลงเบื้องต้นว่าส่วนแบ่งย่อยแต่ละส่วนจะต้องมีความเท่าเทียมกัน

ดังนั้น จากแนวคิดดังกล่าวมักเกิดปัญหาในทางปฏิบัติเพราะนักทดสอบไม่สามารถจะหาเครื่องมือวัดที่มีคุณสมบัติความเท่าเทียมกันอย่างสมบูรณ์ได้นอกจากแบบทดสอบมาตรฐาน ตัวอย่างเช่น เราต้องการศึกษาความเที่ยงในการตรวจคำตอบแบบเรียงความ การจะหาผู้ตรวจที่มีความเท่าเทียมกันเป็นสิ่งที่ทำได้ยากมาก ไม่ว่าจะพิจารณาค่าความแปรปรวนหรือค่าสหสัมพันธ์ระหว่างผู้ตรวจ ก็ตาม ดังนั้น การใช้สูตรที่มีข้อตกลงเบื้องต้นเกี่ยวกับความเท่าเทียมหรือคู่ขนานกันจึงเป็นการปฏิบัติโดยการฝืนข้อตกลง จึงทำให้นักทดสอบคิดพัฒนาทฤษฎีที่จะหาความเที่ยงโดยไม่ยึดมั่นในข้อตกลงของความเท่าเทียมกันขึ้น และแนวคิดหนึ่งได้แก่ทฤษฎีการอ้างอิงสรุป

ก่อนที่จะกล่าวถึงรายละเอียดของทฤษฎีการอ้างอิงสรุป ผู้วิจัยจะขอกล่าวถึงแนวคิดพื้นฐานของการหาค่าความเที่ยงตามทฤษฎีการทดสอบแบบประเพณีนิยมเสียก่อน ทั้งนี้เพื่อให้ท่านผู้อ่านเข้าใจความคล้ายคลึงและความแตกต่างของทฤษฎีทั้งสองได้ดียิ่งขึ้น ดังนี้

ข. แนวคิดพื้นฐานของการหาค่าความเที่ยงตามทฤษฎีการทดสอบแบบประเพณีนิยม

1. การหาค่าความเที่ยงตามแบบประเพณีนิยม

ในทฤษฎีการทดสอบแบบประเพณีนิยม (Classical Test Theory) นั้นเชื่อว่า คะแนนที่เป็นผลการวัดที่สังเกตได้ (observed score :X) นั้น ส่วนหนึ่งเป็นคะแนนความสามารถที่แท้จริง (true score :T) และอีกส่วนหนึ่งเป็นความคลาดเคลื่อน (error :E) กล่าวคือ

$$X = T + E$$

ทฤษฎีนี้ถือว่า

1) ส่วนของความคลาดเคลื่อนนี้มีเพียงส่วนเดียว (single error component) ไม่สามารถจะแยกแหล่งของความคลาดเคลื่อนออกไปได้

2) ส่วนคะแนนความสามารถที่แท้จริง หมายถึง ค่าที่คาดหวังของค่าสังเกตจากการวัดที่มีความเสมอเหมือนกัน (equivalent) ที่เป็นไปได้ทั้งหมด

3) คะแนนความสามารถที่แท้จริงกับคะแนนความคลาดเคลื่อนไม่สัมพันธ์กัน ดังนั้น

$$\sigma_x^2 = \sigma_T^2 + \sigma_E^2$$

4) นิยามค่าสัมประสิทธิ์ความเที่ยง (reliability coefficient) ว่าคืออัตราส่วนของความแปรปรวนของคะแนนจริงกับความแปรปรวนของค่าที่สังเกตได้

$$\rho^2_{XT} = \sigma^2_T/\sigma^2_X = \sigma^2_T/(\sigma^2_T+\sigma^2_E)$$

การประมาณค่าอัตราส่วนความแปรปรวนหรือค่าสัมประสิทธิ์ความเที่ยงก็ใช้ค่าสหสัมพันธ์ระหว่างแบบวัดที่คู่ขนานกันหรือให้การคำนวณหาค่าความสอดคล้องภายใน (internal consistency)

2. การหาค่าความเที่ยงโดยการวิเคราะห์การแปรปรวน

เทคนิควิธีการวิเคราะห์ความแปรปรวนเป็นวิธีการที่คิดขึ้นโดย Sir Ronald Fisher นักสถิติศาสตร์ ผู้มีชื่อเสียงชาวอังกฤษ แต่เดิมการวิเคราะห์ความแปรปรวนนี้ใช้ในการทดสอบเพื่อหาความแตกต่างระหว่างคะแนนของกลุ่มต่าง ๆ แต่ต่อมา Cureton ได้นำมาใช้แทนค่าสัมประสิทธิ์ความเที่ยงเป็นคนแรก และ Pearson เป็นบุคคลแรกที่อธิบายค่าสัมประสิทธิ์สหสัมพันธ์

ในรูปของอัตราส่วนของความแปรปรวน แนวคิดในการนำการวิเคราะห์ความแปรปรวนมาใช้ในการประเมินความเที่ยงเป็นดังนี้

ให้ X_{pi} แทน คะแนนสังเกตของนักเรียน p ที่สอบแบบทดสอบ i (หรือตรวจโดยผู้ตรวจ I)

สามารถเขียนในแบบเชิงเส้นได้ดังนี้

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

เมื่อให้

M_p แทนคะแนนเอกภพ (universe score) คือคะแนนจริงตามความหมายของทฤษฎีการทดสอบแบบประเพณีนิยม

แต่ในที่นี้หมายถึง ค่าเฉลี่ยของคะแนนผลของการทดสอบซ้ำ ๆ เป็นจำนวนอนันต์

M_i แทนค่าเฉลี่ยของแบบทดสอบ i คำนวณจากผู้สอบทุกคนในประชากรของผู้สอบ

M แทนค่าเฉลี่ยรวม (grand mean) ของนักเรียนทุกคนในทุกเงื่อนไขการทดสอบ

e_{pi} แทนความคลาดเคลื่อนเชิงสุ่ม (sampling error) มีค่าเฉลี่ยเป็น 0 ค่าความแปรปรวนเท่ากันและเป็นอิสระต่อผลอื่น ๆ ในตัวแบบ หมายถึงในการวัดครั้งหนึ่ง ๆ จะมีค่าเบี่ยงเบนไปจากค่าเฉลี่ยทั้งด้านบวกและด้านลบ เมื่อรวมกันแล้วความคลาดเคลื่อนจะมีค่าเป็น 0

จากแบบและข้อตกลงข้างต้นสามารถประมาณค่าความแปรปรวนของแหล่งต่าง ๆ แล้วคำนวณค่าความเที่ยงตามนิยาม โดยคำนวณจากอัตราส่วนความแปรปรวนของคะแนนจริง ต่อความแปรปรวนของคะแนนสังเกต เรียกอัตราส่วนนี้ว่า สหสัมพันธ์ภายในชั้น (intraclass correlation)

สำหรับความเที่ยงซึ่งใช้ค่าสัมประสิทธิ์สหสัมพันธ์ชนิดนี้ สามารถประมาณค่าได้จากกำลังสองของความสัมพันธ์ระหว่างคะแนนที่สังเกตได้ กับคะแนนจริง ซึ่งเท่ากับอัตราส่วนระหว่างความแปรปรวนของคะแนนจริง กับค่าความแปรปรวนของคะแนนที่สังเกตได้ ดังที่ได้กล่าวมาแล้ว

3. หลักการพื้นฐานที่สำคัญของทฤษฎีการอ้างอิงสรุป

ก. ข้อตกลงเบื้องต้น

ข้อตกลงเบื้องต้นของทฤษฎีการอ้างอิงสรุปโดยทั่วไปมีข้อตกลงดังนี้

1. ต้องระบุเอกภพที่ต้องการอ้างอิงให้ชัดเจนจนสามารถบอกได้ว่ามีเงื่อนไขใดบ้าง

2. เงื่อนไขการวัดเป็นอิสระต่อกัน กล่าวคือ คะแนนของนักเรียนที่ตอบข้อทดสอบข้อ i ถูกหรือผิด ไม่ขึ้นกับการตอบข้ออื่น

3. คะแนนสังเกต (X_{pi}) เป็นค่าการวัดในมาตราอันตรภาค (interval scale)

ข. มโนมติเบื้องต้น (Basic Concepts)

การที่จะเข้าใจหลักการและการประยุกต์ใช้ทฤษฎีการอ้างอิงสรุปได้ดีนั้นผู้อ่านจำเป็นจะต้องทำความเข้าใจแนวคิดเบื้องต้นรวมทั้งคำศัพท์บางคำที่เกี่ยวข้องกับทฤษฎีดังกล่าวนี้ก่อน ซึ่งจะมีรายละเอียดต่อไปนี้

1. **เงื่อนไขหรือสถานการณ์** (condition) หมายถึง การวัดที่ทำให้เราได้ค่าสังเกตแต่ละค่า สถานการณ์อาจได้แก่ แบบทดสอบ สิ่งเร้า ผู้สังเกต ผู้ตรวจให้คะแนน และโอกาสที่สังเกต เป็นต้น

2. **ปัจจัยประกอบ** (facets) หมายถึง ชุดของเงื่อนไขการวัดที่เป็นชนิดเดียวกัน

ตัวอย่างเช่นอาจารย์สมิทธิ์ ได้พัฒนาวิธีวัดความสามารถในการเขียนโดยให้เขียนเรียงความ 2 เรื่อง และให้มีผู้ตรวจให้คะแนน 3 คน ในกรณีนี้การวัดจะมี 2 ปัจจัยประกอบ คือ

1). หัวข้อเรียงความ ซึ่งมี 2 เงื่อนไข คือ 2 เรื่อง

2). ผู้ตรวจให้คะแนนซึ่งมี 3 เงื่อนไข คือ มีผู้ตรวจ 3 คน

3. **เอกภพ** (universe) หมายถึง กลุ่มเงื่อนไขการวัดทั้งหมดที่สามารถหามาได้ เอกภพมี 2 ขนาด คือ เอกภพขนาดจำกัด (finite universe) และเอกภพขนาดไม่จำกัด (infinite universe)

4. **เอกภพของสิ่งสังเกตที่ยอมรับได้** (universe of admissible observations)หมายถึง ผลร่วมผสม (combination) ระหว่างกลุ่มเงื่อนไขการวัดที่เป็นไปได้ทั้งหมดที่ผู้ทดสอบสามารถทำการสังเกตหรือวัดค่าได้ กรณีของอาจารย์สมิทธิ์ก็จะประกอบด้วยเรียงความเรื่องที่ 1 กับผู้ตรวจคนที่ 1 เรียงความ เรื่องที่ 2 กับผู้ตรวจคนที่1 และเรียงความเรื่องที่ 2 กับผู้ตรวจคนที่ 3 เป็นต้น รวมแล้วทั้งหมดได้ 6 สถานการณ์ ในกรณีที่เป็นแบบผสม (cross design)

5. **เอกภพของการอ้างอิงสรุป** (universe of generalization) หมายถึง กลุ่มเงื่อนไขการวัดทั้งหมดของปัจจัยประกอบที่ผู้วิจัยต้องการสรุปอ้างอิงผลการวัดจากกลุ่มตัวอย่างของเงื่อนไขการวัดเหล่านั้นไปยังกลุ่มเงื่อนไขการวัดทั้งหมดของปัจจัยประกอบนั้น ๆ

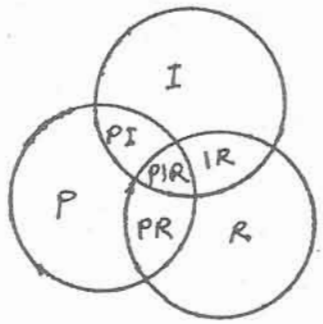
6. **คะแนนเอกภพ** (universe score) หมายถึงค่าเฉลี่ยของค่าสังเกตในเอกภพของการอ้างอิงสรุป เนื่องจากเอกภพของการอ้างอิงสรุปมีได้หลายเอกภพ ดังนั้น คะแนนเอกภพจึงมีได้หลายค่า

ค. กระบวนการวิเคราะห์เพื่อหาค่าสัมประสิทธิ์การอ้างอิงสรุป

การวิเคราะห์หาค่าสัมประสิทธิ์อ้างอิงสรุปสามารถทำได้ดังนี้ (Cronbach and Others 1992 : 35-42) คือ

ขั้นที่ 1 เรียกว่า G (Generalizability) Study เป็นการประมาณค่าความแปรปรวนของแหล่งต่าง ๆ ภายใต้เงื่อนไขการวัดจริงที่ยอมรับได้ เริ่มด้วยการกำหนดสิ่งที่ถูกวัด (object of measurment) เป็นคน (นักเรียน) หรือกลุ่มคน เช่น นักเรียนทั้งห้อง (class) หรือทั้งโรงเรียน ให้ปัจจัยอื่น ๆ ภายใต้การสังเกตเป็นปัจจัยประกอบ เช่น ข้อทดสอบ โอกาสที่สอบ และ ผู้สอบ เป็นต้น แล้วกำหนดความสัมพันธ์ระหว่างปัจจัยประกอบเป็นแบบผสม (cross design) หรือแบบแฝง (nest design) แล้วใช้วิธีวิเคราะห์ความแปรปรวน (Analysis of Variance : ANOVA) ที่สอดคล้องกับแบบที่กำหนดเพื่อหาความสัมพันธ์ของปัจจัยประกอบ โดยประมาณค่า mean square แล้วประมาณค่าความแปรปรวนของแหล่งต่าง ๆ ค่าที่คำนวณได้เรียกว่าค่าประมาณความแปรปรวนของคะแนนเอกภพ (estimated universe score variance) ซึ่งสอดคล้องกับค่าความแปรปรวนของคะแนนจริง (true score variance) ตามทฤษฎีการทดสอบแบบประเพณีนิยม และเรียกค่าประมาณความแปรปรวนของปัจจัยประกอบอื่น ๆ ว่าค่าประมาณความแปรปรวนของความคลาดเคลื่อน (estimated error score variance) การวิเคราะห์ในขั้น G Study ใช้ค่าขนาดของกลุ่มตัวอย่างที่ศึกษาจริงของแต่ละปัจจัยประกอบ เพื่อให้ผู้อ่านเข้าใจการวิเคราะห์ในขั้น G Study ดียิ่งขึ้นจึงขอเสนอตัวอย่างของการวิเคราะห์ตามที่ Brennan (1983: 2-8) ได้อธิบายไว้ดังนี้

สมมุติว่าอาจารย์สมิทธิ์ต้องการที่จะหาวิธีการประเมินผลความสามารถในการเขียนของนักเรียน กระบวนการจะต้องเริ่มจากระบุข้อทดสอบเรียงความที่จะใช้วัด พร้อมกับกำหนดผู้ตรวจที่มีความเชี่ยวชาญ ถึงจุดนี้อาจารย์สมิทธิ์ยังไม่ปักใจว่าจะใช้วิธีการวัดอย่างใดอย่างหนึ่งเป็นการเฉพาะ กล่าวคือ ยังไม่กำหนดแน่นอนลงไปว่าเป็นข้อทดสอบข้อใด ใครจะเป็นคนตรวจ เพียงแต่บรรยายปัจจัยประกอบที่เขาเอง หรือนักวิจัยคนอื่น ๆ สนใจ ปัจจัยประกอบ หมายถึง ชุดของเงื่อนไขการวัดที่คล้ายคลึงกัน (similar conditions of measurement) ข้อทดสอบแต่ละข้อเป็นเงื่อนไขการวัดอย่างหนึ่ง (condition) ที่สามารถยอมรับได้ของปัจจัยประกอบด้านข้อทดสอบ (item facet) และผู้ตรวจแต่ละคน คือเงื่อนไขการวัดที่ยอมรับได้เงื่อนไขหนึ่งของปัจจัยประกอบด้านผู้ตรวจ (rater facet) ดังนั้นเอกภพของสิ่งสังเกตที่ยอมรับได้ (universe of admissible observation) จึงประกอบด้วยผลร่วมผสม (combination) ระหว่างข้อทดสอบและผู้ตรวจ จากนั้นอาจารย์สมิทธิ์จะต้องหันมาพิจารณาความสัมพันธ์ระหว่างปัจจัยประกอบด้านผู้ตรวจ และข้อทดสอบที่ตนเองสนใจ ถ้ายอมรับว่าในการตรวจคำตอบนั้น ผู้ตรวจ (r) แต่ละคนจะตรวจข้อทดสอบ (i) ทุกข้อเหมือนกัน เอกภพของการสังเกตที่ยอมรับได้จะเป็นแบบ "ผสม" (cross) และใช้สัญลักษณ์ว่า ixr (อ่านว่า i ผสมกับ r) จากนั้นอาจารย์สมิทธิ์จะต้องเก็บรวบรวมและวิเคราะห์ข้อมูลโดยการสุ่มนักเรียนมาจำนวน n_p คน ใช้ข้อทดสอบจำนวน n_i ข้อ เมื่อสอบเสร็จก็สุ่มผู้ตรวจที่มีความเชี่ยวชาญมาตรวจให้คะแนนแล้วคำนวณค่าความแปรปรวนของปัจจัยประกอบต่าง ๆ วิธีการดังกล่าวนี้คือ การทำ G Study แบบของการศึกษา (disign) มีชื่อว่า pxixr แบบนี้มีแหล่งความแปรปรวนทั้งหมด 7 แหล่ง จึงอาจเรียกว่า แบบ 7 (VII Design) เขียนแทนด้วยแผนภูมิ Venn ได้ดังภาพที่ 1 ต่อไปนี้



แผนภูมิที่ 1: แสดงองค์ประกอบความแปรปรวนของแบบ pxixr เมื่อปัจจัยประกอบที่ศึกษาทั้งหมดเป็นแบบจำลองสุ่ม (randon model)

ตามแผนภูมิที่ 1 จะพบว่ามีผลหลัก (main effects) อยู่ 3 ค่า คือ

1) ผลของผู้สอบ (p)

2) ผลของข้อทดสอบ (i)

3) ผลของผู้ตรวจ (r)

ผลร่วมหรืออันตรกริยาสองระดับมี 3 ค่า คือ

1) ผลร่วมของผู้สอบและข้อทดสอบ (pi)

2) ผลร่วมของผู้สอบและผู้ตรวจ (pr)

3) ผลร่วมของข้อทดสอบและผู้ตรวจ (ir)

นอกจากนี้จะมีผลร่วมกันของผู้สอบ ข้อทดสอบ และผู้ตรวจ (pir)

ค่าความแปรปรวนของแหล่งต่าง ๆ เหล่านี้คำนวณจาก mean square และ mean square คำนวณโดยใช้ ANOVA แบบ Factorial Design คือ pxixr ในการประมาณค่าความแปรปรวนจาก mean square นั้นใช้สูตรเฉพาะอย่างที่สอดคล้องกับแบบที่ต้องการศึกษา

สูตรในการประมาณค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อปัจจัยประกอบที่ศึกษาเป็นแบบจำลองสุ่มทั้งหมด และกำหนดแบบความสัมพันธ์ระหว่างปัจจัยประกอบเป็นแบบผสม ใช้สูตรดังต่อไปนี้ (Cronbach and Others. 1972 :43) ดังนี้

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

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

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

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

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

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

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

โครงสร้างของสมการนี้สัมพันธ์กับแผนภูมิข้างต้น กล่าวคือ ในวงกลม p ประกอบด้วยพจน์ (term) ต่าง ๆ 4 พจน์ มี p, pi, pr และ pir ,e เป็นพจน์ทางขวามือของสูตรในการหาค่า EMS_p ในพื้นที่ pi ซึ่งเป็นส่วนร่วมระหว่างวงกลม p และ i ประกอบด้วย 2 พจน์คือ pi และ pir, e ก็คือพจน์ที่ปรากฏอยู่ทางขวามือของสูตรในการหาค่า EMS_{pi} นั่นเอง ดังนั้นจึงสามารถสร้างสูตรหาค่า EMS ได้โดยอาศัยแผนภูมิ โดยใช้ n_p เป็นตัวคูณควบถ้าองค์

ประกอบนั้นไม่รวม p และดูณ n_i ถ้าองค์ประกอบนั้นไม่มี i รวมอยู่ด้วย และดูณ n_r ถ้าองค์ประกอบนั้นไม่มี r ร่วมอยู่ด้วย

ขั้นที่ 2 เรียกว่า D (Decision) Study ในขั้นนี้เน้นที่การประมาณค่า การใช้ และการตีความค่าความแปรปรวนภายใต้วิธีการวัดที่ผู้ศึกษากำหนดขึ้นอย่างมีเหตุผล เพื่อประกอบการตัดสินใจ และมีแนวคิดที่สำคัญดังนี้

1. เอกภพของการอ้างอิง (universe of generalization) เป้าหมายสำคัญของ D Study ได้แก่ การกำหนดลักษณะเฉพาะของเอกภพของการอ้างอิงที่ผู้ศึกษาต้องการอ้างอิงถึง ซึ่งอาจจะประกอบด้วยเงื่อนไขทั้งหมดในเอกภพของการสังเกตที่ยอมรับได้หรืออาจเป็นเซตย่อยของเอกภพการสังเกตที่ยอมรับได้ (Brennan 1983 : 3)

2. ขนาดของตัวอย่างของ D Study (D Study sample sizes) จำนวนเงื่อนไขของปัจจัยประกอบใน D Study สามารถกำหนดแตกต่างจากจำนวนเงื่อนไขของ G Study ได้และใช้สัญลักษณ์(′) แทนขนาดตัวอย่างของ D Study เช่น n'_i และ n'_r แทนจำนวนข้อทดสอบและจำนวนผู้ตรวจ

3. โครงสร้างแบบของ D Study (D Study design structure) นอกจากระบุขนาดของตัวอย่างใน D Study แล้ว จะต้องระบุรูป (form) โครงสร้างของแบบหรือความสัมพันธ์ของปัจจัยประกอบที่ศึกษา เช่นอาจารย์สมิทธิ์อาจต้องการตัดสินใจว่าในการสอบนั้นนักเรียนทุกคนต้องทำข้อทดสอบเหมือนกันทั้ง n_i ข้อ และผู้ตรวจทั้ง n_r คนต้องตรวจทุก ๆ ข้อ แบบดังกล่าวนี้เป็น p x I x R ให้สังเกตว่าตัวอักษรตัวใหญ่แทนปัจจัยประกอบใน D Study ซึ่งในแบบของ G Study ใช้อักษรตัวเล็ก ถ้าอาจารย์สมิทธิ์ทำ D Study ตามแบบ p x I x R แล้ว D Study ของเขาจะตรงกับ G Study แต่ไม่จำเป็นต้องทำเช่นนั้นเสมอไป เขาสามารถตัดสินใจว่าผู้สอบทุกคนทำข้อทดสอบทุกข้อแต่ผู้ตรวจแต่ละคนอาจตรวจคำตอบต่างข้อกัน ดังนั้นแบบของ D Study จะเป็น p x (I : R) ซึ่ง : อ่านว่า "แฝงอยู่ภายใน' (nested within)

4. การประมาณค่าความแปรปรวนในขั้น D Study การวิเคราะห์ข้อมูลในขั้น D Study ขึ้นอยู่กับการตัดสินใจของนักวัดผลหรือนักวิจัย ดังนั้นจึงต้องมีการประมาณค่าความแปรปรวนขึ้นมาใหม่อีกครั้งโดยอาศัยผลจากการประมาณค่าในขั้น G Study เป็นฐาน และให้สอดคล้องกับแบบและขนาดของตัวอย่างที่ต้องการตัดสินใจ เช่น ถ้าแบบเป็น p x I x R ที่ข้อทดสอบและผู้ตรวจเป็นปัจจัยประกอบแบบจำลองสุ่ม จากตัวอย่าง สมมุติว่าอาจารย์สมิทธิ์ต้องการอ้างอิงไปยังเอกภพของข้อทดสอบและผู้ตรวจพร้อมกัน เมื่อการสอบของเขาต้องการใช้ $n'_i = 6$ และ $n'_r = 2$ ก็สามารถคำนวณค่าความแปรปรวนใน D Study ได้ดังนี้

สมมุติว่าค่าประมาณของความแปรปรวนใน G Study ของปัจจัยประกอบต่าง ๆ เป็นดังนี้

$\hat{\sigma}^2(p) = 0.30$, $\hat{\sigma}^2(i) = 0.25$, $\hat{\sigma}^2(r) = 0.10$, $\hat{\sigma}^2(pi) = 0.37$

$\hat{\sigma}^2(pr) = 0.50$, $\hat{\sigma}^2(ir) = 0.25$ และ $\hat{\sigma}^2(pir) = 1.00$

ดังนั้น การหาค่าความแปรปรวนใน D Study ทำได้ง่าย ๆ โดยหารค่าความแปรปรวนที่ได้จาก G Study ของผลต่าง ๆ ซึ่งเขียนในรูปสัญลักษณ์ทั่วไปว่า $\sigma^2(\alpha)$ ด้วย $n_i = 6$ ถ้า α เป็นผลที่มี i แต่ไม่มี r และหาร $\sigma^2(\alpha)$ 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

$\hat{\sigma}^2(p) = 0.30$, $\hat{\sigma}^2(I) = 0.04$, $\hat{\sigma}^2(R) = 0.05$, $\hat{\sigma}^2(pI) = 0.06$

$\hat{\sigma}^2(pR) = 0.25$, $\hat{\sigma}^2(IR) = 0.02$ และ $\hat{\sigma}^2(pIR) = 0.08$

5. ประมาณค่าความแปรปรวนของความคลาดเคลื่อน 2 ชนิด คือ

1). ความแปรปรวนของความคลาดเคลื่อนสัมบูรณ์ (absolute error variance) ซึ่งใช้สัญลักษณ์ว่า $\hat{\sigma}^2(\Delta)$ ซึ่งเป็นค่าความแปรปรวนของความแตกต่างระหว่างคะแนนสังเกตกับคะแนนเอกภพของผู้สอบ ค่านี้คำนวณจากผลบวกของค่าความแปรปรวนอื่น ๆ ทั้งหมดยกเว้นค่าความแปรปรวนของคะแนนเอกภพ เช่น

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

2). ความแปรรวนของความคลาดเคลื่อนสัมพัทธ์ (relative error variance) ซึ่งเขียนแทนด้วยสัญลักษณ์ว่า $\hat{\sigma}^2(\delta)$ ค่านี้คำนวณได้จากการรวมปัจจัยประกอบที่มีความแปรปรวนของสิ่งที่วัดร่วมอยู่เข้าด้วยกัน เช่นจากตัวอย่างข้างต้นจะได้ความแปรปรวนนี้ดังนี้

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

6. ประมาณค่าดัชนีความเชื่อถือ และค่าสัมประสิทธิ์การอ้างอิงสรุป จากสูตรต่อไปนี้ (Brennan 1981 : 68)

$$\Phi(\lambda_o) = \hat{\sigma}^2(\pi) + (\mu-\lambda)^2 / [\hat{\sigma}^2(\pi) + (\mu-\lambda)^2 + \hat{\sigma}^2(\Delta)]$$

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

สมการแรกเป็นค่าดัชนีความเชื่อถือ หรือบางที่เรียกว่าค่าสัมประสิทธิ์ที่มีลักษณะคล้ายกับค่าความเที่ยง (reliability-like coefficient) เพราะว่าแนวคิดของค่าดังกล่าวคล้ายกับแนวคิดของค่าความเที่ยงของแบบทดสอบอิงกลุ่ม แต่พัฒนาการของสูตรอาศัยทฤษฎีการทดสอบต่างกัน จากสูตรดังกล่าวจะเห็นได้ว่าค่าดัชนีความเชื่อถือ คือ "อัตราส่วนของความแปรรวนของคะแนนเอกภพของสิ่งที่วัด [$\hat{\sigma}^2(\pi)$] และค่ากำลังสองสองของความแตกต่างของ $(\mu-\lambda)^2$ กับค่าดังกล่าวแล้วทั้งหมดและค่าความแปรปรวนสัมบูรณ์ $\hat{\sigma}^2(\Delta)$]" ค่า $\Phi(\lambda_o)$ จะมีค่าน้อยที่สุดเมื่อคะแนนจุดตัด (λ_o) เท่ากับค่าเฉลี่ยของการสอบ และจะมีค่าเท่ากับค่า KR_{21} แต่เมื่อความแตกต่างของค่าทั้งสองมากขึ้นหรือน้อยลง ค่า $\Phi(\lambda_o)$ จะเพิ่มขึ้น ส่วนค่า $\hat{\sigma}^2(\Delta)$ จะคงเดิม การที่ค่า $\Phi(\lambda_o)$ เพิ่มขึ้นเมื่อคะแนนจุดตัดต่างจากคะแนนเฉลี่ยแสดงว่า ผู้สอบมีคะแนนต่างจากค่าคะแนนเฉลี่ยยิ่งมากก็ยิ่งสามารถถูกแบ่งให้อยู่ในกลุ่มผู้รู้แล้ว (mastered group) หรือกลุ่มผู้ยังไม่รู้ (non-mastered group) ได้ถูกต้องมากยิ่งขึ้น ส่วน $\hat{\sigma}^2(\Delta)$ มีค่าคงเดิมเพราะเป็นค่าความคลาดเคลื่อนที่ไม่ได้รับผลกระทบจากคะแนนจุดตัด (Brennan and Kane 1977 : 281)

ส่วนสูตรที่ 2 เป็นค่าความเที่ยงตามแนวคิดของทฤษฎีการอ้างอิงสรุปที่อาศัยค่า $\hat{\sigma}^2(\pi)$ และ $\hat{\sigma}^2(\delta)$ เท่านั้น และจะมีค่าเท่ากับค่า KR_{20} หรือค่า Cronbach α (Brennan and Kane 1977 : 280) และค่าดังกล่าวนี้เรียกว่าค่าสัมประสิทธิ์การอ้างอิงสรุป

ง. ความหมายของค่าสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือ

ขนาดของค่าสัมประสิทธิ์การอ้างอิงสรุปแสดงถึงระดับความเที่ยงในการอ้างอิงสรุปผลการวัดจากกลุ่มตัวอย่างของปัจจัยประกอบที่ต้องการอ้างอิงผลการวัดไปยังเอกภพของปัจจัยประกอบนั้น ค่าสัมประสิทธิ์การอ้างอิงสรุปขึ้นอยู่กับค่าความคลาดเคลื่อนในการวัด กล่าวคือ ถ้าการวัดมีความคลาดเคลื่อนสูงสัมประสิทธิ์การอ้างอิงสรุปจะมีค่าต่ำ และในทางกลับกัน ถ้าค่าความคลาดเคลื่อนในการวัดต่ำค่าสัมประสิทธิ์การอ้างอิงสรุปจะมีค่าสูง โดยปกติค่าสัมประสิทธิ์การอ้างอิงสรุปจะมีค่าตั้งแต่ 0 ถึง 1

จากทฤษฎีการอ้างอิงสรุป ค่าสัมประสิทธิ์การอ้างอิงสรุปมีความหมายดังต่อไปนี้

1. ค่าสัมประสิทธิ์การอ้างอิงสรุปเป็นดัชนีที่สามารถอธิบายความแม่นยำของการวัดเช่นเดียวกับค่าสัมประสิทธิ์ความเที่ยงของทฤษฎีการทดสอบแบบประเพณีนิยม (Brennan 1983 :5) สามารถใช้คำนวณช่วงความเที่ยงของคะแนนเอกภพ หรือใช้ในสมการถดถอยในการประมาณค่าคะแนนเอกภพ และใช้ในการปรับแก้ค่าสหสัมพันธ์ที่ลดลงอันเนื่องจากความคลาดเคลื่อนได้

2.ค่าสัมประสิทธิ์การอ้างอิงสรุปเป็นค่าประมาณของค่าเฉลี่ยของสหสัมพันธ์ระหว่างค่าการวัดที่สุ่มมาจากเอกภพรายคู่ เช่น ค่าสัมประสิทธิ์การอ้างอิงสรุปเมื่ออ้างไปยังชุดข้อทดสอบ(แบบทดสอบ) ซึ่งประกอบด้วยข้อทดสอบ 20 ข้อ มีค่าเป็น 0.83 หมายความว่า ถ้าเราสุ่มนักเรียนจากประชากรหนึ่ง สมมุติเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ของอำเภอหนึ่งมาทำการทดสอบ สุ่มแบบทดสอบมาทีละฉบับ ๆ ละ 20 ข้อที่ไม่ซ้ำกันเลย ค่าเฉลี่ยของสัมประสิทธิ์สหสัมพันธ์ระหว่างแบบทดสอบที่สุ่มมาจะมีค่าเป็น 0.83

3. ค่าสัมประสิทธิ์การอ้างอิงสรุปเป็นค่ากำลังสองของค่าสหสัมพันธ์ระหว่างคะแนนเอกภพกับคะแนนสังเกต (Brennan and Kane, 1979 :40) เช่น ค่าสัมประสิทธิ์การอ้างอิงสรุปของแบบทดสอบคณิตศาสตร์เรื่องสมการของชั้นประถมศึกษาปีที่ 6 ยกกำลังสองมีค่าเป็น 0.90 ถ้าถอดรากที่ 2 จะได้ค่าสหสัมพันธ์ระหว่างคะแนนเอกภพเรื่องสมการกับคะแนนสังเกต

4. ค่าสัมประสิทธิ์การอ้างอิงสรุป สามารถอธิบายในรูปอัตราส่วนระหว่างความแปรปรวนของคะแนนเอกภพกับคะแนนสังเกต (Brennan 1983 :5) เช่น ค่าสัมประสิทธิ์การอ้างอิงสรุปเป็น 0.90 แสดงว่าความแตกต่างที่วัดได้ร้อยละ 90 เป็นความแตกต่างเนื่องมาจากคะแนนเอกภพ อีกเพียงร้อยละ 10 เป็นความแตกต่างเนื่องมาจากความคลาดเคลื่อน

5. ข้อแตกต่างระหว่างค่าความเที่ยงกับค่าสัมประสิทธิ์การอ้างอิงสรุป ถึงแม้ว่าค่าสัมประสิทธิ์การอ้างอิงสรุปจะมีความหมายเช่นเดียวกับความเที่ยงตามทฤษฎีทดสอบแบบประเพณีนิยม แต่ก็มีประเด็นที่ต่างกันดังนี้ คือ

5.1 การวัดแต่ละครั้งมีค่าสัมประสิทธิ์การอ้างอิงสรุปได้มากกว่า 1 ค่า ขึ้นอยู่กับเอกภพของการอ้างอิงและความสัมพันธ์ของปัจจัยประกอบในแต่ละแบบการวัด

5.2 การอ้างอิงไปยังเอกภพใดจะต้องระบุและอธิบายเอกภพนั้นให้ชัดเจนและต้องสุ่มเงื่อนไขนั้นมาศึกษาด้วย เช่น ถ้าสุ่มเวลามาก็อ้างอิงไปยังเอกภพของเวลา ถ้าสุ่มนักเรียนมาก็อ้างอิงไปยังเอกภพของนักเรียน หรือสุ่มแบบทดสอบมา ก็อ้างอิงไปยังเอกภพของแบบทดสอบ

5.3 ค่าสัมประสิทธิ์การอ้างอิงสรุปสามารถบอกถึงความเป็นเอกพันธ์ (Homogeneity) ของเอกภพได้ด้วย ซึ่งหมายถึงเป็นตัวแทนความเหมือนกันของเอกภพนั้น

จ. การปรับแบบการวัดเพื่อให้เกิดประโยชน์สูงสุด

การปรับปรุงแบบการวัดเพื่อให้เกิดประโยชน์สูงสุดควรกระทำเมื่อผลการวัดครั้งแรกให้ค่าความเที่ยงในการสรุปอ้างอิงต่ำ หรือยังไม่เป็นที่พอใจ หรือความคลาดเคลื่อนในการวัดมีค่าสูง ทั้งนี้เพื่อลดความคลาดเคลื่อนให้น้อยลง รวมทั้งเพื่อปรับปรุงแก้ไขความตรง (validity) และความเที่ยงในการวัด โดยการกำจัดความลำเอียงในการวัด ซึ่งสามารถทำได้ในลักษณะต่าง ๆ กัน เช่น

1. เพิ่มระดับของปัจจัยประกอบที่ต้องการสรุปอ้างอิงผลการวัดให้มีจำนวนเพิ่มขึ้น จากตัวอย่างที่ผ่านมาหากเพิ่มจำนวนข้อทดสอบและจำนวนผู้ตรวจ ค่าสัมประสิทธิ์การอ้างอิงสรุปก็จะเพิ่มขึ้นด้วย

2. เปลี่ยนวิธีการสุ่มปัจจัยประกอบของสิ่งที่ถูกวัด คือการปรับเปลี่ยนแบบการวัดนั่นเอง ก็จะได้ค่าสัมประสิทธิ์ที่เหมาะสมยิ่งขึ้น

3. นิยามประชากรของสิ่งที่ถูกวัดใหม่ หรือการนิยามเอกภพของการสรุปอ้างอิงใหม่ วิธีการนี้รวมถึงการเลือกปัจจัยประกอบใหม่มาศึกษาโดยการเก็บรวบรวมข้อมูลใหม่ หรือการตัดปัจจัยประกอบบางอย่างออกไป

2. การประยุกต์ทฤษฎีการอ้างอิงสรุปใช้ในการวัดผลการศึกษา

ก. งานวิจัยที่นำทฤษฎีการอ้างอิงสรุปไปใช้

ในปี 1983 Macready (1983 : 149-157) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุปในการประเมินความยากและความเป็นเอกพันธ์ (homogeneity) ของข้อทดสอบอิงปริเขต (domain-referenced items)ที่ใช้ในการวินิจฉัย โดยมีปัจจัยประกอบที่ศึกษาประกอบด้วยปริเขตการคูณจำนวนเต็ม (d) ห้องเรียน (c) จำนวนหลักของตัวคูณ (n) ข้อทดสอบซึ่งแฝงอยู่ในปริเขตและจำนวนหลัก (i:(dxn)) และนักเรียนซึ่งแฝงอยู่ในชั้นเรียน (s:c) จากแบบการวิเคราะห์ G Study แบบ (s:c)x(i:(dxn)) โดยให้ห้องเรียน นักเรียน และข้อทดสอบเป็นปัจจัยประกอบแบบจำลองสุ่ม ให้ปริเขตและจำนวนหลักของตัวคูณเป็นปัจจัยประกอบแบบแบบจำลองคงที่ เขาพบว่าแบบ (s:c)x(i:(dx n)),e มีความแปรปรวนที่มีค่ามากที่สุด คือ 47 % อีก 4 แหล่งมีค่ารองลงไป ได้แก่ ห้องเรียน นักเรียนซึ่งแฝงอยู่ในห้องเรียน ปริเขต และผลร่วมระหว่างนักเรียนกับปริเขต คือ (s:c) x d โดยทั้ง 4 แหล่งมีค่ารวมกันถึง 51 % เฉพาะแหล่งความแปรปรวนสุดท้ายหมายถึงความยากของข้อทดสอบในแต่ละปริเขตสำหรับนักเรียนแต่ละคนมีค่าไม่เท่ากัน แหล่งความแปรปรวนอื่น ๆ ที่เหลือมีค่าน้อยมาก โดยเฉพาะจำนวนหลักของตัวคูณ Macready สรุปว่าไม่ควรจะใช้ตัวคูณ 4 ตำแหน่ง ใช้เพียง 3 ตำแหน่งก็พอ เพราะให้ค่าความยากไม่แตกต่างกัน อีกแหล่งหนึ่งที่มีค่าน้อยคือแบบ i:(dxn) แสดงว่าความยากของข้อทดสอบในทุกปริเขตที่มีตำแหน่งตัวคูณเท่ากันมีค่าเท่า ๆ กัน เมื่อตรวจดูความยากรายปริเขต พบว่าเกือบทุกปริเขตมีข้อทดสอบที่มีความยากเท่าเทียมกัน มีเพียงปริเขตที่ 15 ที่ค่อนข้างจะแตกต่างกัน Macready เสนอว่าควรจะแยกให้เป็นปริเขตย่อย หรือนำไปรวมกับปริเขตอื่น จากการประมาณค่าสัมประสิทธิ์การอ้างอิงสรุปของข้อทดสอบแต่ละปริเขต โดยใช้แบบการวิเคราะห์ 2 แบบ คือ sxi และ exr เมื่อ r หมายถึงการสอบซ้ำพบว่า ค่าสัมประสิทธิ์การอ้างอิงสรุปของข้อทดสอบหนึ่งข้อมีค่าอยู่ระหว่าง 0.338 ถึง 0.606

ต่อมาในปี 1984 Ibrahim (1984 :499-A) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุป เพื่อประมาณค่าความแปรปรวนที่มีต่อการประเมินวัตถุประสงค์ทางการศึกษา (The Rating of Evaluational Goals) โดยสุ่มตัวอย่างครู 80 คน และนักศึกษา 80 คน ในประเทศซูดาน ประเมินวัตถุประสงค์ทางการศึกษา 2 ชนิด คือ วัตถุประสงค์ที่สำคัญจริง ๆ และวัตถุประสงค์ตามที่คาดหวัง ปัจจัยประกอบในการศึกษามีผู้ประเมิน กลุ่มผู้ประเมิน จำนวนครั้งของการประเมิน ถิ่นที่อยู่ของผู้ประเมิน ชนิดของวัตถุประสงค์ สถานที่ทำงานของผู้ประเมิน และเพศของผู้ประเมิน พบว่า ปัจจัยประกอบที่มีผลต่อการประเมินมากที่สุด ได้แก่ ผู้ประเมินและกลุ่มผู้ประเมิน ส่วนสถานที่ทำงานของผู้ประเมิน มีผลเล็กน้อยเท่านั้น ส่วนปัจจัยประกอบอื่นที่เหลือไม่มีผลต่อการประเมินเลย

.ในอีก 2 ปีต่อมา O'Brien (1986 :21-27) ใช้ทฤษฎีการอ้างอิงสรุปในการประมาณค่าความเที่ยงของตัวแปรระดับโรงเรียน 16 ตัว ซึ่งเป็นค่าเฉลี่ยหรือร้อยละของกลุ่มตัวอย่างนักเรียน ตัวแปรแบ่งเป็น 5 กลุ่ม คือ ค่าเฉลี่ยของสถานภาพทางสังคมและเศรษฐกิจของครอบครัว ค่าเฉลี่ยของผลสัมฤทธิ์ทางการเรียน ร้อยละของนักเรียนที่มีบิดาหรือมารดาอาศัยอยู่ด้วย ร้อยละของคุณภาพของห้องสมุดและการเรียนการสอนของโรงเรียน และร้อยละของนักเรียนที่ยอมรับกฎระเบียบของโรงเรียนกลุ่มตัวอย่าง คือ โรงเรียน 1,122 โรง และนักเรียนชั้นปีที่ 2 โรงเรียนละ 36 คน ดำเนินการวิจัยโดยให้กลุ่มตัวอย่างตอบคำถามของตัวแปรแต่ละตัว จากการศึกษาพบว่า ความเที่ยงในการตอบแบบทดสอบถามเพิ่มขึ้นตามการเพิ่มจำนวนผู้ตอบคำถามจากแต่ละโรงเรียน จำนวนข้อคำถาม และจำนวนโรงเรียน ข้อค้นพบนี้แสดงให้เห็นว่า จำนวนผู้ประเมิน จำนวนข้อคำถามของเครื่องมือประเมินและจำนวนสถานที่ทำงานของผู้ประเมินมีผลต่อความเที่ยงของการประเมิน

ในปี 1987 Webb, Herman and Cabello (1987 : 130) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุปในการวัดแบบอิงปริเขต (domain-referenced test) เพื่อการสอบแบบวินิจฉัยในการวัดด้านภาษาเรื่องสรรพนาม โดยได้เสนอวิธีการ

พัฒนาและวิเคราะห์การใช้ประโยชน์ของแบบทดสอบวินิจฉัยสำหรับครูด้วยการเชื่อมโยงระหว่างการวัดผลแบบอิงปริเขตและทฤษฎีการอ้างอิงสรุป เพื่อศึกษาว่าความสามารถด้านเนื้อหากลุ่มใดควรนำเสนอในสัณฐานคะแนน (profile) ของนักเรียน การศึกษาแบ่งเป็น 4 ขั้น คือ

(1) กำหนดลักษณะเฉพาะของปริเขตและสร้างแบบทดสอบ

(2) เลือกกลุ่มเนื้อหาที่ควรเสนอในสัณฐานคะแนน

(3) กำหนดจำนวนข้อทดสอบที่จำเป็นต่อความเที่ยงของการวัด และ

(4) คำนวณค่าความแม่นยำของสัณฐานคะแนน ในขั้นที่ 2 เป็นขั้นที่เริ่มใช้ทฤษฎีการอ้างอิงสรุป แบบที่ใช้วิเคราะห์ใน G Study คือ s x i และใช้ความคลาดเคลื่อนแบบสัมบูรณ์คำนวณค่าสัมประสิทธิ์การอ้างอิงสรุป ทั้งแบบตัวแปรเดียว (univariate) และแบบตัวแปรพหุ (multivariate) ปัจจัยประกอบที่ใช้ศึกษาประกอบด้วยกลุ่มเนื้อหาย่อยของเรื่องสรรพนาม มี (1) rule ได้แก่ nominative, direct object, indirect object of preposition (2) form ได้แก่ relative, nonrelative (3) number ได้แก่ singular, plural (4) embeddedness ได้แก่ sentence, paragraph แต่ละปัจจัยประกอบ"ผสม" (cross) กัน แต่ข้อทดสอบ (i) แฝงอยู่ในปัจจัยประกอบอื่น สิ่งที่ถูกวัดคือ นักเรียน ซึ่งสุ่มมาจากนักเรียนเกรด 6 จำนวน 128 คน ใช้เนื้อหาเป็นปัจจัยประกอบแบบแบบจำลองคงที่ แต่ข้อทดสอบเป็นปัจจัยประกอบแบบจำลองสุ่ม ผลการวิเคราะห์ความแปรปรวนพบว่า มี 3 ปัจจัยประกอบที่มีผลต่อคะแนนมากที่สุด คือ form, embeddedness และ rule มีเพียงบางปัจจัยประกอบมีความสัมพันธ์กับความแตกต่างระหว่างนักเรียน ปัจจัยประกอบใดที่ไม่มีอันตรกริยากับนักเรียน แสดงว่ามีผลคงที่สำหรับนักเรียนทุก ๆ คน ได้แก่ ผลหลักของ form ผลร่วมกันระหว่าง form-embeddedness และ rule-form ผลร่วมระหว่าง F-E แสดงว่าแต่ละ form ของ pronoun ในแต่ละ embeddednes ข้อทดสอบมีความยากไม่เท่ากัน แต่ก็ยังถือว่าผลดังกล่าวมีค่าคงเส้นคงวาในนักเรียนที่ทำการทดสอบทุกคน คือ ทุกคนตอบถูกหรือผิดคล้อยตามกันดังนั้น ถ้าต้องการนำเสนอสัณฐานคะแนนก็ให้อยู่ในรูปสัณฐานของกลุ่ม (group profile) ไม่จำเป็นต้องเสนอรายบุคคล (individuals' profile) ผลของปัจจัยประกอบที่สัมพันธ์กับความแตกต่างระหว่างบุคคล Webb, และคณะ แนะว่า ควรนำเสนอในรูปสัณฐานคะแนนรายบุคคล ได้แก่ rule และ form ดังนั้น จากกลุ่มเนื้อหาทั้งหมด 32 กลุ่ม ควรจะนำเสนอในรูปสัณฐานคะแนนรายบุคคลเพียง 8 กลุ่มเท่านั้น จำนวนข้อทดสอบที่เหมาะสมในแต่ละกลุ่มเป็น 8 ข้อ และพบว่าค่าE ของข้อทดสอบ 8 ข้อ สำหรับเนื้อหา 8 กลุ่ม มีค่าอยู่ระหว่าง 0.35 - 0.75

ในประเทศไทยยังมีการประยุกต์ทฤษฎีการอ้างอิงสรุปในการวิจัยน้อยมาก แต่ก็ได้มีการเสนอบทความทางวิชาการเกี่ยวกับทฤษฎีนี้บ้าง เช่น บทความของ จักกฤษณ์ สำราญใจ (2529 : 36-47) สำหรับงานวิจัยที่ประยุกต์ทฤษฎีนี้กับการประเมินผลนั้น แดง กลางท่าได่ (2531) ได้ประยุกต์ทฤษฎีการอ้างอิงสรุปในการหาความเที่ยง (reliability) ของการประเมินความตรงเชิงเนื้อหา โดยให้ผู้เชี่ยวชาญทุกคนประเมินความตรง(ความเหมาะสม)ของข้อทดสอบทุกข้อว่า วัดตรงจุดประสงค์เพียงใด ผู้เชี่ยวชาญแฝง (nested) อยู่ในโรงเรียน และแบบการศึกษา G Study เป็นแบบ i x (r:s) เมื่อ i คือ ปัจจัยประกอบด้านข้อทดสอบ r คือ ปัจจัยประกอบด้านผู้ตรวจ และ s คือปัจจัยประกอบด้านโรงเรียน ผลการศึกษาพบว่า แหล่งความแปรปรวนที่มีอิทธิพลต่อความเที่ยง ได้แก่ ความแปรปรวนของข้อทดสอบ ความแปรปรวนของผู้เชี่ยวชาญซึ่งอยู่ในโรงเรียน และผลร่วมระหว่างข้อทดสอบกับโรงเรียน และขนาดของกลุ่มตัวอย่างที่จะเป็นตัวแทนของสมาชิกทั้งหมดในเอกภพของปัจจัยประกอบ ได้แก่ข้อทดสอบอย่างน้อย 9 ข้อ และผู้เชี่ยวชาญไม่เกิน 9 คนต่อโรงเรียน ซึ่งสุ่มจากโรงเรียนอย่างน้อย 7 โรงเรียน จึงจะทำให้สัมประสิทธิ์การอ้างอิงสรุปมีค่าอย่างน้อยเป็น 0.80

ดังนั้น จะเห็นว่าทฤษฎีการอ้างอิงสรุปสามารถนำไปประยุกต์เข้ากับการวัดและประเมินผลในหลาย ๆ สถานการณ์ได้ดี นับตั้งแต่การพัฒนาข้อทดสอบจนถึงการตีความคะแนนในแบบต่าง ๆ

ข. ตัวอย่างการประยุกต์ใช้ทฤษฎีการอ้างอิงสรุป

สมมุติว่าเราต้องการสร้างแบบทดสอบเพื่อวัดความเข้าใจในการอ่าน เราอาจออกแบบวิธีการวัดให้มีรายละเอียดดังนี้ (จักรกฤษณ์ สำราญใจ, 2529)

1. เลือกบทความที่มีความยาวเท่ากัน 3 เรื่อง ซึ่งเขียนโดยผู้เขียน 3 คน
2. นำบทความที่ได้มาสร้างเป็นแบบ cloze test โดยตัดทุก ๆ คำที่ 7 ออก แล้วให้ผู้ตอบเติมคำที่ถูกต้อง
3. นำแบบทดสอบนี้ไปสอบนักเรียน 100 คนจำนวน 2 ครั้ง

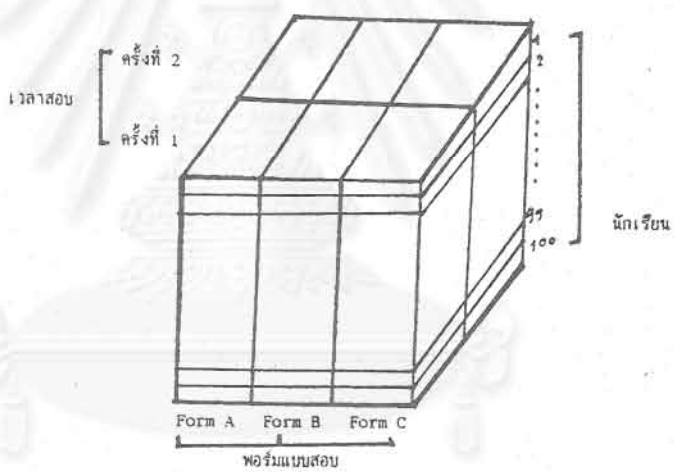
ในการประยุกต์ใช้ทฤษฎีอ้างอิงสรุปจากแบบข้างต้นนี้จะมี 3 ปัจจัยประกอบคือ

(1) ผู้ตอบคือนักเรียน $n_s = 100$ (จำนวนนักเรียนทั้งหมด 100 คน)

(2) ชุดของแบบทดสอบ $n_f = 3$ (แบบทดสอบ 3 ชุด)

(3) เวลาในการสอบ $n_m = 2$ (ครั้ง 1, ครั้ง 2)

แบบการทดลองการใช้แบบทดสอบเพื่อวัดความเข้าใจในการอ่าน 3 ชุด 2 ครั้งจำนวนผู้สอบ 100 คน จะมีรายละเอียดดังภาพข้างล่างนี้



แผนภูมิที่ 2: แบบการทดลองใช้แบบทดสอบวัดความเข้าใจการอ่าน

เมื่อทำการวิเคราะห์ความแปรปรวนของข้อมูลจากแหล่งต่าง ๆ จากแบบการทดลองดังกล่าวจะได้ผลการทดลองโดยการวิเคราะห์ความแปรปรวนจากแหล่งต่าง ๆ ดังต่อไปนี้

1. ผลของผู้สอบ คือนักเรียน (S)
2. ผลของชุดแบบทดสอบ (F)
3. ผลของเวลา (M)
4. ผลร่วมของนักเรียนกับชุดแบบทดสอบ (S x F)
5. ผลร่วมของนักเรียนกับเวลา (S x M)
6. ผลร่วมของชุดแบบทดสอบกับเวลา (F x M)
7. ผลร่วมของนักเรียน ชุดแบบทดสอบ และเวลา (S x F x M)

การกำหนดเอกภพของการอ้างอิงสรุป (universe of generalization) และการกำหนดปัจจัยประกอบในการอ้างอิง (facet of generalization) รวมทั้งกำหนดปัจจัยประกอบที่ต่างกัน(facet of differentation) นั้น การ

กำหนดจะขึ้นอยู่กับปัญหาของผู้ศึกษา สำหรับกรณีตัวอย่างนี้ปัญหาที่เป็นไปได้ทั้งหมดสามารถกำหนดได้ 12 แบบ ดังตารางแสดงต่อไปนี้

ตารางที่ 1: แสดงรายละเอียดของลักษณะของปัญหาที่เป็นไปได้

model ที่	facet of diff	facet of gen rand. fix	ลักษณะปัญหา	เอกภพของการอ้างอิงสรุป
1	S	F M	ความเข้าใจในการอ่านของนักเรียนในการสอบ 2 ครั้ง สำหรับแบบทดสอบชุดใด ๆ	ชุดของการสอบ
2	S	M F	ความเข้าใจในการอ่านของนักเรียนในแบบทดสอบ 3 ชุดนี้ สำหรับการสอบครั้งใด ๆ	เวลาที่สอบ
3	S	F&M -	ความเข้าใจในการอ่านของนักเรียนในการสอบครั้งใด ๆ และแบบทดสอบชุดใด ๆ	ชุดและเวลาที่สอบ
4	F	S M	ความยากของแบบทดสอบ 3 ชุดในการสอบ 2 ครั้งนี้ สำหรับนักเรียนคนใด ๆ	นักเรียน
5	F	M S	ความยากของแบบทดสอบ 3 ชุดนี้ สำหรับนักเรียนกลุ่มนี้กับการสอบครั้งใด ๆ	เวลาที่สอบ
6	F	S&M -	ความยากของแบบทดสอบ 3 ชุดนี้สำหรับนักเรียนคนใด ๆ และการสอบครั้งใด ๆ	นักเรียนและเวลาที่สอบ
7	M	S F	ผลของการสอบอ่านเกี่ยวกับแบบทดสอบ 3 ชุดนี้ สำหรับนักเรียนคนใด ๆ	นักเรียน
8	M	F S	ผลของการสอบอ่านสำหรับนักเรียนกลุ่มนี้ในการเข้าใจการอ่านบทความใด ๆ	ชุดของแบบทดสอบ
9	M	S&F -	ผลของการสอบอ่านสำหรับนักเรียนคนใดๆ ในการเข้าใจในการอ่านบทความใด ๆ	นักเรียนและชุดของแบบทดสอบ
10	S&F	M -	ความเข้าใจในการอ่านบทความในแต่ละบทของนักเรียนกลุ่มนี้ในการสอบครั้งใด ๆ	เวลาที่สอบ
11	S&M	F -	ความเข้าใจในการอ่านของนักเรียนกลุ่มนี้ในการสอบแต่ละครั้งใน 2 ครั้งนี้ ต่อบทความใด ๆ	ชุดของแบบทดสอบ
12	F&M	S -	ความยากของแบบทดสอบทั้ง 3 ชุดในการสอบแต่ละครั้งใน 2 ครั้ง สำหรับนักเรียนคนใด ๆ	นักเรียน

3. การสอบแบบเรียงความ

ก.นิยามของแบบหรือข้อทดสอบเรียงความ

Stalnaker (อ้างถึงใน Coffman 1971 : 271) ให้นิยามข้อทดสอบเรียงความว่า หมายถึง ข้อคำถามซึ่งต้องการให้ผู้ตอบเขียนคำตอบขึ้นเอง โดยธรรมชาติเป็นคำตอบที่ไม่ใช่จะถูกต้องเพียงคำตอบเดียว หรือแบบเดียว

ความถูกต้องและคุณภาพของคำตอบต้องได้รับการตัดสินโดยผู้มีความรู้และทักษะในเนื้อหาที่ถามเป็นอย่างดี ลักษณะเด่นของข้อทดสอบเรียงความ คือ การให้อิสระแก่ผู้ตอบในการแสดงความคิด และคำตอบถูกไม่ใช่มีเพียงคำตอบเดียว ไม่สามารถตรวจโดยวิธีตรวจนับอย่างง่าย แม้แต่ผู้เชี่ยวชาญก็ไม่อาจชี้ถูกหรือผิดได้อย่างเด็ดขาด เพียงแต่สามารถจะพิจารณาถึงระดับ (degree) ของคุณภาพคำตอบได้

Coffman (1971 : 271) ได้ให้นิยามการสอบเรียงความ (essay examination) ว่า หมายถึงการสอบที่ใช้ข้อทดสอบแบบเรียงความ ตั้งแต่หนึ่งข้อขึ้นไปกับกลุ่มนักเรียน ภายใต้สถานการณ์อย่างหนึ่ง เพื่อเก็บรวบรวมข้อมูลในการประเมินผล คำถามแบบเรียงความแตกต่างจากคำถามแบบตอบสั้น ๆ ตรงที่ต้องใช้ผู้เชี่ยวชาญตัดสินไม่ใช่การตรวจอย่างง่ายโดยใช้ตัวเฉลยแบบเดียวกับแบบทดสอบปรนัย แตกต่างจากรายงานที่ทำส่งครูในการเรียนวิชาต่างๆ ตรงที่การดำเนินการสอบ จะต้องดำเนินให้เป็นมาตรฐานเดียวกัน และต้องการข้อทดสอบที่เป็นตัวแทนผลการสอบที่ต้องการ

สำหรับ Tuckman (อ้างถึงใน เยาวดี วิบูลย์ศรี 2528 : 122) ให้ความหมายของแบบทดสอบเรียงความว่าเป็นแบบทดสอบที่ให้ผู้สอบได้แสดงความสามารถในการประยุกต์ความรู้ การวิเคราะห์ และประเมินผลความรู้ ดังนั้นข้อทดสอบของแบบทดสอบเรียงความต้องเป็นข้อคำถามที่ให้โอกาสผู้สอบได้สร้างและเรียบเรียงคำตอบในรูปบูรณาการความรู้ตามขอบข่ายความรู้ที่กว้าง ทั้งนี้ เพื่อให้ผู้สอบเรียงความสามารถวัดกระบวนการคิดในระดับสูงตามแนวคิดเรื่องสารบบวิชาณ (taxonomy) ของ Bloom และคณะ (Bloom and Others 1965) ได้ คือ ระดับการเรียนรู้ขั้นวิเคราะห์ สังเคราะห์ และการประเมินผลเป็นส่วนใหญ่ หรือวัดในระดับการนำไปใช้แก้ปัญหาในสถานการณ์ต่าง ๆ เป็นต้น

จากนิยามต่าง ๆที่กล่าวมาจะเห็นได้ว่า ข้อทดสอบแบบเรียงความเป็นข้อคำถามที่ผู้ตอบจะต้องเขียนคำตอบด้วยตนเอง ผู้ตอบมีอิสระในการใช้ความสามารถด้านการเขียน เรียบเรียง ประยุกต์ความรู้ ความคิดได้กว้างขวาง คำตอบที่ถูกต้องจะมีหลายคำตอบ ผู้ที่จะตัดสินคุณภาพคำตอบจะต้องเป็นผู้เชี่ยวชาญในเรื่องที่ถามเป็นอย่างดี ดังนั้นข้อทดสอบแบบเรียงความจึงเหมาะที่จะใช้วัดพฤติกรรมความรู้ความคิดที่มีความซับซ้อน เช่น การสังเคราะห์และประเมินผล เป็นต้น

ข. ความเที่ยงของแบบทดสอบเรียงความ

ปัญหาที่สำคัญของการทดสอบแบบเรียงความ คือ ความเที่ยงของแบบทดสอบต่ำ การที่แบบทดสอบมีความเที่ยงต่ำแสดงว่า ค่าที่วัดได้มีส่วนผสมที่เป็นคะแนนจริงน้อย แต่มีความคลาดเคลื่อนมาก (Coffman 1971 : 276) งานวิจัยส่วนมากหาความเที่ยงของแบบทดสอบเรียงความ โดยการหาค่าสหสัมพันธ์ระหว่างผู้ตรวจหลายคนหรือผู้ตรวจคนเดียวกันแต่ตรวจต่างโอกาส ดังนั้น ค่าสหสัมพันธ์ที่ได้จึงเป็นค่าที่แสดงความสอดคล้องกัน (agree) ระหว่างผู้ตรวจ ไม่ใช่ความเที่ยงของเครื่องมือโดยตรง ในหัวข้อนี้มีนักวัดผลได้ให้ความคิดเห็น ตลอดทั้งผลงานวิจัยดังนี้

Coffman (1972 : 277) กล่าวว่า ปัญหาความสอดคล้องระหว่างผู้ตรวจเป็นเรื่องค่อนข้างซับซ้อน เกิดจากสาเหตุสำคัญที่ไม่สามารถหลีกเลี่ยงได้ 3 ประการ คือ

1) ผู้ตรวจต่างคนมีแนวโน้มจะให้คะแนนเรียงความชิ้นเดียวกันแตกต่างกัน

2) ผู้ตรวจคนเดียวกันมีแนวโน้มจะให้คะแนนเรียงความชิ้นเดียวกันแตกต่างกันตามโอกาสที่ตรวจ

3) ความไม่สอดคล้องกันจะยิ่งมากขึ้น เมื่อข้อทดสอบเปิดโอกาสให้ผู้ตอบมีอิสระในการตอบมากขึ้น สาเหตุของความไม่สอดคล้องแต่ละด้านเป็นเรื่องที่ค่อนข้างซับซ้อน ความแตกต่างของคะแนนแต่ละคนอาจได้รับอิทธิพลต่อไปนี้

ก. ผู้ตรวจแต่ละคนมีความเข้มงวดแตกต่างกัน เช่น บางคนชอบให้คะแนนมาก แต่บางคนชอบให้คะแนนน้อย

ข. การกระจายของคะแนนหรือพิสัยของคะแนนที่ได้รับจากผู้ตรวจแต่ละคนมีความแตกต่างกัน บางคนให้คะแนนใกล้ ๆ กับค่าเฉลี่ย บางคนให้คะแนนมีพิสัยกว้างคือ เกือบเต็มและเกือบเป็นศูนย์ เป็นต้น

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

นอกจากนี้ ในปี 1951 Finlayson (1951 : 126-134) ได้ศึกษาความเที่ยงของแบบทดสอบเรียงความโดยผู้ตรวจ 4 คน ตรวจเรียงความ 2 ชุด พบว่า มีค่าอยู่ระหว่าง 0.636 - 0.957 โดยมีค่าเฉลี่ยเป็น 0.810 แต่เมื่อเทียบกับการประเมินครั้งแรกโดยผู้ตรวจชุดเดียวกัน พบว่า ค่าสหสัมพันธ์อยู่ระหว่าง 0.610 -0.798 ค่าเฉลี่ยเป็น 0.687 ค่าสหสัมพันธ์ระหว่างผู้ตรวจรายคู่อยู่ระหว่าง 0.591 ถึง 0.770 ค่าเฉลี่ยเป็น 0.703 จะเห็นว่า ค่าความเที่ยงของเรียงความเรื่องเดียวกันโดยผู้ตรวจชุดเดียวกัน มีค่าแตกต่างกันไปเมื่อโอกาสการตรวจเปลี่ยนไป

ในปี 1971 Coffman (1971 : 278) ได้ตรวจสอบเอกสารงานวิจัยที่เกี่ยวกับแบบทดสอบเรียงความ พบค่าความเที่ยงอยู่ระหว่าง 0.35 - 0.98 เขาได้ให้ความเห็นเชิงวิจารณ์ไว้ว่า การตัดสินว่าความเที่ยงของแบบทดสอบหรือข้อคำถามแบบเรียงความควรมีค่าเป็นเท่าไรเป็นเรื่องที่ตอบยาก เพราะค่าความเที่ยงจะเปลี่ยนไปตามปัจจัยที่เกี่ยวข้อง ถ้าประเมินกลุ่มนักเรียนขนาดใหญ่ที่มีพื้นฐานการเรียนการสอนแตกต่างกันมาก ความเที่ยงของการตรวจจะต่ำ แต่จะมีค่าสูงขึ้นเมื่อตรวจกลุ่มเล็กที่มีพื้นความรู้ใกล้เคียงกัน ถ้าผู้ตรวจมีความแตกต่างกันมากความเที่ยงก็จะต่ำ นอกจากนี้เนื้อหาวิชาเป็นอีกปัจจัยหนึ่ง วิชาใดมีเนื้อหาเป็นกฎเกณฑ์ตายตัว เช่น คณิตศาสตร์ เคมี จะมีความเที่ยงสูงกว่าวิชาภาษา

อนึ่ง นอกจากผู้สอบ ผู้ตรวจ และเนื้อหาวิชาแล้ว De Gruijter (1980 : 245-261) ยังพบว่าลักษณะของคำถามก็เป็นอีกปัจจัยหนึ่งที่มีอิทธิพลต่อความเที่ยง เพราะคำว่า "คำถาม" ในการสอบแบบเรียงความ มีความหมายรวมไปถึงบริบททั้งหมดที่ปรากฏอยู่ในคำถามด้วย เช่น แนวทางที่ต้องการให้ผู้สอบตอบ รวมทั้งแนวทางที่กำหนดในการให้คะแนนของผู้ตรวจดังนั้น การเปลี่ยนแปลงคำชี้แจงในคำถามจะทำให้การตอบและการตรวจเปลี่ยนไป คำถามที่ดีต้องระบุประเด็นที่มุ่งวัดอย่างเด่นชัด การปล่อยให้ผู้ตอบคิดหาแนวทางตอบโดยอิสระมากเท่าใด ความเที่ยงก็จะยิ่งต่ำลงไปด้วย

ต่อมาในปี 1985 Block (1985 : 41-52) ได้ศึกษาผลการประเมินเรียงความ ทั้งโดยวิธีใช้ผู้ตรวจหลายคนและคนเดียวตรวจหลายครั้ง เพื่อดูว่าการตรวจแต่ละครั้งหรือแต่ละคนได้ค่าคะแนนจริงตรงกันหรือไม่ โดยให้ผู้ตรวจ 16 คน ตรวจคำตอบเรียงความ 105 ชิ้น 2 ครั้ง พบว่า ผู้ตรวจคนเดียวกันตรวจต่างกัน 2 โอกาสได้ค่าคะแนนจริงตรงกัน แต่ผู้ตรวจต่างคนจะไม่ได้คะแนนจริงตรงกัน ค่าประมาณของสหสัมพันธ์ระหว่างคะแนนจริงของผู้ตรวจหลายคน มีค่าอยู่ระหว่าง 0.415 - 0.910

โดยทฤษฎีแล้วการเพิ่มข้อคำถามหลายข้อในแบบทดสอบหนึ่ง ๆ ทำให้ความเที่ยงเพิ่มขึ้น แบบทดสอบเรียงความก็เช่นกัน แต่มีประเด็นควรคำนึงอยู่อย่างหนึ่ง คือ ในคำถามข้อใดข้อหนึ่งข้อที่ถามยาวกว่าจะมีความ

เที่ยงสูงกว่า แต่การให้เวลาทำข้อทดสอบมากกว่าไม่ได้ทำให้ความเที่ยงสูงกว่าข้อที่ให้เวลาน้อยกว่า (Coffman 1971 : 280)

นอกเหนือจากปัจจัยดังกล่าวมานั้น ความหลากหลายของเทคนิควิธีทางสถิติที่ใช้ในการคำนวณค่าความเที่ยงของแบบทดสอบเรียงความก็เป็นอีกปัจจัยหนึ่งที่มีผลต่อค่าความเที่ยง งานวิจัยบางเรื่องหาค่าความเที่ยงโดยวิธีหาค่าสหสัมพันธ์แบบเปียร์สัน (Pearson product-moment correlation)ระหว่างคะแนนที่ได้จากการตรวจ 2 ชุด วิธีนี้จะได้ค่าความเที่ยงสูงกว่าความเป็นจริง เพราะว่าค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของคะแนนสองชุดนี้จะถูกเหมาว่ามีความเท่าเทียมกัน ซึ่งจริง ๆ แล้วอาจไม่ใช่ นอกจากนี้ การหาค่าความเที่ยงจากข้อมูลชุดเดียวกันโดยอาศัยสูตรการคำนวณที่ต่างกัน มักจะได้คำตอบที่แตกต่างกันด้วย ทั้งนี้เพราะแต่ละสูตรมีข้อตกลงเบื้องต้นที่แตกต่างกัน

แต่อย่างไรก็ตาม การจะเลือกแบบทดสอบเรียงความไปใช้ โดยพิจารณาที่ความเที่ยงนี้ขึ้นอยู่กับสถานการณ์การสอบวัดและการตัดสินใจในการใช้ผลการประเมิน ถ้าเป็นการตัดสินใจเกี่ยวกับการเรียนการสอนในชั้นเรียน ครูอาจใช้แบบทดสอบหรือข้อคำถามที่มีความเที่ยงต่ำได้ แต่ถ้าเป็นการสอบคัดเลือก ควรใช้แบบทดสอบที่มีความเที่ยงสูง อีกประการหนึ่ง ถ้าเวลาในการสอบและตรวจมีจำกัด ครูควรต้องเลือกเอาอย่างใดอย่างหนึ่ง ระหว่างคำถามหลาย ๆ ข้อเพื่อให้ได้คะแนนที่มีความเที่ยงสูงกับคำถามน้อยข้อแต่สามารถวัดพฤติกรรมการเรียนรู้ขั้นสูงได้ดี(Coffman 1971 : 280-281)

จากที่กล่าวมาแล้วจะเห็นได้ว่า แบบทดสอบเรียงความที่ใช้ทั่วไปมีพิสัยของความเที่ยงกว้าง มีค่าจากค่อนข้างต่ำคือ 0.25 จนถึงสูงถึง 0.98 ทั้งนี้ เนื่องจากธรรมชาติของการวัดแบบนี้ต้องใช้วิจารณญาณส่วนตัวของผู้ตรวจซึ่งถือว่าเป็นผู้เชี่ยวชาญตัดสิน ไม่สามารถทำเฉลยแจงนับถูกผิดได้อย่างสมบูรณ์เช่นเดียวกับข้อทดสอบแบบปรนัย นอกจากนั้นยังมีปัจจัยอื่น ๆ อีกมากมายตามที่ได้กล่าวมาแล้ว แต่อย่างไรก็ตาม นักวัดผลได้พยายามหาวิธีการควบคุมแหล่งความคลาดเคลื่อนเหล่านั้นเพื่อทำให้ความเที่ยงเพิ่มขึ้น ซึ่งสามารถทำได้หลายอย่าง

ค. ข้อเสนอแนะในการตรวจเพื่อให้มีความเที่ยงสูง

เนื่องจากการให้คะแนนเรียงความมีความเป็นอัตนัยสูง ดังนั้นครูควรต้องสามารถหาเหตุผลมาอธิบายสนับสนุนการให้คะแนนของตนให้ได้เมื่อมีผู้ซักถาม ไม่เช่นนั้นผลการประเมินก็จะไร้ความหมาย เมื่อครูสามารถให้เหตุผลได้แสดงว่า ครูมีเกณฑ์การให้คะแนน ผลของการที่มีเกณฑ์นี้เชื่อกันว่าจะทำให้ผลการวัดมีความเที่ยงสูง นักวัดผลหลายคนให้ข้อเสนอแนะในการตรวจเพื่อให้คะแนนมีความเที่ยงไว้หลายประการ เช่น

Bergman (1981 : 130-131) แนะให้ปฏิบัติดังนี้

1) เตรียมคำถามให้ตอบสั้น ๆ พร้อมทั้งเตรียมโครงสร้างคำตอบ หรือตัวแบบคำตอบ และตรวจคำตอบในแต่ละประเด็น ว่ามีหรือไม่มี ถูกหรือไม่ถูก

2) กำหนดเค้าโครงคำตอบที่ต้องการ วิธีนี้ช่วยให้ครูมีเกณฑ์ในการตรวจ บางครั้งอาจจะใช้คำตอบของนักเรียนแทนการเขียนขึ้นเองก็ได้

3) กำหนดประเด็นที่จะให้คะแนนในแต่ละข้อให้เด่นชัด เพื่อทำให้การตรวจมีความยุติธรรม และง่าย

4) ตรวจแยกทีละประเด็น เช่น การสะกด ไวยากรณ์ ลีลาการเขียน ความชัดเจนของแนวคิด สี่ประเด็นแรก เป็นประเด็นสำคัญในการประเมินความสามารถในการเขียน ส่วนแนวคิดเป็นประเด็นสำคัญของการประเมินเรียงความวัดเนื้อหา

5) ตรวจทีละข้อครบทุกคนเพื่อทำให้ครูสามารถกำหนดเกณฑ์ที่เหมาะสมในแต่ละข้อได้ และสามารถขจัดความคลาดเคลื่อนแบบ halo effects ได้

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

7) ตรวจหลายครั้งหรือใช้ผู้ตรวจหลายคนแล้วใช้คะแนนเฉลี่ย เช่น ถ้าครู 2 คนตรวจ ก็อาจจะหาค่าเฉลี่ยของทั้งสองคน หรือให้เป็น 2 เกรด เช่น A/B หรือ C/D เป็นต้น

ส่วน Cochran and Weideman (อ้างถึงใน Hopkins and Stanley 1981 : 223) ได้เสนอวิธีตรวจเรียงความให้มีความเที่ยงสูงขึ้น ซึ่งใช้เวลาฝึกฝนเพียง 10 นาที แต่พบว่าค่าความเที่ยงสูงถึง 0.80 และ 0.90 ดังนี้

1. ก่อนตรวจคำตอบใด ๆ ให้ศึกษาเนื้อหาในหนังสือเรียนที่เกี่ยวข้องกับคำถาม รวมทั้งสมุดโน้ตย่อที่ใช้เรียนในวิชานั้น

2. แจกแจงประเด็นสำคัญที่ควรจะตอบในแต่ละข้อ กำหนดคะแนนแต่ละประเด็นจนครบ ถือเป็นคะแนนขั้นต่ำ (minimum score) ถ้าผู้สอบคนใดตอบเพิ่มเติมนอกเหนือจากที่เตรียมไว้ ควรให้คะแนนเพิ่มเติมอีก เรียกว่า คะแนนพิเศษ (extra score) คะแนนพิเศษจะแตกต่างกันไป แต่ไม่เกินคะแนนสูงสุดที่กำหนดไว้ของข้อนั้น

3. สุ่มงานเขียนมาจำนวนหนึ่งแล้วอ่านแบบผ่าน ๆ เพื่อจะได้กำหนดคุณภาพของคำตอบที่ต้องการได้

4. ตรวจแต่ละข้อจนครบทุกคนแล้วจึงตรวจข้ออื่น ซึ่งทำให้เกิดผลดี 2 ประการ คือ 1.) ทำให้คะแนนที่ได้จากการเปรียบเทียบมีความถูกต้องและยุติธรรมมากขึ้น และ 2.) ทำให้ผู้ตรวจมีตัวแบบในการตรวจเพียงตัวแบบเดียว ทำให้ประหยัดเวลาในการตรวจและมีความถูกต้องแม่นยำมากขึ้น

5. อ่านคำตอบให้จบครั้งหนึ่งก่อนแล้วจึงตรวจพิจารณารายละเอียดอีกครั้ง พยายามจดบันทึกประเด็นที่ตอบผิดที่ตรวจพบแล้วแก้ไขสั้น ๆ จดประเด็นที่ผู้สอบไม่ได้กล่าวถึง รวมทั้งค่าคะแนนในแต่ละประเด็น วิธีนี้จะสามารถให้คะแนนตามเกณฑ์ขั้นต่ำได้ถูกต้อง ถ้ามีประเด็นพิเศษเพิ่มเติม ให้รวมกับคะแนนขั้นต่ำ ขั้นนี้อาจใช้เรียงความต้นแบบ (essay model) เป็นเกณฑ์เปรียบเทียบ จะทำให้การตรวจมีความเที่ยงเพิ่มขึ้น

6. ควรใช้ผู้ตรวจมากกว่าหนึ่งคน การมีผู้ตรวจ 2 คน แม้จะตรวจโดยให้อ่านอย่างรวดเร็วก็ยังดีกว่าผู้ตรวจคนเดียว

สำหรับ Linvall and Nitko (1975 : 51-52) ได้เสนอแนะวิธีตรวจเพื่อให้มีความเที่ยงสูงดังนี้

1. ถ้าข้อทดสอบมีหลายข้อให้ตรวจข้อแรกทุกคนก่อน แล้วจึงตรวจข้อที่สองในลักษณะเดิมวิธีนี้ทำให้ผู้ตรวจสามารถกำหนดเกณฑ์เดียวกันได้และสามารถลด halo effect ของคำตอบแต่ละข้อที่จะมีต่อกันได้

2.ถ้าเป็นไปได้ควรตรวจโดยไม่ต้องดูชื่อนักเรียน เพื่อเป็นการลดอิทธพลทรงกลด (halo effect) ซึ่งเกิดจากอิทธิพลของความประทับใจที่เกิดจากการรู้จักนักเรียน

3. ตรวจซ้ำคำตอบที่ตรวจในตอนแรก เพื่อตรวจดูว่าได้ใช้เกณฑ์เดียวกันอย่างสม่ำเสมอหรือไม่

นอกจากนี้ Kubiszyn & Borich (1981 : 100-101) ยังกล่าวถึง วิธีปรับปรุงความเที่ยงของการตรวจดังนี้

1. เขียนข้อทดสอบให้ดี เพราะข้อทดสอบที่ไม่ดีเป็นแหล่งความคลาดเคลื่อนของความเที่ยงอย่างหนึ่ง เช่น ข้อทดสอบที่ไม่กำหนดความยาวทำให้การตรวจขาดความเที่ยงได้ โดยทั่วไปเรียงความที่ไม่จำกัดความยาวมักมีความเที่ยงต่ำกว่าที่จำกัดความยาว เนื่องจากนักเรียนเมื่อยล้า และเกณฑ์การตรวจของผู้ตรวจจะเปลี่ยนไปเมื่อตรวจหลาย ๆ คำตอบ

2. ใช้ข้อทดสอบแบบจำกัดคำตอบหลาย ๆ ข้อ แทนคำถามแบบขยายคำตอบเพียงข้อเดียว แต่ถ้ามีความจำเป็นต้องใช้ข้อทดสอบแบบขยายคำตอบ ควรปฏิบัติตามข้อ 3. ต่อไปนี้

3. ใช้คู่มือการตรวจที่กำหนดเกณฑ์ไว้ล่วงหน้าในการประเมินทุกประเภท ปัจจัยสำคัญคือ เกณฑ์ ถ้าครูไม่สามารถกำหนดเกณฑ์ที่เกี่ยวข้องได้ก่อน ความเที่ยงการตรวจจะลดลงทันที การที่ครูขาดเกณฑ์การตรวจในแต่ละข้อ อาจก่อให้เกิดผลเสียดังนี้

3.1 หลังจากตรวจไปหลาย ๆ คำตอบ เกณฑ์อาจเปลี่ยน ครูอาจให้คะแนนเข้มงวดขึ้น หรือปล่อยคะแนนมากขึ้น ทั้ง ๆ ที่คุณภาพของคำตอบไม่ได้เปลี่ยนไป

3.2 ความสามารถที่จะควบคุมเกณฑ์ให้คงที่จะเปลี่ยนไป เนื่องจากความล้า ถูกรบกวน หรือกรอบความคิดเปลี่ยน ฯลฯ

4. ใช้คู่มือการตรวจอย่างคงเส้นคงวา

5. ปิดชื่อนักเรียนก่อนตรวจให้คะแนน

6. ให้คะแนนคำถามเดียวกันจนครบทุกคนแล้วจึงเริ่มข้อใหม่

7. ปิดคะแนนข้อที่ตรวจเสร็จก่อนในขณะที่ตรวจข้อที่เหลือ

8. พยายามตรวจซ้ำอีกครั้ง ถ้าพบว่าคะแนนการตรวจซ้ำต่างจากครั้งก่อนให้ใช้คำแนนเฉลี่ยแทน

อนึ่ง นอกจากคำแนะนำต่าง ๆ ดังกล่าวมาแล้ว Ebel and Frishbie (1986 : 134-135) ได้เสนอแนะการตรวจเพื่อให้มีความเที่ยงสูงขึ้นดังนี้

1. ตรวจโดยใช้วิธีวิเคราะห์ หรือวิธีประเมินรวม แต่ต้องกำหนดประเด็นที่ต้องการให้คะแนนไว้ล่วงหน้าอย่างชัดเจน ถ้าเป็นวิธีประเมินรวมควรใช้วิธีจัดกลุ่มคุณภาพงานเขียน (sorting) เป็น 3 กลุ่ม ให้มีร้อยละดังนี้ เก่ง 25 % ปานกลาง 50 % และพวกอ่อนอีก 25 % หรือแบ่งเป็น 5 กอง ให้มีร้อยละดังนี้ อ่อนที่สุด 5 % อ่อน 25 % ปานกลาง 40 % เก่ง 25 % และเก่งที่สุด 5 %

2. ตรวจทีละข้อคำถาม ไม่ควรตรวจให้เสร็จเป็นคน ๆ

3. ปิดชื่อผู้ตอบไว้ก่อนตรวจ

4. ใช้คนตรวจมากกว่า 2 คน โดยตรวจเป็นอิสระจากกัน

4. ผลการวิจัยที่เกี่ยวข้องกับการเรียนการสอนและการสอบแบบเรียงความในประเทศ

ในประเทศไทย งานวิจัยเกี่ยวกับการเรียนการสอนและการสอบเรียงความภาษาอังกฤษมีจำนวนน้อยมาก จากการศึกษางานวิจัยเชิงอภิวิเคราะห์และการสังเคราะห์งานวิจัยที่เกี่ยวข้องกับการเรียนการสอนภาษาอังกฤษในระหว่าง พ.ศ. 2515 - 2530 ของสุพัฒน์ สุกมลสันต์ (2532: 76-79) ในปี พ.ศ. 2532 พบว่า ในจำนวนงานวิจัยในช่วงเวลาดังกล่าว 335 เรื่อง มีงานวิจัยเกี่ยวกับการเขียนเพียง 13 เรื่อง หรือเพียงร้อยละ 4 เท่านั้น 'งานวิจัยดังกล่าวแล้วเกี่ยวกับ 1.) ความสามารถในการใช้เครื่องหมายวรรคตอนและเครื่องผูกพันรูปเรื่อง (cohesion) 2.) ปัญหาในการเขียน 3.) วิธีสอนการเขียน 4.) ผลของการใช้สื่อการสอนเขียน และ 5.) ความสัมพันธ์ของทักษะการเขียนกับปัจจัยอื่น

ดังนั้น จะเห็นได้ว่า ก่อนปี พ.ศ. 2530 ยังไม่ปรากฏว่ามีงานวิจัยเกี่ยวกับการตรวจให้คะแนนสอบเรียงความภาษาอังกฤษในประเทศไทย แต่ภายหลังปี พ.ศ. 2530 พบว่ามีงานวิจัยจำนวนหนึ่งเกี่ยวกับการตรวจแก้ข้อผิดในงานเขียนภาษาอังกฤษของนักเรียนและนักศึกษาไทย

ในปี ค.ศ. 1987 Ladda Chabtanom (อ้างถึงใน อัจฉรา วงศ์โสธร และคณะ 2536: 17-20) ได้ทำการศึกษาเกี่ยวกับการตรวจแก้ข้อผิดในงานเขียนภาษาอังกฤษของนักศึกษาวิทยาลัยครูไทยและพบว่าประเด็นที่น่าสนใจหลายอย่าง เช่น นักศึกษาจำนวนไม่น้อยที่เขียนผิดมากจนผู้สอนไม่สามารถแก้ไขได้ และไม่มีเวลาเพียงพอที่จะ

แก้ไขได้อย่างละเอียด หรือให้นักเรียนได้พบและปรึกษาเป็นรายบุคคลได้ ผู้สอนบางคนไม่ได้ตรวจแก้ไขงานของนักศึกษาเพราะไม่มั่นใจในการใช้ภาษา และนักศึกษาบางคนเขียนถูกต้องตามหลักไวยากรณ์แต่ไม่ได้ใจความ เป็นต้น

อนึ่ง ผู้วิจัยยังได้พบว่า วิธีการตรวจงานเขียนวิชาภาษาอังกฤษของอาจารย์ในวิทยาลัยครูมีหลายวิธี คือ

1. ตรวจที่ผิดทุกแห่ง
2. ตรวจแก้ไขโดยเขียนคำหรือเครื่องหมายวรรคตอนที่ถูกต้องให้
3. เขียนระบุที่ผิดและให้กฎเกณฑ์การใช้ที่ถูกต้อง
4. เขียนคำหรือเครื่องหมายวรรคตอนที่ถูกต้องให้พร้อมกฎเกณฑ์การใช้
5. เขียนระบุว่าที่ผิดเป็นการผิดประเภทใด
6. เขียนระบุว่าผิดอย่างไร และเขียนคำหรือเครื่องหมายวรรคตอนที่ถูกต้องให้
7. เขียนวิจารณ์
8. เขียนเครื่องหมายกากบาทที่ข้างหน้าเส้นกั้นหน้าเท่าจำนวนที่ผิดในแต่ละบรรทัด แล้วให้นักศึกษาดูเองว่าผิดที่ใดและแก้ไขเอง
9. ใช้สัญลักษณ์ในการแก้
10. ตรวจแก้เฉพาะที่ผิดที่เห็นว่าสำคัญ ไม่แก้ทุกแห่ง
11. อธิบายและตอบข้อซักถามของนักศึกษาเป็นรายบุคคล
12. ให้นักศึกษาอ่านเรียงความของตนเองหน้าชั้น เพื่อให้เพื่อน ๆ ช่วยกันแก้ไข
13. ให้นักศึกษาตรวจแก้ไขงานเขียนเองโดยดูจากรายการ (checklist) ที่ผู้สอนแจกให้
14. ให้นักศึกษาจับคู่และแลกเปลี่ยนงานเขียนกันเพื่อตรวจแก้งานกันเอง
15. ให้นักศึกษาแบ่งเป็นกลุ่มละ 4-5 คนแล้วช่วยกันตรวจงานเขียนของสมาชิกในกลุ่ม

นอกจากนี้ผู้วิจัยยังได้สำรวจความคิดเห็นของอาจารย์ที่สอนเรียงความภาษาอังกฤษแล้วพบว่า การแก้ไขงานของนักศึกษามีความสำคัญยิ่งในการสอนเรียงความ และชอบให้ผู้เรียนแก้ไขงานเองและแก้ไขงานกับเพื่อน และพบว่าความคิดเห็นนี้ขัดแย้งกับความเป็นจริง เพราะอาจารย์ยังนิยมแก้ไขงานให้แก่นักศึกษาเป็นส่วนใหญ่ แต่ก็ไม่แน่ใจว่าวิธีนี้เป็นวิธีที่ดีที่สุดในการสอนเรียงความ

ต่อมาในไป พ.ศ. 2533 มาลี สาตจีนพงษ์ (อ้างถึงใน อัจฉรา วงศ์โสธร และคณะ 2536: 20-21) ได้ทำการศึกษาเกี่ยวกับการตรวจแก้งานเขียนของครูสอนภาษาอังกฤษในระดับชั้นมัธยมศึกษาปีที่ 3 ในเขตกรุงเทพมหานคร ของโรงเรียนรัฐบาล โรงเรียนราษฎร์ และโรงเรียนสาธิต จำนวน 85 คน โดยวิธีการสำรวจและพบประเด็นที่น่าสนใจหลายอย่าง เช่น

1. ปัญหาที่สำคัญในการตรวจแก้งานเขียนของนักเรียนคือ มีนักเรียนจำนวนมากเกินไป ครูไม่มีเวลาเพียงพอที่จะตรวจแก้ไขได้อย่างละเอียด
2. ข้อผิดพลาดที่ครูได้แก้ไขให้นักเรียนแล้ว นักเรียนก็ยังทำผิดในลักษณะเดิมอีก เพราะนักเรียนไม่ค่อยสนใจการแก้ไขของครูเท่ากับคะแนนที่ได้รับ
3. ครูส่วนใหญ่เป็นผู้ตรวจแก้งานเขียนให้นักเรียนมากกว่าให้นักเรียนตรวจแก้ด้วยตนเอง หรือให้นักเรียนมีส่วนร่วมในการตรวจแก้
4. ครูใช้วิธีขีดฆ่า วงกลม หรือขีดเส้นใต้คำหรือข้อความที่ผิด แล้วเขียนคำหรือข้อความที่ถูกต้องให้ อธิบายข้อผิดต่าง ๆ ให้นักเรียนฟัง และขีดฆ่าคำหรือข้อความที่ไม่จำเป็นออก

5. เห็นว่าข้อผิดพลาดในงานเขียนของนักเรียนเป็นสิ่งที่ครูต้องแก้ไข และควรแก้ไขทุกเรื่อง

ในอีก 1 ปีต่อมาคือในปี พ.ศ. 2534 พิมพรรณ เรืองปราชญ์(อ้างถึงใน อัจฉรา วงศ์โสธร และคณะ 2536: 22-23) ได้ทำการสำรวจความคิดเห็นของอาจารย์สอนภาษาอังกฤษจำนวน 11 คนและนักศึกษาในสถาบันเทคโนโลยีราชมงคลอีกจำนวน 665 คนเกี่ยวกับการตรวจงานเขียนภาษษอังกฤษ และพบประเด็นที่น่าสนใจหลายอย่าง เช่น

ก. วิธีตรวจงานเขียนที่อาจารย์คิดว่าเหมาะสม คือ

1. ยกที่ผิดที่พบในงานเขียนของนักศึกษามาอธิบายพร้อม ๆ กันทั้งชั้น
2. ตรวจที่ผิดทุก ๆ ลักษณะในทันทีที่พบ
3. เขียนคำติชม
4. อธิบายที่ผิดให้นักศึกษาฟังเป็นรายบุคคล
5. เลือกตรวจเฉพาะที่สำคัญ ๆ

ข. วิธีที่อาจารย์ใช้ตรวจงานเขียนที่บ่อยที่สุด คือ

1. ตรวจที่ผิดทุก ๆ ลักษณะในทันทีที่พบ
2. เขียนคำติชม
3. ให้นักศึกษาตรวจงานเองโดยอาจารย์ค่อยแนะนำ
4. ยกที่ผิดที่พบในงานเขียนของนักศึกษามาอธิบายพร้อม ๆ กันทั้งชั้น
5. เลือกตรวจเฉพาะที่สำคัญ ๆ
6. อธิบายที่ผิดให้นักศึกษาฟังเป็นรายบุคคล

ดังนั้น จากงานวิจัยทั้ง 3 เรื่องซึ่งเพิ่งกล่าวถึงข้างต้นจะเห็นได้ว่า ครูและอาจารย์ที่สอนภาษษอังกฤษโดยมากนิยมตรวจงานเขียนของผู้เรียนโดยวิธีวิเคราะห์ (analytical method) ด้วยวิธีต่าง ๆ กัน โดยพยายามตรวจแก้ข้อผิดพลาดในงานของผู้เขียนทุกลักษณะให้มากที่สุดในทันที่พบ และมักจะพบว่าข้อผิดพลาดที่ได้แก้ไขให้กับผู้เรียนแล้ว ผู้เรียนมักจะทำผิดพลาดอีกเหมือนเดิม

นอกจากนี้ ยังมีงานวิจัยเกี่ยวกับวิธีแก้ไขข้อผิดที่เหมาะสมที่สุดในงานเขียนเรียงความภาษาอังกฤษของนิสิต/นักศึกษาในระดับหาวิทยาลัยล่าสุดเรื่องหนึ่งในปี พ.ศ. 2536 ของอัจฉรา วงศ์โสธร และคณะ (2536: 549-591) การศึกษานี้เป็นการวิจัยเชิงสำรวจ โดยใช้นิสิต/นักศึกษาคณะต่าง ๆ ของมหาวิทยาลัยเกษตรศาสตร์ ธรรมศาสตร์ ธุรกิจบัณฑิตย์ และจุฬาลงกรณ์ รวมจำนวน 177 คนเป็นพลวิจัยในการเขียนเรียงความคนละ 8 ชิ้น และตรวจแก้งาน 4 วิธี คือ

1. วิธีขีดฆ่าที่ผิดแล้วเขียนแก้ไขให้
2. วิธีขีดเส้นใต้ที่ผิด และนำข้อผิดร่วมกันมาอธิบายในชั้น
3. วิธีให้สัญลักษณ์กำกับที่ผิด และ
4. วิธีเรียกผู้เรียนมาพบ และ/หรือเขียนคำอธิบายให้เป็นรายบุคคล

การวิจัยครั้งนี้พบว่า การตรวจงานเขียนเรียงความภาษาอังกฤษที่เหมาะสมที่สุดสำหรับนักศึกษาในระดับอุดมศึกษาบางแห่งในเขตกรุงเทพมหานครคือ วิธีที่ 1 รองลงมาคือวิธีที่ 4 ส่วนวิธีที่ 2 และ 3 มีความหมาะสมเป็นลำดับที่ 3 เท่า ๆ กันสำหรับการพัฒนาการสอนเขียนในระยะสั้น แต่สำหรับการพัฒนาระยะยาวแล้ว วิธีที่ 3 ดีเป็นลำดับที่ 3 และวิธีที่ 2 ดีเป็นลำดับที่ 4 ส่วนลำดับที่ 1 และ 2 ยังคงเหมือนเดิม คณะผู้วิจัยพบว่า วิธีที่ 1 มี

ประสิทธิภาพที่สุดในการปรับปรุงการเขียนของนักศึกษาด้าน กลไกการเขียน การใช้ภาษา การใช้ศัพท์ การเรียบเรียงความ และการใช้เนื้อความ โดยเฉพาะอย่างยิ่งเหมาะสมสำหรับนิสิต/นักศึกษาที่เรียนเก่ง

ดังนั้น จากงานวิจัยต่าง ๆ ที่ได้กล่าวมาแล้วข้างต้นจะเห็นได้ว่า การวิจัยที่เกี่ยวข้องกับการเขียนเรียงความและทฤษฎีการอ้างอิงสรุปในประเทศไทยยังมีน้อยมาก นอกจากงานวิจัยของแดง กลางท่าได่ (2531) ดังได้กล่าวมาแล้ว ในปี 2533 ไพรัตน์ วงษ์งาม (2533 :109-112) ได้ศึกษาหาค่าสัมประสิทธิ์การอ้างอิงสรุปของแบบทดสอบอัตนัยชนิดเรียงความภาษาไทยพบว่า การตรวจที่ผู้ตรวจทำตามคำชี้แจงของผู้วิจัยและผู้ตรวจไม่รู้ผลการเรียนของผู้ตอบทำให้ความแปรปรวนของผู้ตรวจลดลงจากการตรวจโดยใช้ประสบการณ์เดิมของตนเองถึงร้อยละ 50 สำหรับวิธีประเมินรวม และร้อยละ 25 สำหรับวิธีวิเคราะห์ ซึ่งแสดงว่าเกณฑ์ในการให้คะแนนมีความสำคัญต่อความเที่ยงในการตรวจมาก และค่าอ้างอิงสรุปของแบบทดสอบเรียงความที่อ้างอิงไปยังเอกภพของผู้ตรวจอย่างเดียวมีสูงกว่าค่าที่อ้างอิงไปยังเอกภพของข้อทดสอบและผู้ตรวจพร้อมกัน แต่ยังไม่พบว่ามีงานวิจัยเกี่ยวกับการให้คะแนนสอบการเขียนเรียงความภาษาอังกฤษที่เหมาะสมที่สุดในเชิงของความเที่ยงของผลการให้คะแนน จึงเป็นเรื่องที่ผู้วิจัยมีความสนใจมากที่จะทำการศึกษาครั้งนี้

บทที่ 3

วิธีดำเนินการวิจัย

เพื่อให้บรรลุวัตถุประสงค์ของการวิจัยครั้งนี้ ทั้ง 2 ข้อ ผู้วิจัยได้ดำเนินการวิจัยดังต่อไปนี้

ก. การประยุกต์ใช้ทฤษฎีการอ้างอิงสรุปเพื่อการตรวจข้อทดสอบแบบเรียงความ

1. เอกภพของการวิจัย (ประชากร)

ตามทฤษฎีการอ้างอิงสรุปเอกภพของการวิจัย (universe) หมายถึง ประชากรทั้งหมดของปัจจัยประกอบแต่ละอย่าง (Cardinet and Others, 1981: 183) ซึ่งสามารถแบ่งออกได้ 2 อย่างคือ

1. **เอกภพของสิ่งสังเกตที่ยอมรับได้** (universe of admissible observation) ซึ่งหมายถึง สถานการณ์หรือเงื่อนไขผสมที่เป็นไปได้ระหว่างปัจจัยประกอบด้านข้อทดสอบ ผู้ตรวจ และผู้สอบตามเงื่อนไขที่มีอยู่จริง กล่าวคือ

ก. ข้อทดสอบแบบเรียงความจำนวนอนันต์ ($N = \infty$)

ข. ผู้ตรวจข้อทดสอบจำนวน 2 ประเภท ($N = 2$) คือ

1. ผู้ที่เป็นกรรมการของแบบทดสอบ CU - TEP และมีประสบการณ์ตรงในการตรวจแบบเรียงความที่นำมาวิจัย หรือเรียกว่า **"ผู้มีประสบการณ์ตรง"**

2. ผู้ที่เป็นอาจารย์สอนภาษาอังกฤษและไม่เคยมีประสบการณ์ในการตรวจแบบเรียงความที่นำมาศึกษาโดยตรง หรือเรียกว่า **"ผู้มีประสบการณ์อ้อม"**

ค. ผู้สอบจำนวนอนันต์ ($N = \infty$)

2. **เอกภพของการอ้างอิงสรุป** (universe of generalizability) ซึ่งหมายถึงสถานการณ์หรือเงื่อนไขผสมที่อาจเป็นไปได้สำหรับประชากรของแต่ละปัจจัยประกอบ ที่ผู้วิจัยต้องการอ้างอิงสรุป (generalize) เช่น ผู้ตรวจที่มีประสบการณ์ทางอ้อมคนใดก็ตามตรวจข้อทดสอบข้อใดก็ตามของผู้สอบคนใดก็ตาม เป็นต้น เอกภพของการอ้างอิงสรุปมีประเภทและขนาดดังนี้

ก. ข้อทดสอบเป็นแบบจำลองสุ่ม (random model) และมีขนาด 3 ข้อ ($n = 3 < N = \infty$)

ข. ผู้ตรวจข้อสอบเป็นแบบจำลองคงที่ (fixed model) และมีขนาด 2 ประเภท ($n = 2$ $N = 2$)

ค. ผู้สอบเป็นแบบจำลองสุ่ม (random model) และมีขนาด 200 คน(($n = 200 < N = \infty$)

อนึ่ง ในการศึกษาครั้งนี้ ผู้วิจัยต้องการศึกษาทั้งในกรณีแบบจำลองคงที่และแบบจำลองผสม (mixed model) ของตัวแปรต่าง ๆ

2. กลุ่มตัวอย่างและการได้มาซึ่งตัวอย่าง

ก. กลุ่มตัวอย่าง

วัตถุประสงค์ที่สำคัญของทฤษฎีการอ้างอิงสรุป คือ การประมาณค่าความแปรปรวนของปัจจัยประกอบที่นำมาศึกษาและประมาณค่าสัมประสิทธิ์การอ้างอิงสรุป (generalizability coefficient) อย่างปราศจากความลำเอียง (bias) ดังนั้น จึงต้องการสุ่มตัวอย่างขนาดใหญ่เพื่อให้สามารถสรุปค่าที่มุ่งแสวงหาได้อย่างมั่นใจ Smith (Smith, 1978:346) ได้เสนอว่าในการศึกษาปัจจัยประกอบ 3 อย่าง กลุ่มตัวอย่างควรมีขนาด $n_p \times n_i \times n_r$ อย่างน้อย

1000 เงื่อนไข (หรือสถานการณ์) ดังนั้น ในการวิจัยครั้งนี้ผู้วิจัยจึงกำหนดให้กลุ่มตัวอย่างมีขนาด n_p (ผู้สอบ) x n_i (ข้อทดสอบ) x n_r (ผู้ตรวจข้อสอบ) = 200 x 3 x 2 = 1,200 เงื่อนไข (หรือสถานการณ์)

ข. การสุ่มเงื่อนไข (หรือสถานการณ์)

ตัวอย่างเงื่อนไขหรือสถานการณ์ที่เป็นแบบสุ่ม (random design) ผู้วิจัยได้มาจากการสุ่มปัจจัยประกอบที่เกี่ยวข้อง ดังนี้

1. ข้อทดสอบ (n_i) แบบทดสอบ CU - TEP แต่ละฉบับมีข้อทดสอบเรียงความ 3 ข้อ แต่ละข้อมีวัตถุประสงค์เฉพาะแตกต่างกัน ผู้วิจัยจึงสุ่มอย่างง่ายได้แบบทดสอบมาฉบับหนึ่ง และได้เลือกใช้ข้อทดสอบเรียงความทั้ง 3 ข้อ ของแบบทดสอบที่สุ่มได้มา

2. ผู้สอบ (n_p) ในแต่ละปี สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย ได้ให้บริการทดสอบด้วยแบบทดสอบ CU - TEP แก่นิสิต นักศึกษา และบุคคลทั่วไปทั้งภาคเอกชนและภาครัฐบาลจำนวนมาก สำหรับการวิจัยครั้งนี้ผู้สอบสุ่มอย่างง่าย ได้ผู้สอบเป็นนิสิต นักศึกษา และบุคคลทั่วไปที่ประสงค์จะสอบชิงทุนของรัฐบาลไทยไปศึกษาต่อต่างประเทศที่จัดสอบผ่านทางสำนักงานกรรมการข้าราชการพลเรือน (ก.พ.) ประจำปี พ.ศ. 2537 (สถาบันภาษาเป็นผู้จัดสอบภาษาอังกฤษให้) ผู้เข้าสอบทั้งหมดเฉพาะการสอบครั้งนี้มี 343 คน ผู้วิจัยได้ทำการสุ่มตัวอย่างอย่างง่ายได้ตัวอย่างมาทำการศึกษา 200 คน

อนึ่ง สำหรับตัวอย่างที่เป็นแบบคงที่ (fixed design) นั้น ผู้วิจัยได้เลือกใช้คะแนนผลการตรวจข้อทดสอบของกรรมการแบบทดสอบ CU - TEP ที่มีอยู่แล้ว กรรมการชุดนี้มี 6 คน ตรวจข้อทดสอบแบบวิธีประเมินรวม (holistic method) การตรวจข้อทดสอบแต่ละข้อจะตรวจโดยกรรมการ 2 คน (ปกตินำคะแนนมารวมกันและหาค่าเฉลี่ยเป็นคะแนนรายข้อ) เวลาตรวจจะอาศัยแถบระดับความสามารถ (ability band) เป็นเกณฑ์ในการตรวจให้คะแนน กรรมการหรือผู้ตรวจประเภทนี้เป็นผู้มีประสบการณ์ในการตรวจข้อทดสอบเรียงความที่นำมาศึกษาครั้งนี้โดยตรง เนื่องจากเคยตรวจมาแล้วหลายครั้ง

ส่วนผู้ตรวจข้อทดสอบที่มีประสบการณ์ในการตรวจข้อทดสอบ CU - TEP ทางอ้อม ได้แก่ อาจารย์สอนภาษาอังกฤษของสถาบันภาษา จำนวน 6 คน ที่ทำการตรวจข้อทดสอบ CU - TEP ที่นำมาศึกษาโดยวิธีประเมินรวม และอาศัยแถบระดับความสามารถเดียวกันกับที่กรรมการประเภทแรกใช้ และตรวจข้อทดสอบแต่ละข้อโดยกรรมการ 2 คน ผู้ตรวจประเภทนี้ไม่เคยมีประสบการณ์ในการตรวจข้อทดสอบเรียงความของแบบทดสอบ CU - TEP มาก่อน

3. การกำหนดความสัมพันธ์ของปัจจัยประกอบที่มุ่งศึกษา

ก. ขั้นการศึกษาเพื่ออ้างอิงสรุป (generalizability study หรือ G Study) ซึ่งเป็นขั้นการศึกษาความแปรปรวนของปัจจัยประกอบต่าง ๆ ที่เกี่ยวข้อง ผู้วิจัยกำหนดให้ความสัมพันธ์ของปัจจัยประกอบต่าง ๆ เป็นแบบ P x I x r ซึ่งเป็นแบบผสม (crossed design) เพราะต้องการศึกษาทุก ๆ เงื่อนไขที่เป็นไปได้ ซึ่งเกิดจากผลคูณของ P (ผู้สอบ) x I (ข้อทดสอบ) x R (ผู้ตรวจข้อสอบ) ซึ่งหมายความว่า *"ผู้สอบทุกคนสอบข้อทดสอบทุกข้อ และผลการสอบได้รับการตรวจให้คะแนนโดยผู้ตรวจทั้ง 2 ประเภททุกคน"*

ข. ขั้นการศึกษาเพื่อตัดสินใจ (decision study หรือ D Study) ซึ่งเป็นขั้นการประมาณค่าความแปรปรวนของปัจจัยประกอบต่าง ๆ ที่เกี่ยวข้อง เพื่อใช้ประกอบการตัดสินใจเลือกแบบจำลองหรือเงื่อนไขที่เหมาะสมสำหรับนำไปประยุกต์ใช้ต่อไป ในขั้นนี้ผู้วิจัยกำหนดให้ความสัมพันธ์ของปัจจัยประกอบต่าง ๆ เป็นแบบผสม

และแบบแฝง (nested within) เฉพาะในสถานการณ์ที่อาจเป็นไปได้ตามความเป็นจริงเท่านั้น (Brenman, 1983:30) คือ

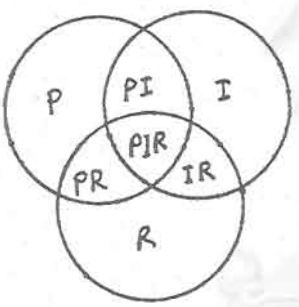
1. แบบ P x I x R ซึ่งหมายความว่า ผู้สอบทุกคนสอบข้อทดสอบทุกข้อและผลการสอบได้รับตรวจให้คะแนนโดยผู้ตรวจทุกประเภททุกคน

2. แบบ P x (I : R) ซึ่งหมายความว่า ผู้สอบทุกคนสอบข้อทดสอบทุกข้อแต่ผู้ตรวจแต่ละประเภทตรวจข้อทดสอบบางข้อของคนทุกคน (เช่น ผู้ตรวจที่มีประสบการณ์โดยตรงตรวจข้อที่ 3 แต่ผู้ตรวจที่มีประสบการณ์อ้อมตรวจข้อ 1-2 เป็นต้น)

3. แบบ I x (P : R) ซึ่งหมายความว่า ผู้สอบทุกคนสอบข้อทดสอบทุกข้อแต่ผู้ตรวจบางประเภทตรวจข้อทดสอบทุกข้อของผู้สอบบางคน (เช่น ผู้ตรวจแต่ละประเภทตรวจข้อทดสอบทุกข้อของผู้สอบที่แบ่งเป็นกลุ่มย่อยของกลุ่ม เป็นต้น)

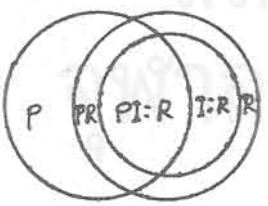
อนึ่ง ความสัมพันธ์ของปัจจัยประกอบทั้ง 3 แบบจำลอง อาจแสดงด้วยแผนผังเวนน์ (Venn Diagram) ดังนี้

1. แบบ P x I x R



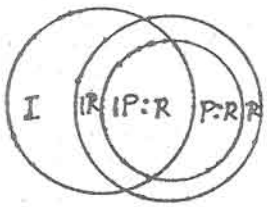
$$
\begin{aligned}
X_{PIR} = \; & \mu \\
& + (\mu_P - \mu) \\
& + (\mu_I - \mu) \\
& + (\mu_R - \mu) \\
& + (\mu_{PI} - \mu_P - \mu_I + \mu) \\
& + (\mu_{PR} - \mu_P - \mu_R + \mu) \\
& + (\mu_{IR} - \mu_I - \mu_R + \mu) \\
& + (X_{PIR} - \mu_{PI} - \mu_{PR} - \mu_{IR} \\
& \quad + \mu_P + \mu_I + \mu_R - \mu)
\end{aligned}
$$

2. แบบ P x (I : R)



$$
\begin{aligned}
X_{PIR} = \; & \mu \\
& + (\mu_P - \mu) \\
& + (\mu_R - \mu) \\
& + (\mu_{IR} - \mu) \\
& + (\mu_{PR} - \mu_P - \mu_R + \mu) \\
& + (X_{PIR} - \mu_{PR} - \mu_{IR} + \mu_R)
\end{aligned}
$$

3. แบบ I x (P : R)



$$
\begin{aligned}
X_{PIR} = \quad & \mu \\
& + (\mu_I - \mu) \\
& + (\mu_R - \mu) \\
& + (\mu_{RI} - \mu) \\
& + (\mu_{IR} - \mu_I - \mu_R + \mu) \\
& + (X_{PIR} - \mu_{IR} - \mu_{PR} + \mu_P)
\end{aligned}
$$

4. เครื่องมือที่ใช้ในการวิจัย

ก. ลักษณะของเครื่องมือ

เครื่องมือที่ใช้ในการวิจัย ได้แก่ ข้อทดสอบเรียงความ 3 ข้อ ของแบบทดสอบ CU - TEP เพื่อทดสอบความรู้ความสามารถทางด้านการเขียนของผู้สอบ โดยแต่ละข้อมีวัตถุประสงค์ในการทดสอบ ดังนี้

ข้อที่ 1 มุ่งทดสอบความสามารถในการเขียนข้อข้อความสั้น ๆ ขนาด 30-50 คำ ที่อาจจำเป็นต้องใช้ในชีวิตประจำวันได้

ข้อที่ 2 มุ่งทดสอบความสามารถในการเขียนจดหมายส่วนตัวยาวประมาณ 50-90 คำ เพื่อโต้ตอบกับผู้อื่นในเรื่องเกี่ยวกับชีวิตประจำวันได้

ข้อที่ 3 มุ่งทดสอบความสามารถในการเขียนแสดงความคิดเห็นเกี่ยวกับเรื่องใดเรื่องหนึ่งที่อาจเกิดขึ้นในชีวิตประจำวัน ในขนาด 250-300 คำได้

ข. การพัฒนาเครื่องมือการสอบ

ข้อทดสอบเรียงความเป็นข้อทดสอบที่จัดสร้างขึ้นโดยคณะกรรมการสร้างและพัฒนาแบบทดสอบ CU - TEP ซึ่งเดิมเป็นโครงการวิจัยเพื่อสร้างและพัฒนาแบบทดสอบวัดสมิทธิภาพทั่วไปทางภาษาอังกฤษของนิสิตระดับชั้นปีที่ 3-4 ของจุฬาลงกรณ์มหาวิทยาลัย และบุคคลที่สนใจทั่วไปที่ต้องการทดสอบความรู้ความสามารถของตนเอง โครงการวิจัยดังกล่าวเริ่มดำเนินการตั้งแต่ปี พ.ศ. 2532 จนถึงปี พ.ศ. 2535 แบบทดสอบ CU - TEP ประกอบด้วย

1. ข้อทดสอบการฟังเข้าใจความ 35 ข้อ (40 นาที)
2. ข้อทดสอบการอ่านเข้าใจความ 60 ข้อ (30 นาที)
3. ข้อทดสอบการเขียนแบบปรนัย 20 ข้อ และแบบเรียงความ 3 ข้อ (70 นาที)

แบบทดสอบ CU - TEP มีอยู่หลายชุด แต่ละชุดได้รับการทดลอง และใช้จริงหลายครั้ง ผลการทดสอบได้รับการวิเคราะห์ข้อทดสอบรายข้อ และปรับปรุงแก้ไขให้เป็นแบบทดสอบที่มีคุณภาพดียิ่งขึ้นเรื่อย ๆ ยกตัวอย่างเช่น แบบทดสอบชุด 7, 8, และ 9 มีค่าความเที่ยง (reliability) ระหว่าง 0.89-0.91 และผลการสอบมีความตรงร่วมสมัย (concurrent validity) กับแบบทดสอบที่มีชื่อเสียง เช่น TOEFL (Test of English as a Foreign Language) และ TOEIC (Test of English for International Communication) ในระดับสูง (r_{xy} = 0.75 - 0.85) ทำให้สามารถเทียบเคียงคะแนนของผลการสอบของแบบทดสอบดังกล่าวได้ (รายงานผลสังเขปเกี่ยวกับการสอบ *CU - TEP* ปีที่ 3, 2535 : 3)

อนึ่ง สำหรับการวิจัยครั้งนี้ ผู้วิจัยใช้เฉพาะข้อทดสอบเรียงความ 3 ข้อ ของแบบทดสอบ CU - TEP ชุด 7 เท่านั้น

5. การเก็บรวบรวมข้อมูล

ก. สุ่มกระดาษคำตอบของผู้สอบ

ดังได้กล่าวมาแล้วว่าในแต่ละปี สถาบันภาษาให้บริการการทดสอบด้วยแบบทดสอบ CU - TEP แก่นิสิต นักศึกษา และบุคคลทั่วไปทั้งของภาคเอกชนและภาครัฐบาลจำนวนมาก สำหรับการวิจัยครั้งนี้ผู้วิจัยสุ่มอย่างง่ายได้กระดาษคำตอบเฉพาะที่เป็นแบบทดสอบเรียงความของผู้สอบที่เป็นนิสิต นักศึกษา และบุคคลทั่วไป จำนวน 200 ฉบับ(คน) จากผู้เข้าสอบทั้งหมด 343 คน ซึ่งสมัครสอบรับทุนของรัฐบาลไทยไปศึกษาต่อต่างประเทศประจำปีการศึกษา 2535 โดยสถาบันภาษาจัดสอบภาษาอังกฤษให้

ข. ตรวจให้คะแนน

1. .ให้ผู้ตรวจข้อสอบที่มีประสบการณ์ตรวจโดยตรงทำการตรวจให้คะแนนกระดาษคำตอบเรียงความดังกล่าวแล้ว โดยผู้ตรวจ 2 คนทำการตรวจทุกข้อด้วยวิธีประเมินผลรวมโดยอาศัยแถบระดับความสามารถเป็นเกณฑ์ ผู้ตรวจคนหนึ่งตรวจกระดาษคำตอบประมาณ 70 ฉบับและบันทึกคะแนนไว้ในกระดาษต่างหาก แล้วจึงเวียนให้ผู้ตรวจคนที่สองตรวจโดยวิธีการเดียวกัน

2. นำกระดาษคำตอบทั้งจำนวน 200 ฉบับ (ของผู้สอบ 200 คน) ไปให้ผู้ตรวจที่มีประสบการณ์ทางอ้อมตรวจฉบับละ 2 คน ทุกข้อ ด้วยวิธีการเดียวกันกับที่กล่าวแล้วข้างตน แล้วบันทึกคะแนนไว้ในกระดาษต่างหาก

ค. เกณฑ์การให้คะแนน

ผู้ตรวจข้อสอบแต่ละประเภทตรวจให้คะแนนโดยอาศัยเกณฑ์การให้คะแนน ที่พัฒนาขึ้นโดยคณะกรรมการสร้างแบบทดสอบ CU - TEP เกณฑ์ดังกล่าวมีลักษณะเป็นเกณฑ์ระดับความสามารถ (ability band) และแต่ละข้อมีเกณฑ์การให้คะแนนแตกต่างกัน ดังนี้

เกณฑ์การตรวจให้คะแนนของข้อสอบข้อที่ 1

Very Good :	– well-constructed phrase/sentences with an
(17-20)	effective presentation of points
	– very few errors
	– a wide range of words
Good :	– well-constructed phrases/sentences
(13-16)	– basic ideas clearly presented
	– free from basic errors
	– good range of vocabulary
Moderate :	– basic ideas communicated
(9-12)	– accuracy in grammar and punctuation
	– few basic errors
	– average range of vocabulary
Weak :	– basic ideas communicated but inadequate control of
(5-8)	grammar and punctuation
	– limited vocabulary
Poor :	– incoherent
(1-4)	– inadequate knowledge of vocabulary

เกณฑ์การตรวจให้คะแนนของข้อสอบข้อที่ 2

Very Good :	_ well-constructed sentences / paragraph (s) with an
(25-30)	effective presentation of points
	– very few errors
	– a wide range of words
Good :	– well-constructed sentences / paragraph (s)
(19-24)	– basic ideas clearly presented
	– free from basic errors
	– good range of vocabulary
Moderate :	– basic ideas communicated
(13-18)	– accuracy in grammar and punctuation
	– few basic errors
	– average range of vocabulary
Weak :	– basic ideas communicated but inadequate control of
(7-12)	grammar and punctuation
	– limited vocabulary
Poor :	– incoherent
(1-6)	– inadequate knowledge of vocabulary

เกณฑ์การตรวจให้คะแนนของข้อสอบข้อที่ 3

Very Good :	– well-constructed paragraph(s) with an
(41-50)	effective presentation of points
	– very few errors
	– a wide range of words
Good :	– well-constructed paragraph(s)
(31-40)	– basic ideas clearly presented
	– free from basic errors
	– good range of vocabulary
Moderate :	– basic ideas communicated
(21-30)	– accuracy in grammar and punctuation
	– few basic errors
	– average range of vocabulary
Weak :	_basic ideas communicated but inadequate control of
(11-20)	grammar and punctuation
	– limited vocabulary
Poor :	– incoherent
(1-10)	– inadequate knowledge of vocabulary

ง. เตรียมข้อมูลเพื่อการวิเคราะห์

1. นำคะแนนที่ได้จากการตรวจข้อทดสอบของผู้สอบแต่ละคนและแต่ละข้อของผู้ตรวจแต่ละประเภท ซึ่งตรวจ 2 ครั้ง มาเตรียมข้อมูลเพื่อใช้ในการวิเคราะห์ด้วยเครื่องคอมพิวเตอร์สำหรับทดสอบสมมุติฐานข้อที่ 3

2. สำหรับข้อมูลที่ใช้เพื่อการทดสอบสมมุติฐานข้อที่ 1, 2, 4, 5 และ 6 เกิดจากการจำลองข้อมูลเสมือนจริง (generate) จากข้อมูลจริงในข้อที่ 1 ด้วยโปรแกรมคอมพิวเตอร์ใช้งานที่สร้างขึ้นสำหรับการวิจัยครั้งนี้

6. การวิเคราะห์ข้อมูล

1. ขั้นการศึกษาเพื่อเพื่ออ้างอิงสรุป (G-Study) ทำการวิเคราะห์ความแปรปรวนของแหล่งปัจจัยประกอบต่าง ๆ ที่เกี่ยวข้องด้วยวิธี 3 Factorial Design แบบผสม(crossed design) ที่เป็นแบบจำลองสุ่ม (random model) ทั้งหมด คือ P x I x R และทำการศึกษาตามแบบจำลองผสม (mixed model) ด้วย เพื่อให้สอดคล้องกับเงื่อนไขของการวิจัยครั้งนี้ ซึ่งมีปัจจัยประกอบด้านผู้ตรวจข้อทดสอบเป็นแบบจำลองคงที่ (fixed model)

2 ขั้นการศึกษาเพื่อตัดสินใจ (D- Study) ทำการวิเคราะห์เพื่อประมาณค่าความแปรปรวนของปัจจัยประกอบต่าง ๆ ที่เกี่ยวข้องต่อจากการศึกษาเพื่ออ้างอิงสรุปตามแบบที่ต้องการศึกษา ทั้ง 3 แบบ ดังกล่าวแล้ว คือ แบบ P x I x R แบบ Px (I : R) และแบบ I x (P : R) ทั้งในกรณีที่เป็นแบบจำลองผสม และแบบจำลองสุ่มโดยอาศัยแนวคิดของทฤษฎีอ้างอิงสรุปของ Brennan (Brennan, 1983 : 33-77)

3 ประมาณค่าสัมประสิทธิ์การอ้างอิงสรุป (Generalizability Coefficient)
หรืออาจเรียกว่า ค่าสัมประสิทธิ์ที่คล้ายกับค่าความเที่ยง (realiability-like coefficient) และคำนวณหาดัชนีความเชื่อมั่น (index of dependability) ของการตรวจให้คะแนนของผู้ตรวจข้อทดสอบแต่ละประเภทในเงื่อนไขต่าง ๆ ตามวิธีของ Brennan และ Kane (อ้างถึงใน Brennan, 1983 : 108)

4. คำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างผลการตรวจให้คะแนนข้อทดสอบเรียงความ โดยผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงกับผู้ที่ประสบการณ์ตรวจทางอ้อม โดยใช้คะแนนรวมของผลการตรวจข้อทดสอบทั้ง 2 ครั้ง ของผู้ตรวจทุกคน

7. สถิติที่ใช้ในการวิเคราะห์ข้อมูล

1. การประมาณค่าความแปรปรวนของปัจจัยประกอบที่ศึกษาแบบ 3 Factorial Design ใช้สูตรในการคำนวนตามลำดับดังนี้

1.1 คำนวนผลบวกของกำลังสอง (SS) โดยใช้สูตรจากการวิเคราะห์ความแปรปรวนแบบแฟคทอเรียลสุ่มสมบูรณ์ 3 ตัวแปร คือ CRF-Pqr (อุทุมพร ทองอุไทย, 2523 : 311-314) เมื่อกำหนดให้จำนวนซ้ำหรือจำนวนข้อมูลในแต่ละ Cell เป็น 1 (n = 1)

หลักการและสูตรในการคำนวน เป็นดังนี้

ขั้นที่ 1 เขียนตารางสรุป ABCS, ABC, AB, AC, BC เพื่อแจกแจงหรือรวมคะแนนตามผลร่วมผสม (combination) ของตัวแปร

ขั้นที่ 2 คำนวนค่าต่าง ๆ ดังนี้

(1) $\sum_{1}^{N} ABCS$

(2) $\sum_{1}^{N} (ABCS)^2 = [ABCS]$

(3) $\frac{\left(\sum_{1}^{N} ABCS\right)^2}{npqr} = [X]$

(4) $\sum_{1}^{p} \frac{\left(\sum_{1}^{q} A\right)^2}{npr} = [A]$

(5= $\sum_{1}^{q} \frac{\left(\sum_{1}^{p} B\right)^2}{npr} = B]$

(6) $\sum_{1}^{r} \frac{\left(\sum_{1}^{p} C\right)^2}{npq} = [C]$

(7) $\sum_{1}^{p}\sum_{1}^{q} \frac{(AB)^2}{nr} = [AB]$

(8) $\sum_{1}^{p}\sum_{1}^{r} \frac{(AC)^2}{nq} = [AC]$

(9) $\sum_{1}^{q}\sum_{1}^{r} \frac{(BC)^2}{np} = [BC]$

(10) $\sum_{1}^{p}\sum_{1}^{q}\sum_{1}^{r} \frac{(ABC)^2}{n} = [ABC]$

ขั้นที่ 3 แทนค่าในสูตรผลบวกของกำลังสอง (SS) แต่ละค่าดังนี้

SS_{Total} = [ABCS] - [X]

SS_A = [A] - [X]

SS_B = [B] - [X]

SS_C = [C] - [X]

SS_{AB} = [AB]- [A] -[B]+[X]

SS_{AC} = [AC] - [A] - [C] + [X]

SS_{BC} = [BC] - [B] - [C] + [X]

SS_{ABC} = [ABC] - [AB] - [AC] - [BC] + [A] + [B] + [C] - [X]

$SS_{W..cell}$ = [ABCS] - [ABC]

1.2 คำนวณค่าเฉลี่ยกำลังสอง (MS) โดยสูตร (อุทุมพร ทองอุไทย 2523 : 120)

$$MS = \frac{SS(\alpha)}{df(\alpha)}$$

เมื่อ df คือ ขั้นแห่งความเป็นอิสระที่สอดคล้องกับขนาดของปัจจัยประกอบ

1.3 การประมาณค่าความแปรปรวนของแบบจำลองกลุ่ม

คำนวนความแปรปรวน เมื่อปัจจัยประกอบทุกตัวเป็นแบบจำลองสุ่มโดยใช้สูตรของ Tourneur (อ้างถึงใน Cardinet and Others, 1983 : 25) ดังนี้

$$\sigma^2(\alpha) = \frac{1}{f(\alpha)}\left[MS(\alpha) + \sum_{i=1}^{i=j} (-1)^i \right]$$

เมื่อ α คือ ปัจจัยประกอบใด ๆ

$f(\alpha)$ คือ ผลคูณของขนาดกลุ่มตัวอย่างที่ไม่ปรากฏอยู่ใน α

j คือ จำนวนนิพจน์ (expression) ที่นำมารวมกันภายในเครื่องหมายแสดงผลรวม

i คือ ลำดับที่ (rank) ของนิพจน์จาก 1 ถึง j ภายในวงเล็บปีกกา

จากความสัมพันธ์ของปัจจัยประกอบแบบ P x I x R จะหาคำตอบค่าความแปรปรวนได้จากสูตรต่อไปนี้

$$\hat{\sigma}^2(pir) = MS(pir)$$

$$\hat{\sigma}^2(pi) = \frac{1}{n_r}[MS(pi) - MS(pir)]$$

$$\hat{\sigma}^2(pr) = \frac{1}{n_i}[MS(pr) - MS(pir)]$$

$$\hat{\sigma}^2(ir) = \frac{1}{n_p}[MS(ir) - MS(pir)]$$

$$\hat{\sigma}^2(r) = \frac{1}{n_p n_i}[MS(ir) - MS(pr) + MS(pir)]$$

$$\hat{\sigma}^2(i) = \frac{1}{n_p n_r}[MS(i) - MS(ir) - MS(pi) + MS(pir)]$$

$$\hat{\sigma}^2(p) = \frac{1}{n_i n_r}[MS(p) - MS(pi) - MS(pr) + MS(pir)]$$

1.4 การประมาณค่าความแปรปรวนของแบบจำลองผสม เมื่อกำหนดให้ปัจจัยประกอบบางตัวเป็นแบบจำลองคงที่ แต่ปัจจัยประกอบที่เหลือเป็นแบบจำลองสุ่มใช้สูตร คำนวณโดยอาศัยความแปรปรวนที่คำนวณได้ในหัวข้อ 1.3 โดยใช้สูตรดังนี้ (Brennan, 1983 : 51)

$$\hat{\sigma}^2(\alpha IM) = [\hat{\sigma}^2(\beta)/F(\beta I\alpha)]$$

เมื่อ $\sigma^2(\alpha IM)$ คือ ค่าประมาณความแปรปรวนของ α ในแบบจำลองผสม

β คือ ค่าปัจจัยประกอบใด ๆ ที่อย่างน้อยจะต้องมีตัวกำกับซึ่งแทนผลของ α อยู่ด้วย

และ $F(\beta I\alpha)$ คือ ผลคูณของขนาดของเอกภพ (N) ของตัวกำกับทุกตัว ใน β แต่ไม่อยู่ใน I

แต่ถ้า $\alpha = \beta$, $F(\beta I\alpha) = 1$

2. ประมาณค่าความแปรปรวนในขั้นการสรุปอ้างอิง

2.1 หาค่าคาดหวังของความแปรปรวน (expectancies of variance) เป็นค่าปรับแก้ค่าความแปรปรวนในขั้น G study ที่มีปัจจัยประกอบเป็นสุ่ม แต่มีจำนวนจำกัด หรือมีจำนวนคงที่อยู่ในตัวกำกับแรก การปรับใช้สูตรดังนี้ (Cardinet and Others, 1983 : 34-35)

$$E^2(\alpha) = \frac{(N_f - I)\,\hat{\sigma}^2(\alpha)}{N_f}$$

เมื่อ $E^2(\alpha)$ คือ ค่าคาดหวังของความแปรปรวนของ α เมื่อ α มีตัวกำกับแรกของปัจจัยประกอบที่เป็นแบบจำลองคงที่ หรือเป็นแบบจำลองสุ่มแต่มีขนาดจำกัดอยู่ด้วย

N_f คือ ขนาดของเอกภพของปัจจัยประกอบที่เป็นแบบจำลองคงที่ หรือเป็นแบบจำลองสุ่มแต่มีขนาดจำกัด

2.2 ประมาณค่าความแปรปรวนของคะแนนเอกภพ $\hat{\sigma}^2(\tau)$ ใช้หลักการดังนี้ (Cardinet and Allal, 1983 : 35-36)

2.2.1 หาค่าปัจจัยประกอบของ active variance ซึ่งมีตัวกำกับของสิ่งที่ถูกวัดอยู่ในตัวกำกับแรกทุก ๆ ค่า

2.2.2 หาค่าผลบวกของปัจจัยประกอบตามข้อ 1 เมื่อ active variance คือ ความแปรปรวนที่มีผลต่อการสรุปอ้างอิง ได้แก่ ความแปรปรวนของปัจจัยประกอบทุกตัวยกเว้นความแปรปรวนที่มีปัจจัยประกอบแบบจำลองคงที่ (I^F) ปนอยู่ในตัวกำกับ ค่าความแปรปรวนที่ไม่มีผลต่อการอ้างอิงสรุปนี้เรียกว่า Passive variance

2.3 คำนวณความแปรปรวนของความคลาดเคลื่อนสัมบูรณ์ $\hat{\sigma}^2(\Delta)$ โดยดำเนินการดังนี้

2.3.1 หาค่าประมาณความแปรปรวนของปัจจัยประกอบต่าง ๆ ดังนี้ (Brennan, 1983 : 57)

$$\hat{\sigma}^2(\bar{\alpha}) = \hat{\sigma}^2(\alpha)/d\ (\bar{\alpha})$$

เมื่อ $d(\bar{\alpha}) = 1$ ถ้า $\alpha = p$ ซึ่งเป็นปัจจัยประกอบของสิ่งที่ถูกวัด

แต่มีค่าเท่ากับผลคูณของขนาดตัวอย่างใน D study ที่อยู่ใน α ยกเว้น p

2.3.2 ปรับแก้ความแปรปรวนของปัจจัยประกอบที่เป็นแบบจำลองสุ่มที่มีขนาดจำกัด โดยการคูณ (N-n)/(N-1) ซึ่งเรียกว่า finite population correction (Cardinet and Allal, 1983: 37-38)

2.3.3 จากค่าประมาณความแปรปรวนของปัจจัยประกอบใน 2.3.1 หักค่าความแปรปรวนของสิ่งที่ถูกวัดออกไป แล้วหาผลบวกของความแปรปรวนที่เหลือ โดยการคูณด้วย (N-n) /(N-1) เข้ากับความแปรปรวนของปัจจัยประกอบที่มีขนาดจำกัด

3. การประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง โดยใช้สูตรต่อไปนี้ (Cardinet and Allal, 1983 : 40)

$$\hat{\Sigma\rho}^2(\Delta) = \frac{\hat{\sigma}^2(\tau)}{\hat{\sigma}^2(\tau) + \hat{\sigma}^2(\Delta)}$$

เมื่อ $\Sigma\rho^2(\Delta)$ คือ ค่าสัมประสิทธิ์การอ้างอิงสรุป

$\sigma^2(\tau)$ คือ ความแปรปรวนของคะแนนเอกภพ

$\sigma^2(\Delta)$ คือ ความแปรปรวนของความคลาดเคลื่อนแบบสัมบูรณ์

4. คำนวณหาดัชนีความเชื่อถือ (index of dependability) ของจุดตัดระหว่าง 0%-100% เมื่อมีพิสัย (range) เท่ากับ 10% ตามสูตรและวิธีคำนวณของ Brenan และ Kane (Brenen and Kane, 1977 : 277-283) คือ

$$\Phi(\lambda_0) = \frac{\frac{(\bar{X}_{Pi}-C)^2}{n_p n_i} + \frac{(n_p-1)MS(P)}{n_p n_i} - \frac{(n_p-1)MS(PI)}{n_p n_i} + \frac{MS(I)}{n_p n_i}}{(\bar{X}_{Pi}-C)^2 + \frac{(n_p-1)MS(P)}{n_p n_i}}$$

ในเมื่อ $\bar{X}_{Pi}$ = ค่าเฉลี่ยของค่าเฉลี่ยรายข้อของบุคคล

C = จุดตัด (เกณฑ์)

n_0 = จำนวนผู้สอบ

n_i = จำนวนข้อทดสอบ

MS(P) = ค่า mean of square ของผู้สอบ

MS(I) = ค่า mean of square ของข้อสอบ

MS(PI) = ค่า mean of square ของ interaction

5. ทดสอบความแตกต่างของค่าดัชนีความเชื่อถือโดยเฉลี่ยระหว่างผู้ตรวจต่างประเภท โดย t-test แบบกลุ่มตัวอย่างไม่เป็นอิสระต่อกัน (correlated t-test) เมื่อ n มีขนาดน้อยกว่า 30 โดยสูตรต่อไปนี้ (Downie and Heath, 1974 : 131)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{SE_X^{\,2} + SE_X^{\,2} - 2r_{XY}\, SE_X SE_X}} \; ; df = (n-1)$$

ในเมื่อ $\bar{X}_1, \bar{X}_2$ = ค่าเฉลี่ยของคะแนนชุดที่ 1 และ 2

SE_X = SD / n-1

r_{XY} = ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าดัชนีความเชื่อถือ 2 ชุด

n = จำนวนคู่ของค่าดัชนีความเชื่อถือ

6. ทดสอบความแตกต่างของค่าสัมประสิทธิ์การอ้างอิงสรุประหว่างผู้ตรวจต่างประเภท เมื่อกลุ่มตัวอย่างไม่เป็นอิสระต่อกัน โดย t-test (Fildt, 1987 : 99) คือ

$$t = \frac{r_{tt1} - r_{tt2}\sqrt{(N-2)}}{\sqrt{4\,(1-r_{tt1})(1-r_{tt2})(1-r^{2}_{xy})}} \; ; df = N-2$$

ในเมื่อ r_{tt1} = ค่าสัมประสิทธิ์การอ้างอิงสรุปค่าที่ 1

r_{tt2} = ค่าสัมประสิทธิ์การอ้างอิงสรุปค่าที่ 2

r_{XY} = ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างผลการทดสอบ

N = จำนวนคน

7. การคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคะแนนหรือระหว่างค่าดัชนีความเชื่อถือ หรือค่าสัมประสิทธิ์การอ้างอิงสรุป โดยวิธี Pearson product-moment correlation (Downie and Heath, 1974 : 92)

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X^2)]\,[N\Sigma Y^2 - (\Sigma Y)^2]}}$$

ในเมื่อ X = คะแนนหรือค่าในชุด X

Y = คะแนนหรือค่าในชุด Y

N = จำนวนคู่ของคะแนนหรือค่าต่าง ๆที่ทดสอบ

ข. การพัฒนาโปรแกรมคอมพิวเตอร์เพื่อใช้ในการคำนวณค่าต่าง ๆ

1. เขียนโปรแกรม

ครั้งแรกผู้วิจัยได้เขียนโปรแกรมคอมพิวเตอร์เพื่อใช้คำนวณค่าต่าง ๆ ที่เกี่ยวข้องสำหรับสถานการณ์หรือเงื่อนไขต่าง ๆ ด้วยภาษา FORTRAN 77 เพื่อใช้กับเครื่องไมโครคอมพิวเตอร์ แต่เนื่องจากโปรแกรมมีขนาดยาวมากทำให้ไม่สามารถจะ compile โปรแกรมในเครื่องไมโครคอมพิวเตอร์ได้ กล่าวคือ โปรแกรมใช้หน่วยความจำมากกว่า 640 Kb ซึ่งเป็นข้อจำกัดของ compiler ของภาษา FORTRAN ในเครื่องไมโครคอมพิวเตอร์ ดังนั้น ต่อมาผู้เขียนจึงได้เขียนโปรแกรมด้วยภาษา FORTRAN สำหรับใช้กับเครื่องคอมพิวเตอร์ขนาดใหญ่ (mainframe) แทน

อนึ่ง ในการเขียนโปรแกรม ผู้วิจัยอาศัยแนวคิดและสูตรต่าง ๆ ที่เกี่ยวข้องในการคำนวณจากตำราของ Brennan (Brennan, 1979 : 54-58 ; Brennan, 1983 : 7-90, 142-153)

2. การทดลองใช้โปรแกรม

ผู้วิจัยได้อาศัยข้อมูลและผลลัพธ์การคำนวณหาค่าดัชนีความเชื่อถือและค่าสัมประสิทธิ์การอ้างอิงสรุปของ Brennan (Brennan, 1979 : 54-58 ; Brennan, 1983 : 142-153) เป็นเครื่องมือทดสอบความถูกต้องในการในการคำนวณของโปรแกรม ปรากฏว่า โปรแกรมที่เขียนขึ้นสามารถทำงานได้ผลไม่แตกต่างจากผลลัพธ์ดังกล่าว

3. การปรับปรุงโปรแกรม

ผู้วิจัยได้ทำการปรับปรุงแก้ไข โปรแกรมในรายละเอียดต่าง ๆ เพื่อให้สามารถใช้งานได้ง่ายขึ้นจนได้ผลเป็นที่น่าพอใจ

บทที่ 4
การวิเคราะห์ข้อมูล

เพื่อให้สามารถตอบวัตถุประสงค์ของการวิจัยและการสมมุติฐานของการวิจัยที่ได้ระบุไว้ในบทที่ 1 ผู้วิจัยจึงทำการวิเคราะห์ข้อมูลตามลำดับดังนี้

ก. การวิเคราะห์ข้อมูล (สำหรับวัตถุประสงค์ข้อที่ 1)

1. เมื่อกำหนดให้ผู้ตรวจข้อทดสอบทุกประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน หรือ P x I x R

1.1 ในกรณีที่เป็นแบบจำลองผสม (mixed model)

λ_o	$R_1(\Phi)$	$R_2(\Phi)$	$R_1(\Sigma\rho^2)$	$R_2(\Sigma\rho^2)$
0%	0.9715	0.9620	0.84055	0.72610
10%	0.9711	0.9616		
20%	0.9708	0.9611		
30%	0.9704	0.9606		
40%	0.9699	0.9601		
50%	0.9696	0.9596		
60%	0.9692	0.9592		
70%	0.9688	0.9587		
80%	0.9684	0.9581		
90%	0.9679	0.9576		
100%	0.9675	0.9571		
$\bar{X}$	0.9696	0.9596		
S.D.	.001	0.002		
n	11	11	200	200
r_{xy}	0.999		0.8012	
t	104.98*		6.4508*	

* $p < 0.05$

ตารางที่ 2 : การทดสอบความแตกต่างของค่า $\Phi(\lambda_o)$ และ $\Sigma\rho^2$ ในกรณีของ P x I x R เมื่อเป็นแบบจำลองผสม

จากตารางที่ 2 จะพบว่า โดยเฉลี่ยแล้วค่าดัชนีความเชื่อถือของแบบทดสอบที่ทำการตรวจโดยผู้ตรวจที่มีประสบการณ์ตรวจข้อทดสอบเรียงความโดยตรงและผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อมมีค่าเท่ากับ 0.9696 และ 0.9596 ซึ่งนับว่าสูงมาก และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าดังกล่าว ณ จุดตัดของคะแนนระหว่าง 0%-100% ที่มีพิสัย 10% เท่ากับ 0.999 ซึ่งนับว่าสูงมาก ค่าดัชนีความเชื่อถือโดยเฉลี่ยแตกต่างกันอย่างมีนัยสำคัญ ($p = 0.5$) กล่าวคือ การตรวจโดยผู้ตรวจข้อทดสอบเรียงความโดยผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือสูงกว่าดัชนีความเชื่อถือของผู้ตรวจข้อทดสอบเรียงความที่มีประสบการณ์ในการตรวจทางอ้อม [$\Phi(\lambda_o)_1 = 0.9696 > \Phi(\lambda_o)_2 = 0.9596$]

นอกจากนี้ จากตารางที่ 2 จะเห็นได้ว่าค่าสัมประสิทธิ์การอ้างอิงสรุปที่เกิดจากการตรวจข้อทดสอบของผู้ตรวจที่มีประสบการณ์โดยตรงสูงกว่าค่าที่เกิดจากผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อมอย่างมีนัยสำคัญ กล่าวคือ $\Sigma\rho^2 = 0.84055 > \Sigma\rho^2 = 0.72610$ ($p = 0.05$)

ดังนั้น อาจสรุปได้ว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงจำนวน 2 คน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบทุกคนจะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ $\Sigma\rho^2$ และ $\Phi(\lambda_o)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ตรงและ fixed (คงที่) > $\Sigma\rho^2$ และ $\Phi(\lambda_o)$ ของ p x I x R เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed

อนึ่ง ผลของการศึกษาดังกล่าวแล้วสอดคล้องกับสมมุติฐานในการวิจัยข้อที่ 1 ที่ระบุไว้ในบทที่ 1 ทุกประการ

1.2 ในกรณีที่เป็นแบบจำลองสุ่ม (random model)

λ_0	$R_1(\Phi)$	$R_2(\Phi)$	$R_1(\Sigma\rho^2)$	$R_2(\Sigma\rho^2)$
0%	0.9259	0.9236	0.66086	0.52018
10%	0.9249	0.9227		
20%	0.9241	0.9218		
30%	0.9231	0.9209		
40%	0.9221	0.9199		
50%	0.9212	0.9190		
60%	0.9202	0.9181		
70%	0.9192	0.9171		
80%	0.9182	0.9161		
90%	0.9171	0.9151		
100%	0.9675	0.9141		
$\bar{X}$	0.9211	0.9189		
S.D.	0.003	0.003		
n	11	11	200	200
r_{xy}	0.999		0.8012	
t	68.99*		4.1003*	

* $p < 0.05$

ตารางที่ 3 : การทดสอบความแตกต่างของค่า $\Phi(\lambda_0)$ และ $\Sigma\rho^2$ ในกรณีของ P x I x R เมื่อเป็นแบบจำลองสุ่ม

จากตารางที่ 3 พบว่าโดยเฉลี่ยแล้วค่าดัชนีความเชื่อถือของแบบทดสอบที่ทำการตรวจโดยผู้ตรวจที่มีประสบการณ์ตรวจข้อทดสอบเรียงความโดยตรง และผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อมมีค่าเท่ากับ 0.9211 และ 0.9189 ซึ่งนับว่าสูงมาก และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าดังกล่าว ณ จุดตัดของคะแนนระหว่าง 0%-100% ที่มีพิสัย 10% เท่ากับ 0.999 ซึ่งนับว่าสูงมาก ค่าดัชนีความเชื่อถือโดยเฉลี่ยแตกต่างกันอย่างมีนัยสำคัญ ($p = 0.5$) โดยผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือสูงกว่าค่าดัชนีความเชื่อถือของผู้ตรวจข้อทดสอบโดยตรงที่มีประสบการณ์ตรวจข้อทดสอบเรียงความที่ใช้ในการศึกษาครั้งนี้ทางอ้อม $[\Phi(\lambda_0)_1 = 0.9211 > [\Phi(\lambda_0)_2 = 0.9189]$

นอกจากนี้จากตารางที่ 3 พบว่าค่าสัมประสิทธิ์การอ้างอิงสรุปที่เกิดจากการตรวจข้อทดสอบของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงสูงกว่าค่าที่เกิดจากผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อมอย่างมีนัยสำคัญ กล่าวคือ $\Sigma\rho^2 = 0.66086 > \Sigma\rho^2 = 0.52018$ ($p = 0.5$)

ดังนั้น อาจสรุปได้ว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ $\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ตรง และ random (สุ่มหรือไม่คงที่) > $\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ทางอ้อม และ random

อนึ่ง ผลการศึกษาดังกล่าวแล้วสอดคล้องกับสมมุติฐานในการวิจัยข้อที่ 4 ที่ระบุไว้ในบทที่ 1 ทุกประการ แต่เป็นที่น่าสังเกตว่า ค่าสัมประสิทธิ์การอ้างอิงสรุปที่ได้จากแบบจำลองสุ่ม (random model) ในตารางที่ 3 มีค่าต่ำกว่าค่าดังกล่าวที่ได้จากแบบจำลองผสม (mixed model) ในตารางที่ 2 ซึ่งมีปัจจัยประกอบ (facet) บางชนิดเป็นแบบคงที่ (fixed) และบางชนิดเป็นแบบสุ่ม (random)

2. เมื่อกำหนดให้ผู้ตรวจบางประเภทตรวจข้อทดสอบบางข้อของผู้สอบทุกคน หรือ P x (R:I)

2.1 ในกรณีที่เป็นแบบจำลองผสม (mixed model)

λ_0	$R_1(\Phi)$	$R_2(\Phi)$	$R_1(\Sigma\rho^2)$	$R_2(\Sigma\rho^2)$
0%	0.9715	0.9620	0.84055	0.72610
10%	0.9711	0.9616		
20%	0.9708	0.9611		
30%	0.9704	0.9606		
40%	0.9699	0.9601		
50%	0.9696	0.9597		
60%	0.9692	0.9592		
70%	0.9688	0.9587		
80%	0.9684	0.9581		
90%	0.9679	0.9576		
100%	0.9675	0.9571		
$\bar{X}$	0.9696	0.9596		
S.D.	0.001	0.002		
n	11	11	200	200
r_{xy}	0.999		0.8012	
t	104.98*		6.4508*	

* $p < 0.05$

ตารางที่ 4 : การทดสอบความแตกต่างของค่า $\Phi(\lambda_0)$ และ $\Sigma\rho^2$ ในกรณีของ P x (R:I) เมื่อเป็นแบบจำลองผสม

จากตารางที่ 4 จะสังเกตเห็นได้ว่า ผลการศึกษาที่เกิดขึ้นเหมือนกับผลที่ได้จากตารางที่ 2 ทุกประการ แม้ว่าเงื่อนไขของการศึกษาของปัจจัยประกอบจะแตกต่างกันก็ตาม ทำให้สามารถสรุปได้ว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงจำนวน 2 คน เมื่อตรวจข้อทดสอบบางข้อของผู้สอบทุกคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ $\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R:I) เมื่อ n_r = มีประสบการณ์ตรง และ fixed > $\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ p x (R:I) เมื่อ n_r = มีประสบการณ์ทางอ้อมและ fixed

ผลการศึกษาดังกล่าวสอดคล้องกับสมมุติฐานของการวิจัยข้อที่ 2 ทุกประการ

2.2 ในกรณีที่เป็นแบบจำลองสุ่ม (random model)

λ_0	$R_1(\Phi)$	$R_2(\Phi)$	$R_1(\Sigma\rho^2)$	$R_2(\Sigma\rho^2)$
0%	0.9259	0.9238	0.66086	0.52018
10%	0.9249	0.9230		
20%	0.9241	0.9221		
30%	0.9231	0.9212		
40%	0.9222	0.9203		
50%	0.9212	0.9193		
60%	0.9202	0.9184		
70%	0.9192	0.9174		
80%	0.9182	0.9164		
90%	0.9171	0.9154		
100%	0.9161	0.9144		
$\bar{X}$	0.9211	0.9192		
S.D.	0.003	0.003		
n	11	11	200	200
r_{xy}	0.999		0.8012	
t	51.25*		4.1003*	

* $p < 0.05$

ตารางที่ 5 : การทดสอบความแตกต่างของค่า $\Phi(\lambda_0)$ และ $\Sigma\rho^2$ ในกรณีของ P x (R=I) เมื่อเป็นแบบจำลองสุ่ม

จากตารางที่ 5 จะสังเกตเห็นได้ว่าผลการศึกษาที่เกิดขึ้นเหมือนกับผลที่ได้จากตารางที่ 3 ทุกประการ แม้ว่าเงื่อนไขของการศึกษาของปัจจัยประกอบจะแตกต่างกันก็ตาม ทำให้สามารถสรุปได้ว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบบางข้อของผู้สอบทุกคนจะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ $\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ p x (R:I) เมื่อ n_r = มีประสบการณ์ตรง และ random < $\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ p x (R:I) เมื่อ n_r = มีประสบการณ์ทางอ้อมและ random

ผลการศึกษาดังกล่าวสอดคล้องกับสมมุติฐานของการวิจัยข้อที่ 5 ทุกประการ

3. เมื่อกำหนดให้ผู้ตรวจบางประเภทตรวจข้อทดสอบทุกข้อของผู้สอบบางคน หรือ I x (P:R)

3.1 ในกรณีที่เป็นแบบจำลองผสม (mixed model)

λ_0	$R_1(\Phi)$	$R_2(\Phi)$	$R_1(\Sigma\rho^2)$	$R_2(\Sigma\rho^2)$
0%	0.9999	0.9999	0.99971	0.99991
10%	0.9999	0.9999		
20%	0.9999	0.9999		
30%	0.9999	0.9999		
40%	0.9999	0.9999		
50%	0.9999	0.9999		
60%	0.9999	0.9999		
70%	0.9999	0.9999		
80%	0.9999	0.9999		
90%	0.9999	0.9999		
100%	0.9999	0.9999		
$\overline{X}$	0.9999	0.9999		
S.D.	0.00	0.00		
n	11	11	200	200
r_{xy}	1.00		0.8012	
t	0.00		-14.5553*	

* $p < 0.05$

ตารางที่ 6 : การทดสอบความแตกต่างของค่า $\Phi(\lambda_0)$ และ $\Sigma\rho^2$ ในกรณีของ I x (P : R) เมื่อเป็นแบบจำลองผสม

จากตารางที่ 6 เป็นที่น่าสังเกตว่าดัชนีความเชื่อถือของผู้ตรวจข้อทดสอบเรียงความที่มีประสบการณ์การตรวจข้อสอบโดยตรง และผู้ที่มีประสบการณ์การตรวจทางอ้อม ณ จุดตัดแต่ละจุดสูงมาก คือ 0.9999 เท่ากัน ทำให้ค่าเฉลี่ยของค่าดังกล่าวสูงมากเท่ากับ 0.9999 ทั้ง 2 กรณี และไม่แตกต่างกันอย่างมีนัยสำคัญ (p = 05)

ส่วนค่าสัมประสิทธิ์การอ้างอิงสรุปนั้น ปรากฏว่า การตรวจข้อทดสอบโดยผู้ที่มีประสบการณ์ตรวจทางอ้อมมีค่าดังกล่าวสูงกว่าค่าของผู้ตรวจที่มีประสบการณ์โดยตรงอย่างมีนัยสำคัญ กล่าวคือ 0.99991 > 0.99971 (p =0.05) แม้ว่าค่าดังกล่าวจะมีขนาดสูงมากและใกล้เคียงกันมากก็ตาม

ดังนั้น จึงอาจสรุปได้ว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงจำนวน 2 คน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบบางคนจะมีสัมประสิทธิ์การอ้างอิงสรุปต่ำกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม แต่จะมีค่าดัชนีความเชื่อถือโดยเฉลี่ยไม่แตกต่างกันอย่างมีนัยสำคัญ กล่าวคือ $\Sigma\rho^2$ ของ I x (P:R) เมื่อ n_r = มีประสบการณ์ตรง และ fixed < $\Sigma\rho^2$ ของ I x (P:R) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed แต่ $\Phi(\lambda_o)_1 = \Phi(\lambda_o)_2$ เมื่อ p =0.05

ข้อค้นพบดังกล่าวข้างต้นขัดแย้งกับสมมุติฐานในการวิจัยข้อที่ 3 ทั้งกรณีของค่าสัมประสิทธิ์การอ้างอิงสรุปและค่าดัชนีความเชื่อถือ

3.2 ในกรณีที่เป็นแบบจำลองสุ่ม (random model)

λ_o	$R_1(\Phi)$	$R_2(\Phi)$	$R_1(\Sigma\rho^2)$	$R_2(\Sigma\rho^2)$
0%	0.8265	0.8308	0.44340	0.31235
10%	0.8244	0.8288		
20%	0.8222	0.8268		
30%	0.8200	0.8249		
40%	0.8177	0.8229		
50%	0.8155	0.8208		
60%	0.3132	0.8187		
70%	0.8108	0.8166		
80%	0.8084	0.8144		
90%	0.8060	0.8123		
100%	0.8035	0.8100		
$\bar{X}$	0.8153	0.8207		
S.D.	0.008	0.007		
n	11	11	200	200
r_{xy}	0.999		0.8012	
t	-19.97*		2.4306*	

* p <0.05

ตารางที่ 7 : การทดสอบความแตกต่างของค่า $\Phi(\lambda_o)$ และ $\Sigma\rho^2$ ในกรณีของ I x (P : R) เมื่อเป็นแบบจำลองสุ่ม

จากตารางที่ 7 พบว่า โดยเฉลี่ยแล้วค่าดัชนีความเชื่อถือของผู้ตรวจข้อสอบเรียงความที่มีประสบการณ์ในการตรวจข้อทดสอบโดยตรงต่ำกว่าค่าดังกล่าวของผู้ตรวจที่มีประสบการณ์ในการตรวจทางอ้อมอย่างมีนัยสำคัญ คือ 0.8153 < 0.8207 (p = 0.5) แต่ละค่าสัมประสิทธิ์การอ้างอิงสรุปของผู้ตรวจข้อทดสอบเรียงความที่มีประสบการณ์ในการตรวจข้อทดสอบโดยตรงสูงกว่าค่าดังกล่าวของผู้ตรวจที่มีประสบการณ์ในการตรวจทางอ้อมอย่างมีนัยสำคัญ คือ 0.44340 > 0.31235 (p = 0.5)

ดังนั้น จึงอาจกล่าวสรุปได้ว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบบางคน จะมีสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม แต่จะมีค่าดัชนีความเชื่อถือโดยเฉลี่ยต่ำกว่าอย่างมีนัยสำคัญ กล่าวคือ $\Sigma\rho^2$ ของ I x (P:R) เมื่อ n_r = มีประสบการณ์ตรงและ random > $\Sigma\rho^2$ ของ I x (P:R) เมื่อ n_r = มีประสบการณ์ทางอ้อมและ random แต่ $\Phi(\lambda_0)_1 < \Phi(\lambda_0)_2$ เมื่อ p =0.05

ข้อค้นพบดังกล่าวข้างต้นในส่วนที่เกี่ยวกับค่าสัมประสิทธิ์การอ้างอิงสรุป ($\Sigma\rho^2$) สอดคล้องกับสมมุติฐานในการวิจัยข้อที่ 6 แต่ส่วนที่เกี่ยวกับค่าดัชนีความเชื่อถือ [$\Phi(\lambda_0)$] ขัดแย้งกับสมมุติฐานดังกล่าว

ข. การตรวจสอบโปรแกรม

1. การตรวจสอบผลการคำนวณหาค่าดัชนีความเชื่อถือ

ผู้วิจัยได้อาศัยข้อมูล วิธีการคำนวณ และผลของการคำนวณหาค่าดัชนีของความเชื่อถือของ Brennan (Brennan,1979:54-58) เพื่อศึกษาเปรียบเทียบความถูกต้องในการคำนวณของโปรแกรมที่ได้เขียนขึ้น และผลของการเปรียบเทียบ ปรากฏดังนี้

จุดตัด (λ_0)	$\Phi(\lambda_0)$ จาก Brennan	$\Phi(\lambda_0)$ จาก โปรแกรม	หมายเหตุ
40%	0.807	0.80625	ในกรณีของ PxI
50%	0.754	0.75351	แบบ fixed model
60%	0.760	0.75947	เมื่อ P=10 และ
70%	0.817	0.81693	I=12 (รายละเอียดดู
80%	0.875	0.87456	ได้จากภาคผนวก
90%	0.915	0.91443	ก.)
$\bar{X}$	0.82133	0.82085	
S.D.	0.06352	0.06352	
r_{xy}	0.999		
t	0.11948		

ตารางที่ 8 : การเปรียบเทียบค่า $\Phi(\lambda_0)$ ที่ได้จากโปรแกรม และจากการคำนวณโดย Brennan เมื่อใช้ข้อมูลเดียวกัน

จากตารางที่ 8 พบว่า ค่าดัชนีความเชื่อถือที่คำนวณได้จากโปรแกรมที่เขียนขึ้นกับค่าที่ได้จากการคำนวณโดย Brennan มีความสัมพันธ์กันสูงมาก คือ r_{xy} = 0.999 และค่าที่ได้โดยเฉลี่ยแตกต่างกันอย่างไม่มีนัยสำคัญ (p = 0.5) แสดงว่า โปรแกรมที่เขียนขึ้นสามารถคำนวณค่าดัชนีความเชื่อถือได้อย่างถูกต้อง

2. การตรวจสอบผลการคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุป

ผู้วิจัยได้อาศัยข้อมูล วิธีการคำนวณ และผลการคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุปของ Brennan (Brennan, 1983:7-90, 142-153) เพื่อศึกษาเปรียบเทียบความถูกต้องในการคำนวณของโปรแกรมที่ได้เขียนขึ้น และผลของการเปรียบเทียบ ปรากฏดังนี้

model	$\Sigma\rho^2$ จาก Brennan	$\Sigma\rho^2$ จาก โปรแกรม	หมายเหตุ
	0.32562	0.32562	รายละเอียดดูได้จาก
P x (I:R)	0.44790	0.44790	ภาคผนวก ก.
I = random	0.51199	0.51199	
R = random	0.55144	0.55144	
	0.26622	0.26622	
I : P:R	0.37177	0.37177	
I = random	0.42830	0.42830	
R = random	0.46370	0.46370	
	0.45398	0.45398	
P x (I:R)	0.62446	0.62446	
I = random	0.71382	0.71382	
R =fixed	0.76882	0.76882	
$\bar{X}$	0.49400	0.49400	
S.D.	0.15034	0.15034	
r_{xy}	1.00		
t	0.00		

ตารางที่ 9 : การเปรียบเทียบค่า $\Sigma\rho^2$ ที่ได้จากโปรแกรม และจากการคำนวณโดย Brennan เมื่อใช้ข้อมูลเดียวกัน

จากตารางที่ 9 พบว่าค่าสัมประสิทธิ์การอ้างอิงสรุปที่คำนวณได้จากโปรแกรมที่เขียนขึ้นกับค่าที่คำนวณได้โดย Brennan เท่ากันทุกค่า สำหรับเงื่อนไขต่าง ๆ จึงทำให้ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าทั้งสองสมบูรณ์ คือ $r_{xy} = 1.00$ และค่าสัมประสิทธิ์การอ้างอิงสรุปโดยเฉลี่ยเท่ากัน รวมทั้งค่า S.D. ก็เท่ากันด้วย จนทำให้ค่าสัมประสิทธิ์การอ้างอิงสรุปโดยเฉลี่ยเท่ากันทุกประการ จึงแสดงว่าโปรแกรมที่เขียนขึ้นคำนวณค่าดังกล่าวได้อย่างถูกต้อง

บทที่ 5

สรุปผล อภิปรายผล และข้อเสนอแนะ

ความสำคัญและที่มาของปัญหาการวิจัย

แบบทดสอบเรียงความเป็นเครื่องมือการทดสอบที่มีความสำคัญมาก เพราะสามารถวัดความรู้ความสามารถในระดับสูง เช่น การวิเคราะห์ สังเคราะห์ และการประเมินค่าได้ แต่ว่าแบบทดสอบแบบเรียงความมักมีความเที่ยงในการทดสอบต่ำ และเสียเวลาในการตรวจมาก ปัจจัยที่ทำให้แบบทดสอบชนิดมีความเที่ยงต่ำเกิดจากหลายสาเหตุ เช่น คำถามไม่ครอบคลุมเนื้อหา ความไม่จำเพาะเจาะจงของคำถาม และความเป็นอัตนัยของการตรวจให้คะแนน เป็นต้น ดังนั้นจึงทำให้นักทดสอบแสวงหาวิธีการที่จะทำให้ความเที่ยงในการทดสอบสูงขึ้น เช่น การสร้างเกณฑ์ในการตรวจให้คะแนน และการเพิ่มจำนวนข้อทดสอบให้มากขึ้น หรือเพิ่มจำนวนผู้ตรวจข้อทดสอบขึ้น เป็นต้น แต่วิธีต่างๆ เหล่านี้มักก่อให้เกิดปัญหาอื่นตามมา เช่น ทำให้เสียเวลาในการตรวจเพิ่มขึ้น และต้องเพิ่มบุคลากรในการตรวจขึ้น เป็นต้น

ในปัจจุบันนี้มีทฤษฎีใหม่เกี่ยวกับการทดสอบและประเมินผลชื่อทฤษฎีการอ้างอิงสรุป (Generalizability Theory) ซึ่งสามารถวิเคราะห์ความแปรปรวนจากปัจจัยประกอบหลายแห่ง (multifacet analysis of variance) ที่เกี่ยวกับผลของคะแนนที่ได้จากทดสอบ เช่น จำนวนข้อทดสอบ จำนวนผู้ตรวจข้อทดสอบ จำนวนผู้สอบข้อทดสอบ เวลาในการสอบ และการได้มาของปัจจัยประกอบ (facet) เหล่านี้ว่า สามารถเป็นแบบจำลองคงที่ (fixed model) หรือแบบจำลองสุ่ม (random model) หรือรูแบบผสม (mixed model) ก็ได้ ทั้งนี้เพื่อใช้คำนวณ หาค่าสัมประสิทธิ์การอ้างอิงสรุป (generalizability coefficient) ซึ่งมีลักษณะคล้ายกับค่าความเที่ยงของแบบทดสอบอย่างหนึ่ง แต่จะถูกต้องมากยิ่งขึ้นเนื่องจากสามารถนำเอาปัจจัยประกอบต่างๆ ที่เกี่ยวข้องกับการทดสอบมาร่วมในการคำนวณหาค่าดังกล่าวได้ด้วย

ดังนั้น ผู้วิจัยจึงมีความสนใจที่จะทำการประยุกต์แนวความคิดของทฤษฎีการอ้างอิงสรุปมาใช้ในการตรวจให้คะแนนแบบทดสอบเรียงความวิชาภาษาอังกฤษของสถาบันภาษาจุฬาลงกรณ์มหาวิทยาลัย ที่ให้บริการทดสอบแก่นิสิต นักศึกษาและบุคคลทั่วไป คือ แบบทดสอบวัดสมิทธิภาพทั่วไปทางภาษาอังกฤษ ของจุฬาลงกรณ์มหาวิทยาลัย (Chulalongkorn University Test of English Proficiency : CU-TEP) และเพื่อเขียนโปรแกรมคอมพิวเตอร์เพื่อใช้คำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุปสำหรับเงื่อนไขต่างๆ ด้วย

วัตถุประสงค์ของการวิจัย

1. เพื่อประยุกต์แนวความคิดของทฤษฎีการอ้างอิงสรุปมาใช้ในการตรวจให้คะแนนการสอบแบบเรียงความภาษาอังกฤษในหลายๆเงื่อนไข ทั้งนี้เพื่อศึกษาว่าเงื่อนไขใดทำให้การตรวจมีความเที่ยงมากที่สุด
2. เพื่อสร้างโปรแกรมคอมพิวเตอร์สำหรับใช้ในการคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุปในเงื่อนไขต่างๆ ในการทดสอบ

ขอบเขตของการวิจัย

1. ปัจจัยประกอบ (facet) มีเพียง 3 อย่าง คือ ผู้สอบ ข้อทดสอบ และผู้ตรวจข้อทดสอบ
2. เงื่อนไขที่เป็นไปได้ในทางปฏิบัติไม่เกิน 6 อย่าง
3. จุดตัดของคะแนนอยู่ระหว่าง 0.0%-100% เมื่อมีพิสัย = 10%

4. การตรวจข้อทดสอบเป็นแบบสังเคราะห์ โดยอาศัยแถบระดับความสามารถ (ability band) เป็นเกณฑ์ในการให้คะแนนผลการทดสอบ

ตัวแปรในการวิจัย

ก. ตัวแปรต้น ได้แก่ ปัจจัยประกอบ 3 อย่าง คือ

1. ข้อทดสอบ(I)จำนวน 3 ข้อ ซึ่งเป็นแบบสุ่ม (random design)
2. ผู้ตรวจข้อทดสอบ(R)จำนวน 2 ประเภท ซึ่งเป็นแบบคงที่ (fixed design)
3. ผู้สอบ(P)จำนวน 200 คน ซึ่งเป็นแบบสุ่ม (random design)

ข. ตัวแปรตาม ได้แก่

1. ค่าสัมประสิทธิ์การอ้างอิงสรุป ($\sum\rho^2$) ของการตรวจคะแนนแต่ละเงื่อนไข
2. ค่าดัชนีความเชื่อถือ [$\Phi(\lambda_0)$] ของผลการตรวจคะแนนแต่ละเงื่อนไข

สมมุติฐานในการวิจัย

สมมุติฐานที่ 1

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงจำนวน 2 คน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ตรง และ fixed >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed

สมมุติฐานที่ 2

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงจำนวน 2 คน เมื่อตรวจข้อทดสอบบางข้อของผู้สอบทุกคนจะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ตรง และ fixed >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed

สมมุติฐานที่ 3

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงจำนวน 2 คน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบบางคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ตรง และ fixed >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ fixed

สมมุติฐานที่ 4

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ตรง และ random >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x I x R เมื่อ n_r = มีประสบการณ์ทางอ้อม และ random

สมมุติฐานที่ 5

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบบางข้อของผู้สอบทุกคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ตรง และ random >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ P x (R : I) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ random

สมมุติฐานที่ 6

ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์การตรวจโดยตรงทุกคน เมื่อตรวจข้อทดสอบทุกข้อของผู้สอบบางคน จะมีสัมประสิทธิ์การอ้างอิงสรุปและดัชนีความเชื่อถือโดยเฉลี่ยสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์การตรวจทางอ้อม กล่าวคือ

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ตรง และ random >

$\Sigma\rho^2$ และ $\Phi(\lambda_0)$ ของ I x (P : R) เมื่อ n_r = มีประสบการณ์ทางอ้อม และ random

วิธีดำเนินการวิจัย

1. เอกภพของการวิจัย (จำนวนประชากร)

1.1 เอกภพของสิ่งสังเกตที่ยอมรับได้ (universe of admissible observation) ได้แก่ เงื่อนไขผสมที่เป็นไปได้ระหว่างปัจจัยประกอบด้าน ข้อทดสอบ ผู้ตรวจข้อทดสอบ และผู้สอบตามเงื่อนไขที่มีอยู่จริง ได้แก่

1.) ข้อทดสอบจำนวนอนันต์

2.) ผู้ตรวจข้อทดสอบ 2 ประเภท คือ ผู้มีประสบการณ์ตรวจข้อทดสอบ CU-TEP โดยตรง และผู้ที่มีประสบการณ์ตรวจทางอ้อม

3.) ผู้สอบจำนวนอนันต์

1.2) เอกภพของการอ้างอิงสรุป (universe of generalizability) ได้แก่ เงื่อนไขหรือสถานการณ์ผสมที่อาจเป็นไปได้ของแต่ละปัจจัยประกอบที่ต้องการอ้างอิงสรุป (generalize) ซึ่งในการวิจัยนี้ได้แก่ 6 เงื่อนไข ที่ระบุไว้ในสมมุติฐานการวิจัย กล่าวคือ

1.) PxIxR เมื่อ R = fixed และ P,I = random

2.) PxIxR เมื่อ I,P,R =random

3.) Px(R:I) เมื่อ R = fixed และ P,I = random

4.) Px(R:I) เมื่อ I,P,R = random

5.) Ix(P:R) เมื่อ R = fixed และ P,I = random

6.) Ix(P:R) เมื่อ I,P,R = random

2. ตัวอย่าง

ตัวอย่างได้แก่ ผลคูณของปัจจัยประกอบที่ทำให้เกิดเงื่อนไขหรือสถานการณ์ คือ (n_i x n_p x n_r) = 3x2x200 หรือ = 1,200 เงื่อนไข หรือสถานการณ์

3. เครื่องมือในการวิจัย

ได้แก่ ข้อทดสอบ CU-TEP ซึ่งเป็นแบบทดสอบแบบเรียงความภาษาอังกฤษจำนวน 3 ข้อ ที่คณะกรรมการสร้างและพัฒนาแบบทดสอบวัดสมิทธิภาพทั่วไปทางภาษาอังกฤษของสถาบันภาษาจุฬาลงกรณ์มหาวิทยาลัยสร้างและพัฒนาขึ้น เพื่อใช้ทดสอบความสามารถทางด้านการเขียนของนิสิต นักศึกษา และบุคคลที่สนใจทั่วไป

4. การเก็บรวบรวมข้อมูล

1. สุ่มกระดาษคำตอบของแบบทดสอบ CU-TEP จำนวน 200 ฉบับ จากจำนวน 343 ฉบับ จากผู้สอบที่เป็นนิสิต นักศึกษา และผู้สนใจทั่วไปที่สมัครสอบชิงทุนรัฐบาลเพื่อไปศึกษาต่อ ณ ต่างประเทศ ประจำปี 2535 ของคณะกรรมการข้าราชการพลเรือน (ก.พ.)

2. .ให้ผู้ตรวจที่มีประสบการณ์ตรวจข้อทดสอบเรียงความของแบบทดสอบ CU-TEP ที่มีประสบการณ์ตรวจโดยตรงทำการตรวจข้อทดสอบแต่ละข้อ โดยวิธีประเมินผลรวม (holistic method) ด้วยอาศัยแถบระดับความสามารถ (ability band) เป็นเกณฑ์ ผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงนี้มีทั้งหมด 6 คน แต่ละคนตรวจทุกข้อและตรวจประมาณ 70 ฉบับ และบันทึกคะแนนแต่ละข้อไว้ แล้วจึงเวียนให้อีกผู้หนึ่งตรวจโดยวิธีเดียวกัน

3. ให้ผู้ตรวจข้อทดสอบเรียงความที่ไม่เคยตรวจข้อทดสอบ CU-TEP มาก่อน แต่เป็นอาจารย์สอนภาษาอังกฤษที่เคยตรวจข้อทดสอบเรียงความมาก่อนทำการตรวจข้อทดสอบแต่ละฉบับ 2 คน โดยอาศัยวิธีการเดียวกันกับที่กล่าวมาแล้วข้างต้น ผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อมนี้มีทั้งหมด 6 คน แต่ละคนตรวจทุกข้อและตรวจประมาณ 70 ฉบับเช่นเดียวกัน

4. นำคะแนนที่ได้จากการตรวจข้อทดสอบของผู้สอบแต่ละคนและแต่ละข้อของผู้ตรวจแต่ละประเภทซึ่งตรวจ 2 ครั้ง มาเตรียมข้อมูลเพื่อใช้ในการวิเคราะห์ด้วยเครื่องคอมพิวเตอร์สำหรับทดสอบสมมุติฐานข้อที่ 3

5. สำหรับข้อมูลที่ใช้เพื่อการทดสอบสมมุติฐานข้อที่ 1, 2, 4, 5 และ 6 เกิดจากการจำลองข้อมูลเสมือนจริง (generate) จากข้อมูลจริงในข้อที่ 1 ด้วยโปรแกรมคอมพิวเตอร์ใช้งานที่สร้างขึ้นสำหรับการวิจัยครั้งนี้

5. การสร้างและพัฒนาโปรแกรมคอมพิวเตอร์

1. ศึกษาแนวคิดของทฤษฎีการอ้างอิงสรุปจากหนังสือที่เกี่ยวข้อง คือ Elements of Generalizability Theory (Brennan, 1983) และ Some Applications of Generalizability Theory to the Dependability of Domain-Referenced Tests (Brennan, 1979)

2. เขียนโปรแกรมคอมพิวเตอร์ขึ้นด้วยภาษา FORTRAN 77 ตามสูตร และแนวคิดในหนังสือดังกล่าวแล้วข้างต้น (Brennan, 1979 : 54-58 ; Brennan, 1983 : 7-90,142-153) โดยมุ่งที่จะใช้โปรแกรมดังกล่าวกับเครื่องไมโครคอมพิวเตอร์ แต่เนื่องจากโปรแกรมมีขนาดยาวมาก จึงเกินขีดจำกัดของโปรแกรมเปลี่ยนรหัสโปรแกรมให้เป็นภาษาเครื่อง (compiler) ดังนั้น จึงต้องเปลี่ยนไปเขียนโปรแกรมด้วยภาษา FORTRAN เพื่อใช้กับเครื่องคอมพิวเตอร์ขนาดใหญ่ (mainframe) แทน

3. ทำการทดสอบการทำงานของโปรแกรม โดยอาศัยข้อมูลและผลลัพธ์ที่แสดงไว้ในหนังสือทั้ง 2 เล่มดังกล่าว เป็นเครื่องตรวจสอบความถูกต้อง แล้วปรับปรุงแก้ไขโปรแกรมเพื่อให้สามารถทำงานได้ผลถูกต้องเหมือนกับที่แสดงไว้ในหนังสือดังกล่าวแล้ว

6. การวิเคราะห์ข้อมูล

1. ใช้โปรแกรมที่สร้างขึ้นเพื่อคำนวณหาค่าดัชนีความเชื่อถือ และค่าสัมประสิทธิ์การอ้างอิงสรุปของปัจจัยประกอบต่างๆ ในเงื่อนไขต่างๆ ที่ต้องการศึกษา

2. ทดสอบความแตกต่างของค่าดัชนีความเชื่อถือและค่าสัมประสิทธิ์การอ้างอิงสรุปที่ได้จากการตรวจให้คะแนนของผู้
ตรวจที่มีประสบการณ์ตรวจโดยตรงกับผู้ตรวจข้อทดสอบที่มีประสบการณ์ตรวจทางอ้อมด้วย t-test

3. ทดสอบความแตกต่างของค่าดัชนีความเชื่อถือและค่าสัมประสิทธิ์การอ้างอิงสรุปที่ได้จากโปรแกรมที่เขียนขึ้นกับค่าต่างๆ ดังกล่าวที่ปรากฏในหนังสือทั้ง 2 เล่ม ดังกล่าว แล้วด้วย t-test เพื่อตรวจสอบความถูกต้องของโปรแกรม

สรุปผลการวิจัย

ก. ด้านการประยุกต์ใช้แนวคิดทฤษฎีการอ้างอิงสรุปเพื่อการตรวจข้อทดสอบเรียงความ

1. ในกรณีที่กำหนดให้ผู้ตรวจข้อทดสอบทุกประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน หรือ [P×I×R]

1.1 เมื่อ R = fixed และ P,I = random ปรากฏว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง มีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญ กล่าวคือ $\Phi(\lambda_0) = 0.9696 > \Phi(\lambda_0) = 0.9596$ และ $\Sigma\rho^2 = 0.84055 > \Sigma\rho^2 = 0.72610$ ($p=0.05$)

1.2 เมื่อ I,P,R = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญ เช่นเดียวกันกับข้อ 1.1 กล่าวคือ $\Phi(\lambda_0) = 0.9211 > \Phi(\lambda_0) = 0.9189$ และ $\Sigma\rho^2 = 0.66086 > \Sigma\rho^2 = 0.52018$ ($p=0.05$)

2. ในกรณีที่กำหนดให้ผู้ตรวจข้อทดสอบบางประเภทตรวจข้อทดสอบบางข้อของผู้สอบทุกคน หรือ [Px(R:I)]

2.1 เมื่อ R = fixed และ P,I = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง มีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญ กล่าวคือ $\Phi(\lambda_0) = 0.9696 > \Phi(\lambda_0) = 0.9596$ และ $\Sigma\rho^2 = 0.84055 > \Sigma\rho^2 = 0.72610$ ($p=0.05$)

2.2 เมื่อ I,P,R = random ปรากฏว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง มีดัชนีความเชื่อถือโดยเฉลี่ยและสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าผลการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ทางอ้อมอย่างมีนัยสำคัญเช่นเดียวกับข้อ 2.1 กล่าวคือ $\Phi(\lambda_0) = 0.9211 > \Phi(\lambda_0) = 0.9192$ และ $\Sigma\rho^2 = 0.66086 > \Sigma\rho^2 = 0.52018$ ($p=0.05$)

3. เมื่อกำหนดให้ผู้ตรวจบางประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบบางคน หรือ [I x(P:R)]

3.1 เมื่อ R = fixed และ P,I = random ปรากฏว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยเท่ากับค่าดังกล่าวของผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อมและมีสัมประสิทธิ์การอ้างอิงสรุปต่ำ

กว่าอย่างมีนัยสำคัญ กล่าวคือ $\Phi(\lambda_0)=0.9999 > \Phi(\lambda_0)=0.9999$ และ $\Sigma\rho^2=0.99971 > \Sigma\rho^2=0.99991$ $(p=0.05)$

3.2 เมื่อ I,P,R = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยต่ำกว่าค่าดังกล่าวของผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ทางอ้อมแต่มีสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าอย่างมีนัยสำคัญ กล่าวคือ $\Phi(\lambda_0)=0.8153 < \Phi(\lambda_0)=0.8207$ และ $\Sigma\rho^2=0.44340 > \Sigma\rho^2=0.31235$ $(p=0.05)$

ดังนั้น ผลของการวิจัยครั้งนี้สอดคล้องกับสมมุติฐานในการวิจัย 4 ข้อ คือ ข้อที่ 1-2 และ 4-5 แต่ขัดแย้งกับสมมุติฐานการวิจัย 1 ข้อ คือข้อที่ 3 และขัดแย้งกับบางส่วนของสมมุติฐาน 1 ข้อ คือ ข้อที่ 6 และในกรณีที่กำหนดให้ผู้ตรวจบางประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบบางคน หรือ [I x(P:R)] เมื่อ R = fixed และ P,I = random (เงื่อนไขตามสมมุติฐานที่ 5) ปรากฏว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง และผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อม มีดัชนีความเชื่อถือและมีสัมประสิทธิ์การอ้างอิงสรุปโดยเฉลี่ยสูงที่สุด เมื่อเปรียบเทียบกับเงื่อนไขอื่น ๆ อีก 5 เงื่อนไข

ข. ด้านการสร้างโปรแกรมคอมพิวเตอร์เพื่อใช้งาน

ปรากฏว่า โปรแกรมที่สร้างขึ้นสามารถคำนวณหาค่าสัมประสิทธิ์การอ้างอิงสรุป $(\Sigma\rho^2)$ และค่าดัชนีความเชื่อถือ $[\Phi(\lambda_0)]$ สำหรับเงื่อนไขต่างๆ ได้อย่างถูกต้องเมื่อเปรียบเทียบกับค่าต่างๆ ดังกล่าว จากข้อมูลและผลลัพธ์ของการคำนวณจากหนังสือที่ใช้ในการอ้างอิงดังกล่าวแล้ว

<u>การอภิปรายผล</u>

เนื่องจากผลการวิจัยครั้งนี้ขัดแย้งกับสมมุติฐานการวิจัย 1 ข้อ คือข้อที่ 3 และขัดแย้งกับบางส่วนของสมมุติฐาน 1 ข้อ คือ ข้อที่ 6 ประเด็นของความขัดแย้งกับสมมุติฐานการวิจัยเป็นเรื่องที่น่าสนใจมากดังนี้

ก. เมื่อกำหนดให้ผู้ตรวจบางประเภทตรวจข้อทดสอบทุกข้อของผู้สอบบางคน คือ I x(P:R) เมื่อ R = fixed และ P,I = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยเท่ากับค่าดังกล่าวของผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ทางอ้อมแทนที่จะสูงกว่าและมีสัมประสิทธิ์การอ้างอิงสรุปต่ำกว่าอย่างมีนัยสำคัญแทนที่จะสูงกว่า

การที่ $\Phi(\lambda_0)_1=\Phi(\lambda_0)_2$ อาจเป็นเพราะว่า

1. การให้ผู้ตรวจข้อทดสอบบางประเภทตรวจข้อทดสอบทุกข้อของผู้สอบบางคนทำให้ผู้ตรวจแต่ละประเภทตรวจข้อทดสอบน้อยฉบับแต่ว่าแต่ละฉบับต้องตรวจทุกข้อ การตรวจเช่นนี้อาจทำให้ผู้ตรวจแต่ละประเภทไม่ต้องรีบเร่งมาก จึงอาจมีเวลามากพอที่จะตรวจได้อย่างละเอียด ส่วนการตรวจนั้นผู้ตรวจแต่ละคนอาจใช้วิธีการต่างกัน เช่นอาจตรวจทีละข้อของแต่ละฉบับจนครบหรือตรวจทุกๆ ข้อของแต่ละฉบับก็ได้

2. ผู้ตรวจที่ไม่มีประสบการณ์ในการตรวจแบบทดสอบเรียงความของ CU-TEP โดยตรงเป็นผู้ที่มีประสบการณ์ในการตรวจแบบทดสอบเรียงความภาษาอังกฤษมาก่อนเป็นอย่างดี เพราะทุกคนเป็นอาจารย์สอนภาษาอังกฤษที่มีความรู้และความสามารถมาก ดังนั้น หากมีเวลาในการตรวจมากพอและเข้าใจเกณฑ์ในการตรวจ

ข้อทดสอบเป็นอย่างดีก็อาจสามารถตรวจข้อทดสอบเรียงความได้ดีมากเท่ากับผู้ตรวจที่มีประสบการณ์ในการตรวจข้อทดสอบ CU-TEP โดยตรงได้

อนึ่ง เป็นที่น่าสังเกตว่า ค่า $\Phi(\lambda_0)_1$ และ $\Phi(\lambda_0)_2$ ที่เกิดจากการตรวจให้คะแนนของผู้ตรวจที่มีประสบการณ์ตรวจข้อทดสอบ CU-TEP โดยตรงและผู้ที่มีประสบการณ์ตรวจทางอ้อมในเงื่อนไขต่างๆ มีขนาดใกล้เคียงกันมาก แม้ว่าโดยมากแล้วค่าต่างๆ ของผู้ตรวจที่มีประสบการณ์ในการตรวจโดยตรงจะสูงกว่าเล็กน้อยก็ตาม ดังผลการเปรียบเทียบในตาราง ต่อไปนี้

No.	model	fixed	random	$\Phi(\lambda_0)$		$\Sigma\rho^2$	
				ประสบการณ์ตรง	ประสบการณ์อ้อม	ประสบการณ์ตรง	ประสบการณ์อ้อม
1.	PxIxR	R	P,I	0.9696	0.9596	0.84055	0.72610
2.	PxIxR	-	I,P,R	0.9211	0.9189	0.66086	0.52018
3.	Px(R:I)	R	P,I	0.9696	0.9596	0.84055	0.72610
4.	Px(R:I)	-	I,P,R	0.9211	0.9192	0.66086	0.52018
5.	Ix(P:R)	R	P,I	0.9999	0.9999	0.99971	0.99991
6.	Ix(P:R)	-	I,P,R	0.8153	0.8207	0.44340	0.313235

ตารางที่ 10: สรุปค่า $\Phi(\lambda_0)$ และ $\Sigma\rho^2$ ของการให้คะแนนในเงื่อนไขต่างๆ

จากตารางที่ 10 แสดงให้เห็นว่าโดยทั่วๆไปแล้วผู้ตรวจข้อทดสอบเรียงความที่มีประสบการณ์ในการตรวจโดยตรงและผู้ที่มีประสบการณ์ในการตรวจทางอ้อมมีความสามารถในการตรวจใกล้เคียงกันมาก แต่ว่าความคงเส้นคงวาของผลการตรวจให้คะแนนของผู้ที่มีประสบการณ์ในการตรวจโดยตรงมีมากกว่าจึงทำให้ค่า $\Phi(\lambda_0)$ โดยเฉลี่ยในเงื่อนไขต่างๆ สูงกว่าผลการตรวจให้คะแนนของผู้ที่มีประสบการณ์ในการตรวจทางอ้อมเล็กน้อยแม้ว่าจะแตกต่างกันอย่างมีนัยสำคัญก็ตามและที่น่าสนใจมากก็คือค่าดัชนีความเชื่อถือ ณ จุดตัดต่างๆ ของผู้ตรวจทั้ง 2 ประเภท มีความสัมพันธ์กันสูงมากคือ r_{xy} อยู่ระหว่าง 0.999-1.00 (ดูตารางที่ 2-7)

ดังนั้น จึงอาจกล่าวโดยสรุปได้ว่า ความรู้ความสามารถในการตรวจให้คะแนนข้อทดสอบเรียงความของผู้ตรวจทั้ง 2 ประเภท ใกล้เคียงกันมาก แต่ว่าผู้ตรวจที่มีประสบการณ์ในการตรวจโดยตรงมีความคงเส้นคงวาในการให้คะแนนมากกว่าผู้ตรวจที่มีประสบการณ์ในการตรวจทางอ้อมเล็กน้อย แต่ในกรณีที่กำหนดให้ผู้ตรวจบางประเภทตรวจข้อทดสอบทุกข้อของผู้สอบบางคนจะทำให้ผลการตรวจให้คะแนนของผู้ตรวจทั้ง 2 ประเภท ไม่แตกต่างกันอย่างมีนัยสำคัญ จึงทำให้ค่าดัชนีความเชื่อถือ ณ จุดตัดต่างๆ และโดยเฉลี่ยเท่าเทียมกัน

อนึ่ง สำหรับค่าสัมประสิทธิ์การอ้างอิงสรุปนั้น ตามทฤษฎีแล้วมีค่าเท่ากับค่า KR_{20} ซึ่งเป็นค่าความเที่ยงที่บ่งบอกถึงความคงเส้นคงวาภายใน (internal consistency) ของการให้คะแนนอย่างหนึ่ง (Brennan and Kane, 1977 : 281) และการที่ค่าดังกล่าวของผู้ตรวจที่มีประสบการณ์ในการตรวจข้อทดสอบ CU-TEP โดยตรง มีค่าต่ำกว่าของผู้ตรวจที่มีประสบการณ์ในการตรวจข้อทดสอบ CU-TEP ทางอ้อมอย่างมีนัยสำคัญ คือ $\Sigma\rho_1^2=0.99971 < \Sigma\rho_2^2=0.99991$ $(p=0.05)$ จะเห็นได้ว่าขนาดของค่าทั้ง 2 ดังกล่าวแตกต่าง

กันน้อยมาก คือ เพียง 0.00020 เท่านั้น จึงไม่มีความสำคัญมากนัก แต่การที่ค่าทั้ง 2 แตกต่างกันอย่างมีนัยสำคัญ เป็นเพราะว่าการทดสอบครั้งนี้มีกลุ่มตัวอย่างซึ่งได้แก่ ปัจจัยประกอบด้านผู้สอบจำนวนมาก คือ $n_p = 200$ คน จึงทำให้มีผลโดยตรงกับการเปรียบเทียบความแตกต่าง ซึ่งสามารถสังเกตได้จากสูตรที่ใช้ในการเปรียบเทียบต่อไปนี้ คือ

$$t = \frac{r_{tt1} - r_{tt2}\ (N-2)}{\sqrt{4\,(1-r_{tt1})(1-r_{tt2})(1-r^2_{xy})}};\ df = N-2$$

ดังนั้นด้วยเหตุผลต่างๆ ดังกล่าวแล้วข้างต้น จึงกล่าวได้ว่าผู้ตรวจให้คะแนนข้อทดสอบเรียงความทั้ง 2 ประเภท มีความสามารถในการตรวจให้คะแนนในเงื่อนไข Ix(P:R) เมื่อ R = fixed และ P,I = random แตกต่างกันไม่มากนัก เมื่อพิจารณาหาค่า $\Phi(\lambda_0)$, $\Phi(\lambda_0)$, r_{xy} และ $\Sigma\rho^2$

ข. เมื่อกำหนดให้ผู้ตรวจบางประเภทตรวจข้อทดสอบทุกข้อของผู้สอบบางคน คือ Ix(P:R) เมื่อ I,P,R = random ปรากฏว่าผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรงมีดัชนีความเชื่อถือโดยเฉลี่ยต่ำกว่าค่าดังกล่าวของผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อม แต่ว่ามีสัมประสิทธิ์การอ้างอิงสรุปสูงกว่าอย่างมีนัยสำคัญ

การที่ $\Phi(\lambda_0)_1 < \Phi(\lambda_0)_2$ อาจเป็นเพราะว่า

1. ความรู้ความสามารถของผู้ตรวจที่มีประสบการณ์ในการตรวจข้อทดสอบแบบเรียงความโดยตรง หรือทางอ้อมซึ่งมีอยู่จำนวนมากแตกต่างกันทำให้ผู้ตรวจข้อทดสอบแต่ละประเภทที่สุ่มได้มีความรู้ความสามารถไม่ทัดเทียมกัน จึงอาจเป็นเหตุให้คะแนนที่ได้จากการตรวจข้อทดสอบมีความคงเส้นคงวาแตกต่างกัน แต่ในกรณีที่ R = fixed นั้นผู้ตรวจแต่ละประเภทมีจำนวน
จำกัดและได้รับการคัดสรรมาแล้วเป็นอย่างดีว่ามีความรู้ความสามารถหรือประสบการณ์ในการตรวจข้อทดสอบแบบเรียงความมาก่อน จึงอาจทำให้ความรู้ความสามารถแตกต่างกันไม่มากนัก ดังได้กล่าวมาแล้ว และข้อสังเกตดังกล่าวนี้น่าจะเป็นจริง เพราะจากข้อมูลในตาราง จะสังเกตเห็นได้ชัดว่าทุกเงื่อนไขที่ R และ I,P มีลักษณะเป็น random แล้วค่าดัชนีความเชื่อถือโดยเฉลี่ยและค่าสัมประสิทธิ์การอ้างอิงสรุปของผู้ตรวจแต่ละประเภทต่ำกว่าค่าดังกล่าวเมื่อเงื่อนไขเป็น R = fixed และ I,R = random

2. ค่าดัชนีความเชื่อถือ $[\Phi(\lambda)]$ เป็นค่าความเที่ยงอย่างหนึ่งซึ่งเมื่อจุดตัด (Cutting point) เท่ากับค่าเฉลี่ยของคะแนนสอบจะทำให้ค่าความเชื่อถือนี้เท่ากับค่าความเที่ยงแบบ KR_{21} ซึ่งเป็นค่าความเที่ยงโดยประมาณ และเมื่อจุดตัดและคะแนนเฉลี่ยแตกต่างกันจะทำให้ค่าดัชนีความเชื่อถือเพิ่มมากขึ้น (Brennan and Kane,1977:281) แต่ทว่าค่าสัมประสิทธิ์การอ้างอิงสรุปเป็นค่าความเที่ยงอีกอย่างหนึ่งที่มีค่าเท่ากับความเที่ยงแบบ KR_{20} ซึ่งเป็นค่าความเที่ยงที่ถูกต้องมากที่สุด เนื่องจากเป็นค่าที่ได้จากการคำนวณผลการสอบแต่ละข้อ ไม่ใช่อาศัยค่าเฉลี่ยของผลการสอบเช่นการคำนวณหาค่า KR_{21}

ดังนั้น การที่ค่าเฉลี่ยของดัชนีความเชื่อถือและค่าสัมประสิทธิ์การอ้างอิงสรุปของกลุ่มผู้ตรวจทั้ง 2 ประเภท ไม่สอดคล้องตามกันจึงไม่ใช่เรื่องที่ผิดปกติมากโดยเฉพาะเมื่อความแตกต่างของแต่ละค่ามีไม่มากนัก

ค. ในกรณีที่กำหนดให้ผู้ตรวจบางประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบบางคน หรือ [I x(P:R)] เมื่อ R = fixed และ P,I = random (เงื่อนไขตามสมมุติฐานที่ 5) ปรากฏว่า ผลการตรวจให้คะแนนแบบทดสอบเรียงความของผู้ตรวจที่มีประสบการณ์ตรวจโดยตรง และผู้ตรวจที่มีประสบการณ์ตรวจทางอ้อม มีดัชนีความเชื่อถือและมีสัมประสิทธิ์การอ้างอิงสรุปโดยเฉลี่ยสูงที่สุด เมื่อเปรียบเทียบกับเงื่อนไขอื่น ๆ อีก 5 เงื่อนไข อาจเป็นเพราะว่า

1. ผู้ตรวจข้อทดสอบทั้งสองประเภทเป็นบุคคลที่มีประสบการณ์ในการตรวจข้อทดสอบแบบเรียงความมาแล้วอย่างดี และเป็นบุคคลกลุ่มคัดสรรว่าเป็นผู้ที่มีความรู้ความสามารถทางภาษาอังกฤษมาแล้วเป็นอย่างดี ดังนั้นเมื่อตรวจข้อทดสอบทุกข้อของผู้สอบบางคนทำให้ภาระงานในการตรวจน้อยกว่าการที่จะต้องตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน จึงมีโอกาสที่จะทำให้ผลการตรวจมีความคงที่และสอดคล้องกันมากกว่าการตรวจของเงื่อนไขอื่น ๆ

2. การตรวจให้คะแนนข้อทดสอบเรียงความแต่ละข้อโดยวิธีประเมินผลรวม (wholistic method) ตามแถบระดับความสามารถ (ability band) ซึ่งเป็นวิธีที่ใช้ในการวิจัยครั้งนี้ง่ายต่อการทำความเข้าใจและปฏิบัติสำหรับผู้ที่มีประสบการณ์ในการสอนภาษาอังกฤษมาแล้วเป็นอย่างดี จึงทำให้มีความเข้าใจตรงกันได้ง่ายและมีผลทำให้คะแนนการตรวจมีความแตกต่างกันน้อย (มีความคงที่และสอดคล้องกันมาก)

ดังนั้น การตรวจข้อทดสอบตามเงื่อนไขดังกล่าวแล้วจึงมีดัชนีความเชื่อถือและมีสัมประสิทธิ์การอ้างอิงสรุปโดยเฉลี่ยสูงที่สุด เมื่อเปรียบเทียบกับเงื่อนไขอื่น ๆ อีก 5 เงื่อนไข

ข้อเสนอแนะ

ก. เพื่อการนำไปใช้

1. การตรวจข้อทดสอบเรียงความภาษาอังกฤษควรจะต้องตรวจโดยผู้ตรวจที่มีความรู้ความสามารถในการตรวจ ซึ่งอาจเกิดจากมีประสบการณ์ในการตรวจโดยตรง หรือเกิดจากการเข้าใจเกณฑ์การตรวจเป็นอย่างดี รวมทั้งมีพื้นความรู้ทางภาษาอังกฤษเป็นอย่างดีด้วย หากจำเป็นต้องอาศัยผู้ตรวจอื่นที่ไม่มีความรู้หรือความสามารถดังกล่าว บุคคลเหล่านั้นควรได้รับการฝึกฝนการตรวจให้คะแนนข้อทดสอบเรียงความมาก่อนด้วยจึงจะได้ผลดี

2. วิธีการตรวจให้คะแนนข้อทดสอบเรียงความที่ดีที่สุด ในกรณีที่มีปัจจัยประกอบ 3 อย่าง คือ ผู้ตรวจข้อทดสอบ และผู้สอบ ควรที่จะกำหนดให้ผู้ตรวจเป็นผู้ที่มีประสบการณ์ในการตรวจโดยตรง หรือทางอ้อมและมีพื้นความรู้ความสามารถดี ได้รับการคัดสรรมาแล้วและมีจำนวนจำกัด (fixed) โดยให้ผู้ตรวจบางประเภทตรวจข้อทดสอบทุกข้อ ของผู้สอบบางคน กล่าวคือ Ix(P:R) เมื่อ R = fixed และ P,I = random เช่นแบ่งกระดาษคำตอบของผู้สอบออกเป็นกลุ่มย่อยหลายๆ กลุ่ม แล้วให้ผู้ตรวจที่มีความสามารถดังกล่าวตรวจ โดยผู้ตรวจแต่ละคนตรวจข้อทดสอบทุกข้อของผู้สอบเสร็จแล้วให้ผู้ตรวจอีกคนหนึ่งทำการตรวจซ้ำ และนำคะแนนของผู้ตรวจทั้ง 2 คน มารวมกัน

3. หากไม่จำเป็น ไม่ควรสุ่มผู้ตรวจที่คิดว่ามีความรู้ความสามารถทางภาษาอังกฤษเท่านั้น มาเป็นกรรมการตรวจข้อทดสอบเรียงความ เพราะจะทำให้ความเที่ยงในการตรวจต่ำ ซึ่งเกิดจากการให้คะแนนไม่คงเส้นคงวาเท่าที่ควร

4. ในอนาคตถ้ามีโปรแกรมที่สามารถเปลี่ยนโปรแกรมภาษา FORTRAN ที่มีขนาดใหญ่ เป็นภาษาเครื่องคอมพิวเตอร์ได้ (FORTRAN COMPILER) ควรที่จะทำการเปลี่ยน โปรแกรมดังกล่าว เพื่อให้สามารถใช้งานได้

กับเครื่องไมโครคอมพิวเตอร์ ซึ่งจะสะดวกกว่าการใช้เครื่องคอมพิวเตอร์ขนาดใหญ่ แต่โปรแกรมต้องมีการปรับปรุงแก้ไขบ้างเล็กน้อย ขณะนี้ผู้วิจัยทราบว่าบริษัท Microsoft กำลังพัฒนา Compiler ดังกล่าวอยู่ คาดว่าในอนาคตอันใกล้นี้โปรแกรมดังกล่าวคงจะนำออกวางตลาดได้

อนึ่ง สำหรับท่านที่เชี่ยวชาญภาษาคอมพิวเตอร์อย่างอื่น ท่านก็อาจจะเปลี่ยนภาษา FORTRAN เป็นภาษาอื่นที่เหมาะกับการคำนวณก็ได้ เช่น ภาษา BASIC และ Quick BASIC เป็นต้น แต่ผู้วิจัยไม่ทราบว่าทั้ง 2 ภาษา นี้ โปรแกรม สำหรับเปลี่ยนรหัสเป็นภาษาเครื่อง (compiler) มีขีดจำกัดมากน้อยเพียงใด ดังนั้นท่านที่สนใจควรต้องศึกษาเรื่องนี้ก่อนล่วงหน้าด้วย

ข. เพื่อการวิจัยต่อไป

1. ควรทำการศึกษาเรื่องในทำนองเดียวกันนี้อีก แต่ควรเพิ่มเงื่อนไขของการทดสอบให้มากขึ้น รวมทั้งควรเพิ่มจำนวน
ปัจจัยประกอบ (facet) ที่เกี่ยวข้องให้มากขึ้น เช่นเวลาในการสอบ เวลาในการตรวจให้คะแนน ความรู้ความสามารถของผู้ตรวจ และวิธีการตรวจให้คะแนน เป็นต้น เพื่อศึกษาว่าเงื่อนไขใดบ้างจะทำให้ความเที่ยงในการให้คะแนน (คือค่า $\Phi(\lambda_0)$ และ $\Sigma\rho^2$) สูงอยู่ในระดับที่ยอมรับได้ และสะดวกในทางปฏิบัติ ทั้งนี้เพื่อเป็นการกระตุ้นให้นักทดสอบนำข้อทดสอบแบบเรียงความมาใช้ให้มากยิ่งขึ้น

2. ควรทำการศึกษาด้วยว่าในกรณีที่เป็นข้อทดสอบเรียงความภาษาอื่นที่ไม่ใช่ภาษาอังกฤษ ข้อค้นพบต่างๆ จากการศึกษาครั้งนี้รวมทั้งที่ได้เสนอแนะไว้ในข้อที่ 1 ข้างต้นนี้จะเปลี่ยนแปลงไปอย่างไรหรือไม่ ทั้งนี้เพื่อจะได้มีผู้สนใจนำผลการศึกษาไปใช้ให้เหมาะสมกับการเรียนการสอน หรือการทดสอบแต่ละภาษาต่อไป

บรรณานุกรม

โกวิท ประวาลพฤกษ์ และ สมศักดิ์ สินธุรเวชญ์. **การประเมินในชั้นเรียน.** วัฒนาพานิช, 2527.

จักรกฤษณ์ สำราญใจ. "Generalizability" **วิธีวิจัยวิทยาจบับที่** 3 กันยายน-ธันวาคม 2529.

แดง กลางท่าไค่. "การประยุกต์ทฤษฎีการสรุปอ้างอิงในการหาความเชื่อมั่นของการประเมินความตรงเชิงเนื้อหา." **ปริญญานิพนธ์ศึกษาศาสตร์มหาบัณฑิต** ภาควิชาการวัดและประเมินผลการศึกษา บัณฑิตวิทยาลัย มหาวิทยาลัยขอนแก่น, 2531.

ฝ่ายวิจัยและวางแผน. **รายงานผลสังเขปการสอบ CU-TEP ปีที่ 3.** สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย, 2535.

ไพรัตน์ วงษ์นาม. "สัมประสิทธิ์การอ้างอิงสรุปของแบบทดสอบความเรียง". **ปริญญานิพนธ์ครุศาสตร์ดุษฎีบัณฑิต** บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2533.

เยาวดี วิบูลย์ศรี. **หลักการวัดผลและการสร้างข้อสอบ.** ภาควิชาวิจัยการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2528.

สมบูรณ์ ชิตพงศ์. "สมรรถภาพสมองที่ส่งผลต่อความสามารถในการเขียนเรียงความ." **ปริญญานิพนธ์ศึกษาศาสตร์มหาบัณฑิต** บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, 2511.

สุพัฒน์ สุกมลสันต์. "การอภิวิเคราะห์และการสังเคราะห์งานวิจัยที่เกี่ยวกับการเรียนการสอนภาษาอังกฤษในระหว่างปี พ.ศ. 2515 - 2530". **รายงานการวิจัย.** สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย, 2532.

อัจฉรา วงศ์โสธร และคณะ. "การวิจัยเพื่อหาวิธีการแก้ไขข้อผิดในการเขียนภาษาอังกฤษของนักศึกษาไทยที่เหมาะสมที่สุด". **รายงานการวิจัย** สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย, 2536.

อุทุมพร ทองอุไทย. **แผนวิเคราะห์ข้อมูลพฤติกรรมศาสตร์.** โรงพิมพ์และทำยกเจริญผล, 2523.

Anastasi, A. **Psychological Testing.** 3rd ed. New York : Crowell Collier and Macmillan, 1968.

Bergman, J. **Understanding Educational Measurement and Evaluation.** Boston : Houghton Mifflin Company, 1981.

Block, H. "Estimating the Reliability, validity and Invalidity of Essey Rating." **Journal of Educational Measurement.** 22 (Spring 1985) : 44-52.

Brennan, R. L. **Elements of Generalizability Theory.** Iowa : The American College Testing Program, 1983.

_____. **Some Applications of Generalizability Theory to the Dependability of Domain-Referenced Tests** Iowa: Research and Development Division, The American College Testing Program, 1979.

----------. **Some Statistical Procedures for Domain-referenced Testing: A Handbook for Practitioners** Techical Bulletin No. 38. Iowa: ACT Publication, 1979.

Brennan, R. L. and Kane, M. T. "An Index of Dependability for Mastery Test." **Journal of Educational Measurement.** 14(fall 1977) : 277-289.

----------. "Generalizability Theory : A Review. **"New Directions for Testing and Measurement** 4(1979) : 33-51.

Cardinet, J. and Allal, L. "Estimations of generalizability parameters." **New Directions for Testing and Measurement** 18(1983) : 17-48.

Cardinet, J., Tournuer. Y. and Allal, L. "Extention of generalizability Theory and its Application in educational measurement." **Journal of Educational Measurement** 18(Winter 1981) : 183-204.

————. "The Symmetry of Generalizability Theory : Applications to Educational Measurement." **Journal of Educational Measurement** 13(Summer 1976) : 119-135.

Chase, C. I. "The Impact of Achievement Expectations and Handwriting Quality of Scoring Essay Tests." **Journal of Educational Measurement** 16(Spring 1979) : 39-42.

Coffman, W. E. "Essay Examinations." In **Educational Measurement.** 2nd ed. pp. 271-280. Edited by Thorndike, R. L. Washington D.C. : American Council on Education, 1971.

————."On the Validity of EssayTests of Achievement." **Journal of Educational Measurement** 3(Summer 1966) : 151-156.

Coffman, W. E. and Kurfman, D. A. "A Comparison of Two Methods of Reading Essay Examinations." **American Educational Research Journal.** 5(1968) : 99-107.

Cronbach, L. J. and Others. **The Dependability of Behavioral Measurement : Theory of Generalizability for Score and Profiles.** New York : John Wiley & Sons, 1972.

Cronbach, L. J., Rajaratnam, N. and Gleser, G. C. "Theory of Generalizability : A Liberalization of Reliability Theory." **The British Journal of Statistical Psychology** 16(1963) : 137-163.

Cronbach, L. J. and Gleser, G. "The Signal/Noise Ratio in Comparison of Reliability Coefficient." **Educational and Psychological Measurement** XXIV (1964) : 467-479.

De Gruijter, D. N. M., "The Essay Examination'" In **Psychometrics for Educational Debates.** pp. 145-262. Edited by Van der Kamp, L. J. Th., Langerach, W. F. and De Gruijter, D. N. M. New York : Willey & Sons, 1980.

De Gruijter, D. N. M. and Van der Kamp, L. J. Th. **Statistical Methods in Psychological and Educational Testing.** Abladsserdam : Offsetdrukkerij Kanters B. V., 1984.

Downie, N.M. and Heath, R.W. Basic Statistical Methods. New York: Harper and Row Publishers, 1974.

Ebel, R. L. **Essentials of Educational Measurement.** Englewood Cliffs, N. J. : Prentice-Hall, 1972.

Ebel, R. L. and Frisbie, D. A. **Essential of Educational Measurements.** 4th ed. Englewood Cliffs, N. J. : Prentice-Hall, 1986.

Feldt, L.S., Woodruff, D.J. and Salih, Fathi A. " Statistical Inference for Coefficient Alpha" **Applied Psychological Measurement.** 11(1987): 93-103.

Finlayson, D.S. "The Reliability of the Marking of Essays." **British Journal of Educational Psychology** 21 (1951) : 126-134.

Godshalk,F.I., Swineford, F. and Coffman, W.E. **The Measurement of Writing Ability.** New York : College Entrance Examination Board, 1966.

Hales, G. W. and Tokar, E. "The Effect of the Quality of Preceding Responses on the Grades Assigned to Subsequent Responses to an Essay Question." **Journal of Educational Measurement** 12 (Summer 1975) : 115-117.

Hopkins, K. D. and Stanley, J. C. **Educational and Psychological Measurement and Evaluation.** Englewood Cliffs, N.J. : Prentice- Hall, 1981.

Hughes, D. C., Keeling, E. and Tuck, B.F. "The Influence of Context Position and Scoring Method on Essay Scoring." **Journal of Educational Measurement.** 17(summer 1980) : 131-135.

Ibrahim, A. M. "The Dependability of Need Assessment Data : An Application of Generalizability Theory to the Ratings of Educational Goals." **Dissertation Abstracts International.** 45(1984) : 499-A.

Kubiszyn, T. and Borich, G. **Educational Testing and Measurement Classroom Application and Practice.** Dallas : Scott Foresman and Company, 1981.

Lindvall, C. M. and Nitko, A. J. **Measuring Pupil Achievement and Aptitude.** 2nd ed. New York : Harcourt Brace Jovanovich, 1975.

Macready, G. B. "The Use of Generalizability Theory for Assesing Relations among Items within Domains in Diagnostic Testing." **Applied Psychological Measurement** 7(Spring 1983) : 149-157.

Magnusson, D. **Test Theory.** London : Addison-Wesley Publishing Company, 1967.

Marshall, J.C."Composition Errors and Essay Examination Grades Reexamined." **American Educational Research Journal.** 4(1967) : 375-386.

Marshall, J. C. and Powers, J. C. "Writing Neatness Composition Errors and Essay Grades." **Journal of Educational Measurement** 6(Summer 1969) : 97-101.

Mehrens, W. A. and Ebel, R. L. **Principles of Educational and Psychological Measurement.** Chicago : R and Mcnally & Company, 1966.

O'Brien, R. M. and Jones, B. "The Reliability of School-level Aggregate Variables : An Application of Generalizability Theory." **Journal of Research and Development in Education** 20(1986) : 21-27.

Scannell, D. P. and Marshall, J. C. "The Effect of Selected Composition Error on Grades Assigned to Essay Examinations." **American Educational Research Journal.** 3(1966) : 125-130.

Smith, P. L. "Sampling Errors of Variance Components in Small Sample Multifacet Generalizability Studies." **Journal of Educational Statistics** 3(Winter 1978) : 319-346.

Stanley, J. C. and Hopkins, K. D. **Educational Measurement and Evaluation.** Englewood Cliffs, N. J. : Prentice-Hall, 1972.

Thorndike, R. L. and Hargen, E. P. Measurement and Evaluation in Psychology and Education. 4th ed. New York : John Wiley & Sons, 1977.

Van der Kamp, L. J. Th. "Genneralizability and Educational Measurement." In **Advances in Psychological and Educational Measurement.** pp. 173-184. Edited by De Gruijter, D. N. M., and Van der Kamp, L. J. Th. New York : John Wiley & Sons, 1976.

Wallapa Devahastin and Pateep Methakunavudhi. "Achievement in Written Composition in Thailand (Grade 11)." Bangkok, Chulalongkorn University, 1986.

Webb, N., Herman, J. and Cabello, B. "A Domain Referenced Approach to Diagnostic Testing Using Generalizability Theory." **Journal of Educational Measurement** 24(Summer 1987) : 119-130.

ภาคผนวก ก. ผลการทำงานของโปรแกรมที่เขียนขึ้น

CONTROL CARD INPUT LISTING

```
COLUMN  1111111111222222222233333333344444444445555555556666666666777777777 8
12345678901234567890123456789012345678901234567890123456789012345678901234567890

GSTUDY    P x I design -- fixed model
OPTIONS   RECORDS 2
EFFECT    * P 10
EFFECT    + I 12
FORMAT    (12F2.0)
PROCESS
```

G STUDY **P x I design -- fixed model**

ANOVA TABLE

(** = INFINITE) P I
SAMPLE SIZE 10 12
UNIVERSE SIZE **** ****

EFFECT	DEGREES OF FREEDOM	SUMS OF SQUARES FOR MEAN SCORES	SUMS OF SQUARES FOR SCORE EFFECTS	MEAN SQUARES	(QF = QUASI F RATIO) F STATISTIC	F-TEST DEGREES OF FREEDOM NUMERATOR	DENOMINATOR
P	9	43.08333	7.87500	.87500			
+				6.97183	9	99	
I	11	44.70000	9.49167	.86288			
+				6.87525	11	99	
PI	99	65.00000	12.42500	.12551			
MEAN		35.20833					
TOTAL	119		29.79167				

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE	
UNIVERSE SCORE	.06246	.24992	.03113	
EXPECTED OBSERVED SCORE	.07292	.27003	.03109	
LOWER CASE DELTA	.01046	.10227	.00147	GENERALIZABILITY COEFFICIENT = .85657 (5.97183)
UPPER CASE DELTA	.01660	.12885	.00312	PHI = .78999 (3.76172)
MEAN	.01344	.11592		

DBAR = .541667 PHI (LAMBDA = DBAR) = .74699 (2.95247)

PHI (LAMBDA = .0000) = .95375 (20.62357)	PHI (LAMBDA = .1000) = .93631 (14.70114)
PHI (LAMBDA = .2000) = .90895 (9.98327)	PHI (LAMBDA = .3000) = .86613 (6.46996)
PHI (LAMBDA = .4000) = .80625 (4.16122)	PHI (LAMBDA = .5000) = .75351 (3.05703)
PHI (LAMBDA = .6000) = .75947 (3.15741)	PHI (LAMBDA = .7000) = .81693 (4.46236)
PHI (LAMBDA = .8000) = .87456 (6.97186)	PHI (LAMBDA = .9000) = .91443 (10.68593)
PHI (LAMBDA = 1.0000) = .93978 (15.60456)	

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

CONTROL CARD INPUT LISTING

```
COLUMN  1111111111222222222233333333344444444455555555566666666667777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
COMMENT
COMMENT   First set of D-Study control cards- TRY RUN #1
COMMENT
DSTUDY    #1 -- P x (I:R) design -- I,R - random
DEFFECT   $ P
DEFFECT    R 3
DEFFECT   I:R 1 2 3 4
ENDDSTUDY
```

D STUDY **#1 -- P x (I:R) design -- I,R - random**

D STUDY DESIGN NUMBER 001-004

OBJECT OF MEASUREMENT : P FACETS : R I:R

G STUDY POPULATION SIZE : INFINITE G STUDY UNIVERSE SIZES : INFINITE INFINITE

D STUDY POPULATION SIZE : INFINITE D STUDY UNIVERSE SIZES : INFINITE INFINITE

D STUDY SAMPLE SIZE : 10 D STUDY SAMPLE SIZES : 3 4

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	FOR MEAN SCORES STANDARD ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	FOR MEAN SCORES STANDARD ESTIMATES	STANDARD ERRORS
P	0.47315	1.0000	1	0.47315	0.38558	0.47315	1.0000	1	0.47315	0.38558
R	0.32515	1.0000	3	0.10838	0.14600	0.32515	1.0000	3	0.10838	0.14600
I:R	0.64753	1.0000	12	0.05396	0.03162	0.64753	1.0000	12	0.05396	0.03162
PR	0.55957	1.0000	3	0.18652	0.12554	0.55957	1.0000	3	0.18652	0.12554
PI:R	2.38025	1.0000	12	0.19835	0.03079	2.38025	1.0000	12	0.19835	0.03079

QFM = QUADRATIC FORM

	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE	
UNIVERSE SCORE	.47315	.68786	.38558
EXPECTED OBSERVED SCORE	.85802	.92630	.36586
LOWER CASE DELTA	.38488	.62038	.12171
UPPER CASE DELTA	.54722	.73974	.17935
MEAN	.24815	.49814	

GENERALIZABILITY COEFFICIENT = .55144 (1.22935)

PHI = .46370 (.86464)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

CONTROL CARD INPUT LISTING

```
COLUMN  1111111111222222222233333333344444444455555555556666666666777777777 8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
COMMENT
COMMENT   Second set of D-Study control cards- TRY RUN #2
COMMENT
DSTUDY    #2 -- I:P:R design -- I,R - random
DEFFECT   $ P
DEFFECT    R:P 3
DEFFECT    I:P:R 1 2 3 4
ENDDSTUDY
```

D STUDY **#2 -- I:P:R design -- I,R - random**

D STUDY DESIGN NUMBER 002-004

OBJECT OF MEASUREMENT : P FACETS : R:P I:P:R

G STUDY POPULATION SIZE : INFINITE G STUDY UNIVERSE SIZES : INFINITE INFINITE

D STUDY POPULATION SIZE : INFINITE D STUDY UNIVERSE SIZES : INFINITE INFINITE

D STUDY SAMPLE SIZE : 10 D STUDY SAMPLE SIZES : 3 4

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE SAMPLING CORRECTIONS	D STUDY FREQUENCIES	FOR MEAN SCORES ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE SAMPLING CORRECTIONS	D STUDY FREQUENCIES	FOR MEAN SCORES ESTIMATES	STANDARD ERRORS
EFFECT										
P	0.47315	1.0000	1	0.47315	0.38558	0.47315	1.0000	1	0.47315	0.38558
R:P	0.88472	1.0000	3	0.29491	0.18418	0.88472	1.0000	3	0.29491	0.18418
I:P:R	3.02778	1.0000	12	0.25231	0.04193	3.02778	1.0000	12	0.25231	0.04193

QFM = QUADRATIC FORM

	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE	VARIANCE
UNIVERSE SCORE	.47315	.68786	.38558
EXPECTED OBSERVED SCORE	1.02037	1.01013	.39265
LOWER CASE DELTA	.54722	.73974	.17935
UPPER CASE DELTA	.54722	.73974	.17935
MEAN	.10204	.31943	

GENERALIZABILITY COEFFICIENT = .46370 (.86464)

PHI = .46370 (.86464)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

CONTROL CARD INPUT LISTING

```
COLUMN  1111111111222222222233333333344444444455555555566666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
COMMENT
COMMENT   Third set of D-Study control cards - TRY RUN #3
COMMENT
DSTUDY    #3 -- P x (I:R) -- I - random, R - fixed
DEFFECT   $ P
DEFFECT    R 3 / 3
DEFFECT    I:R 1 2 3 4
ENDDSTUDY
```

D STUDY **#3 -- P x (I:R) -- I - random, R - fixed**

D STUDY DESIGN NUMBER 003-004

OBJECT OF MEASUREMENT : P FACETS : R I:R
G STUDY POPULATION SIZE : INFINITE G STUDY UNIVERSE SIZES : INFINITE INFINITE
D STUDY POPULATION SIZE : INFINITE D STUDY UNIVERSE SIZES : 3 INFINITE
D STUDY SAMPLE SIZE : 10 D STUDY SAMPLE SIZES : 3 4

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
	VARIANCE COMPONENTS					VARIANCE COMPONENTS				
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR- RECTIONS	D STUDY SAMPLING FRE- QUENCIES	FOR MEAN SCORES STANDARD ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR- RECTIONS	D STUDY SAMPLING FRE- QUENCIES	FOR MEAN SCORES STANDARD ESTIMATES	STANDARD ERRORS
P	0.47315	1.0000	1	0.47315	0.38558	0.65967	1.0000	1	0.65967	0.36716
R	0.32515	1.0000	3	0.10838	0.14600	0.32515QFM	0.0000	3	-------	-------
I:R	0.64753	1.0000	12	0.05396	0.03162	0.64753	1.0000	12	0.05396	0.03162
PR	0.55957	1.0000	3	0.18652	0.12554	0.55957	0.0000	3	-------	-------
PI:R	2.38025	1.0000	12	0.19835	0.03079	2.38025	1.0000	12	0.19835	0.03079

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	.65967	.81220	.36716
EXPECTED OBSERVED SCORE	.85802	.92630	.36586
LOWER CASE DELTA	.19835	.44537	.03079
UPPER CASE DELTA	.25231	.50231	.04193
MEAN	.13976	.37385	

GENERALIZABILITY COEFFICIENT = .76882 (3.32573)
PHI = .72333 (2.61448)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

#1 -- P x (I:R) design -- I,R - random

SUMMARY OF D STUDY RESULTS FOR SET OF CONTROL CARDS NO. 001

D STUDY DESIGN NO	INDEX= UNIV.=	$P INF.	R INF.	I INF.	VARIANCES: UNIVERSE SCORE	OBSERVED SCORE	EXPECTED CASE DELTA	LOWER CASE DELTA	UPPER MEAN	GEN. COEF.	PHI
001-001	10	3	1		.47315	1.45309	.97994	1.30417	.46954	.32562	.26622
001-002	10	3	2		.47315	1.05638	.58323	.79954	.32194	.44790	.37177
001-003	10	3	3		.47315	.92414	.45099	.63133	.27275	.51199	.42839
001-004	10	3	4		.47315	.85802	.38488	.54722	.24815	.55144	.46370

#2 -- I:P:R design -- I,R - random

SUMMARY OF D STUDY RESULTS FOR SET OF CONTROL CARDS NO. 002

D STUDY DESIGN NO	INDEX= UNIV.=	$P INF.	R INF.	I INF.	VARIANCES: UNIVERSE SCORE	OBSERVED SCORE	EXPECTED CASE DELTA	LOWER CASE DELTA	UPPER MEAN	GEN. COEF.	PHI
002-001	10	3	1		.47315	1.77731	1.30417	1.30417	.17773	.26622	.26622
002-002	10	3	2		.47315	1.27269	.79954	.79954	.12727	.37177	.37177
002-003	10	3	3		.47315	1.10448	.63133	.63133	.11045	.42839	.42839
002-004	10	3	4		.47315	1.02037	.54722	.54722	.10204	.46370	.46370

#3 -- P x (I:R) -- I - random, R - fixed

SUMMARY OF D STUDY RESULTS FOR SET OF CONTROL CARDS NO. 003

D STUDY DESIGN NO	INDEX= UNIV.=	$P INF.	R 3	I INF.	VARIANCES: UNIVERSE SCORE	OBSERVED SCORE	EXPECTED CASE DELTA	LOWER CASE DELTA	UPPER MEAN	GEN. COEF.	PHI
003-001	10	3	1		.65967	1.45309	.79342	1.00926	.36115	.45398	.39527
003-002	10	3	2		.65967	1.05638	.39671	.50463	.21356	.62446	.56658
003-003	10	3	3		.65967	.92414	.26447	.33642	.16436	.71382	.66226
003-004	10	3	4		.65967	.85802	.19835	.25231	.13976	.76882	.72333

ภาคผนวก ข. ผลการวิเคราะห์ข้อมูล

ก. การวิเคราะห์ข้อมูล (สำหรับวัตถุประสงค์ข้อที่ 1)

1. เมื่อกำหนดให้ผู้ตรวจข้อทดสอบทุกประเภท ตรวจข้อทดสอบทุกข้อของผู้สอบทุกคน หรือ P x I x R

1.1.1 ในกรณีที่เป็นแบบจำลองผสม (mixed model) และผู้ตรวจมีประสบการณ์โดยตรง

D STUDY # 1.1 — P x I x R DESIGN -- MIXED MODEL

D STUDY DESIGN NUMBER 001-001

OBJECT OF MEASUREMENT :	P	FACETS :	I R
G STUDY POPULATION SIZE :	INFINITE	G STUDY UNIVERSE SIZES :	INFINITE 2
D STUDY POPULATION SIZE :	INFINITE	D STUDY UNIVERSE SIZES :	INFINITE 2
D STUDY SAMPLE SIZE :	200	D STUDY SAMPLE SIZES :	3 2

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS
P	29.27036	1.0000	1	29.27036	3.49572	29.27036	1.0000	1	29.27036	3.49572
I	0.86624	1.0000	3	0.28875	0.22381	0.86624	1.0000	3	0.28875	0.22381
R	0.00000QFM	0.0000	2	-------	-------	0.00000QFM	0.0000	2	-------	-------
PI	16.65709	1.0000	3	5.55236	0.39261	16.65709	1.0000	3	5.55236	0.39261
PR	0.00000	0.0000	2	-------	-------	0.00000	0.0000	2	-------	-------
IR	3.44779	0.0000	6	-------	-------	3.44779	0.0000	6	-------	-------
PIR	56.81221	0.0000	6	-------	-------	56.81221	0.0000	6	-------	-------

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	29.27036	5.41021	3.49572
EXPECTED OBSERVED SCORE	34.82272	5.90108	3.47360
LOWER CASE DELTA	5.55236	2.35635	.39261
UPPER CASE DELTA	5.84111	2.41684	.45022
MEAN	.46286	.68034	

GENERALIZABILITY COEFFICIENT = .84055 (5.27169)
PHI = .83364 (5.01109)

DBAR = 13.050000

PHI (LAMBDA = .0000) =	.97150 (34.08769)	PHI (LAMBDA = DBAR) =	.83142 (4.93185)
PHI (LAMBDA = .2000) =	.97076 (33.20088)	PHI (LAMBDA = .1000) =	.97113 (33.64257)
PHI (LAMBDA = .4000) =	.96999 (32.32775)	PHI (LAMBDA = .3000) =	.97038 (32.76260)
PHI (LAMBDA = .6000) =	.96920 (31.46833)	PHI (LAMBDA = .5000) =	.96960 (31.89633)
PHI (LAMBDA = .8000) =	.96838 (30.62260)	PHI (LAMBDA = .7000) =	.96879 (31.04375)
PHI (LAMBDA = 1.0000) =	.96752 (29.79056)	PHI (LAMBDA = .9000) =	.96795 (30.20487)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

1.1.2 ในกรณีที่เป็นแบบจำลองผสม (mixed model) และผู้ตรวจมีประสบการณ์ทางอ้อม

```
D STUDY                          # 1.1 -- P x I x R DESIGN -- MIXED MODEL

                                 D STUDY DESIGN NUMBER 001-001

  OBJECT OF MEASUREMENT :     P                  FACETS :     I         R
G STUDY POPULATION SIZE : INFINITE    G STUDY UNIVERSE SIZES : INFINITE      2
D STUDY POPULATION SIZE : INFINITE    D STUDY UNIVERSE SIZES : INFINITE      2
    D STUDY SAMPLE SIZE :   200         D STUDY SAMPLE SIZES :     3        2
```

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS
P	18.77335	1.0000	1	18.77335	2.62723	18.77335	1.0000	1	18.77335	2.62723
I	7.45434	1.0000	3	2.48478	1.78204	7.45434	1.0000	3	2.48478	1.78204
R	0.22078QFM	0.0000	2	-------	-------	0.22078QFM	0.0000	2	-------	-------
PI	21.24524	1.0000	3	7.08175	0.50076	21.24524	1.0000	3	7.08175	0.50076
PR	0.00000	0.0000	2	-------	-------	0.00000	0.0000	2	-------	-------
IR	0.65820	0.0000	6	-------	-------	0.65820	0.0000	6	-------	-------
PIR	61.41097	0.0000	6	-------	-------	61.41097	0.0000	6	-------	-------

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	18.77335	4.33282	2.62723
EXPECTED OBSERVED SCORE	25.85510	5.08479	2.57907
LOWER CASE DELTA	7.08175	2.66116	.50076
UPPER CASE DELTA	9.56653	3.09298	1.85039
MEAN	2.61405	1.61680	

GENERALIZABILITY COEFFICIENT = .72610 (2.65095)
PHI = .66244 (1.96240)

DBAR = 15.039167

PHI (LAMBDA = .0000) = .96202 (25.33164)
PHI (LAMBDA = .2000) = .96110 (24.70700)
PHI (LAMBDA = .4000) = .96014 (24.09072)
PHI (LAMBDA = .6000) = .95915 (23.48280)
PHI (LAMBDA = .8000) = .95813 (22.88324)
PHI (LAMBDA = 1.0000) = .95707 (22.29205)

PHI (LAMBDA = DBAR) = .62814 (1.68915)
PHI (LAMBDA = .1000) = .96157 (25.01827)
PHI (LAMBDA = .3000) = .96063 (24.39781)
PHI (LAMBDA = .5000) = .95965 (23.78571)
PHI (LAMBDA = .7000) = .95865 (23.18197)
PHI (LAMBDA = .9000) = .95760 (22.58660)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

1.2.1 ในกรณีที่เป็นแบบจำลองสุ่ม (random model) และผู้ตรวจมีประสบการณ์โดยตรง

D-STUDY # 1.2 -- P x I x R DESIGN -- RANDOM MODEL

D STUDY DESIGN NUMBER 004-001

OBJECT OF MEASUREMENT :	P	FACETS :	I	R
G STUDY POPULATION SIZE :	INFINITE	G STUDY UNIVERSE SIZES :	INFINITE	2
D STUDY POPULATION SIZE :	INFINITE	D STUDY UNIVERSE SIZES :	INFINITE	INFINITE
D STUDY SAMPLE SIZE :	200	D STUDY SAMPLE SIZES :	3	2

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS
P	29.27036	1.0000	1	29.27036	3.49572	29.27036	1.0000	1	29.27036	3.49572
I	0.86624	1.0000	3	0.28875	0.22381	0.86624	1.0000	3	0.28875	0.22381
R	0.00000QFM	0.0000	2	-------	-------	0.00000	1.0000	2	0.00000	0.44020
PI	16.65709	1.0000	3	5.55236	0.39261	16.65709	1.0000	3	5.55236	0.39261
PR	0.00000	0.0000	2	-------	-------	0.00000	1.0000	2	0.00000	0.67975
IR	3.44779	0.0000	6	-------	-------	3.44779	1.0000	6	0.57463	0.43982
PIR	56.81221	0.0000	6	-------	-------	56.81221	1.0000	6	9.46870	0.66954

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	29.27036	5.41021	3.49572
EXPECTED OBSERVED SCORE	44.29143	6.65518	3.47558
LOWER CASE DELTA	15.02107	3.87570	.40978
UPPER CASE DELTA	15.88444	3.98553	.46548
MEAN	1.08484	1.04155	

GENERALIZABILITY COEFFICIENT = .66086 (1.94862)
PHI = .64822 (1.84271)

DBAR = 13.050000

PHI (LAMBDA =	.0000) =	.92590	(12.49575)
PHI (LAMBDA =	.2000) =	.92407	(12.16964)
PHI (LAMBDA =	.4000) =	.92217	(11.84857)
PHI (LAMBDA =	.6000) =	.92021	(11.53254)
PHI (LAMBDA =	.8000) =	.91818	(11.22155)
PHI (LAMBDA =	1.0000) =	.91608	(10.91559)

PHI (LAMBDA =	DBAR) =	.63956	(1.77441)
PHI (LAMBDA =	.1000) =	.92499	(12.33207)
PHI (LAMBDA =	.3000) =	.92313	(12.00848)
PHI (LAMBDA =	.5000) =	.92120	(11.68993)
PHI (LAMBDA =	.7000) =	.91920	(11.37641)
PHI (LAMBDA =	.9000) =	.91714	(11.06794)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

1.2.2 ในกรณีที่เป็นแบบจำลองสุ่ม (random model) และผู้ตรวจมีประสบการณ์ทางอ้อม

D STUDY # 1.2 -- P x I x R DESIGN -- RANDOM MODEL

D STUDY DESIGN NUMBER 004-001

OBJECT OF MEASUREMENT : P FACETS : I R
G STUDY POPULATION SIZE : INFINITE G STUDY UNIVERSE SIZES : INFINITE 2
D STUDY POPULATION SIZE : INFINITE D STUDY UNIVERSE SIZES : INFINITE INFINITE
D STUDY SAMPLE SIZE : 200 D STUDY SAMPLE SIZES : 3 2

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	ESTIMATES	STANDARD ERRORS
P	18.77335	1.0000	1	18.77335	2.62723	18.77335	1.0000	1	18.77335	2.62723
I	7.45434	1.0000	3	2.48478	1.78204	7.45434	1.0000	3	2.48478	1.78204
R	0.22078QFM	0.0000	2	-------	-------	0.22078	1.0000	2	0.11039	0.24902
PI	21.24524	1.0000	3	7.08175	0.50076	21.24524	1.0000	3	7.08175	0.50076
PR	0.00000	0.0000	2	-------	-------	0.00000	1.0000	2	0.00000	0.76934
IR	0.65820	0.0000	6	-------	-------	0.65820	1.0000	6	0.10970	0.11381
PIR	61.41097	0.0000	6	-------	-------	61.41097	1.0000	6	10.23516	0.72374

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	18.77335	4.33282	2.62723
EXPECTED OBSERVED SCORE	36.09026	6.00752	2.59224
LOWER CASE DELTA	17.31691	4.16136	.56466
UPPER CASE DELTA	20.02178	4.47457	1.88159
MEAN	2.88532	1.69862	

GENERALIZABILITY COEFFICIENT = .52018 (1.08411)
PHI = .48391 (.93765)

DBAR = 15.039167

PHI (LAMBDA = .0000) = .92361 (12.09006)
PHI (LAMBDA = .2000) = .92182 (11.79160)
PHI (LAMBDA = .4000) = .91998 (11.49714)
PHI (LAMBDA = .6000) = .91808 (11.20667)
PHI (LAMBDA = .8000) = .91611 (10.92020)
PHI (LAMBDA = 1.0000) = .91407 (10.63773)

PHI (LAMBDA = DBAR) = .44244 (.79354)
PHI (LAMBDA = .1000) = .92272 (11.94033)
PHI (LAMBDA = .3000) = .92091 (11.64387)
PHI (LAMBDA = .5000) = .91904 (11.35141)
PHI (LAMBDA = .7000) = .91710 (11.06294)
PHI (LAMBDA = .9000) = .91510 (10.77847)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

2. เมื่อกำหนดให้ผู้ตรวจบางประเภทตรวจข้อทดสอบบางข้อของผู้สอบทุกคน หรือ P x (R:I)

2.1.1 ในกรณีที่เป็นแบบจำลองผสม (mixed model) และผู้ตรวจมีประสบการณ์โดยตรง

```
D STUDY                          # 3.1 -- P x (R:I) -- MIXED MODEL

                                 D STUDY DESIGN NUMBER 003-001

  OBJECT OF MEASUREMENT :     P                  FACETS :     I        R:I
G STUDY POPULATION SIZE : INFINITE    G STUDY UNIVERSE SIZES : INFINITE     2
D STUDY POPULATION SIZE : INFINITE    D STUDY UNIVERSE SIZES : INFINITE     2
  D STUDY SAMPLE SIZE :   200           D STUDY SAMPLE SIZES :   3        2
```

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS
P	29.27036	1.0000	1	29.27036	3.49572	29.27036	1.0000	1	29.27036	3.49572
I	0.86624	1.0000	3	0.28875	0.22381	0.86624	1.0000	3	0.28875	0.22381
R:I	3.44779	0.0000	6	-------	-------	3.44779	0.0000	6	-------	-------
PI	16.65709	1.0000	3	5.55236	0.39261	16.65709	1.0000	3	5.55236	0.39261
PR:I	56.81221	0.0000	6	-------	-------	56.81221	0.0000	6	-------	-------

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	29.27036	5.41021	3.49572
EXPECTED OBSERVED SCORE	34.82272	5.90108	3.47360
LOWER CASE DELTA	5.55236	2.35635	.39261
UPPER CASE DELTA	5.84111	2.41684	.45022
MEAN	.46286	.68034	

GENERALIZABILITY COEFFICIENT = .84055 (5.27169)
PHI = .83364 (5.01109)

```
DBAR = 13.050000
PHI (LAMBDA =  .0000) =  .97150 ( 34.08769)
PHI (LAMBDA =  .2000) =  .97076 ( 33.20088)
PHI (LAMBDA =  .4000) =  .96999 ( 32.32775)
PHI (LAMBDA =  .6000) =  .96920 ( 31.46833)
PHI (LAMBDA =  .8000) =  .96838 ( 30.62260)
PHI (LAMBDA = 1.0000) =  .96752 ( 29.79056)

PHI (LAMBDA =  DBAR) =  .83142 (  4.93185)
PHI (LAMBDA =  .1000) =  .97113 ( 33.64257)
PHI (LAMBDA =  .3000) =  .97038 ( 32.76260)
PHI (LAMBDA =  .5000) =  .96960 ( 31.89633)
PHI (LAMBDA =  .7000) =  .96879 ( 31.04375)
PHI (LAMBDA =  .9000) =  .96795 ( 30.20487)
```

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

2.1.2 ในกรณีที่เป็นแบบจำลองผสม (mixed model) และผู้ตรวจมีประสบการณ์ทางอ้อม

D STUDY # 3.1 -- P x (R:I) -- MIXED MODEL

D STUDY DESIGN NUMBER 003-001

OBJECT OF MEASUREMENT :	P	FACETS :	I	R:I	
G STUDY POPULATION SIZE :	INFINITE	G STUDY UNIVERSE SIZES :	INFINITE	2	
D STUDY POPULATION SIZE :	INFINITE	D STUDY UNIVERSE SIZES :	INFINITE	2	
D STUDY SAMPLE SIZE :	200	D STUDY SAMPLE SIZES :	3	2	

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS
P	18.77335	1.0000	1	18.77335	2.62723	18.77335	1.0000	1	18.77335	2.62723
I	7.45434	1.0000	3	2.48478	1.78204	7.45434	1.0000	3	2.48478	1.78204
R:I	0.87898	0.0000	6	-------	-------	0.87898	0.0000	6	-------	-------
PI	21.24524	1.0000	3	7.08175	0.50076	21.24524	1.0000	3	7.08175	0.50076
PR:I	61.41097	0.0000	6	-------	-------	61.41097	0.0000	6	-------	-------

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	18.77335	4.33282	2.62723
EXPECTED OBSERVED SCORE	25.85510	5.08479	2.57907
LOWER CASE DELTA	7.08175	2.66116	.50076
UPPER CASE DELTA	9.56653	3.09298	1.85039
MEAN	2.61405	1.61680	

GENERALIZABILITY COEFFICIENT = .72610 (2.65095)
PHI = .66244 (1.96240)

DBAR = 15.039167

PHI (LAMBDA = .0000) = .96202 (25.33164)
PHI (LAMBDA = .2000) = .96110 (24.70700)
PHI (LAMBDA = .4000) = .96014 (24.09072)
PHI (LAMBDA = .6000) = .95915 (23.48280)
PHI (LAMBDA = .8000) = .95813 (22.88324)
PHI (LAMBDA = 1.0000) = .95707 (22.29205)

PHI (LAMBDA = DBAR) = .62814 (1.68915)
PHI (LAMBDA = .1000) = .96157 (25.01827)
PHI (LAMBDA = .3000) = .96063 (24.39781)
PHI (LAMBDA = .5000) = .95965 (23.78571)
PHI (LAMBDA = .7000) = .95865 (23.18197)
PHI (LAMBDA = .9000) = .95760 (22.58660)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

2.2.1 ในกรณีที่เป็นแบบจำลองสุ่ม (random model) และผู้ตรวจมีประสบการณ์โดยตรง

D STUDY # 3.2 -- P x (R:I) -- RANDOM MODEL

D STUDY DESIGN NUMBER 006-001

OBJECT OF MEASUREMENT :	P	FACETS :	I R:I
G STUDY POPULATION SIZE :	INFINITE	G STUDY UNIVERSE SIZES :	INFINITE 2
D STUDY POPULATION SIZE :	INFINITE	D STUDY UNIVERSE SIZES :	INFINITE INFINITE
D STUDY SAMPLE SIZE :	200	D STUDY SAMPLE SIZES :	3 2

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	VARIANCE COMPONENTS FOR MEAN SCORES ESTIMATES	VARIANCE COMPONENTS FOR MEAN SCORES STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR-RECTIONS	D STUDY SAMPLING FRE-QUENCIES	VARIANCE COMPONENTS FOR MEAN SCORES ESTIMATES	VARIANCE COMPONENTS FOR MEAN SCORES STANDARD ERRORS
P	29.27036	1.0000	1	29.27036	3.49572	29.27036	1.0000	1	29.27036	3.49572
I	0.86624	1.0000	3	0.28875	0.22381	0.86624	1.0000	3	0.28875	0.22381
R:I	3.44779	0.0000	6	-------	-------	3.44779	1.0000	6	0.57463	0.29327
PI	16.65709	1.0000	3	5.55236	0.39261	16.65709	1.0000	3	5.55236	0.39261
PR:I	56.81221	0.0000	6	-------	-------	56.81221	1.0000	6	9.46870	0.44807

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	29.27036	5.41021	3.49572
EXPECTED OBSERVED SCORE	44.29143	6.65518	3.50238
LOWER CASE DELTA	15.02107	3.87570	.59574
UPPER CASE DELTA	15.88444	3.98553	.69819
MEAN	1.08484	1.04155	

GENERALIZABILITY COEFFICIENT = .66086 (1.94862)
PHI = .64822 (1.84271)

DBAR = 13.050000

PHI (LAMBDA =	.0000) =	.92590	(12.49575)
PHI (LAMBDA =	.2000) =	.92407	(12.16964)
PHI (LAMBDA =	.4000) =	.92217	(11.84857)
PHI (LAMBDA =	.6000) =	.92021	(11.53254)
PHI (LAMBDA =	.8000) =	.91818	(11.22155)
PHI (LAMBDA =	1.0000) =	.91608	(10.91559)

PHI (LAMBDA =	DBAR) =	.63956	(1.77441)
PHI (LAMBDA =	.1000) =	.92499	(12.33207)
PHI (LAMBDA =	.3000) =	.92313	(12.00848)
PHI (LAMBDA =	.5000) =	.92120	(11.68993)
PHI (LAMBDA =	.7000) =	.91920	(11.37641)
PHI (LAMBDA =	.9000) =	.91714	(11.06794)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

2.2.2 ในกรณีที่เป็นแบบจำลองสุ่ม (random model) และผู้ตรวจมีประสบการณ์ทางอ้อม

D STUDY # 3.2 -- P x (R:I) -- RANDOM MODEL

D STUDY DESIGN NUMBER 006-001

OBJECT OF MEASUREMENT : P
FACETS : I R:I
G STUDY POPULATION SIZE : INFINITE
G STUDY UNIVERSE SIZES : INFINITE 2
D STUDY POPULATION SIZE : INFINITE
D STUDY UNIVERSE SIZES : INFINITE INFINITE
D STUDY SAMPLE SIZE : 200
D STUDY SAMPLE SIZES : 3 2

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS
P	18.77335	1.0000	1	18.77335	2.62723	18.77335	1.0000	1	18.77335	2.62723
I	7.45434	1.0000	3	2.48478	1.78204	7.45434	1.0000	3	2.48478	1.78204
R:I	0.87898	0.0000	6	-------	-------	0.87898	1.0000	6	0.14650	0.10587
PI	21.24524	1.0000	3	7.08175	0.50076	21.24524	1.0000	3	7.08175	0.50076
PR:I	61.41097	0.0000	6	-------	-------	61.41097	1.0000	6	10.23516	0.49027

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	18.77335	4.33282	2.62723
EXPECTED OBSERVED SCORE	36.09026	6.00752	2.62526
LOWER CASE DELTA	17.31691	4.16136	.70080
UPPER CASE DELTA	19.94819	4.46634	1.91653
MEAN	2.81173	1.67682	

GENERALIZABILITY COEFFICIENT = .52018 (1.08411)
PHI = .48483 (.94111)

DBAR = 15.039167

PHI (LAMBDA = .0000) = .92389 (12.13835)
PHI (LAMBDA = .2000) = .92211 (11.83880)
PHI (LAMBDA = .4000) = .92028 (11.54325)
PHI (LAMBDA = .6000) = .91838 (11.25171)
PHI (LAMBDA = .8000) = .91642 (10.96418)
PHI (LAMBDA = 1.0000) = .91439 (10.68066)

PHI (LAMBDA = DBAR) = .44449 (.80015)
PHI (LAMBDA = .1000) = .92301 (11.98807)
PHI (LAMBDA = .3000) = .92120 (11.69052)
PHI (LAMBDA = .5000) = .91934 (11.39698)
PHI (LAMBDA = .7000) = .91741 (11.10744)
PHI (LAMBDA = .9000) = .91541 (10.82192)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

3. เมื่อกำหนดให้ผู้ตรวจบางประเภทตรวจข้อทดสอบทุกข้อของผู้สอบบางคน หรือ I x (P:R)

3.1.1 ในกรณีที่เป็นแบบจำลองผสม (mixed model) และผู้ตรวจมีประสบการณ์โดยตรง

```
D STUDY                          # 2.1 -- I x (P:R) -- MIXED MODEL

                                 D STUDY DESIGN NUMBER 002-002

  OBJECT OF MEASUREMENT :     I                  FACETS :    P:R        R
G STUDY POPULATION SIZE : INFINITE    G STUDY UNIVERSE SIZES : INFINITE     2
D STUDY POPULATION SIZE : INFINITE    D STUDY UNIVERSE SIZES : INFINITE     2
  D STUDY SAMPLE SIZE :     3         D STUDY SAMPLE SIZES :   200        2
```

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS
P:R	0.00000	1.0000	400	0.00000	0.00938	0.00000	1.0000	400	0.00000	0.00220
I	29.27036	1.0000	1	29.27036	0.67144	29.27036	1.0000	1	29.27036	3.75077
R	0.86624QFM	0.0000	2	-------	-------	0.86624QFM	0.0000	2	-------	-------
PI:R	3.44779	1.0000	400	0.00862	0.01047	3.44779	1.0000	400	0.00862	0.00660
IR	73.46930	0.0000	2	-------	-------	73.46930	0.0000	2	-------	-------

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	29.27036	5.41021	3.75077
EXPECTED OBSERVED SCORE	29.27898	5.41101	.00440
LOWER CASE DELTA	.00862	.09284	.00660
UPPER CASE DELTA	.00862	.09284	3.75079
MEAN	9.75966	3.12405	

GENERALIZABILITY COEFFICIENT = .99971 (********)
PHI = .99971 (********)

```
      DBAR = 13.050000
PHI (LAMBDA =  .0000) =  .99995 (*********)
PHI (LAMBDA =  .2000) =  .99995 (*********)
PHI (LAMBDA =  .4000) =  .99995 (*********)
PHI (LAMBDA =  .6000) =  .99995 (*********)
PHI (LAMBDA =  .8000) =  .99995 (*********)
PHI (LAMBDA = 1.0000) =  .99995 (*********)

PHI (LAMBDA =  DBAR) =  .99956 (*********)
PHI (LAMBDA =  .1000) =  .99995 (*********)
PHI (LAMBDA =  .3000) =  .99995 (*********)
PHI (LAMBDA =  .5000) =  .99995 (*********)
PHI (LAMBDA =  .7000) =  .99995 (*********)
PHI (LAMBDA =  .9000) =  .99995 (*********)
```

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

3.1.2 ในกรณีที่เป็นแบบจำลองผสม (mixed model) และผู้ตรวจมีประสบการณ์ทางอ้อม

D STUDY # 2.1 -- I x (P:R) -- MIXED MODEL

D STUDY DESIGN NUMBER 002-002

OBJECT OF MEASUREMENT :	I		FACETS :	P:R	R
G STUDY POPULATION SIZE :	INFINITE	G STUDY UNIVERSE SIZES :		INFINITE	2
D STUDY POPULATION SIZE :	INFINITE	D STUDY UNIVERSE SIZES :		INFINITE	2
D STUDY SAMPLE SIZE :	3	D STUDY SAMPLE SIZES :		200	2

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS
P:R	0.22078	1.0000	400	0.00055	0.00761	0.22078	1.0000	400	0.00055	0.00125
I	18.77335	1.0000	1	18.77335	5.34613	18.77335	1.0000	1	18.77335	3.04465
R	7.45434QFM	0.0000	2	-------	-------	7.45434QFM	0.0000	2	-------	-------
PI:R	0.65820	1.0000	400	0.00165	0.01149	0.65820	1.0000	400	0.00165	0.90171
IR	82.65621	0.0000	2	-------	-------	82.65621	0.0000	2	-------	-------

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	18.77335	4.33282	3.04465
EXPECTED OBSERVED SCORE	18.77500	4.33301	.00159
LOWER CASE DELTA	.00165	.04056	.00171
UPPER CASE DELTA	.00220	.04688	3.04467
MEAN	6.25889	2.50178	

GENERALIZABILITY COEFFICIENT = .99991 (********)
PHI = .99988 (********)

DBAR = 15.039167

PHI (LAMBDA = .0000) = .99999 (*********)
PHI (LAMBDA = .2000) = .99999 (*********)
PHI (LAMBDA = .4000) = .99999 (*********)
PHI (LAMBDA = .6000) = .99999 (*********)
PHI (LAMBDA = .8000) = .99999 (*********)
PHI (LAMBDA = 1.0000) = .99999 (*********)

PHI (LAMBDA = DBAR) = .99982 (*********)
PHI (LAMBDA = .1000) = .99999 (*********)
PHI (LAMBDA = .3000) = .99999 (*********)
PHI (LAMBDA = .5000) = .99999 (*********)
PHI (LAMBDA = .7000) = .99999 (*********)
PHI (LAMBDA = .9000) = .99999 (*********)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

3.2:1 ในกรณีที่เป็นแบบจำลองสุ่ม (random model) และผู้ตรวจมีประสบการณ์โดยตรง

D STUDY # 2.2 -- I x (P:R) -- RANDOM MODEL

D STUDY DESIGN NUMBER 005-002

OBJECT OF MEASUREMENT :	I	FACETS :	P:R R
G STUDY POPULATION SIZE :	INFINITE	G STUDY UNIVERSE SIZES :	INFINITE 2
D STUDY POPULATION SIZE :	INFINITE	D STUDY UNIVERSE SIZES :	INFINITE INFINITE
D STUDY SAMPLE SIZE :	3	D STUDY SAMPLE SIZES :	200 2

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS
P:R	0.00000	1.0000	400	0.00000	0.00938	0.00000	1.0000	400	0.00000	0.00220
I	29.27036	1.0000	1	29.27036	0.67144	29.27036	1.0000	1	29.27036	3.75077
R	0.86624QFM	0.0000	2	-------	-------	0.86624	1.0000	2	0.43312	0.33572
PI:R	3.44779	1.0000	400	0.00862	0.01047	3.44779	1.0000	400	0.00862	0.00660
IR	73.46930	0.0000	2	-------	-------	73.46930	1.0000	2	36.73465	2.09317

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	29.27036	5.41021	3.75077
EXPECTED OBSERVED SCORE	66.01363	8.12488	2.09314
LOWER CASE DELTA	36.74327	6.06162	2.09313
UPPER CASE DELTA	37.17639	6.09724	3.56642
MEAN	22.43766	4.73684	

GENERALIZABILITY COEFFICIENT = .44340 (.79662)
PHI = .44051 (.78734)

DBAR = 13.050000

PHI (LAMBDA = .0000) =	.82653 (4.76472)	PHI (LAMBDA = DBAR) =	.15526 (.18379)
PHI (LAMBDA = .2000) =	.82223 (4.62539)	PHI (LAMBDA = .1000) =	.82440 (4.69479)
PHI (LAMBDA = .4000) =	.81779 (4.48820)	PHI (LAMBDA = .3000) =	.82003 (4.55653)
PHI (LAMBDA = .6000) =	.81319 (4.35317)	PHI (LAMBDA = .5000) =	.81551 (4.42042)
PHI (LAMBDA = .8000) =	.80844 (4.22029)	PHI (LAMBDA = .7000) =	.81084 (4.28646)
PHI (LAMBDA = 1.0000) =	.80352 (4.08956)	PHI (LAMBDA = .9000) =	.80600 (4.15466)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

3.2.2 ในกรณีที่เป็นแบบจำลองสุ่ม (random model) และผู้ตรวจมีประสบการณ์ทางอ้อม

```
D STUDY                    # 2.2 -- I x (P:R) -- RANDOM MODEL

                           D STUDY DESIGN NUMBER 005-002

   OBJECT OF MEASUREMENT :      I                    FACETS :    P:R        R
G STUDY POPULATION SIZE : INFINITE    G STUDY UNIVERSE SIZES : INFINITE      2
D STUDY POPULATION SIZE : INFINITE    D STUDY UNIVERSE SIZES : INFINITE  INFINITE
  D STUDY SAMPLE SIZE :      3        D STUDY SAMPLE SIZES :    200        2
```

	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
				VARIANCE COMPONENTS FOR MEAN SCORES					VARIANCE COMPONENTS FOR MEAN SCORES	
EFFECT	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE CORRECTIONS	D STUDY SAMPLING FREQUENCIES	ESTIMATES	STANDARD ERRORS
P:R	0.22078	1.0000	400	0.00055	0.00761	0.22078	1.0000	400	0.00055	0.00125
I	18.77335	1.0000	1	18.77335	5.34613	18.77335	1.0000	1	18.77335	3.04465
R	7.45434QFM	0.0000	2	-------	-------	7.45434	1.0000	2	3.72717	2.67307
PI:R	0.65820	1.0000	400	0.00165	0.01149	0.65820	1.0000	400	0.00165	0.00171
IR	82.65621	0.0000	2	-------	-------	82.65621	1.0000	2	41.32811	2.29746

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE
UNIVERSE SCORE	18.77335	4.33282	3.04465
EXPECTED OBSERVED SCORE	60.10311	7.75262	2.29743
LOWER CASE DELTA	41.32975	6.42882	2.29741
UPPER CASE DELTA	45.05747	6.71249	3.82802
MEAN	23.76209	4.87464	

GENERALIZABILITY COEFFICIENT = .31235 (.45423)
PHI = .29411 (.41665)

DBAR = 15.039167

PHI (LAMBDA = .0000) =	.83077 (4.90901)	PHI (LAMBDA = DBAR) = -.12450 (-.11072)
PHI (LAMBDA = .2000) =	.82688 (4.77639)	PHI (LAMBDA = .1000) = .82884 (4.84248)
PHI (LAMBDA = .4000) =	.82287 (4.64554)	PHI (LAMBDA = .3000) = .82489 (4.71075)
PHI (LAMBDA = .6000) =	.81872 (4.51647)	PHI (LAMBDA = .5000) = .82081 (4.58079)
PHI (LAMBDA = .8000) =	.81444 (4.38917)	PHI (LAMBDA = .7000) = .81660 (4.45260)
PHI (LAMBDA = 1.0000) =	.81002 (4.26365)	PHI (LAMBDA = .9000) = .81225 (4.32619)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

ภาคผนวก ค. โปรแกรมคอมพิวเตอร์ที่เขียนขึ้นใช้งาน

หมายเหตุ

ในงานวิจัยเล่มนี้ผู้ขียนไม่ได้รวม Source Program มาด้วยเพราะมีขนาดยาวมาก คือยาว 165 หน้า หรือประมาณ 6,900 บรรทัด และจุประมาณ 1.3 Mb ผู้ที่สนใจจะศึกษาโปรแกรม โปรดศึกษาได้ในฉบับที่อยู่ในห้องสมุดสถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย หรือที่ผู้วิจัย