การประเมินความเชื่อมั่นระหว่างผู้ประเมินโดยใช้สถิติแคปปา
Evaluation of Inter-Rater Reliability Using Kappa Statistics

ประสพชัย พสุนนท์ [1]

บทคัดย่อ

ความเชื่อมั่นระหว่างผู้ประเมิน (Inter-Rater Reliability: IRR) เป็นการวัดความเชื่อมั่นแบบความสอดคล้องภายในของเครื่องมือวิจัยประเภทหนึ่ง การวัดความเชื่อมั่นด้วย IRR เป็นการพิจารณาของผู้ประเมินตั้งแต่ 2 คนขึ้นไป โดยที่ผู้ประเมินแต่ละคนมีความเป็นอิสระต่อกัน และค่าความเชื่อมั่นจึงขึ้นกับความคงเส้นคงวาของผลการพิจารณาจากผู้ประเมิน และขึ้นอยู่กับสถิติที่ใช้คำนวณความสอดคล้อง สำหรับการวิเคราะห์ IRR มีสถิติในการคำนวณแบ่งตามระดับของข้อมูล บทความนี้ ผู้เขียนเน้นสถิติของ IRR ในการวิเคราะห์ข้อมูลเชิงกลุ่ม มีวัตถุประสงค์เพื่อนำเสนอสถิติแคปปาและสถิติฟลีสแคปปาสำหรับข้อมูลระดับนามบัญญัติ นอกจากนี้ ยังได้นำเสนอสถิติแคปปาถ่วงน้ำหนักสำหรับข้อมูลระดับอันดับ การวิเคราะห์ IRR มีประโยชน์ต่อผู้วิจัยเพราะเป็นอีกทางเลือกในการประเมินความเชื่อมั่นของเครื่องมือวิจัยให้มีความเหมาะสมกับบริบทการวิจัย โดยเป็นการวัดความเที่ยงตรงที่ต้องอาศัยผู้ประเมินในการพิจารณาความสอดคล้องในระหว่างวัตถุประสงค์ของการประเมินเครื่องมือวิจัย และเพื่อให้สะดวกต่อการนำไปใช้ในบทความได้ให้แนวทางและตัวอย่างการคำนวณด้วยโปรแกรมสำเร็จรูป และโปรแกรม Excel

คำสำคัญ : ความเชื่อมั่นระหว่างผู้ประเมิน สถิติแคปปา

Abstract

Inter-Rater Reliability (IRR) is internal consistency reliability measurement of research tool. Using IRR requires 2 or more raters. Each rater is independent so that the reliability relies on consistency of evaluation results from raters and also on statistics used to calculate consistency. As the IRR analysis, there are statistics for calculation divided by scale of data. In this article, the writer focused on IRR statistics to analyze categorical data. Its objective was to propose Kappa statistics and Fleiss's Kappa Statistics for nominal scale, and also Weighted Kappa for ordinal scale. IRR analysis gives an advantage to researchers as an alternative to evaluate reliability of research tool so that it fits into research context. This measurement requires raters to consider a consistency during evaluating research tool. For the convenience of its use, the article included guidelines and examples of calculation with statistical program and Excel programs.

Keywords : Inter-Rater Reliability, Kappa Statistics

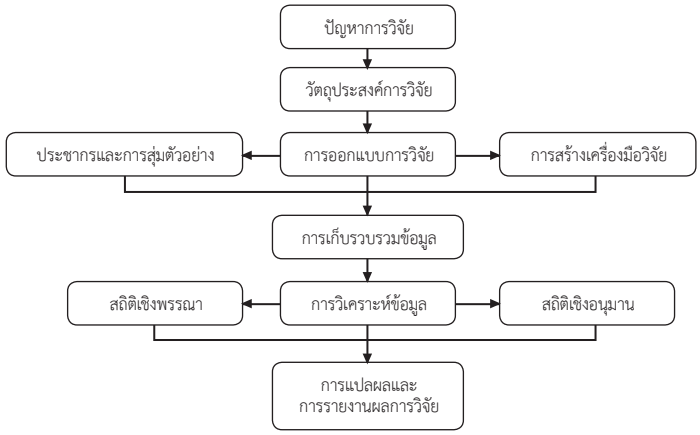
[1] รองศาสตราจารย์ ประจำคณะวิทยาการจัดการ มหาวิทยาลัยศิลปากร วิทยาเขตสารสนเทศเพชรบุรี

1. บทนำ

การวิจัย เป็นวิธีการทางวิทยาศาสตร์ในการแสวงหาความรู้และความจริงที่มีความน่าเชื่อถือและเป็นที่ยอมรับในปัจจุบัน ผลการวิจัยที่มีความถูกต้องและน่าเชื่อถือมีส่วนประกอบจากหลายส่วน เช่น การตีความปัญหาการวิจัย การออกแบบการวิจัย ข้อมูลในการวิจัย การวิเคราะห์ข้อมูล เป็นต้น นอกจากนี้ องค์ประกอบประการหนึ่งที่มีส่วนสำคัญต่อการวิจัย คือ เครื่องมือการวิจัย ซึ่งเป็นอุปกรณ์หรือสิ่งที่ผู้วิจัยใช้เป็นสื่อในการเก็บรวบรวมข้อมูลทั้งข้อมูลเชิงปริมาณหรือเชิงคุณภาพ

ในกรณีที่การวิจัยเป็นการวิจัยทางวิทยาศาสตร์ เครื่องมือที่ใช้ในการวิจัยอาจจะเป็นอุปกรณ์ที่ใช้ในการชั่ง ตวง หรือวัดเพื่อวิเคราะห์ทางการแพทย์ การเกษตร คอมพิวเตอร์ เคมี ภูมิศาสตร์ ฯลฯ การวิจัยลักษณะนี้มักเป็นการวิจัยและพัฒนาหรือการวิจัยเชิงทดลองด้วยวัสดุ อุปกรณ์ หรือสิ่งของต่างๆ ในแบบแผนการทดลองทางวิทยาศาสตร์ ต่างจากการวิจัยทางพฤติกรรมศาสตร์และสังคมศาสตร์ โดยเฉพาะการวิจัยที่ใช้ระเบียบวิธีวิจัยเชิงปริมาณ (Quantitative Research Methodology) นั้น ต้องอาศัยการวัดค่าตัวแปรเพื่อนำไปวิเคราะห์ข้อมูลในทางพฤติกรรม การวัดดังกล่าวเป็นการวัดทางอ้อม เป็นการวัดที่ไม่สมบูรณ์และไม่มีเครื่องมือวัดมาตรฐาน ดังนั้น การวัดในการวิจัยทางพฤติกรรมศาสตร์และสังคมศาสตร์ จึงเป็นการวัดที่ไม่มีกฎเกณฑ์ตายตัวและหน่วยการวัดมีความไม่เท่ากัน (เช่น ความสุข ความพึงพอใจ ความเครียด เป็นต้น) (พิษณุ ฟองศรี, 2552) การวิเคราะห์ผลจากการวัดจึงนิยมใช้การเปรียบเทียบข้อมูลทั้งภายในกลุ่มหรือต่างกลุ่ม เครื่องมือวิจัยจำเป็นจึงต้องมีคุณสมบัติความเที่ยงตรง (Validity) และความเชื่อมั่น (Reliability) (Drost, 2011) เครื่องมือวิจัยลักษณะนี้ ได้แก่ แบบสอบถาม แบบทดสอบ แบบสัมภาษณ์ แบบประเมินพฤติกรรม แบบประเมินผลงาน เป็นต้น

โดยทั่วไป ขั้นตอนการวิจัยเชิงปริมาณเริ่มต้นจากปัญหาการวิจัยจนกระทั่งการเขียนรายงานผลการวิจัย แสดงดังภาพที่ 1 ดังนั้น ผลการวิจัยจะขาดความน่าเชื่อถือหากเครื่องมือการวิจัยไม่มีความเที่ยงตรงและความเชื่อมั่น กล่าวได้ว่าความเชื่อมั่นเป็นสาเหตุสำคัญประการหนึ่งที่ส่งผลต่อคุณภาพของเครื่องมือการวิจัย เพราะความเชื่อมั่นของเครื่องมือการวิจัยจะช่วยให้ผู้วิจัยสามารถชั่ง ตวง และวัด ได้ตรงตามวัตถุประสงค์ของการวิจัยเช่นเดียวกับการวิจัยทางวิทยาศาสตร์ ดังนั้น หากเครื่องมือการวิจัยไม่มีความเชื่อมั่น จะทำให้การแปลผลและรายงานผลการวิจัยมีความคลาดเคลื่อนไปจากที่ควรจะเป็นการวัดความเชื่อมั่นของเครื่องมือวิจัยนั้น แบ่งออกเป็น 3 ประเภท คือ ความเชื่อมั่นแบบคงที่ (Stability Reliability) ความเชื่อมั่นแบบสมมูล (Equivalent Reliability) และความเชื่อมั่นแบบความสอดคล้องภายใน (Internal Consistency Reliability) โดยที่การวัดความเชื่อมั่นแบบคงที่และแบบสมมูลนั้น จะใช้วิธีการทดสอบซ้ำ (Test-retest Method) ด้วยเครื่องมือเดิม และวิธีการทดสอบโดยใช้เครื่องมือวัดคู่ขนาน (Parallel Method) ตามลำดับ ส่วนวิธีการที่นิยมใช้ในการวัดความเชื่อมั่นแบบความสอดคล้องภายในที่พบมากในการวิจัย คือ วิธีการครอนบาค (Cronbach Method) ซึ่งเป็นการพิจารณาในรูปค่าสัมประสิทธิ์ที่มีค่าระหว่าง 0 – 1 อย่างไรก็ตาม ความเชื่อมั่นแบบความสอดคล้องภายในนั้นมีหลายวิธีการ ผู้วิจัยสามารถเลือกใช้ให้เหมาะสมกับบริบทของการวิจัยได้



ภาพที่ 1 ขั้นตอนการวิจัย

แนวทางหนึ่งของวิธีการวัดความเชื่อมั่นแบบความสอดคล้องภายใน คือ ความเชื่อมั่นระหว่างผู้ประเมิน (Inter-Rater Reliability: IRR) ซึ่งเป็นการวัดความเชื่อมั่นของเครื่องมือวิจัยที่ต้องอาศัยผู้ประเมิน ผู้เชี่ยวชาญ หรือผู้ทรงคุณวุฒิให้คะแนน ค่าความเชื่อมั่นในลักษณะนี้ จึงขึ้นกับความคงเส้นคงวาของผลการพิจารณาของผู้ประเมิน (Rosenthal and Rosnow, 1991) เช่น ให้แพทย์ 2 คน ประเมินอาการผู้ป่วยโรคหัวใจ 1 ราย โดยประเมินพร้อมกันด้วยแบบประเมินเดียวกัน สามารถคำนวณความเชื่อมั่นจากการสังเกตระหว่างผู้ประเมินได้จาก (1) ถ้าค่าที่ได้มีค่าระหว่าง 0.8 ถึง 1 แสดงว่ามีความเชื่อมั่นระหว่างผู้ประเมินสูง การประเมินด้วย (1) ถือเป็น IRR ประเภทหนึ่ง (Gisev Bell and Chen, 2013) นอกจากนี้ IRR ยังมีชื่อเรียกอีกอย่างหนึ่งว่า ความเชื่อมั่นระหว่างการสังเกต (Inter-Observer Reliability) นั่นคือ ความเชื่อมั่นระหว่างผู้ประเมินและความเชื่อมั่นจากการสังเกต เป็นการประเมินในลักษณะแบบเดียวกัน

$$\text{ความเชื่อมั่นจากการสังเกต} = \frac{\text{จำนวนที่ประเมินเหมือนกัน}}{(\text{จำนวนที่ประเมินเหมือนกัน}+\text{จำนวนที่ประเมินต่างกัน})} \quad \text{--------(1)}$$

การประเมินความเชื่อมั่นด้วย IRR ยังไม่เป็นที่แพร่หลายในการวิจัยทางพฤติกรรมศาสตร์และสังคมศาสตร์มากนัก เห็นได้จากบทความวิจัย รายงานการวิจัย หรือวิทยานิพนธ์ในทางพฤติกรรมศาสตร์และสังคมศาสตร์ในประเทศไทย แทบไม่มีการประเมินความเชื่อมั่นด้วย IRR ทั้งนี้ อาจเป็นเพราะความไม่คุ้นเคย IRR เนื่องจากในหนังสือตำรา หรืองานวิจัยในอดีตก็นิยมวัดความเชื่อมั่นด้วยวิธีการครอนบาค โดยไม่ได้มีเนื้อหาของ IRR นำเสนอไว้หรือมีรายละเอียดของ IRR ค่อนข้างน้อย นอกจากนี้ อาจมีสาเหตุการการขาดความเข้าใจในการเลือกใช้สถิติสำหรับวิเคราะห์ IRR

ส่วนใหญ่ IRR พบในการวิจัยที่เป็นการประยุกต์รวมกันระหว่างการวิจัยด้านสังคมศาสตร์และด้านการแพทย์ สาธารณสุข และการพยาบาล เพื่อใช้ในการประเมินความเชื่อมั่นที่มีความไว (Sensitivity) และให้ความสำคัญต่อผลการประเมินโดยอาศัยผู้ประเมิน บทความนี้ มีวัตถุประสงค์ในการนำเสนอวิธีการวัดความเชื่อมั่นระหว่างผู้ประเมิน (IRR) เพื่อเป็นทางเลือกหนึ่งสำหรับผู้วิจัยในการเลือกใช้วิธีการประเมินความเชื่อมั่นของเครื่องมือการวิจัย

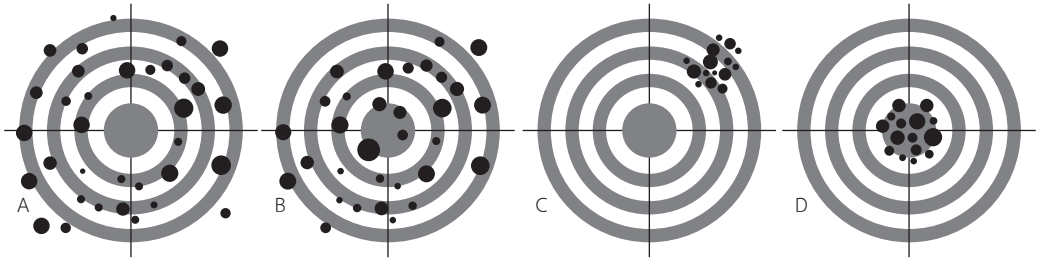
สถิติแคปปา (Kappa Statistics) เป็นสถิติที่นิยมใช้ในการประเมิน IRR สำหรับข้อมูลระดับนามบัญญัติ (Nominal Scale) และข้อมูลระดับอันดับ (Ordinal Scale) ในบางกรณี สำหรับการนำเสนอเนื้อหาในบทความนี้ เริ่มต้นจากการทำความเข้าใจแนวคิดของความเชื่อมั่นที่มีความสัมพันธ์กับความเที่ยงตรง จากนั้น แสดงความเชื่อมโยงของ IRR และการประเมินความเชื่อมั่นแบบความสอดคล้องภายใน ก่อนที่จะเป็นการทำความเข้าใจการประเมิน IRR ด้วยสถิติแคปปา กรณีมี 2 ผู้ประเมินและประเมิน 2 ประเภท พร้อมเกณฑ์ในการพิจารณาระดับความสอดคล้อง สถิติแคปปาสำหรับกรณีมี 2 ผู้ประเมินและประเมินตั้งแต่ 3 ประเภทพร้อมแสดงตัวอย่างการคำนวณด้วยโปรแกรมสำเร็จรูป รวมไปถึงการคำนวณความคลาดเคลื่อนมาตรฐาน (Standard Error) ถัดจากนั้น ผู้เขียนได้นำเสนอสถิติแคปปาในกรณีที่มีค่าสูญหาย (Missing Data) สถิติแคปปาถ่วงน้ำหนัก (Weighted Kappa) และสถิติฟลีสแคปปา (Fleiss's Kappa Statistic) พร้อมตัวอย่างการคำนวณด้วยโปรแกรม Excel และตอนสุดท้ายของบทความเป็นข้อเสนอแนะในการใช้ IRR

บทความนี้ มุ่งประเด็น (Focus) ที่สถิติแคปปาในการวัดความสอดคล้องหรือประเมินความเชื่อมั่นโดยใช้ผู้ประเมินตั้งแต่ 2 คนขึ้นไป อย่างไรก็ตาม สำหรับข้อมูลระดับอันตรภาค (Interval Scale) และอัตราส่วน (Ratio Scale) สามารถวิเคราะห์ IRR ด้วยสถิติสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) หรือสัมประสิทธิ์สหสัมพันธ์ความสอดคล้อง (Concordance Correlation Coefficient: CCC) ซึ่งจะได้นำเสนอในโอกาสถัดไป เนื่องจากมีรายละเอียดในการทำความเข้าใจค่อนข้างมากและมีความซับซ้อน

2. แนวคิดของความเที่ยงตรงและความเชื่อมั่น

ผู้เขียนใช้เกมปาเป้าลูกดอกในภาพที่ 2 เพื่ออธิบายแนวคิดของความเที่ยงตรงและความเชื่อมั่นของเครื่องมือการวิจัย หากลูกดอกเข้าตรงกลางเป้าแสดงว่าวัดสิ่งที่ต้องการวัดได้ถูกต้องจริง แต่หากพลาดเป้าแสดงว่าวัดสิ่งที่ต้องการวัดได้ไม่ตรงหรือไม่ถูกต้อง รูปย่อย A จะเห็นว่าการวัดไม่มีความเที่ยงตรงและไม่มีความเชื่อมั่น เพราะไม่มีลูกดอกใดที่เข้าบริเวณตรงกลาง และการเข้าเป้ากระจัดกระจาย ในขณะที่รูปย่อย B มีบางลูกดอกเข้าใกล้บริเวณตรงกลาง แต่ลักษณะการเข้าเป้ากระจัดกระจาย

นั่นคือ มีความเที่ยงตรงอยู่บ้างแต่ไม่มีความเชื่อมั่น ส่วนรูปย่อย C การเข้าเป้ามีลักษณะเกาะกลุ่ม แต่ไม่ใช่ตรงกลาง ลักษณะเช่นนี้ คือ เครื่องมือวิจัยสามารถใช้วัดค่าในทิศทางเดียวกัน แต่ไม่ตรงกับสิ่งที่ต้องการวัด (หรือไม่ตรงกับวัตถุประสงค์ของการวิจัย) นั่นคือ มีความเชื่อมั่นแต่ไม่มีความเที่ยงตรง และรูปย่อย D เป็นการวัดที่มีทั้งความเที่ยงตรงและความเชื่อมั่น



ที่มา : ปรับปรุงจาก Crutzen (2014)

ภาพที่ 2 แนวคิดของความเที่ยงตรงและความเชื่อมั่น

ทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory) อธิบายความสัมพันธ์ระหว่างคะแนนที่ได้จากการวัดและความคลาดเคลื่อนในการวัด โดยมีข้อตกลงเบื้องต้น คือ คะแนนสังเกต (X) = คะแนนจริง (T) + ความคลาดเคลื่อนจากการวัด (E) (หรือ Observed Score (X) = True Score (T) + Measurement Error (E)) แทนในรูปสัญลักษณ์ดัง (2) (Hallgren, 2012) โดยทฤษฎีการทดสอบแบบดั้งเดิมให้ความสำคัญต่อความเที่ยงตรงและความเชื่อมั่น ถ้าข้อมูลในการวัดสอดคล้องกับข้อตกลงเบื้องต้น การสรุปผลการวัดจะมีความสมเหตุสมผล

$$X = T + E \text{ --------(2)}$$

เหตุที่ต้องกล่าวถึงทฤษฎีการทดสอบแบบดั้งเดิมนั้น เพราะในการพัฒนาสูตรการประเมินความเชื่อมั่นแบบความสอดคล้องภายในมีพื้นฐานมาจากความสัมพันธ์ของ (2) โดยที่ E(X) = T และ $\rho_{T,E} = 0$ เมื่อ E(X) คือ ค่าคาดหวัง (Expectation) ของ X และ $\rho_{T,E}$ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ของ T และ E

3. ความเชื่อมั่นระหว่างผู้ประเมิน

การประเมินความเชื่อมั่นแบบความสอดคล้องภายใน โดยมากล้วนมีรากฐานจาก (2) อาทิ วิธีการคูเดอร์-ริชาร์ดสัน (Kuder-Richardson Method) ทั้งวิธีการ KR20 และ KR21 วิธีการครอนบาค (Cronbach Method) วิธีการกัตแมน (Guttman Method) วิธีการลิวิงสตัน (Livingston Method) เป็นต้น วิธีการ IRR ก็เช่นเดียวกัน โดยการพิจาณา (2) ในรูปของความแปรปรวน (Variance) แสดงดัง (3)

$$Var(X) = Var(T) + Var(E) \text{ --------(3)}$$

เมื่อกำหนดความแปรปรวนของคะแนนจริง (Var(T)) ต่อความแปรปรวนของคะแนนสังเกต (Var(X)) เมื่อกำจัดความคลาดเคลื่อนจากการวัดจากผู้ประเมิน (Novick, 1966) จะได้ค่าความเชื่อมั่นแสดงความสัมพันธ์ดัง (4)

$$\text{ความเชื่อมั่น} = \frac{Var(T)}{Var(X)} = \frac{Var(X) - Var(E)}{Var(X)} = \frac{Var(T)}{Var(T)+Var(E)} \text{ --------(4)}$$

เนื่องจาก T และ E ไม่สามารถวัดค่าได้โดยตรง จึงไม่สามารถคำนวณความเชื่อมั่นโดยตรง ดังนั้น การคำนวณ IRR จึงไม่มีเครื่องมือที่ใช้ในการวัดโดยสมบูรณ์ แนวทางการประมาณค่าคะแนนจริงโดยคำนวณความแปรปรวนร่วม (Covariance) ระหว่างคะแนนสังเกต (X) ที่ได้จากผู้ประเมินตามวัตถุประสงค์การวิจัย (Hallgren, 2012) ซึ่งสอดคล้องกับ (1) อย่างไรก็ตาม IRR นั้น แตกต่างจากการวัดความเที่ยงตรงโดยผู้ทรงคุณวุฒิ โดเฉพาะค่าดัชนีความสอดคล้องของข้อคำถามแต่ละข้อกับวัตถุประสงค์ (Index of Item – Objective Congruence: IOC) ซึ่งนิยมในการวัดความเที่ยงตรงของแบบสอบถาม เพราะการวัดความเที่ยงตรงเป็นการวัดความถูกต้องและเหมาะสมของข้อคำถาม จึงต้องใช้ผู้ที่มีความเชี่ยวชาญในเรื่องนั้นๆ เป็นผู้ประเมินคำถามรายข้อ แล้วคำนวณค่าเฉลี่ยจากผลการประเมินว่าคำถามเหมาะสม (+1) คำถามไม่เหมาะสม (-1) และไม่แน่ใจ (0) โดยทั่วไป ถ้าค่าเฉลี่ยมีค่ามากกว่าหรือเท่ากับ 0.5 แสดงว่ามีความเที่ยงตรง ในขณะที่ IRR เป็นการพิจารณาความสอดคล้องกันระหว่างผู้ประเมินว่ามีความเห็นสอดคล้องกันหรือไม่

ในการวิจัยที่มีการออกแบบการวิจัย โดยใช้ความเห็นของผู้ประเมินเพื่อประเมินความเชื่อมั่น มักจะต้องใช้ IRR ในการวิจัย นอกจากนี้ การประเมิน IRR ยังเป็นอีกหนึ่งทางเลือกในการอธิบายความเห็นของผู้ประเมินตั้งแต่ 2 คนขึ้นไป โดยที่ผู้ประเมินแต่ละคนมีความเป็นอิสระต่อกัน สำหรับสถิติที่ใช้ในการประเมิน IRR นั้น Hallgren (2012) และ Gisev Bell and Chen (2013) ได้นำเสนอไว้ 2 แนวทาง คือ 1) สถิติแคปปา (Kappa Statistics) สำหรับข้อมูลระดับนามบัญญัติ และสถิติแคปปาถ่วงน้ำหนัก (Weighted Kappa) สำหรับข้อมูลระดับเรียงอันดับ และ 2) สัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) สำหรับข้อมูลระดับอันดับ อันตรภาค และอัตราส่วน นอกจากนี้ IRR ยังมีความหมายเดียวกันกับความสอดคล้องระหว่างผู้ประเมิน (Inter-Rater Agreement) โดยใช้สถิติในการวิเคราะห์เหมือนกัน (Gisev Bell and Chen, 2013)

4. สถิติแคปปาสำหรับกรณีมี 2 ผู้ประเมินและประเมิน 2 ประเภท

สถิติแคปปาหรือค่าสัมประสิทธิ์โคเฮนแคปปา (Cohen's Kappa Coefficient) เป็นค่าที่ใช้พิจารณาความเห็นระหว่างผู้ประเมินว่ามีความเห็นสอดคล้องมากหรือน้อยเพียงใด โดยในช่วงแรกของการพัฒนาสถิติใช้ประเมินกับข้อมูลระดับนามบัญญัติหรือข้อมูลเชิงกลุ่ม (Categorical Data) โดยมีผู้ประเมินเพียง 2 คน ก่อนที่จะได้มีการพัฒนาเพื่อประยุกต์ให้สถิติแคปปาสามารถใช้กับข้อมูลระดับเรียงอันดับ และสามารถใช้กรณีที่มีผู้ประเมินมากกว่า 2 คน

Stevens (1946) ให้ความหมายของข้อมูลระดับนามบัญญัติเป็นข้อมูลที่อยู่ระดับต่ำที่สุด ทำหน้าที่ในการแบ่งข้อมูลออกเป็นกลุ่มๆ ตามลักษณะข้อมูล เช่น ศาสนา (พุทธ อิสลาม คริสต์ ฮินดู) เพศ (ชาย หญิง) ภูมิภาค (เหนือ กลาง อีสาน ใต้ ตะวันออก ตะวันตก) เป็นต้น จะถือว่าแต่ละกลุ่มมีความเท่าเทียมกัน หลักเกณฑ์การจำแนก คือ ความเหมือนหรือความแตกต่างของแต่ละข้อมูล ความครอบคลุม ความครบถ้วน สมบูรณ์ และความเป็นอิสระต่อกัน บางครั้งเรียกข้อมูลระดับนี้ว่าข้อมูลเชิงกลุ่ม ส่วนข้อมูลระดับอันดับ เป็นระดับการวัดที่สูงกว่าระดับนามบัญญัติ คือ สามารถแบ่งข้อมูลเป็นกลุ่มๆ และเรียงลำดับกันเองตามลักษณะของข้อมูลโดยธรรมชาติ บอกได้ว่าข้อมูลของกลุ่มหนึ่งสูงกว่าอีกกลุ่มหนึ่ง แต่ไม่สามารถหาระยะห่างระหว่างกลุ่มได้ เช่น การวัดระดับการศึกษา (มัธยมศึกษาตอนปลาย ปริญญาตรี ปริญญาโท ปริญญาเอก) การวัดทางทัศนคติ (มากที่สุด มาก ปานกลาง น้อย และน้อยที่สุด) ตำแหน่งทางวิชาการ (ศาสตราจารย์ รองศาสตราจารย์ ผู้ช่วยศาสตราจารย์ อาจารย์) เป็นต้น (ประสพชัย พสุนนท์, 2555)

สำหรับการวัดความสอดคล้องกันระหว่างความเห็นของผู้ประเมิน สมมติว่ามีผู้ประเมิน A และ B โดยผู้ประเมินทั้งสองได้จำแนกข้อมูลประชากร (Population) ขนาด N ซึ่งเป็นชุดเดียวกัน ออกเป็น 2 ประเภท คือ ประเภท 1 และประเภท 2 โดยเป็นเหตุการณ์ที่ไม่เกิดร่วมกัน แสดงดังตารางที่ 1 เมื่อ N_{ij} แทนจำนวนข้อมูลที่ผู้ประเมิน A เห็นว่าข้อมูลควรเป็นประเภท i ส่วนผู้ประเมิน B เห็นว่าข้อมูลชุดเดียวกันนั้นควรเป็นประเภท j (เมื่อ i = 1 , 2; j = 1 , 2) และ $N_{.1} = N_{11} + N_{21}$, $N_{.2} = N_{12} + N_{22}$, $N_{1.} = N_{11} + N_{12}$, $N_{2.} = N_{21} + N_{22}$ และ $N = N_{11} + N_{12} + N_{21} + N_{22}$

ตารางที่ 1 การแจกแจงความถี่ข้อมูลประชากร 2 ประเภท ของผู้ประเมิน A และ B

ผู้ประเมิน A	ผู้ประเมิน B		รวม
	ประเภท 1	ประเภท 2	
ประเภท 1	N_{11}	N_{12}	$N_{1.}$
ประเภท 2	N_{21}	N_{22}	$N_{2.}$
รวม	$N_{.1}$	$N_{.2}$	N

ตารางที่ 2 การแจกแจงความน่าจะเป็นข้อมูลประชากร 2 ประเภท ของผู้ประเมิน A และ B

ผู้ประเมิน A	ผู้ประเมิน B		รวม
	ประเภท 1	ประเภท 2	
ประเภท 1	$\pi_{11} = N_{11}/N$	$\pi_{12} = N_{12}/N$	$\pi_{1.} = N_{1.}/N$
ประเภท 2	$\pi_{21} = N_{21}/N$	$\pi_{22} = N_{22}/N$	$\pi_{2.} = N_{2.}/N$
รวม	$\pi_{.1} = N_{.1}/N$	$\pi_{.2} = N_{.2}/N$	1

จากตารางที่ 1 สามารถสร้างตารางแจกแจงความน่าจะเป็นได้ดังตารางที่ 2 โดยที่ π_{ij} แทนความน่าจะเป็นของเหตุการณ์ที่ผู้ประเมิน A เห็นว่าข้อมูลเป็นประเภท i ส่วนผู้ประเมิน B เห็นว่าข้อมูลเป็นประเภท j กล่าวคือ π_{11} และ π_{22} แทนความน่าจะเป็นที่ผู้ประเมินเห็นสอดคล้องกัน ส่วน π_{12} และ π_{21} แทนความน่าจะเป็นที่ผู้ประเมินเห็นไม่สอดคล้องกัน Cohen (1960) ได้เสนอพารามิเตอร์ (Parameter) K_C สำหรับข้อมูลประชากรดัง (5) เพื่อใช้ประเมินความสอดคล้องระหว่างผู้ประเมิน

$$K_C = \frac{\theta_a - \theta_e}{1 - \theta_e} \text{ --------(5)}$$

โดยที่ $\theta_a = \pi_{11} + \pi_{22}$ และ $\theta_e = \pi_{1.}\pi_{.1} + \pi_{2.}\pi_{.2}$ ในทางปฏิบัติ จะประมาณพารามิเตอร์ K_C ด้วยตัวประมาณ (Estimator) หรือสถิติ $\hat{K}_C$ ดัง (6) เมื่อ $\theta_a \approx P_a$ และ $\theta_e \approx P_e$ หรือใช้ข้อมูลตัวอย่าง (Sample) แทนข้อมูลประชากร เมื่อแทนขนาดตัวอย่างด้วย n และ $P_{ji} = \frac{n_{ij}}{n}$ นั่นคือ จะประมาณ N_{ij} และ π_{ij} ด้วย n_{ij} และ P_{ij} ตามลำดับ เมื่อ $P_a = P_{11} + P_{22}$ และ $P_e = P_{1.}P_{.1} + P_{2.}P_{.2}$

$$\hat{K}_C = \frac{P_a - P_e}{1 - P_e} \text{ --------(6)}$$

จาก (5) และ (6) ความเป็นไปได้ของค่าสถิติ K_C และ $\hat{K}_C$ จะมีค่าระหว่าง -1.0 ถึง 1.0 ซึ่งทั้ง K_C และ $\hat{K}_C$ มีคุณสมบัติเหมือนกันต่างกันเพียงเป็นพารามิเตอร์ (ข้อมูลประชากร) หรือตัวประมาณ (ข้อมูลตัวอย่าง) หากค่า $\hat{K}_C$ ที่เป็นลบนั้น แสดงถึงความเห็นระหว่างผู้ประเมินมีความสอดคล้องกันในลักษณะตรงกันข้าม (Hallgren, 2012) อย่างไรก็ตาม ในทางทฤษฎีมักพิจารณาค่า $\hat{K}_C$ ควรมีค่าตั้งแต่ 0.0 แต่ไม่เกิน 1.0 ($0.0 \leq K_C \leq 1.0$) ตัวอย่างการคำนวณ $\hat{K}_C$ จะใช้ข้อมูลในตารางที่ 3 พบว่า

$P_a = \frac{35}{100} + \frac{40}{100} = 0.75$ และ $P_e = \left(\frac{40}{100}\frac{55}{100}\right) + \left(\frac{60}{100}\frac{45}{100}\right) = 0.49$

ดังนั้น $\hat{K}_C = \frac{0.75 - 0.49}{1 - 0.49} = 0.51$

ตารางที่ 3 ข้อมูลตัวอย่างการคำนวณ $\hat{K}_C$ ของผู้ประเมิน A และ B

ผู้ประเมิน A	ผู้ประเมิน B		รวม
	ประเภท 1	ประเภท 2	
ประเภท 1	35	20	55
ประเภท 2	5	40	45
รวม	40	60	100

5. เกณฑ์การพิจารณาระดับความสอดคล้องของสถิติแคปปา

ตารางที่ 4 เป็นเกณฑ์การพิจารณาระดับความสอดคล้องของสถิติแคปปาตามแนวทางของ Landis and Koch (1977) แต่ Krippendorff (1980) กลับเสนอแนวทางในการพิจารณาระดับความสอดคล้องที่ค่อนข้างมีความอนุรักษ์นิยมกว่า Landis and Koch (1977) โดยระบุค่า $\hat{K}_C$ ที่น้อยกว่า 0.67 หมายถึง ยังไม่มีความสอดคล้องระหว่างผู้ประเมิน ถ้า $0.67 \leq \hat{K}_C \leq 0.80$ หมายถึง ความสอดคล้องระหว่างผู้ประเมินยังคงมีความกำกึ่งหรือมีสภาพไม่แน่นอน และถ้า $\hat{K}_C$ มีค่ามากกว่า 0.80 หมายถึง ความสอดคล้องระหว่างผู้ประเมินเป็นที่สามารถยอมรับได้ในทางปฏิบัติ

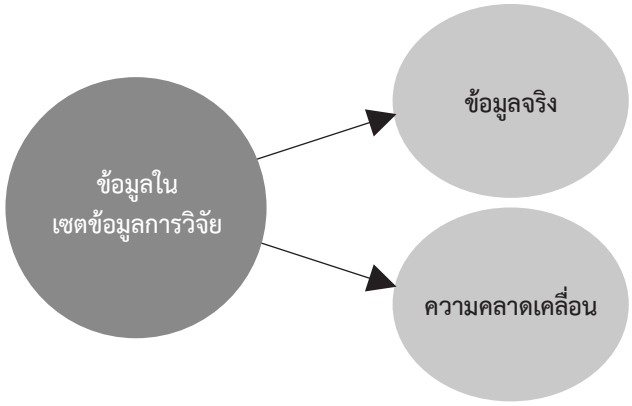
ตารางที่ 4 ระดับความสอดคล้องของสถิติแคปปาตามแนวทางของ Landis and Koch (1977)

ค่าสถิติ Kappa	ระดับความสอดคล้องระหว่างผู้ประเมิน
0.81 – 1.00	ความสอดคล้องดีมาก (Almost Perfect)
0.61 – 0.80	ความสอดคล้องดี (Substantial)
0.41 – 0.60	ความสอดคล้องปานกลาง (Moderate)
0.21 – 0.40	ความสอดคล้องพอใช้ (Fair)
0.00 – 0.20	ความสอดคล้องเล็กน้อย (Slight)
น้อยกว่า 0.00	ไม่มีความสอดคล้อง (Poor)

แนวทางของ Krippendorff (1980) ยังคงได้รับความนิยมในการวิจัยที่ต้องใช้ในการตัดสินใจในประเด็นที่มีความละเอียดอ่อน เช่น การวินิจฉัยโรค การสร้างแบบวัดทางสุขภาพจิต เป็นต้น ทั้งนี้ Hallgren (2012) เสนอให้ผู้วิจัยสามารถเลือกใช้เกณฑ์ของ Landis and Koch (1977) และ Krippendorff (1980) ตามความเหมาะสม โดยอาจพิจารณาจากปัญหาการวิจัย บริบทการวิจัย และวิธีดำเนินการวิจัยประกอบการแปลความหมายของสถิติแคปปา นอกจากนี้ Fleiss Levin and Paik (2003) เสนอเกณฑ์การพิจารณาระดับความสอดคล้องดังตารางที่ 5

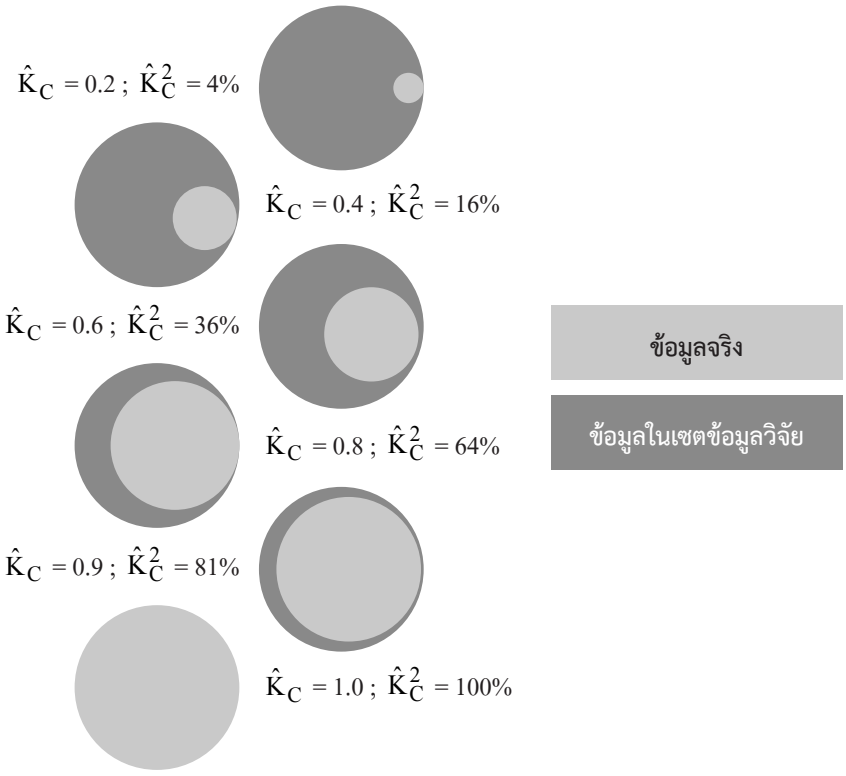
ตารางที่ 5 ระดับความสอดคล้องของสถิติแคปปาตามแนวทางของ Fleiss Levin and Paik (2003)

ค่าสถิติ Kappa	ระดับความสอดคล้องระหว่างผู้ประเมิน
0.75 – 1.00	ความสอดคล้องดีมาก
0.40 – 0.74	ความสอดคล้องดี
0.00 – 0.39	ความสอดคล้องต่ำ



ที่มา : ปรับปรุงจาก McHugh (2012)

ภาพที่ 3 องค์ประกอบของข้อมูลในเซตข้อมูลการวิจัย



ที่มา : ปรับปรุงจาก McHugh (2012)

ภาพที่ 4 ร้อยละความสอดคล้องของสถิติหรือค่า Squared Kappa ($\hat{K}_C^2$)

McHugh (2012) ได้อธิบายถึงองค์ประกอบของข้อมูล (Data) ในเซตข้อมูลการวิจัย (Research Data Set) ที่ประกอบไปด้วยข้อมูลจริงและความคลาดเคลื่อน ดังภาพที่ 3 พร้อมทั้งเชื่อมโยงถึงการตีความสถิติ $\hat{K}_C$ ซึ่งเป็นค่าสัมประสิทธิ์สหสัมพันธ์ที่ไม่สามารถตีความได้โดยตรง การตีความต้องอยู่ในรูปสัมประสิทธิ์การกำหนด (Coefficient of Determination: COD) จึงจะตีความโดยตรงได้ COD ใช้ตีความค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's Correlation) ในการอธิบายความสัมพันธ์ระหว่างตัวแปรการวิจัยในรูปความแปรปรวน McHugh (2012) ได้ประยุกต์แนวคิดดังกล่าว ในการตีความความสัมพันธ์ระหว่างความสอดคล้องระหว่างผู้ประเมินด้วยค่าสถิติ Squared Kappa ($\hat{K}_C^2$) ดังภาพที่ 4 และเกณฑ์ระดับความสอดคล้องของสถิติแคปปา นำเสนอในตารางที่ 6

ตารางที่ 6 ระดับความสอดคล้องของสถิติแคปปาตามแนวทางของ McHugh (2012)

ค่าสถิติ Kappa	ระดับความสอดคล้องระหว่างผู้ประเมิน	ร้อยละของความเชื่อมั่น
0.00 – 0.19	ไม่มีความสอดคล้อง (None)	0 – 3%
0.20 – 0.39	ความสอดคล้องน้อยมาก (Minimal)	4 – 15%
0.40 – 0.59	ความสอดคล้องน้อย (Weak)	16 – 35%
0.60 – 0.79	ความสอดคล้องปานกลาง (Moderate)	36 – 63%
0.80 – 0.89	ความสอดคล้องมาก (Strong)	64 – 80%
0.90 – 1.00	ความสอดคล้องมากที่สุด (Almost Perfect)	81 – 100%

6. สถิติแคปปาสำหรับกรณีมี 2 ผู้ประเมินและประเมินตั้งแต่ 3 ประเภท

ในกรณีที่มีผู้ประเมิน A และ B โดยผู้ประเมินทั้งสองได้จำแนกข้อมูลตัวอย่างขนาด n ซึ่งเป็นชุดเดียวกัน ออกเป็น q ประเภท คือ ประเภทที่ 1 , 2 , ... , q โดยเป็นเหตุการณ์ที่ไม่เกิดร่วมกัน แสดงดังตารางที่ 7 สถิติ Kappa แสดงดัง (6) โดยที่ $P_a = \sum_{k=1}^{q} P_{kk}$ และ $P_e = \sum_{k=1}^{q} P_{k\cdot} P_{\cdot k}$ เมื่อ

$$P_{ij} = \frac{n_{ij}}{n} \quad (i = 1, 2, \ldots, q ; j = 1, 2, \ldots, q)$$

สำหรับตัวอย่างการคำนวณจะใช้โปรแกรมสำเร็จรูป สำหรับรายละเอียดจะกล่าวต่อไปในหัวข้อโปรแกรมสำเร็จรูปสำหรับคำนวณสถิติ $\hat{K}_C$

ตารางที่ 7 การแจกแจงความถี่ข้อมูลตัวอย่าง q ประเภท (q≥3) ของผู้ประเมิน A และ B

ผู้ประเมิน A	ผู้ประเมิน B				
	ประเภท 1	ประเภท 2	...	ประเภท q	รวม
ประเภท 1	n_{11}	n_{12}	...	n_{1q}	$n_{1\cdot}$
ประเภท 2	n_{21}	n_{22}	...	n_{21}	$n_{21\cdot}$
⋮	⋮	⋮	⋱	⋮	⋮
ประเภท q	n_{q1}	n_{q1}	...	n_{qq}	$n_{q\cdot}$
รวม	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot q}$	n

7. ความคลาดเคลื่อนมาตรฐานของสถิติ $\hat{K}_C$

ความคลาดเคลื่อนมาตรฐาน (Standard Error: SE) ของสถิติ $\hat{K}_C$ มีหลายแนวทางในการคำนวณ จากการทบทวนวรรณกรรม สามารถคำนวณได้ดังนี้

1. Fleiss Cohen and Everitt (1969) นำเสนอค่า SE แบบดั้งเดิมของสถิติ $\hat{K}_C$ ดัง (7) เมื่อ

$$A = \sum_{i=1}^{q} P_{ii}\left(1-(P_{i\cdot}+P_{\cdot i})(1-q)\right)^2, \quad B = (1-q)^2$$

$$\sum_{i\neq j}^{q}\sum P_{ij}(P_{i\cdot}+P_{\cdot j})^2 \text{ และ } C = (q - P_e(1 - q))^2$$

โดยที่ตัวอย่างต้องมีขนาดใหญ่ และ q คือ จำนวนของประเภทที่ใช้ประเมินความเชื่อมั่น

$$SE = \frac{\sqrt{A+B-C}}{\sqrt{n}(1-P_e)} \text{ --------(7)}$$

2. Perreault and Leigh (1989) ได้พัฒนาดัชนีความเชื่อมั่น (Index of Reliability) แทนด้วย I_r ดัง (8) เมื่อ $P_a \geq \frac{1}{q}$ แต่ถ้า $P_a < \frac{1}{q}$ จะได้ $I_r = 0$ และดัชนีความเชื่อมั่นมีความสัมพันธ์กับความคลาดเคลื่อนมาตรฐานของสถิติ $\hat{K}_C$ แสดงดัง (9) ภายใต้เงื่อนไขว่า nI_r มากกว่า 5

$$I_r = \sqrt{\left(P_a - \frac{1}{q}\right)\left(\frac{q}{q-1}\right)} \text{ --------(8)}$$

$$SE = \sqrt{\frac{I_r(1-I_r)}{n}} \text{ --------(9)}$$

3. McHugh (2012) ระบุว่าถ้า nP_a และ $n(1 - P_a)$ มีค่ามากกว่า 5 การแจกแจงของ $\hat{K}_C$ จะเข้าใกล้การแจกแจงแบบปกติ (Normal Distribution) และสามารถคำนวณค่า SE ได้ดัง (10)

$$SE = \sqrt{\frac{P_a(1-P_a)}{n(1-P_e)^2}} \text{ --------(10)}$$

ในกรณีที่ทราบค่า SE ของ $\hat{K}_C$ จะสามารถประมาณค่าแบบช่วง (Interval Estimation) ของ K_C ที่ช่วงความเชื่อมั่น 95% (95% Confidence Interval) ของ K_C ได้จาก $\hat{K}_C$ 1.96(SE) อย่างไรก็ตาม การคำนวณ SE จาก (7) (9) และ (10) อยู่บนพื้นฐาน IRR กรณี 2 ผู้ประเมินและประเมิน 2 ประเภท และในกรณี 2 ผู้ประเมินและประเมินตั้งแต่ 3 ประเภทขึ้นไป

8. โปรแกรมสำเร็จรูปสำหรับคำนวณสถิติ $\hat{K}_C$

การทดสอบสมมติฐานของสถิติ $\hat{K}_C$ เพื่อประเมินความเชื่อมั่นระหว่างผู้ประเมินของเครื่องมือการวิจัยมีข้อตกลงเบื้องต้น 3 ประการ (Gisev Bell and Chen, 2013) ดังนี้

1. ผู้ประเมินแต่ละคนมีความเป็นอิสระกัน (Independence)

2. ประเภทของการประเมินต้องมีความเป็นอิสระกัน เป็นเหตุการณ์ที่ไม่เกิดร่วมกัน (Mutually Exclusive Events) และเหตุการณ์ที่เกิดขึ้นรวมแล้วเป็นทั้งหมด (Exhaustive Events) ซึ่งคือ ความเป็นไปได้ของเหตุการณ์ทั้งหมด เมื่อรวมกันแล้วจะได้ปริภูมิตัวอย่าง (Sample Space)

3. รายการหรือหัวข้อที่ประเมินมีความเป็นอิสระกัน

สมมติฐานทางสถิติสำหรับการทดสอบ คือ H_O: เครื่องมือการวิจัยไม่มีความสอดคล้องกันระหว่างผู้ประเมิน (หรือ $H_O : K_C = 0$) ในขณะที่ H_1: เครื่องมือการวิจัยมีความสอดคล้องกันระหว่างผู้ประเมิน (หรือ $H_O : K_C > 0$) สถิติที่ใช้ในการทดสอบอาจจะเป็นสถิติ Z หรือสถิติ t ขึ้นกับลักษณะการแจกแจงของข้อมูลและเงื่อนไขการทดสอบ โดยสถิติทดสอบจะคำนวณจากอัตราส่วนของ $\hat{K}_C$ และ SE

สำหรับโปรแกรมคอมพิวเตอร์ที่ใช้คำนวณ $\hat{K}_C$ มีหลายโปรแกรม โปรแกรมหนึ่งที่นิยมและมีความคุ้นเคยในหมู่นักวิจัย คือ โปรแกรมสำเร็จรูป ผู้เขียนจึงใช้ข้อมูลของ Sim and Wright (2005) เป็นตัวอย่างในการแสดงขั้นตอนการคำนวณสถิติ $\hat{K}_C$ ด้วยโปรแกรมสำเร็จรูป

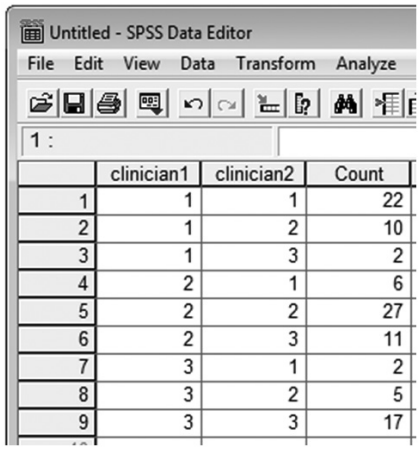
ตารางที่ 8 ข้อมูลของ Sim and Wright (2005)

แพทย์ผู้เชี่ยวชาญท่านที่ 1	แพทย์ผู้เชี่ยวชาญท่านที่ 2			
	อาการ Derangement	อาการ Dysfunctional	อาการ Postural	รวม
อาการ Derangement	22	10	2	34
อาการ Dysfunctional	6	27	11	44
อาการ Postural	2	5	17	24
รวม	30	42	30	102

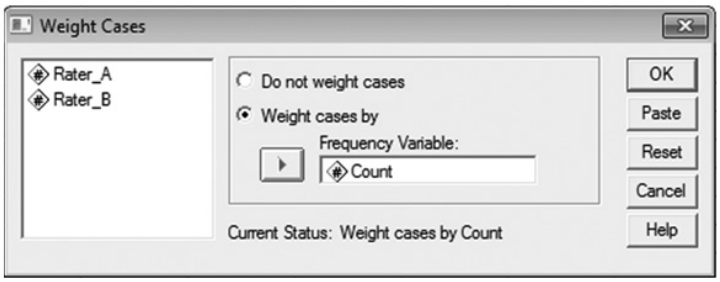
ตารางที่ 8 เป็นข้อมูลของ Sim and Wright (2005) ซึ่งเป็นข้อมูลตัวอย่างการประเมินอาการป่วยด้วยแบบเครื่องมือในการวินิจฉัยโรคจากผู้ป่วยจำนวน 102 คน โดยมีแพทย์ผู้เชี่ยวชาญในการประเมิน 2 คน อาการที่ประเมินแบ่งออกเป็น 3 ประเภท คือ อาการ Derangement อาการ Dysfunctional และอาการ Postural

ขั้นตอนการคำนวณสถิติ $\hat{K}_C$ ของข้อมูลของ Sim and Wright (2005) ด้วยโปรแกรมสำเร็จรูป มีดังนี้

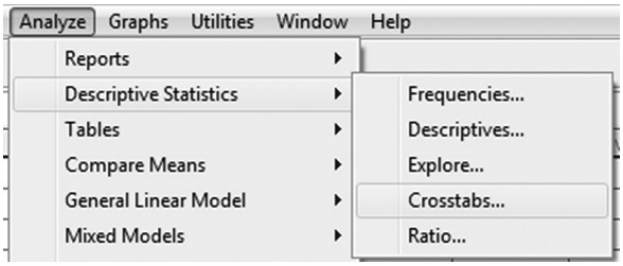
1. จัดเตรียมข้อมูลของ Sim and Wright (2005) ในโปรแกรมสำเร็จรูป ตามภาพที่ 5
2. เลือกเมนู Data >>> Weight cases... จะปรากฏหน้าต่าง (Dialog Box) ดังภาพที่ 6 ให้นำตัวแปร Count ย้ายเข้าไปอยู่ในช่อง Frequency Variable: แล้วคลิก OK
3. เลือกเมนู Analyze >>> Descriptive Statistics >>> Crosstabs... ตามภาพที่ 7 จะปรากฏหน้าต่างดังภาพที่ 8 ให้นำตัวแปร Clinician1 ย้ายเข้าไปอยู่ในช่อง Row(s): และนำตัวแปร Clinician2 ย้ายเข้าไปอยู่ในช่อง Column(s): ดังภาพที่ 8 จากนั้นคลิกที่ปุ่ม Statistics... จะปรากฏหน้าต่าง Crosstabs: Statistic ให้คลิกเครื่องหมาย ✓ ใน ☐ Kappa ดังภาพที่ 9 แล้วคลิก Continue
4. หน้าต่างจะกลับไปยังหน้าต่าง Crosstabs ให้คลิก OK จะได้ผลลัพธ์ดังภาพที่ 10



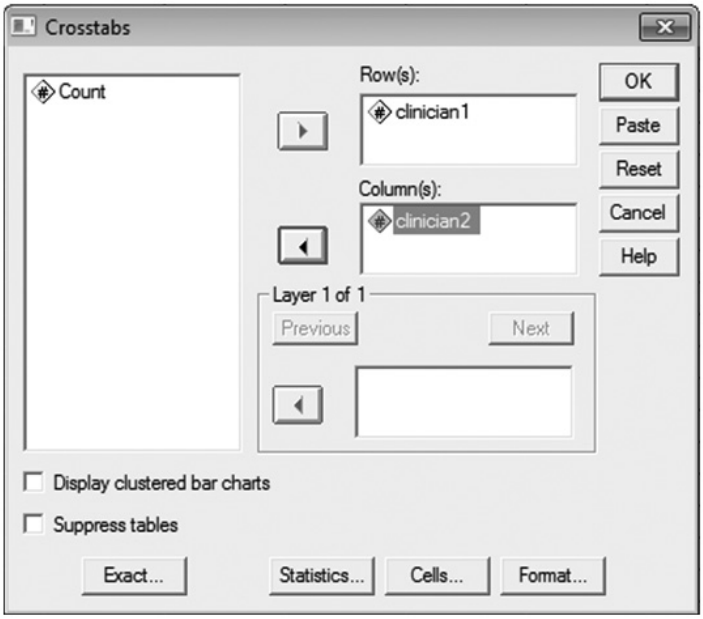
ภาพที่ 5 การจัดเตรียมข้อมูลของ Sim and Wright (2005)



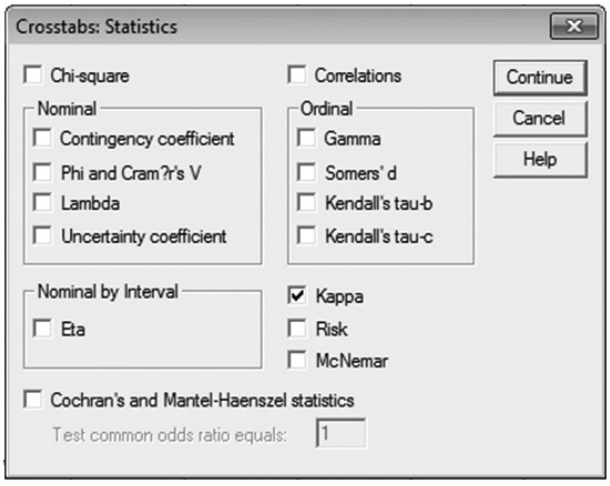
ภาพที่ 6 หน้าต่าง Weight cases



ภาพที่ 7 การสร้าง Crosstabs



ภาพที่ 8 หน้าต่าง Crosstabs



ภาพที่ 9 หน้าต่าง Crosstabs: Statistic

clinician1 * clinician2 Crosstabulation

Count

		clinician2			
		Derangement	Dysfunctional	Postural	Total
clinician1	Derangement	22	10	2	34
	Dysfunctional	6	27	11	44
	Postural	2	5	17	24
Total		30	42	30	102

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.461	.073	6.569	.000
N of Valid Cases		102			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

ภาพที่ 10 ผลลัพธ์ในการคำนวนสถิติ $\hat{K}_C$

จากภาพที่ 10 จะได้ สถิติ $\hat{K}_C$ = 0.461 จากข้อมูลขนาด 102 ตัวอย่าง มีค่า SE = 0.073 ดังนั้น ช่วงความเชื่อมั่น 95% ของ K_C คือ (0.388 , 0.534) นอกจากนี้ Approx. T = 6.569 โดยมีค่า Sig. = 0.000 โดยประมาณ ทำให้ปฏิเสธ $H_O: K_C = 0$ ที่ระดับนัยสำคัญ 0.05 กล่าวคือ เครื่องมือการวิจัยมีความสอดคล้องกันระหว่างผู้ประเมิน (หรือ $K_C > 0$)

อย่างไรก็ตาม แม้ผลการทดสอบความสอดคล้องกันระหว่างผู้ประเมินจะมีนัยทางสถิติก็ตาม ผู้วิจัยอาจจะใช้ผลการทดสอบสมมติฐานประกอบการตัดสินใจถึงความสอดคล้องได้อย่างคร่าวๆ เพราะการวิเคราะห์ IRR ในการวิจัยมักให้ความสำคัญของผู้ประเมิน โดยนิยมใช้เกณฑ์การพิจารณาความสอดคล้องของ Landis and Koch (1977) Fleiss Levin and Paik (1977) Krippendorff (1980) หรือ McHugh (2012) ดังที่ได้กล่าวไว้แล้ว ดังนั้นในกรณีนี้จะได้ว่าความเข้มของเกณฑ์มากที่สุด คือ เกณฑ์ของ Krippendorff (1980) (คือ พบว่ายังไม่มีความสอดคล้อง) รองลงมา คือ เกณฑ์ของ McHugh (2012) (คือ พบว่ามีความสอดคล้องน้อย) เกณฑ์ของ Landis and Koch (1977) (คือ พบว่ามีความสอดคล้องปานกลาง) และเกณฑ์ของ Fleiss Levin and Paik (1977) (คือ พบว่ามีความสอดคล้องดี) ตามลำดับ

9. การคำนวณค่าสถิติ $\hat{K}_C$ กรณีมีค่าสูญหาย

Gwet (2012) ได้แสดงวิธีการคำนวณ $\hat{K}_C$ ในกรณีที่มีค่าสูญหาย (Missing Data) 1 ค่า โดยสมมติว่ามีผู้ประเมิน 2 คน และมีประเภทที่ต้องประเมิน 3 ประเภท คือ ประเภท 1, 2 และ X และ n_{XX} เป็นข้อมูลสูญหายแสดงดังตารางที่ 9 โดยที่การคำนวณเป็นไปตาม (6) เพียงแต่ $P_a = \frac{n_{11}+n_{22}}{n-(n_{.X}+n_{X.})}$ ตัวอย่างข้อมูลการคำนวณจะใช้ข้อมูลในตารางที่ 10 จะได้ $P_a = \frac{34+44}{141-(20+11)}$ $= 0.709$, $P_e = \frac{61}{141}\frac{47}{141}+\frac{69}{141}\frac{74}{141} = 0.401$ และ $\hat{K}_C = \frac{0.709-0.401}{1-0.401} = 0.514$ จากตัวอย่างนี้สามารถนำวิธีการดังกล่าวไปประยุกต์ในการคำนวณ $\hat{K}_C$ ในกรณีจำนวนประเภทที่ต้องประมาณตั้งแต่ 3 ประเภทขึ้นไปและมีค่าสูญหายเพียง 1 ค่า

ตารางที่ 9 การแจกแจงความถี่ข้อมูลตัวอย่าง 3 ประเภทของผู้ประเมิน A และ B กรณีมีข้อมูลสูญหาย

ผู้ประเมิน A	ผู้ประเมิน B			รวม
	ประเภท 1	ประเภท 2	ประเภท X	
ประเภท 1	n_{11}	n_{12}	n_{1X}	$n_{1.}$
ประเภท 2	n_{21}	n_{22}	n_{2X}	$n_{2.}$
ประเภท X	n_{X1}	n_{X2}	n_{XX} *	$n_{X.}$
รวม	$n_{.1}$	$n_{.2}$	$n_{.X}$	n

* ข้อมูลสูญหาย

ตารางที่ 10 การแจกแจงความถี่ผู้ประเมิน A และ B กรณีมีข้อมูลสูญหาย

ผู้ประเมิน A	ผู้ประเมิน B			รวม
	ประเภท 1	ประเภท 2	ประเภท X	
ประเภท 1	34	22	5	61
ประเภท 2	10	44	15	69
ประเภท X	3	8	0*	11
รวม	47	74	20	141

* ข้อมูลสูญหาย

สำหรับกรณีที่มีข้อมูลสูญหายมากกว่า 1 หรือมีเงื่อนไขที่ซับซ้อนอันเกิดจากค่าสูญหายนั้น สามารถจัดการกับปัญหาตามแนวทางของ Adejumo (2005) ซึ่งมิได้นำเสนอปัญหาไว้หลายกรณีพร้อมตัวสถิติแคปปาที่มีการปรับให้สอดคล้องกับแต่ละสถาณการณ์ สามารถอ่านเพิ่มเติมได้จากงานของ Bechger Hemker and Maris (2001) Adejumo (2005) และ Kuchenhoff Augustin and Kunz (2011)

10. สถิติแคปปาถ่วงน้ำหนัก

สถิติแคปปาถ่วงน้ำหนัก (Weighted Kappa) แทนด้วย $\hat{K}_W$ เป็นสถิติที่พัฒนาต่อจาก $\hat{K}_C$ ใช้ในการวัดความสอดคล้องกันระหว่าง 2 ผู้ประเมินในกรณีดังนี้ คือ 1) ต้องการพิจารณาความสอดคล้องบางส่วน 2) ความสำคัญของความคลาดเคลื่อนมีความแตกต่างกัน และ 3) วิเคราะห์ความสอดคล้องของข้อมูลที่เป็นข้อมูลเชิงกลุ่มที่มีน้ำหนักความสำคัญต่างกันหรือเป็นข้อมูลระดับเรียงอันดับ (Jiang and Liu, 2011; Gisev Bell and Chen, 2013) เช่น ผู้ประเมินต้องประเมินข้อมูล 5 ประเภท คือ การต้องการความช่วยเหลือ 5 ระดับ ได้แก่ มากที่สุด มาก ปานกลาง น้อย และน้อยที่สุด (Viera and Garrett, 2005)

สมมติ ผู้ประเมิน A และ B มีวัตถุประสงค์ในการประเมินแบ่งออกเป็น q ประเภท (ดังตารางที่ 7) และมี

น้ำหนักของความไม่สอดคล้อง (Disagreement Weight) แทนด้วย W_{ij} ซึ่งมีความสัมพันธ์กับประเภทของ i และ j (i = 1, 2, ..., q; j = 1, 2, ..., q) และ $W_{ij} = 0$ เนื่องจากไม่มีความไม่สอดคล้องเพราะเป็นค่าความสอดคล้องที่เกิดจากผู้ประเมินทั้งสอง นอกจากนี้ ยังกำหนดให้ $W_{ij} > 0$ (สำหรับ $i \neq j$) ที่เกิดจากความไม่สอดคล้องของ 2 ผู้ประเมินในการพิจารณาประเภทที่ต่างกัน (Fleiss and Cohen, 1973) จากตารางที่ 7 และ W_{ij} สามารถการคำนวณ $\hat{K}_w$ ดัง (11) เมื่อ

$$P_a^* = \frac{1}{n}\sum_{k=1}^{q}\sum_{k=1}^{q} w_{ij}n_{ij} \text{ และ } P_e^* = \frac{1}{n}\sum_{k=1}^{q}\sum_{k=1}^{q} w_{ij}n_{k}.n_{.k}$$

(Jiang and Liu, 2011) เรียก P^*_a และ P^*_e ว่า ค่าเฉลี่ยจากการสังเกตของความไม่สอดคล้อง (Mean Observed of Disagreement) และค่าเฉลี่ยคาดหวังของความไม่สอดคล้อง (Mean Expected of Disagreement) ตามลำดับ นอกจากนี้ สถิติ $\hat{K}_C$ คือ สถิติแคปปาถ่วงน้ำหนักในกรณีที่ $W_{ij} = 1$ สำหรับ $i \neq j$

$$\hat{K}_W = \frac{P_a^* - P_e^*}{1 - P_e^*} \text{ --------(11)}$$

เนื่องจากโปรแกรมสำเร็จรูปไม่มีการคำนวณสถิติแคปปาถ่วงน้ำหนักโดยตรง ดังนั้น การคำนวณ $\hat{K}_w$ จะใช้โปรแกรม Excel แสดงวิธีการคำนวณข้อมูลของ Sim and Wright (2005) ในตารางที่ 8 แสดงดังรูปที่ 11 โดยสมมติ $w_{12} = 5$, $w_{13} = 3$, $w_{21} = 4$, $w_{23} = 1$, $w_{31} = 2$ และ $w_{32} = 2$ จะได้ค่าสถิติ $\hat{K}_w = 0.471$ รายละเอียดสูตรการคำนวณในช่อง (Cell) แสดงดังตาราง 11 ซึ่งเป็นการประยุกต์ตามวิธีการของ Zaiontz (2013)

	A	B	C	D	E	F	G	H	I	J	K
1							Observation				
2			Rater2						Weights		
3			Derangement	Dysfunctional	Postural	ราวม			Derangement	Dysfunctional	Postural
4	Rater1	Derangement	22	10	2	34		Derangement	0	5	3
5		Dysfunctional	6	27	11	44		Dysfunctional	4	0	1
6		Postural	2	5	17	24		Postural	2	2	0
7		ราวม	30	42	30	102					
8											
9							Expectation				
10			Rater2								
11			Derangement	Dysfunctional	Postural	ราวม	Weighted Kappa =	0.471			
12	Rater1	Derangement	10.00	14.00	10.00	34					
13		Dysfunctional	12.94	18.12	12.94	44					
14		Postural	7.06	9.88	7.06	24					
15		ราวม	30	42	30	102					

ภาพที่ 11 การคำนวณสถิติ $\hat{K}_w$

ตารางที่ 11 สูตรการคำนวณสถิติ $\hat{K}_w$ ตามภาพที่ 11

ช่อง	รายละเอียด	สูตรการคำนวณ
C12	$\frac{n_1.n_{.1}}{n} = \frac{34 \times 30}{102}$	=C7*F4/F7
H11	สถิติ $\hat{K}_W$	=1-SUMPRODUCT(C4:E6,I4:K6)/SUMPRODUCT(C12:E14,I4:K6)

11. สถิติแคปปาสำหรับกรณีมี 3 ผู้ประเมินขึ้นไป

Fleiss (1971) ได้พัฒนาสถิติแคปปาสำหรับการประเมินที่มีผู้ประเมินตั้งแต่ 3 คนขึ้นไป เรียกสถิติฟลีสแคปปา (Fleiss's Kappa Statistic) แทนด้วย $\hat{K}_F$ สามารถคำนวณตาม (12) เมื่อ $\bar{P}_a = \frac{1}{r}\sum_{i=1}^{r} z_j$, $\bar{P}_e = \sum_{k=1}^{q} p_j^2$ โดยที่ $z_j = \frac{1}{m(m-1)}\left(\sum_{k=1}^{q} n_{ij}^2 - \sum_{k=1}^{q} n_{ij}\right)$ และ $p_j = \frac{1}{rm}\sum_{i=1}^{r} n_{ij}$

กำหนดให้ r แทนจำนวนของวัตถุประสงค์ที่ประเมิน m แทนจำนวนผู้ประเมิน ($m \geq 3$) และ q แทนจำนวนประเภทที่ประเมิน

$$\hat{K}_F = \frac{\bar{P}_a - \bar{P}_e}{1 - \bar{P}_e} \text{ --------(12)}$$

ตัวอย่างการคำนวณสถิติ $\hat{K}_F$ ใช้โปรแกรม Excel ดังรูปที่ 12 โดยกำหนดผู้ประเมินจำนวน 20 คน มีวัตถุประสงค์ในการประเมิน 10 ข้อ และมีประเภทที่ต้องประเมิน 6 ประเภท จะได้ว่า $\hat{K}_F$ = 0.178 โดยที่รายละเอียดสูตรการคำนวณในช่องแสดงดังตารางที่ 12

	A	B	C	D	E	F	G	H	I	J	K
1	**Fleiss's Kappa**										
2		Categories									
3	Subjects	1	2	3	4	5	6	z_j		r =	10
4	1	0	0	0	0	18	2	0.811		m =	20
5	2	0	2	2	8	5	3	0.226		q =	6
6	3	0	0	6	8	6	0	0.305		$\bar{P}_a$ =	0.339
7	4	0	3	9	8	0	0	0.353		$\bar{P}_e$ =	0.195
8	5	2	2	1	8	2	5	0.216		Fleiss's Kappa =	0.178
9	6	7	7	0	0	5	1	0.274			
10	7	3	2	11	3	1	0	0.326			
11	8	2	5	5	2	6	0	0.195			
12	9	9	8	2	1	0	0	0.342			
13	10	0	1	2	8	9	0	0.342			
14	p_j	0.115	0.150	0.190	0.230	0.260	0.055				

ภาพที่ 12 การคำนวณสถิติ $\hat{K}_F$

ตารางที่ 12 สูตรการคำนวณสถิติ $\hat{K}_F$ ตามภาพที่ 12

ช่อง	รายละเอียด	สูตรการคำนวณ
H4	z_1	=(SUMSQ(B4:G4)-SUM(B4:G4))/(K4*(K4-1))
D14	p_3	=SUM(D4:D13)/(K3*K4)
K6	$\bar{P}_a$	=AVERAGE(H4:H13)
K7	$\bar{P}_e$	=SUMSQ(B14:G14)
K8	สถิติ $\hat{K}_F$	=(K6-K7)/(1-K7)

12. สรุป

12.1 IRR เป็นอีกแนวทางหนึ่งในการประเมินความเชื่อมั่นของเครื่องมือการวิจัย โดยเป็นการพิจารณาความเชื่อมั่นของความสอดคล้องภายในโดยใช้ผู้ประเมินเป็นหลัก วิธีการนี้จะแตกต่างจากวิธีการที่ใช้ค่าสัมประสิทธิ์อื่นๆ ในการคำนวณค่าความเชื่อมั่น เช่น วิธีการครอนบาค ค่า KR20 และค่า KR21 ค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย เป็นต้น ดังนั้น จึงเป็นเรื่องน่าสนใจและเป็นอีกหนึ่งทางเลือกสำหรับผู้วิจัยในการประเมินความเชื่อมั่นของเครื่องมือการวิจัย

12.2 การวัดความเชื่อมั่นของเครื่องมือการวิจัยโดยใช้ IRR ด้วยสถิติที่ใช้ในการพิจารณาความสอดคล้องสำหรับกรณีที่ข้อมูลระดับนามบัญญัติจะนิยมใช้สถิติแคปปา ส่วนกรณีที่เป็นข้อมูลระดับเรียงอันดับจะใช้สถิติแคปปาถ่วงน้ำหนักหรือสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intra-Class Correlation Coefficient: ICC)

12.3 ในกรณีที่มีผู้ประเมินตั้งแต่ 3 คน ต้องใช้สถิติฟลีสแคปปาในการคำนวณ โดยสามารถใช้โปรแกรม Excel ช่วยในการคำนวณค่า จะทำให้การคำนวณค่าง่ายขึ้น เช่นเดียวกับสถิติแคปปาถ่วงน้ำหนัก เพราะโปรแกรม Excel เป็นที่รู้จักกันดีในหมู่ผู้วิจัยทำให้การประยุกต์ใช้ทำได้ง่าย และทั้งสถิติแคปปาถ่วงน้ำหนักและสถิติฟลีสแคปปาไม่มีในโปรแกรมสำเร็จรูป แต่ถ้าเป็นกรณีสถิติแคปปานั้นสามารถใช้โปรแกรมสำเร็จรูป ในการคำนวณเพราะจะมีความสะดวกและคุ้นเคยกับผู้วิจัยส่วนใหญ่

12.4 นอกจากใช้ในการประเมินความเชื่อมั่นของแบบสอบถามแล้ว สถิติแคปปายังสามารถนำไปประยุกต์ใช้ในการประเมินความสอดคล้องในประเด็นต่างๆ เพื่อหาข้อสรุป เช่น ความเห็นพ้องของแพทย์ในการรักษาผู้ป่วย การพิจารณาคุณภาพของโรงแรม 5 ดาว โดยผู้เชี่ยวชาญ การพิจาณาคดีของคณะผู้พิพากษา การคัดเลือกนักเรียนที่มีคุณสมบัติในการตัวแทนของประเทศเพื่อเผยแพร่วัฒนธรรม เป็นต้น

12.5 สถิติแคปปายังคงมีการพัฒนาอย่างต่อเนื่อง เห็นได้จากงานของ Sinha Yimprayoon and Tiensuwan (2006) และพรพิศ ยิ้มประยูร (2555) ในการปรับปรุงสถิติแคปปา (Modified Kappa Statistics) เพื่อใช้ในสถาณการณ์ต่างๆ ได้อย่างเหมาะสม รวมถึงในงานของ Warrens (2013) ที่พัฒนาวิธีการประเมินความเชื่อมั่นในตารางที่ 3 ของสถิติแคปปาถ่วงน้ำหนัก

13. ข้อเสนอแนะ

ในการประเมินความเชื่อมั่นระหว่างผู้ประเมินหรือการประเมินความสอดคล้องระหว่างผู้ประเมิน มีข้อเสนอแนะสำหรับการเลือกใช้สถิติ ดังนี้

1. กรณีที่ข้อมูลระดับนามบัญญัติหรือเป็นข้อมูลเชิงกลุ่ม สถิติที่ใช้ในการวิเคราะห์ IRR คือ 1) สถิติแคปปา (สถิติ $\hat{K}_C$) ใช้ในการประเมินเครื่องมือวิจัยโดยมีผู้ประเมินเพียง 2 ผู้ประเมิน และประเมินได้ตั้งแต่ 2 ประเภท และ 2) สถิติฟลีสแคปปา (สถิติ $\hat{K}_F$) ใช้ในการวิเคราะห์ IRR สำหรับการประเมินเครื่องมือวิจัยที่มีผู้ประเมินตั้งแต่ 3 ผู้ประเมินขึ้นไป

2. สำหรับสถิติแคปปาถ่วงน้ำหนัก สามารถใช้วิเคราะห์ IRR ในกรณีที่ข้อมูลอยู่ในระดับอันดับ โดยเป็นการพิจารณาน้ำหนักของความไม่สอดคล้อง โดยที่การคำนวณสถิติแคปปาถ่วงน้ำหนักและสถิติฟลีสแคปปาสามารถใช้โปรแกรม Excel ของบทความในการประยุกต์เพื่อหาความเชื่อมั่นระหว่างผู้ประเมิน ส่วนสถิติแคปปาสามารถคำนวณพร้อมรายงานผลการทดสอบสมมติฐานได้ด้วยโปรแกรมสำเร็จรูป

3. ในกรณีที่ผู้วิจัยต้องการเครื่องมือวิจัยที่มีความเชื่อมั่นสูง และเลือกใช้การวิเคราะห์ IRR การพิจารณาความเชื่อมั่นระหว่างผู้ประเมินจะไม่นิยมทดสอบสมมติฐานทางสถิติ แต่จะใช้เกณฑ์การพิจารณาในการประเมินความเชื่อมั่นอาทิ เกณฑ์ของ Landis and Koch (1977) เกณฑ์ของ Krippendorff (1980) เกณฑ์ของ Fleiss Levin and Paik (2003) เกณฑ์ของ McHugh (2012) เป็นต้น ที่มีความเหมาะสมกับบริบทของการวิจัยนั้นๆ

14. เอกสารอ้างอิง

ภาษาไทย

ประสพชัย พสุนนท์. (2555). **การวิจัยการตลาด.** กรุงเทพฯ : บริษัท สำนักพิมพ์ท้อป จำกัด.

พรพิศ ยิ้มประยูร. (2555). **"บางลักษณะเชิงสถิติของการวัดความเห็นพ้องต้องกันระหว่างสองผู้ประเมิน"** วารสารวิจัย มสด สาขาวิทยาศาสตร์และเทคโนโลย. 5(2), 139 – 154.

พิษณุ ฟองศรี. (2552). **การสร้างและพัฒนาเครื่องมือวิจัย.** กรุงเทพฯ : บริษัท ด่านสุทธาการพิมพ์จำกัด.

ภาษาอังกฤษ

Adejumo, A. O. (2005). "Effect of Missing Values on the Cohen's Kappa Statistic for Raters Agreement Measurement". **International Journal**

of Pure and Applied Mathematics. 22(1), 13 – 31.

Bechger, T. M., Hemker, B. T., and Maris, G. K. J. (2001) **About the Cluster Kappa Coefficient.** Retrieved April 16, 2014, from http://webcache.googleusercontent.com/search?q=cache:2ragP4KJ5bQJ:www.cito.com/research_and_development/psychometrics/~/media/cito_com/research_and_development/publications/cito_report01_3.ashx+&cd=6&hl=th&ct=clnk&gl=th

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". **Educational and Psychological Measurement.** 20(1), 37-46.

Crutzen, R. (2014). "Time is a Jailer: What do Alpha and its Alternatives Tell Us About Reliability?" **The European Health Psychologist.** 16(2), 70 – 74.

Drost, E. (2011). "Validity and Reliability in Social Science Research". **Education Research and Perpectives.** 38(1), 105 – 123.

Fleiss, J. L. (1971). "Measuring nominal scale agreement among many raters." **Psychological Bulletin.** 76, 378 – 382.

Fleiss, J. L. and Cohen, J. (1973). "The Equivalence of Weighted Kappa and The Intraclass Correlation Coefficient as Measures of Reliability" **Education and Psychological Measurement.** 33, 613 – 619.

Fleiss, J. L., Cohen, J. and Everitt. B. S. (1969). "Large Sample Standard Errors for Kappa and Weighted Kappa". **Psychological Bulletin.** 72, 323–327.

Fleiss, J. L., Levin, B. and Paik, M. C. (2003). **Statistical Methods for Rates and Proportions.** Third Edition. New Jersey: John Wiley & Sons, Inc.

Gisev, N., Bell, J. S. and Chen, T. F. (2013). "Interrater Agreement and Interrater Reliability: Key Concepts, Approaches, and Applications". **Research in Social and Administrative Pharmacy.** 9, 330 – 338.

Gwet, K. L. (2012). **Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.** Gaithersburg, MD: Advanced Analytics, LLC.

Hallgren, K. A., (2012). "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial". **Tutorials in Quantitative Methods for Psychology.** 8(1), 23 – 34.

Jiang, S. and Liu, D. (2011). "On Chance-Adjusted Measures for Accuracy Assessment in Remote Sensing Image Classification" **ASPRS2011 Annual Conference.** Retrieved January 24, 2014, from http://www.asprs.org/a/publications/proceedings/Milwaukee2011/files/Jiang.pdf

Krippendorff, K. (1980). **Content Analysis: An Introduction to Its Methodology.** Beverly Hills, CA : Sage Publications, Inc.

Kuchenhoff, H., Augustin, T., and Kunz, A. (2011). **Partially Identified Prevalence Estimation Under Misclassification Using the Kappa Coefficient.** 7th International Symposium on Imprecise Probability: Theories and Applications, Innsbruck, Austria, 2011 Retrieved April 16, 2014, from http://www.sipta.org/isipta11/proceedings/papers/s031.pdf

Landis, J. R., and Koch, G. G. (1977). "The Measurement of Observer Agreement for Categorical Data". **Biometrics.** 33(1), 159-174.

McHugh, M. L. (2012). "Interrater Reliability: The Kappa Statistic". **Biochemia Medica.** 22(3), 276 - 282.

Novick, M. R. (1966). "The Axioms and Principle Results of Classical Test Theory". **Journal of Mathematical Psychology.** 3, 1-18.

Perreault, W. D. and Leigh, L. E. (1989). "Reliability of Nominal Data Based on Qualitative Judgments". **Journal of Marketing Research.** 26, 135 – 148.

Rosenthal, R. and Rosnow, R. L. (1991). **Essentials of Behavioral Research: Methods and Data Analysis.** New York : McGraw-Hill Publishing.

Sim, J. and Wright, C. C. (2005). "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements". **Physical Therapy.** 85, 257 – 268.

Sinha, B. K., Yimprayoon, P. and Tiensuwan, M. (2006). "Cohen's Kappa Statistic: A Critical Appraisal and Some Modifications" **Calcutta Statistical Association Bulletin.** 58, 151-169.

Viera, A. J. and Garrett, J. M. (2005). "Understanding Interobserver Agreement: The Kappa Statistic" **Research Series.** 37(5), 360 – 363.

Warrens, M. J. (2013). "Weighted Kappas for 3 Tables" **Journal of Probability and Statistics.** Retrieved January 24, 2014, from http://www.hindawi. com/journals/jps/2013/325831/

Zaiontz, C. (2013). **Real Statistics Using Excel.** Retrieved December 2, 2013, from http://www.real-statistics.com/