# Various Methods to Find Content Validity of a Quantitative Research Tool and How to Choose the Appropriate Ones

**Suphat Sukamolson**[1]

## Abstract

Content Validity is the most essential characteristic of a test and any other type of research tool (PTI, 2006). It is an index that a researcher is expected to report explicitly in his/her research report or article; otherwise, the work may seem questionable and untrustworthy. There are many methods to find such an index depending on the typical data characteristics, types of measurement scales, and number of experts in the fields (raters). This paper presents 15 methods for finding various kinds of the mentioned index and their advantages and disadvantages. Suggestions on using some appropriate methods were offered for researchers to use in their research studies.

**Keywords:** Content Validity, Content Validity Selection, Types of Content Validity, New Types of Content Validity

---

[1] Language Center, International College Maejo University, Chiang Mai

E-mail: ssuphatz@gmail.com

Received: 20 September 2023; Revised: 1 February 2024; Accepted: 20 April 2024

# นานาวิธีในการหาความตรงเชิงเนื้อหาของเครื่องมือวิจัยเชิงปริมาณและการเลือกใช้วิธีที่เหมาะสม

**สุพัฒน์ สุกมลสันต์**[2]

**บทคัดย่อ**

ความตรงเชิงเนื้อหาเป็นคุณลักษณะที่สำคัญที่สุดของบททดสอบและเครื่องมือวิจัยประเภทต่างๆ (PTI, 2006) เป็นดัชนีที่นักวิจัยได้รับการคาดหวังให้รายงานอย่างชัดเจนในรายงานการวิจัยหรือบทความของตน มิฉะนั้นงานอาจดูน่าสงสัยและไม่น่าไว้วางใจ มีหลายวิธีในการคำนวณหาดัชนีดังกล่าว ขึ้นอยู่กับลักษณะทั่วไปของข้อมูล ประเภทของมาตรวัด และจำนวนผู้เชี่ยวชาญในสาขา (ผู้ประเมิน) บทความนี้นำเสนอ 15 วิธีในการหาดัชนีประเภทต่างๆ ที่กล่าวถึง รวมถึงข้อดีและข้อเสีย และเสนอแนะแนวทางการใช้วิธีที่เหมาะสมให้นักวิจัยนำไปใช้ในการศึกษาวิจัย

**คำสำคัญ:** ความตรงเชิงเนื้อหา, การเลือกความตรงเชิงเนื้อหา, ประเภทของความตรงเชิงเนื้อหา, ความตรงเชิงเนื้อหาชนิดใหม่

---

[2] ศูนย์ภาษา วิทยาลัยนานาชาติ มหาวิทยาลัยแม่โจ้ เชียงใหม่

**ความนำความตรงเชิงเนื้อหาคืออะไร**

เป็นที่ทราบดีทั่วไปของนักทดสอบ และนักวิจัยว่าคุณลักษณะที่จำเป็นของเครื่องมือการวิจัยอื่นๆ ชนิดต่างๆ เช่น แบบทดสอบถาม แบบประเมิน แบบสัมภาษณ์ และแบบประเมินเชิงมิติสัมพันธ์ (Rubrics) ที่ดีมีอยู่หลายอย่าง แต่ที่มีความสำคัญมากมี 4 อย่างที่เครื่องมือเหล่านี้จำเป็นจะต้องมี คือ ความตรง (Validity) ความเที่ยง (Reliability) ความเป็นปรนัย (Objectivity) และความสะดวกในการใช้ (Usability) แต่ในจำนวนคุณลักษณะดังกล่าวแล้วนี้ ความตรงมีความสำคัญมากที่สุด (Choudhury, 2018) เพราะว่าหากเครื่องมือขาดคุณสมบัตินี้แล้วจะไม่สามารถกล่าวได้ว่า เครื่องมือนั้นนั้นสามารถวัดสิ่งที่ต้องการวัดหรือประเมินได้ในกรณีของการทดสอบ มีนักทดสอบที่มีชื่อเสียงหลายคนให้คำนิยามเรื่องความตรงไว้หลายอย่าง เช่น (อ้างถึงใน Cherry, 2017) “แบบทดสอบมีความตรงเมื่อสามารถวัดสิ่งที่แบบทดสอบนั้นกล่าวว่าสามารถวัดได้” (Garrett, 1964) “แบบทดสอบมีความตรงเมื่อสามารถวัดสิ่งที่แบบทดสอบนั้นควรจะวัดได้” (Ebel, 1972) “แบบทดสอบนั้นสามารถวัดสิ่งที่มุ่งทดสอบได้หรือไม่? ถ้าแบบทดสอบนั้นสามารถทำได้ แบบทดสอบนั้นมีความตรง” (Lado, 1975) “แบบทดสอบไม่สามารถเป็นแบบทดสอบที่ดีได้เว้นแต่ว่าแบบทดสอบนั้นจะมีความตรง สิ่งที่จำเป็นต้องมีของความตรงคือความถูกต้องซึ่งคะแนนชุดหนึ่งของแบบทดสอบนั้นวัดสิ่งที่อ้างว่าสามารถวัดได้” (Abbott & Perkins, 1982) และ “แบบทดสอบชุดหนึ่งเป็นทดสอบที่มีความตรง ถ้าสามารถวัดสิ่งที่มุ่งทำการทดสอบได้อย่างถูกต้อง” (Hughes, 1995) เป็นต้น

แต่อย่างไรก็ตาม นักทดสอบและนักวิจัยจำนวนมากยังมีความเห็นเรื่องความตรงเชิงเนื้อหาแตกต่างกันในรายละเอียดอยู่บ้าง เช่น Laerd (Laerd, 2018) มีความเห็นว่า ความตรงเชิงเนื้อหา คือส่วนต่างๆ ของแบบทดสอบที่มีส่วนเกี่ยวข้องกัน เป็นตัวแทนที่ดีและมีจำนวนเพียงพอของสิ่งที่ต้องการทดสอบ ส่วน Shuttleworth (2009) มีความเห็นว่า ความตรงเชิงเนื้อหาของเครื่องมือในการวัดชนิดหนึ่งประกอบด้วยสิ่งต่างๆ ที่เป็นตัวแทนที่ดีของสภาวะเชิงสันนิษฐาน (Construct) ซึ่งคือสิ่งต่างๆ ที่เครื่องมือนั้นมุ่งทำการวัดหรือประเมินมากน้อยเพียงใด ความตรงชนิดนี้บางครั้งเรียกว่าความตรงเชิงตรรกะ (Logical Validity) หรือความตรงเชิงเหตุผล (Rational Validity) ถ้าแบบทดสอบชุดหนึ่งมีความตรงเชิงเนื้อหาสูงแสดงว่าแบบทดสอบชุดนั้นประกอบด้วย 1) ข้อคำถามที่สามารถวัดเนื้อหาที่ต้องการวัดได้ 2) มีจำนวนข้อคำถามมากอย่างเพียงพอ 3) ข้อคำถามเป็นตัวแทนที่ดีของเนื้อหาที่ต้องการวัดและ 4) ข้อคำถามจะมีคุณลักษณะที่ดีดังนั้น ผู้สร้างแบบทดสอบควรต้องคำนึงถึง 4 ปัจจัยที่เกี่ยวข้อง คือ นิยามของเนื้อหาที่ต้องการทดสอบ (Domain Definition) ความเป็นตัวแทนที่ดีของข้อคำถาม (Domain Representation) ความเกี่ยวข้องของข้อคำถามกับเนื้อหาที่ต้องการทดสอบ (Domain Relevance) และคุณภาพที่ดีของข้อคำถามนั้น (Cronbach & Meehl, 1955; Haynes et al., 1995; Sireci, 1998)

นอกจากนี้บุคคลที่ให้ความหมายของความตรงเชิงเนื้อหาได้อย่างจัดเจนยิ่งขึ้นคือ A.R. Fitzpatrick (Fitzpatrick, 1983) เพราะว่าได้ทำการปริทัศน์แนวคิดของนักทดสอบจำนวนมากแล้วสรุปได้ว่า ความตรงเชิงเนื้อหาประกอบด้วย (1) ความชัดเจนของนิยามของเนื้อหาที่มุ่งทดสอบ (Clarity of domain definition) (2) ความเป็นตัวแทนของเนื้อหาของข้อคำถาม (Content representativeness) (3) ความเกี่ยวข้องกับเนื้อหา (Content relevance) และ (4) คุณภาพเชิงเทคนิคของข้อคำถาม (Technical quality in test items)

ตัวอย่าง เช่น ถ้าแบบทดสอบคัดเลือกครูวิทยาศาสตร์ประกอบด้วยข้อคำถามจำนวนมากทางด้านสาขาฟิสิกส์แล้วคัดเลือกผู้สอบที่ได้คะแนนสูงสุดเป็นครูวิทยาศาสตร์ แต่ว่าครูไม่มีความสามารถในการสอนวิทยาศาสตร์ดีเท่าที่ควรก็เพราะว่าแบบทดสอบชุดนั้นขาดความตรงเชิงเนื้อหาในการวัดความรู้และความสามารถทางวิทยาศาสตร์สาขาอื่น เช่น เคมี ชีวะ และวิทยาศาสตร์ทั่วไป เป็นต้น

อนึ่ง ความตรงเชิงเนื้อหามีธรรมชาติของข้อมูลเชิงคุณลักษณะ (qualitative in nature) ดังนั้นการตรวจสอบความตรงชนิดนี้ต้องอาศัยความคิดเห็นของผู้เชี่ยวชาญในเนื้อหาด้วยคำถามว่า ข้อคำถามแต่ละข้อจำเป็น (essential) มีประโยชน์ (useful) หรือไม่เกี่ยวข้อง (irrelevant) กับสภาวะเชิงสันนิษฐาน (Construct) ที่ต้องการทดสอบมากน้อยเพียงใด คำตอบของคำถามนี้ หากได้มาเป็นเพียงความคิดเห็นถือว่าเป็นความตรงที่เรียกว่า ความตรงเชิงประจักษ์ (Face Validity) แต่หากคำตอบได้มาเป็นค่าสถิติ หรือดัชนี จะเรียกว่าความตรงเชิงเนื้อหา (Shuttleworth, 2009)

ดังนั้น วิธีที่นักทดสอบทั่วไปใช้การหาค่าความตรงเชิงเนื้อหาของข้อคำถาม หรือเครื่องมือวิจัยอื่นๆ โดยการสอบถามความคิดเห็นของผู้เชี่ยวชาญในเนื้อหาว่าข้อคำถามหรือข้อคำถามของเครื่องมือมีความเกี่ยวข้องหรือสอดคล้องกับเนื้อหาที่นักทดสอบหรือผู้วิจัยต้องการทดสอบหรือสอบถามหรือไม่ และมากน้อยเพียงใด ด้วยวิธีการต่างๆ ซึ่งมีจำนวนมาก และค่าที่คำนวณได้ก็มีชื่อเรียกต่างกัน เช่น ค่าดัชนี (Index) และค่าสัมประสิทธิ์ (Coefficient) เป็นต้น แต่ว่าในการแปลความหมายนั้นนักทดสอบหรือผู้วิจัยควรพึงตระหนักว่าแบบทดสอบหรือเครื่องมือวิจัยใดก็ตามที่มีค่าดัชนี (Index) หรือค่าสัมประสิทธิ์ (Coefficient) ของความตรงเชิงเนื้อหาสูง ไม่ได้หมายความโดยอัตโนมัติว่าแบบทดสอบ หรือเครื่องมือวิจัยนั้นมีความตรงเชิงเนื้อหาสูง (จักรกฤษณ์ สำราญใจ, 2554; พิศิษฐ ตัณฑวณิช และพนา จินดาศรี, 2561) ทั้งนี้เพราะว่าความสอดคล้องของข้อคำถาม หรือข้อคำถามของเครื่องมือวิจัยกับเนื้อหาที่ต้องการทดสอบหรือสอบถามเป็นเพียงปัจจัย 1 ใน 4 ของความหมายของความตรงเชิงเนื้อหา ดังนั้น ในการแปลความหมายควรต้องนำปัจจัยที่สำคัญอีก 3 ด้านดังกล่าวแล้วมาประกอบการพิจารณาด้วย

**ความสำคัญของความตรงเชิงเนื้อหา**

จากที่กล่าวมาแล้วข้างต้นจะเห็นได้ว่าความตรงมีความสำคัญมากสำหรับแบบทดสอบ รวมทั้งเครื่องมืออื่นๆ ที่ใช้วัดหรือสอบถาม ทั้งนี้เนื่องจากวัตถุประสงค์แรกของเครื่องมือต่างๆ เหล่านี้ก็คือการวัดหรือทดสอบสิ่งที่มุ่งทดสอบเป็นประการแรก ดังนั้น คะแนนที่ได้จากแบบทดสอบที่ไม่มีความตรงจะเป็นคะแนนที่ไม่มีความหมายเลย นอกจากนี้ความตรงของแบบทดสอบถือว่ามีความสำคัญที่จะใช้เป็นหลักฐานทางกฎหมายได้เป็นอย่างดีหากว่ามีเรื่องการฟ้องร้องกันในศาลเกี่ยวกับผลการทดสอบว่า แบบทดสอบนั้นสามารถวัดสิ่งที่ต้องการทดสอบได้จริงมากน้อยเพียงใด (PTI., 2006) ดังนั้น เราสามารถกล่าวสรุปได้ว่า ความตรงมีความจำเป็นมากสำหรับแบบทดสอบเนื่องจากเป็นสิ่งที่ทำให้เกิดความแน่ใจได้ว่าแบบทดสอบสามารถทดสอบสิ่งที่มุ่งทดสอบได้ถูกต้อง และสามารถนำผลการทดสอบไปใช้ได้อย่างแม่นตรง รวมทั้งการตีความได้ถูกต้องด้วย ดังนั้นในบรรดาความตรงที่มีอยู่หลายชนิดความตรงเชิงเนื้อหานับว่ามีความสำคัญมากที่สุด (Disha, 2018).

อนึ่ง แบบทดสอบที่มีค่าความเที่ยง (Reliability) สูงอาจไม่ใช่แบบทดสอบที่ดีหากว่าไม่มีความตรง (Zamanzadeh et al., 2015) และถ้าแบบทดสอบไม่มีความตรงเชิงเนื้อหา แบบทดสอบนั้นก็ไม่สามารถทดสอบสิ่งที่แบบทดสอบนั้นมุ่งทำการทดสอบได้ (Kleeman, 2018)

**วิธีหาความตรงเชิงเนื้อหา**

เนื่องจากวิธีการหาความตรงเชิงเนื้อหาของข้อคำถามแบบทดสอบ หรือเครื่องมืออื่นทางด้านการวิจัยมีจำนวนมาก แต่ในที่นี้ผู้เขียนต้องการนำเสนอเฉพาะวิธีที่นักทดสอบ หรือนักวิจัยนิยมใช้กันมากเท่านั้น

1. วิธีหาค่าดัชนีความตรงรายข้อ (Item-Content Validity Index: I-CVI) โดยคิดจากอัตราส่วนของผู้เชี่ยวชาญในเนื้อหาเกี่ยวกับความเห็นว่าข้อคำถามสอดคล้องกับเนื้อหาที่ต้องการวัดมากน้อยเพียงใด วิธีหาค่าเฉลี่ยของความเห็นของผู้เชี่ยวชาญเรื่องความสอดคล้องของข้อคำถามกับเนื้อหาได้รับความนิยมติดต่อกันมาหลายทศวรรษแล้ว (Polit & Beck, 2006; Polit, Beck & Oven, 2007) และมีสูตรการคำนวณหลายสูตรที่อาศัยแนวคิดเดียวกัน เช่น

   โดยสูตรต่อไปนี้ (Polit & Beck, 2006)

$$I-CVI=\frac{n_{3.4}}{N_t}$$

   เมื่อ $I-CVI$ = ค่าดรรชนีความตรงรายข้อ

   $n_{3.4}$ = จำนวนผู้เชี่ยวชาญที่มีความเห็นระดับ 3 และ 4

   $N_t$ = จำนวนผู้เชี่ยวชาญทั้งหมด

   ตัวอย่างเช่น กำหนดให้ผู้เชี่ยวชาญแต่ละคนเลือกตอบได้อย่างใดอย่างหนึ่งจาก 4 ตัวเลือกว่าข้อคำถามมีความสอดคล้องกับเนื้อหาที่มุ่งทดสอบในระดับใด คือ 1 = ไม่สอดคล้อง (not relevant) 2 = สอดคล้องบ้างเล็กน้อย (somewhat relevant) 3 = สอดคล้องมาก (quite relevant) และ 4 = สอดคล้องมากอย่างยิ่ง (highly relevant) ถ้ามีผู้เชี่ยวชาญทั้งหมด 5 คน มีอยู่ 4 คน ที่มีความเห็นว่าข้อคำถามข้อหนึ่งสอดคล้องกับเนื้อหาที่มุ่งทดสอบในระดับ 3 หรือ 4 ดังนั้น ข้อคำถามนั้นมีค่า I-CVI = 0.80 จากสูตรดังกล่าว คือ

$$I-CVI=\frac{4}{5}=0.80$$

   เกณฑ์มาตรฐานของ I-CVI มีดังนี้ (Polit & Beck, 2006; Polit, Beck & Oven, 2007) คือ

   - ถ้ามีผู้เชี่ยวชาญ 5 คน หรือน้อยกว่าค่า I-CVI ควรเท่ากับ 1.00
   - ถ้ามีผู้เชี่ยวชาญ 6 คน ค่า I-CVI ไม่ควรน้อยกว่า 0.83 (Polit & Beck, 2006; Polit, Beck & Oven, 2007)
   - ถ้ามีผู้เชี่ยวชาญ 6-8 คน ค่า I-CVI ไม่ควรน้อยกว่า 0.83 (Lynn, 1986)
   - ถ้ามีผู้เชี่ยวชาญอย่างน้อย 9 คน ค่า I-CVI ไม่ควรน้อยกว่า 0.78 (Lynn, 1986)

อนึ่ง ค่า I-CVI ใช้สำหรับหาค่าความตรงเชิงเนื้อหาของแบบทดสอบแบบทดสอบถาม และแบบสำรวจต่างๆ เพื่อการวิจัย และเป็นวิธีที่ได้รับความนิยมมากในวงการพยาบาล และการแพทย์

2. วิธีหาค่าดรรชนีความตรงทั้งฉบับของแบบทดสอบ(Scale-Content Validity Index: S-CVI) โดยคิดจากอัตราส่วนของผู้เชี่ยวชาญในเนื้อหาว่าข้อคำถามที่มีสอดคล้องกับเนื้อหาของแบบทดสอบทั้งหมดมากน้อยเพียงใด โดยสูตรต่อไปนี้ (Polit & Beck, 2006; Polit, Beck & Oven, 2007)

$$S-CVI = \frac{\sum \frac{n_{3,4}}{N_t}}{k}$$

เมื่อ $S-CVI$ = ค่าดรรชนีความตรงทั้งฉบับของแบบทดสอบ (หรือเครื่องมืออื่น)

k = จำนวนข้อคำถามทั้งหมด

ตัวอย่าง เช่น กำหนดให้ผู้เชี่ยวชาญแต่ละคนเลือกตอบอย่างใดอย่างหนึ่งจาก 4 ตัวเลือกว่าข้อคำถามมีความสอดคล้องกับเนื้อหาที่มุ่งทดสอบในระดับใดเช่นเดียวกับการหาค่า I-CVI ดังกล่าวแล้วข้างต้น เช่น ถ้าแบบทดสอบมีทั้งหมด 10 ข้อมีผู้เชี่ยวชาญทั้งหมด 6 คนมีความเห็นว่าข้อคำถามข้อที่ 1-6 มีค่า I-CVI ข้อละ 0.83 และอีก 4 ข้อมีค่า I-CVI ข้อละ 1.00 ดังนั้น ค่า S-CVI = 0.90 จากสูตรดังกล่าว คือ

$$S-CVI = \frac{(0.83+0.83+0.83+0.83+0.83+0.83+1.0+1.0+1.0+1.0)}{10} = 0.90$$

เกณฑ์มาตรฐานของ S-CVI มีดังนี้ (Polit & Beck, 2006; Polit, Beck & Oven, 2007)

- ไม่ว่าจะมีผู้เชี่ยวชาญกี่คนก็ตาม ค่า S-CVI ไม่ควรน้อยกว่า 0.90
- ในระหว่างการพัฒนาแบบทดสอบ ค่า S-CVI ไม่ควรน้อยกว่า 0.80 (Davis, 1992 อ้างจาก (Polit & Beck, 2006; Polit, Beck & Oven, 2007)

อนึ่ง ค่า S-CVI ใช้สำหรับหาค่าความตรงเชิงเนื้อหาทั้งฉบับของแบบทดสอบแบบทดสอบถามและแบบสำรวจต่างๆ เพื่อการวิจัย และเป็นวิธีที่ได้รับความนิยมมากในวงการพยาบาล และการแพทย์

3. วิธีหาค่าดรรชนีความสอดคล้องระหว่างข้อคำถามหรือข้อคำถามกับวัตถุประสงค์เพียงอย่างเดียว หรือเอกมิติ (Unidimensional Item-Objective Congruence Index: U-IOC) โดยคิดจากอัตราส่วนของผู้เชี่ยวชาญในเนื้อหาว่ามีความเห็นเกี่ยวกับข้อคำถาม หรือข้อคำถามแต่ละข้อว่าสอดคล้องกับวัตถุประสงค์หรือเนื้อหาอย่างหนึ่งหรือไม่ด้วยสูตรที่คิดขึ้นเป็นครั้งแรกโดย Rovinelli (1976) ในรายงานวิทยานิพนธ์ปริญญาเอกของเขา และต่อมามีการนำเสนอเผยแพร่ทั่วไปโดย Rovinelli และ Hambleton (1976, 1977) สูตรนี้พัฒนามาจากสูตรคำนวณหาค่าดรรชนีของความเป็นเอกนัยของข้อคำถาม (Index of Homogeneity) ของ Hemphill และ Westie (1950 อ้างถึงใน Rovinelli & Hambleton, 1977) ต่อไปนี้ (Turner & Carlson, 2003; Turner et al., 2002) คือ

$$I_{ik} = \frac{(n-1)\sum_{j=1}^{n} X_{ijk} - \sum_{i=1}^{N}\sum_{j=1}^{n} X_{ijk} + \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

เมื่อ $I_{ik}$ = ค่าดรรชนีความสอดคล้องระหว่างข้อคำถามหรือข้อคำถามข้อที่ k กับวัตถุประสงค์ข้อที่ i

i = จำนวนวัตถุประสงค์ทั้งหมด (i = 1, 2, 3,..., N)

j = จำนวนผู้เชี่ยวชาญทั้งหมด (j = 1, 2, 3,..., n)

ตัวอย่าง เช่น กำหนดให้ผู้เชี่ยวชาญแต่ละคนเลือกตอบอย่างใดอย่างหนึ่งจาก 3 ตัวเลือกว่าข้อสอบมีความสอดคล้องกับเนื้อหาที่มุ่งทดสอบหรือไม่ คือ +1 = สอดคล้อง (relevant) -1 = ไม่สอดคล้อง (not relevant) 0 = ไม่แน่ใจ (questionable) ถ้ามีผู้เชี่ยวชาญทั้งหมด 4 คน มีอยู่ 3 คน ที่มีความเห็นว่าข้อสอบข้อหนึ่งทดสอบเนื้อหาที่มุ่งทดสอบตามวัตถุประสงค์ที่ 2 ข้อสอบนั้นมีค่า $I_{ik}$ = 0.97 จากสูตรดังกล่าวแล้วและมีข้อตกลงเบื้องต้นว่า (Rovinelli, 1976)

1. ถ้าผู้เชี่ยวชาญทุกคนมีความเห็นสอดคล้องกันว่าข้อคำถามข้อใดสามารถวัดสิ่งที่ต้องการจะวัดได้ ค่า $I_{ik}$ = +1 แต่ถ้าทุกคนมีความเห็นสอดคล้องกันว่าข้อคำถามข้อใดไม่สามารถวัดสิ่งที่ต้องการจะวัดได้ ค่า $I_{ik}$= -1
2. ความเห็นของผู้เชี่ยวชาญที่ไม่แน่ใจว่าข้อคำถามข้อใดไม่สามารถวัดสิ่งที่ต้องการจะวัดได้ คือ 0 (ไม่แน่ใจ :questionable) มีความสำคัญมากกว่าความคิดเห็นว่า ข้อคำถามนั้นไม่สามารถวัดสิ่งที่ต้องการจะวัดได้ คือ -1 (ไม่สอดคล้อง : not relevant)
3. ความเห็นของผู้เชี่ยวชาญที่มีค่าต่ำที่สุดคือ -1 (ไม่สอดคล้อง : not relevant)

เกณฑ์มาตรฐานของ $I_{ik}$ มีดังนี้ (Rovinelli & Hambleton, 1977; Hambleton, 1978 อ้างถึงใน Turner & Carlson, 2003)

- ในกรณีที่มีผู้เชี่ยวชาญ 5 คน ค่า $I_{ik}$ ไม่ควรน้อยกว่า 0.80
- ในกรณีทั่วไป ค่า $I_{ik}$ ไม่ควรน้อยกว่า 0.75

ตัวอย่างผลการคำนวณ

เช่น มีผู้เชี่ยวชาญ 3 คน ทำการประเมินข้อคำถาม 5 ข้อ และมีผลการประเมินตามตารางต่อไปนี้

ตารางที่ 1 ตัวอย่างผลการประเมินค่า IOC รายข้อตามสูตรของวิธีที่ 3

| | ความเห็นของผู้เชี่ยวชาญ | | | | |
|---|---|---|---|---|---|
| ผู้เชี่ยวชาญ | ข้อที่ 1 | ข้อที่ 2 | ข้อที่ 3 | ข้อที่ 4 | ข้อที่ 5 |
| คนที่ 1 | 1 | 0 | 1 | 1 | -1 |
| คนที่ 2 | 1 | -1 | 0 | 1 | -1 |
| คนที่ 3 | 1 | 1 | -1 | -1 | 0 |
| IOC | 1.0 | 0.50 | 0.50 | 0.666 | 0.166 |

ดังนั้น ค่า IOC ของข้อคำถามแต่ละข้อจะได้ตามตาราง และค่า IOC โดยเฉลี่ยคือ (1.0 + 0.50+ 0.50 + 0.666 + 0.166)/5 = 0.566 เป็นต้น

4. วิธีหาค่าดรรชนีความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์เนกมิติ (Multidimensional Item-Objective Congruence Index: M-IOC) โดยคิดจากอัตราส่วนของผู้เชี่ยวชาญในเนื้อหาว่ามีความเห็นเกี่ยวกับข้อคำถามหรือข้อสอบอิงเกณฑ์แต่ละข้อมีสอดคล้องกับวัตถุประสงค์หรือเนื้อหาหลายอย่างหรือไม่ โดยสูตรที่ Turner และ Carlson (2003) ซึ่งพัฒนามาจากสูตรของ Rovinelli และ Hambleton (1977) ที่กล่าวมาแล้วข้างต้นต่อไปนี้ (Turner et al., 2002) คือ

$$I_{ik} = \frac{(N + 2p - 2)\sum_{j=1}^{r} X_{ijk} - \sum_{i=1}^{N}\sum_{j=1}^{r} X_{ijk}}{2(N-1)r * p_k}$$

เมื่อ $I_{ik}$ = ค่าดรรชนีความสอดคล้องระหว่างข้อสอบข้อที่ k กับวัตถุประสงค์เนกมิติข้อที่ i

N = จำนวนวัตถุประสงค์ทั้งหมด (i = 1, 2, 3,..., N)

r = จำนวนผู้เชี่ยวชาญ (j = 1, 2, 3,..., r)

$p_k$ = จำนวนวัตถุประสงค์ที่ข้อสอบข้อที่ *k*ทดสอบสำหรับข้อคำถามข้อ k ตามวัตถุประสงค์ชุด i

ตัวอย่าง เช่น กำหนดให้ผู้เชี่ยวชาญแต่ละคนเลือกตอบอย่างใดอย่างหนึ่งจาก 3 ตัวเลือกว่าข้อสอบมีความสอดคล้องกับวัตถุประสงค์หรือเนื้อหาหลายอย่างที่มุ่งทดสอบหรือไม่ คือ +1 = สอดคล้อง (relevant) -1 = ไม่สอดคล้อง (not relevant) 0 = ไม่แน่ใจ (questionable) เช่น มีผู้เชี่ยวชาญทั้งหมด 3 คน (r = 3) ทำการประเมินข้อคำถาม3 ข้อ (k = 3) ว่าวัดตามวัตถุประสงค์ 5 ข้อ/ระดับของ CEFR (Common European Framework of Reference)คือ A1, A2, B1, B2และ C1 หรือไม่ และมากน้อยเพียงใด (N = 5) และถ้าผลการประเมินเป็นดังต่อไปนี้ คือ

ตารางที่ 2 ข้อมูลเพื่อหาค่า M-IOC ตามสูตรของวิธีที่ 4

| r | K | N1 | N2 | N3 | N4 | N5 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | -1 | 0 | -1 | -1 |
| 1 | 2 | -1 | 1 | 0 | -1 | -1 |
| 1 | 3 | 1 | -1 | 1 | -1 | -1 |
| 2 | 1 | -1 | 1 | 0 | -1 | -1 |
| 2 | 2 | -1 | 1 | 0 | -1 | -1 |
| 2 | 3 | -1 | 0 | 1 | -1 | -1 |
| 3 | 1 | 1 | -1 | -1 | 1 | 0 |
| 3 | 2 | -1 | 1 | 0 | -1 | 0 |
| 3 | 3 | -1 | 0 | 0 | -1 | -1 |

ผลการคำนวณจะพบว่า

$I_{ik}$ =0.791 สำหรับข้อคำถามข้อที่ 1 วัดตรงตามวัตถุประสงค์ข้อที่ 1 (A1)

$I_{ik}$ =0.883 สำหรับข้อคำถามข้อที่ 2 วัดตรงตามวัตถุประสงค์ข้อที่ 2 (A2)

$I_{ik}$ =0.666 สำหรับข้อคำถามข้อที่ 3 วัดตรงตามวัตถุประสงค์ข้อที่ 3 (B1)

และไม่มีข้อคำถามข้อใดวัดได้ตามวัตถุประสงค์ข้อที่ 4 และ 5 (B2 และ C1)

อนึ่ง สูตรดังกล่าวนี้สามารถใช้คำนวณหาค่าดรรชนีความสอดคล้องของข้อสอบอิงเกณฑ์กับวัตถุประสงค์เพียงอย่างเดียว (U-IOC) ตามข้อ 3 ดังกล่าวแล้วข้างต้นจะได้ผลลัพธ์เท่ากัน เมื่อ N = 1 และสามารถใช้เกณฑ์ของค่า $I_{ik}$ เช่นเดียวกันกับสูตรที่กล่าวแล้วข้างต้นได้

5. วิธีหาค่าดรรชนีความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์ (Item-Objective Consistency Index: IOC) โดยคิดจากอัตราส่วนของผู้เชี่ยวชาญในเนื้อหาว่ามีความเห็นว่าข้อสอบแต่ละข้อมีสอดคล้องกับวัตถุประสงค์หรือเนื้อหาที่ต้องการทดสอบหรือไม่ โดยสูตรที่ดัดแปลงมาจากสูตรที่ 3 (U-IOC) ต่อไปนี้ (Roid & Haladyna, 1982 อ้างถึงในจักรกฤษณ์ สำราญใจ, 2554) คือ

$$IOC = \frac{(N-1)S_1 - S_2 + S_1}{2(N-1)n}$$

เมื่อ IOC = ค่าดัชนีความคงที่ระหว่างข้อคำถามข้อที่ iกับวัตถุประสงค์ของเนื้อหา

N = จำนวนวัตถุประสงค์

n = จำนวนผู้เชี่ยวชาญ

$S_1$ = ผลรวมของคะแนนผลการตัดสินความสอดคล้องระหว่างข้อสอบข้อที่iกับวัตถุประสงค์ที่กำหนดให้ข้อสอบนั้นวัด ซึ่งคือ $\sum_{j=1}^{n} X_{ijk}$ ในสูตรข้อที่ 3 ข้างต้น

$S_2$= ผลรวมของคะแนนผลการตัดสินความสอดคล้องระหว่างข้อสอบข้อที่iกับวัตถุประสงค์ทุกข้อซึ่งคือ $\sum_{i=1}^{N}\sum_{j=1}^{n} X_{ijk}$ ในสูตรข้อที่ 3 ข้างต้น

ผลของการคำนวณของสูตรที่ 5 จะเท่ากับผลที่ได้จากสูตรที่ 3 พอดี และใช้สำหรับหาดรรชนีความความตรงเชิงเนื้อหาของแบบทดสอบอิงเกณฑ์เช่นเดียวกับสูตรที่ 3 รวมทั้งใช้เกณฑ์ในการแปลผลอย่างเดียวกัน

6. วิธีหาค่าดรรชนีความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์ (Item-Objective Congruence Index) โดยคิดจากอัตราส่วนของผู้เชี่ยวชาญในเนื้อหาว่ามีความเห็นว่าข้อสอบแต่ละข้อมีสอดคล้องกับวัตถุประสงค์หรือเนื้อหาหรือไม่ด้วยสูตรที่พัฒนาโดย Rovinelli and Hambleton มาจากสูตรที่ 3 (อ้างถึงในเยาวดี รางชัยกุล วิบูลย์ศรี, 2556)และผลการคำนวณจะได้เท่ากับสูตรที่ 3 พอดี (พิศิษฐ ตัณฑวณิช และ พนา จินดาศรี, 2561) และคล้ายคลึงกับสูตรที่ 5 มาก คือ

$$I_{io.} = \frac{(M-1)S_o - S_o^{'}}{2N(M-1)}$$

เมื่อ $I_{io}$ = ค่าดัชนีความสอดคล้องของข้อสอบที่ i กับวัตถุประสงค์ที่ O

M = จำนวนวัตถุประสงค์

N = จำนวนผู้เชี่ยวชาญตัดสิน

$S_o$= ผลรวมของคะแนนของผู้เชี่ยวชาญตัดสินข้อสอบที่ iกับวัตถุประสงค์ที่ O ที่มุ่งทดสอบ

$S_o^{'}$ = ผลรวมของคะแนนของผู้เชี่ยวชาญตัดสินข้อสอบที่ iกับวัตถุประสงค์อื่น ยกเว้นวัตถุประสงค์ที่ O

ผลของการคำนวณของสูตรที่ 6 จะเท่ากับผลที่ได้จากสูตรที่ 3 พอดีเช่นเดียวกัน (พิศิษฐ ตัณฑวณิช และ พนา จินดาศรี, 2561) และใช้สำหรับหาดรรชนีความความตรงเชิงเนื้อหาของแบบทดสอบอิงเกณฑ์เช่นเดียวกับสูตรที่ 3 รวมทั้งใช้เกณฑ์ในการแปลผลอย่างเดียวกันเนื่องจากสูตรพัฒนามาจากแนวคิดเดียวกัน

7. วิธีหาค่าดรรชนีความคงที่ระหว่างข้อสอบกับวัตถุประสงค์ (Item-Objective Congruence Index: IOC) โดยคิดจากอัตราส่วนของผู้เชี่ยวชาญในเนื้อหาเกี่ยวกับความเห็นว่าข้อสอบแต่ละข้อสอดคล้องกับเนื้อหาหรือไม่ ในกรณีที่มีวัตถุประสงค์หนึ่งข้อและกำหนดให้ข้อสอบวัดเพียงวัตถุประสงค์เดียวนั้น โดยสูตรต่อไปนี้ (ล้วน สายยศและอังคณา สายยศ, 2539 อ้างถึงใน จักรกฤษณ์ สำราญใจ, 2554) คือ

$$IOC = \frac{\sum R}{N}$$

เมื่อ IOC = ค่าดัชนีความสอดคล้องระหว่างข้อคำถามแต่ละข้อที่ i กับเนื้อหาหรือวัตถุประสงค์

R = ผลรวมของคะแนนความคิดเห็นของผู้เชี่ยวชาญ เมื่อกำหนดให้ 1 = สอดคล้อง -1 = ไม่สอดคล้อง และ 0 = ไม่แน่ใจ

N = จำนวนผู้เชี่ยวชาญในเนื้อหา

เกณฑ์มาตรฐานของค่า IOC ≥ 0.50

สูตรดังกล่าวนี้ได้รับการเผยแพร่อย่างกว้างขวางในวงการศึกษาในประเทศไทย และปรากฏอยู่ในหนังสือหรือตำราเกี่ยวกับการวิจัย หรือทดสอบและประเมินผลจำนวนมากแต่ไม่ปรากฏอยู่ในเอกสารหรือวรรณกรรมทางด้านการวัดและประเมินผลในต่างประเทศเลย จากการศึกษาของจักรกฤษณ์ สำราญใจ รวมทั้ง พิศิษฐ ตัณฑวณิชและ พนา จินดาศรี(จักรกฤษณ์ สำราญใจ, 2554; พิศิษฐ ตัณฑวณิช และ พนา จินดาศรี, 2561) พบว่าสูตรดังกล่าวนี้ล้วน สายยศและอังคณา สายยศ (2539) นำมาเผยแพร่ในประเทศไทยและอ้างว่าดัดแปลงมาจากสูตรของ Rovinelli และ Hambleton (1977) แต่นักวิชาการทั้ง 3 คนไม่สามารถหาหลักฐานยืนยันที่มาและแนวคิดของสูตรนี้ได้ เพราะว่าในเอกสารที่อ้างถึงไม่มีสูตรดังกล่าวนี้นอกจากสูตรในข้อ 3 ดังกล่าวแล้วข้างต้น และในเอกสารนั้นก็ไม่ได้ใช้คำว่า IOC ในสูตร นอกจากนี้สูตรของ Rovinelli และ Hambleton (1977) นั้นกำหนดให้ใช้เพื่อหาค่าความตรงของแบบทดสอบอิงเกณฑ์ (Criterion-Reference Test) แต่ว่าสูตรที่นำเสนอโดยล้วน สายยศ และอังคณา สายยศ (2539) มีผู้นำมาใช้กับแบบทดสอบทุกชนิด รวมทั้งแบบสอบถามด้วย ผู้เขียนเองได้ทดลองใช้สูตรดังกล่าวนี้กับสูตรในข้อที่ 3 ที่อ้างถึง ปรากฏว่าได้ผลแตกต่างกัน เนื่องจากสูตรในข้อ 7 นี้ คือค่าเฉลี่ยทางคณิตศาสตร์ของความคิดเห็นของผู้เชี่ยวชาญในเนื้อหาเท่านั้น (จักรกฤษณ์ สำราญใจ,

2554; พิศิษฐ ตัณฑวณิช และ พนา จินดาศรี, 2561) โดยให้ค่าน้ำหนักความสำคัญของค่า -1 = ไม่สอดคล้อง และ 0 = ไม่แน่ใจเท่ากัน แต่ว่าสูตรของ Rovinelli และ Hambleton (1977) ให้ค่าน้ำหนักดังกล่าวแตกต่างกันตามข้อตกลงเบื้องต้นที่ได้กล่าวถึงแล้ว ด้วยเหตุนี้ ผู้ที่ต้องการใช้สูตรตามหมายเลข 7 ไม่สมควรอย่างยิ่งที่จะอ้างอิงว่าเป็นสูตรของ Rovinelli และ Hambleton (1977) และควรระมัดระวังในการแปลผลจากการคำนวณด้วย (จักรกฤษณ์ สำราญใจ, 2554) หรือควรเลือกใช้สูตรอื่นที่เหมาะสมกว่าเนื่องจากที่มาของสูตรและแนวคิดในการใช้สูตรนี้ไม่ชัดเจน

นอกจากนี้ ค่าเฉลี่ยทางคณิตศาสตร์ของความคิดเห็นของผู้เชี่ยวชาญในเนื้อหา หรือร้อยละของความเห็นที่สอดคล้องกัน (Percentage of agreement/Percent agreement) มีจุดอ่อนที่สำคัญมาก คือ ไม่มีการแก้ไขค่าความผิดพลาดจากการให้ความเห็นที่สอดคล้องกันโดยบังเอิญของผู้เชี่ยวชาญ (Chance agreement) รวมทั้งไม่มีการคำนวณหาค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า (Standard Errorof Estimate) ของค่าที่คำนวณได้เหมือนกับวิธีอื่นๆ (Goodwin, 2009)

8. วิธีหาค่าอัตราส่วนของความตรงเชิงเนื้อหา (Content Validity Ratio: CVR) เป็นวิธีที่คิดขึ้นใช้โดย Lawshe (Lawshe, 1975) และได้รับความนิยมใช้กันอย่างแพร่หลายทางการแพทย์ การพยาบาล การศึกษา จิตวิทยา การพัฒนาองค์กร และการตลาด (Ayre & Scally, 2014) วิธีนี้อาศัยความคิดเห็นของผู้เชี่ยวชาญในเนื้อหาจำนวนหนึ่งในการพิจารณาตัดสินว่าข้อคำถาม หรือเครื่องมือวัดต่างๆ แต่ละข้อมีลักษณะดังต่อไปนี้หรือไม่ กล่าวคือ

    1. จำเป็นอย่างยิ่ง (Essential)
    2. มีประโยชน์แต่ไม่จำเป็นอย่างยิ่ง(Useful but not essential) หรือ
    3. ไม่จำเป็น (Not necessary)

    สูตรในการคำนวณคือ (Lawshe, 1975)

$$CRV = \frac{n_e - (N/2)}{N/2}$$

    เมื่อ CVR = ค่าอัตราส่วนของความตรงเชิงเนื้อหา

    $n_e$ =จำนวนผู้เชี่ยวชาญที่มีความเห็นว่าเครื่องมือข้อที่ เมีความจำเป็นอย่างอย่างยิ่ง

    N = จำนวนผู้เชี่ยวชาญทั้งหมด

    ต่อมา Ayre และ Scally (Ayre & Scally, 2014) ได้มีการปรับแก้สูตรดังกล่าวให้มีความถูกต้องยิ่งขึ้นสำหรับความคิดเห็นที่เป็นแบบ 2 อย่างคือ ใช่ หรือไม่ใช่ (binomial) และการกระจายของความคิดเห็น 2 อย่างที่เป็นโค้งปกติโดยประมาณ (normal approximation) รวมทั้งเพื่อใช้ทดสอบความมีนัยสำคัญของค่าที่คำนวณได้ด้วยสูตรต่อไปนี้ คือ

$$CRV_{Critical} = \frac{[(z\sqrt{N})+1]}{N}$$

    เมื่อ

$$z = \frac{(n_e - N_p - 0.5)}{\sqrt{[N_p(1-p)]}}$$

เกณฑ์ที่ได้จากสูตรเดิมและสูตรที่ปรับปรุงใหม่ของ CVR เป็นดังนี้ (Lawshe, 1975; Ayre & Scally, 2014)

- ควรมีผู้เชี่ยวชาญอย่างน้อย 5 คน
- ถ้ามีผู้เชี่ยวชาญจำนวน 5-7 คน ค่า CRV อย่างน้อยควรจะเท่ากับ 0.999 จึงจะแสดงว่าข้อคำถามมีความตรงเชิงเนื้อหาดีเพียงพอ
- ถ้ามีผู้เชี่ยวชาญจำนวน 8คน ค่า CRV อย่างน้อยควรจะเท่ากับ 0.750 หรือ 0.875
- ถ้ามีผู้เชี่ยวชาญจำนวน 9คน ค่า CRV อย่างน้อยควรจะเท่ากับ 0.778 หรือ 0.889
- ถ้ามีผู้เชี่ยวชาญจำนวน 10คน ค่า CRV อย่างน้อยควรจะเท่ากับ 0.800 หรือ 0.900
- ถ้ามีผู้เชี่ยวชาญจำนวน 11– 40คน ค่า CRVอย่างน้อยควรเท่ากับ 0.636 – 0.300 หรือ 0.818 – 0.650

อนึ่ง การแปลผลของค่า CVR มีความหมายดังนี้ (Lawshe, 1975)

1. ถ้าผู้เชี่ยวชาญมากกว่าครึ่งมีความเห็นว่าข้อคำถามนั้นมีความจำเป็นอย่างยิ่ง (essential) ต่อสิ่งที่วัด ค่า CVR จะเป็นบวก
2. ถ้าผู้เชี่ยวชาญน้อยกว่าครึ่งมีความเห็นว่าข้อคำถามนั้นมีความจำเป็นอย่างยิ่ง (essential) ต่อสิ่งที่วัด ค่า CVR จะเป็นลบ
3. ถ้าผู้เชี่ยวชาญครึ่งหนึ่งมีความเห็นว่าข้อคำถามนั้นมีความจำเป็นอย่างยิ่ง (essential) ต่อสิ่งที่วัด ค่า CVR จะศูนย์
4. ถ้าผู้เชี่ยวชาญทั้งหมดมีความเห็นว่าข้อคำถามนั้นมีความจำเป็นอย่างยิ่ง (essential) ต่อสิ่งที่วัด ค่า CVR = 1.00

อนึ่ง ค่า CVR ใช้สำหรับหาค่าความตรงเชิงเนื้อหาแต่ละข้อของแบบทดสอบแบบอิงเกณฑ์ เช่น แบบทดสอบคัดเลือกบุคคลเข้าทำงานในองค์กร (Personnel Test) ส่วนค่าความตรงของแบบทดสอบทั้งฉบับคือค่าเฉลี่ยของค่า CVR ของแบบทดสอบทุกๆ ข้อที่ได้รับการคัดเลือกว่าเป็นข้อคำถามที่มีค่าความตรงเชิงเนื้อหาตามเกณฑ์ดังกล่าวแล้ว (Lawshe, 1975)

9. วิธีหาค่าสัมประสิทธิ์ความเที่ยงระหว่างผู้ประเมินแบบ Scott's pi ( $\pi$ ) ใช้วัดความสอดคล้องของความเห็นของผู้เชี่ยวชาญในเนื้อหา 2 คนหรือมากกว่าต่อสิ่งที่วัดหรือประเมิน เช่น ข้อคำถามอย่างอื่นที่เป็นมาตรวัดนามบัญญัติ 2 ระดับหรือมากกว่าก็ได้ เป็นวิธีที่คิดขึ้นใช้โดย William A Scottในปี 1955 (Scott, 1955) เพื่อหาค่าความตรงเชิงเนื้อหาและค่าสัมประสิทธิ์ความเที่ยงระหว่างผู้ประเมินของแบบทดสอบ และเครื่องมือวิจัยอื่นๆ ผลการคำนวณของวิธี Scott's pi ( $\pi$ ) จะใกล้เคียงกับผลที่ได้จากวิธี Cohen's kappa ( $\kappa$ ) แต่จะไม่เกิดปัญหาเรื่องค่าที่ขัดแย้งกับความเป็นจริง (paradox) และใช้ได้ในหลายกรณีดังนี้

   1. สิ่งที่ผู้เชี่ยวชาญประเมินเป็นมาตรวัดนามบัญญัติ (nominal scale)
   2. ผู้เชี่ยวชาญในเนื้อหามีจำนวน 2 หรือมากกว่า

3. จำนวนค่าของมาตรวัด (n of categories) อาจมี 2 ระดับหรือมากกว่า
4. ระดับของมาตรวัดมีน้ำหนักแตกต่างกันได้

สูตรในการคำนวณ คือ (Girard, 2017)

$$\pi = \frac{p_o - p_c}{1 - p_c}$$

$$p_o = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^{q} \frac{r_{ik}(r_{ik}^{*} - 1)}{r_i(r_i - 1)}$$

$$r_{ik}^{*} = \sum_{l=1}^{q} w_{kl} r_{il}$$

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ik}}{r_i}$$

$$p_c = \sum_{k,l}^{q} w_{kl} \pi_k \pi_l$$

เมื่อ q = จำนวนระดับของมาตรวัด

n = จำนวนข้อคำถามหรือข้อคำถาม

$r_i$ = จำนวนผู้เชี่ยวชาญที่ประเมินข้อ i ว่าอยู่ในระดับใดก็ได้

$r_{ik}$ = จำนวนผู้เชี่ยวชาญที่ประเมินข้อ i ว่าอยู่ในระดับ k

$r_{il}$ = จำนวนผู้เชี่ยวชาญที่ประเมินข้อ i ว่าอยู่ในระดับ l

$n'$ = จำนวนข้อคำถามหรือข้อคำถามที่ได้รับการประเมินจากผู้เชี่ยวชาญ 2 คนหรือมากกว่า

$w_{kl}$ = ค่าน้ำหนักของข้อที่ผู้เชี่ยวชาญ 2 คนกำหนดให้ของมาตรวัดระดับ k และ l

10. วิธีหาค่าสัมประสิทธิ์ความสอดคล้องแบบ Cohen's kappa ( $\kappa$ ) เป็นวิธีที่คิดขึ้นใช้โดย Jacob Cohen (Cohen, 1960) เพื่อหาค่าความตรงเชิงเนื้อหาของแบบทดสอบ และเครื่องมือเพื่อการประเมินอย่างอื่นด้วย เช่นแบบสอบถามและแบบสัมภาษณ์ เป็นต้น นอกจากนี้ยังใช้เพื่อการหาความเที่ยงระหว่างผู้ประเมินได้ด้วย (Inter-rater Reliability) และได้รับความนิยมใช้กันอย่างแพร่หลายทางการแพทย์ การพยาบาล การศึกษา และจิตวิทยาการศึกษาโดยอาศัยความคิดเห็นของผู้เชี่ยวชาญในเนื้อหาเพียง 2 คน ในการพิจารณาตัดสินว่าข้อคำถาม หรือเครื่องมือวัดต่างๆ ว่ามีลักษณะความสอดคล้องกับสิ่งที่มุ่งทดสอบหรือไม่ และมากน้อยเพียงใด Cohen's kappa มีอยู่ 3 สูตร และใช้ในกรณีที่แตกต่างกัน คือ

ก. **Cohen's kappa ( $\kappa$ ) แบบทั่วไป** สำหรับใช้ในกรณีต่อไปนี้ คือ

1. ข้อมูลเป็นนามบัญญัติ (nominal scale)
2. มาตรวัดมี 2 ระดับ เช่น สอดคล้อง หรือไม่สอดคล้อง ใช่ หรือไม่ใช่ เป็นต้น
3. มาตรวัดแต่ละระดับมีความเป็นอิสระ และไม่ทับซ้อนกัน และ
4. ผู้เชี่ยวชาญในเนื้อหามี 2 คน และมีอิสระในการตัดสินเรื่องความสอดคล้องระหว่างข้อคำถามและสิ่งที่ต้องการวัด

สูตรในการคำนวณคือ (Tang et al., 2015)

$$\kappa = \frac{p_{11} + p_{00} - (p_{1+}p_{+1} + p_{0+}p_{+0})}{1 - (p_{1+}p_{+1} + p_{0+}p_{+0})}$$

เมื่อ k = ค่าสัมประสิทธิ์ความสอดคล้องแบบ Kappa

$p_{11}$ = อัตราส่วนของความเห็นของผู้เชี่ยวชาญที่มีความเห็นสอดคล้องกันทางบวก

$p_{00}$ =อัตราส่วนของความเห็นของผู้เชี่ยวชาญที่มีความเห็นสอดคล้องกันทางลบ

$p_{1+}$ =อัตราส่วนของความเห็นของผู้เชี่ยวชาญที่มีความเห็นในทางบวกในแนวตั้ง

$p_{+1}$ =อัตราส่วนของความเห็นของผู้เชี่ยวชาญที่มีความเห็นในทางบวกในแนวนอน

$p_{0+}$ =อัตราส่วนของความเห็นของผู้เชี่ยวชาญที่มีความเห็นในทางลบในแนวตั้ง

$p_{+0}$ = อัตราส่วนของความเห็นของผู้เชี่ยวชาญที่มีความเห็นในทางลบในแนวนอน

ตัวอย่าง ในการประเมินความตรงเชิงเนื้อหาของข้อสอบ 100 ข้อของผู้เชี่ยวชาญในเนื้อหาวิชา 2 คนมีผลการประเมินดังต่อไปนี้ คือ

ตารางที่ 3 ข้อมูลเพื่อหาค่า Cohen's kappa ( $\kappa$ ) แบบทั่วไป

| ความคิดเห็นของผู้เชี่ยวชาญ | | ผู้เชี่ยวชาญคนที่ 1 | | ความถี่รวมแนวตั้ง | รวม |
|---|---|---|---|---|---|
| | | สอดคล้อง | ไม่สอดคล้อง | | |
| ผู้เชี่ยวชาญคนที่ 2 | สอดคล้อง | 55 ข้อ | 5 ข้อ | 60 ข้อ | $n_{1+}$ |
| | ไม่สอดคล้อง | 7 ข้อ | 33 ข้อ | 40 ข้อ | $n_{0+}$ |
| จำนวนความถี่รวม | รวม | 62 ข้อ | 38 ข้อ | 100 ข้อ | |
| แนวนอน | | $n_{+1}$ | $n_{+0}$ | N | |

ดังนั้น

$$\kappa = \frac{\frac{55}{100} + \frac{33}{100} - (\frac{60}{100} * \frac{62}{100} + \frac{40}{100} * \frac{38}{100})}{1 - (\frac{60}{100} * \frac{62}{100} + \frac{40}{100} * \frac{38}{100})} = 0.748$$

ตารางที่ 4 เกณฑ์ในการแปลผลค่า Cohen's Kappa (McHugh, 2012)

| ขนาดของ Kappa ( $\kappa$ ) | ระดับความสอดคล้อง | ร้อยละของข้อมูลที่น่าเชื่อถือ |
|---|---|---|
| 0.0 – 0.20 | ไม่มีเลย (None) | 0 – 4% |
| 0.21 – 0.39 | น้อยมาก (Minimal) | 4 – 15% |
| 0.40 – 0.59 | น้อย (Weak) | 15 – 35% |
| 0.60 – 0.79 | ปานกลาง (Moderate) | 35 – 63% |
| 0.80 – 0.90 | มาก Strong | 64 – 81% |
| > 0.90 | เกือบสมบูรณ์ (Almost Perfect) | 82 – 100% |

สูตร Kappa Correlation สามารถเขียนได้หลายอย่าง เช่น ใช้อัตราส่วนของความคิดเห็น หรือค่าความถี่ของคะแนนก็ได้ ค่าสูงสุดของ Kappa = 1.00 แต่ค่าที่ต่ำที่สุดอาจน้อยกว่า 0.0 เช่น -1.00ก็ได้ แล้วแต่การกระจายของคะแนนความคิดเห็นของผู้เชี่ยวชาญนอกจากนี้ค่า Kappa จะเท่ากับค่า Scott's pi Correlation Coefficient($\phi$) พอดีถ้าการกระจายของความคิดเห็นของผู้เชี่ยวชาญไม่มีลักษณะเบ้มาก (Cohen, 1960) ค่า Kappa และ pi สามารถคำนวณได้ด้วยโปรแกรมสำเร็จรูป เช่น SPSS (Statistical Package for the Social Sciences) และมีค่าเท่ากับ Cramer's V ด้วย

อนึ่ง ค่า Kappa ใช้สำหรับหาค่าความตรงเชิงเนื้อหาแต่ละข้อของแบบทดสอบหรือเครื่องมือวัดอย่างอื่นทั้งฉบับที่ใช้ผู้เชี่ยวชาญเพียง 2 คนเท่านั้น ซึ่งอาจเรียกว่าค่าสัมประสิทธิ์ของความเที่ยงระหว่างผู้ประเมิน (Inter-rater Reliability Coefficient) และอาจใช้สำหรับผู้เชี่ยวชาญเพียงคนเดียวแต่ทำการประเมินข้อคำถาม หรือเครื่องมืออย่างอื่น 2 ครั้งซึ่งเรียกว่าค่าสัมประสิทธิ์ของความเที่ยงภายในของผู้ประเมินเอง (Intra-rater Reliability Coefficient)

เนื่องจาก Cohen's kappa เป็นวิธีที่นักวิจัยนิยมใช้กันมานานแล้ว ทำให้มีการค้นพบว่ามีจุดอ่อนที่สำคัญหลายอย่าง (Xie, 2013; Ato et al., 2011; Martín & Femia, 2004; Haley et al., 2008) เช่น

1. เกิดค่าขัดแย้งกับความเป็นจริง (paradox) เมื่อความเห็นของผู้เชี่ยวชาญรวมกันด้านใดด้านหนึ่งแตกต่างกันมากหรือเบ้มาก (very skewed) หรือโดดเด่นมากกว่าอีกด้านหนึ่ง (prevalence problem)เช่น ผู้เชี่ยวชาญเห็นสอดคล้องกันด้านบวกทั้งหมด หรือเกือบทั้งหมดจะทำให้ kappa มีค่าต่ำหรือเมื่อผู้เชี่ยวชาญคนหนึ่งมีความเห็นอย่างใดอย่างหนึ่งทั้งหมดทุกกรณี ค่า kappa = 0.0 หรือติดลบ ทำให้การแปลความหมายค่า kappa ยาก
2. เกิดอคติจากผู้เชี่ยวชาญ (rater bias problem)ในกรณีที่ผู้เชี่ยวชาญให้ความเห็นอย่างใดอย่างหนึ่งทุกกรณีต่อข้อสอบหรือข้อคำถามซึ่งจะทำให้ค่า kappa ต่ำมากหรือ = 0.0
3. จำนวนคะแนนรวมของความคิดเห็นด้านใดด้านหนึ่งที่แตกต่างกันมากทำให้ kappa มีค่าผิดจากความเป็นจริงมากกว่าค่าที่ได้การกระจายของความคิดเห็นที่เท่าๆ กัน
4. นอกจากนี้ Zhao (2011อ้างถึงใน Xie, 2013) ยังค้นพบว่า ค่า kappa ยังมีความขัดแย้งกับความเป็นจริง (paradox)อื่นๆ อีก 12 อย่างและมีความเห็นว่าเป็นวิธีที่นักวิจัยไม่ควรนำไปใช้ และ Xie (2013) เสนอแนะว่า ผู้วิจัยควรจะเลือกใช้วิธีอื่นที่ดีกว่า เช่น Gwet's AC1ซึ่งสอดคล้องกับผลการวิจัยของ Wongpakarn et al. (2013) และ Gwet (2002) ที่พบว่า Gwet's AC1 ดีกว่า Cohen's kappa เพราะว่าสามารถแก้ปัญหาค่าที่ขัดแย้งกับความเป็นจริงในข้อที่ 1 ที่สำคัญมากดังกล่าวแล้วข้างต้นได้ และค่าที่คำนวณได้มีความคงเส้นคงวามากกว่า

ข. **Cohen's kappa** ( $\kappa$ ) สำหรับใช้ในกรณีต่อไปนี้ คือ

1. ข้อมูลเป็นนามบัญญัติ (nominal scale)
2. มาตรวัดมีมากกว่า 2 ระดับ เช่น สอดคล้องหรือไม่สอดคล้อง แน่ใจ หรือไม่แน่ใจ เป็นต้น
3. มาตรวัดแต่ละระดับมีความเป็นอิสระ และไม่ทับซ้อนกัน และ

4. ผู้เชี่ยวชาญในเนื้อหามี 2 คน และมีอิสระในการตัดสินเรื่องความสอดคล้องระหว่างข้อคำถามและสิ่งที่ต้องการวัด

สูตรในการคำนวณคือ (Tang et al., 2015)

$$\kappa = \frac{\sum_{i=1}^{k} p_{ij} - \sum_{i=1}^{k} p_{i+} p_{+i}}{1 - \sum_{i=1}^{k} p_{i+} p_{+i}}$$

เมื่อ k = จำนวนระดับของมาตรวัด

ค. **Weighted Cohen's kappa** ($\kappa_w$) สำหรับใช้ดังนี้

1. ข้อมูลเป็นลำดับที่ (ordinal scale) เช่น มากที่สุด มาก ปานกลาง น้อย น้อยที่สุด เป็นต้น
2. ระดับต่างๆ ของข้อมูลมีน้ำหนักแตกต่างกัน
3. มาตรวัดแต่ละระดับมีความเป็นอิสระ และไม่ทับซ้อนกัน และ
4. ผู้เชี่ยวชาญในเนื้อหามี 2 คน และมีอิสระในการตัดสินเรื่องความสอดคล้องระหว่างข้อคำถามและสิ่งที่ต้องการวัด

สูตรในการคำนวณคือ (Tang et al., 2015)

$$\kappa_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{i+} p_{+i}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{i+} p_{+i}}$$

w = ค่าน้ำหนักของระดับมาตรวัดซึ่งมีวิธีคิดได้หลายอย่าง

$$\Delta = \frac{I_o - I_\pi}{1 - I_\pi}$$

$$I_\pi = \sum_{i=1}^{K} \prod_{r=1}^{R} \pi_{ir}$$

11. วิธีหาค่าสัมประสิทธิ์ความสอดคล้องแบบ Fleiss' kappa วิธีนี้ใช้สำหรับมาตรวัดนามบัญญัติหรือมาตรลำดับที่ที่ประเมินโดยผู้เชี่ยวชาญจำนวนมาก เป็นวิธีที่คิดขึ้นใช้โดย Joseph L. Fleiss (Fleiss, 1971) เพื่อหาค่าความตรงเชิงเนื้อหาของแบบทดสอบ และเครื่องมือเพื่อการประเมินอย่างอื่นด้วย เช่น แบบสอบถามและแบบสัมภาษณ์ เป็นต้น และเพื่อการหาความเที่ยงระหว่างผู้ประเมินหลายคนได้ด้วย (Interrater Reliability) โดยผู้พัฒนาสูตรมุ่งแก้ปัญหาที่เกิดขึ้นกับค่า Cohen's kappa เมื่อการกระจายของความคิดเห็นของผู้เชี่ยวชาญมีลักษณะเบ้มาก และใช้ได้ในกรณีดังต่อไปนี้ (Laerd Statistics, 2019; Zaiontz, 2012, 2019) คือ

1. สิ่งที่ผู้เชี่ยวชาญประเมินเป็นมาตรนามบัญญัติ หรือมาตรลำดับที่ที่มี 2 อย่างหรือมากกว่า
2. ผู้เชี่ยวชาญในเนื้อหามีจำนวน 2 คน หรือมากกว่า

3. จำนวนระดับของความสอดคล้องของข้อสอบหรือข้อคำถามแต่ละข้อต้องเท่ากัน และสื่อความหมายเหมือนกัน
4. ผู้เชี่ยวชาญแต่ละคนไม่จำเป็นต้องประเมินข้อสอบหรือข้อคำถามครบทุกข้อก็ได้
5. ข้อสอบหรือข้อคำถามได้มาจากการสุ่มเลือก
6. ผู้เชี่ยวชาญในเนื้อหาได้มาจากการสุ่มเลือก และมีอิสระในการตัดสินเรื่องความสอดคล้องระหว่างข้อคำถามและสิ่งที่ต้องการวัดหรือประเมิน

สูตรในการคำนวณคือ (Zaiontz, 2012, 2019)

$$k = \frac{p_a - p_s}{1 - p_s}$$

เมื่อ k = ค่าสัมประสิทธิ์ความสอดคล้องแบบ Fleiss' kappa

$p_a$ = อัตราส่วนของผู้เชี่ยวชาญที่มีความเห็นสอดคล้องกันหลายลักษณะที่ปรากฏ จากสูตร

$$p_a = \frac{1}{mn(m-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2 - mn\right]$$

$p_s$ = อัตราส่วนของผู้เชี่ยวชาญที่มีความเห็นสอดคล้องกันหลายลักษณะที่คาดไว้ (โดยบังเอิญ) จาก

$$p_s = \sum_{j=1}^{k} q_j^2 \text{ และ } q_j = \frac{1}{nm}\sum_{i=1}^{n} x_{ij}$$

m = จำนวนผู้เชี่ยวชาญในเนื้อหา

n = จำนวนข้อสอบ หรือข้อคำถามในเครื่องมือวัดหรือประเมินผล

k = จำนวนระดับของมาตรวัด

ตัวอย่าง เช่น มีผู้เชี่ยวชาญในเนื้อหาภาษาอังกฤษจำนวน 6 คน ทำการประเมินข้อคำถามจำนวน 20 ข้อ เพื่อตัดสินว่าข้อคำถามเหล่านี้วัดความสามารถทางภาษาอังกฤษตามเกณฑ์ CEFR (Common European Framework of Reference) ที่มีอยู่ 6 ระดับสอดคล้องกันมากน้อยเพียงใด

อนึ่ง สำหรับเกณฑ์การแปลผลของ Fleiss' kappa มีความคล้ายคลึงกับเกณฑ์ของ Cohen's kappa มาก กล่าวคือ

ตารางที่ 5 เกณฑ์ในการแปลผลค่า Fleiss' kappa (Hartling et al., 2012)

| ขนาดของ Fleiss' kappa | ระดับความสอดคล้อง | ร้อยละของข้อมูลที่น่าเชื่อถือ |
|---|---|---|
| <0.0 | ไม่มีเลย (Poor) | 0% |
| 0.00 – 0.20 | มีน้อยมาก (Slight) | 0.0 – 4% |
| 0.21 – 0.40 | น้อย (Fair) | 4 – 16% |
| 0.41 – 0.60 | ปานกลาง (Moderate) | 16 – 36% |
| 0.61 – 0.80 | มาก (Substantial) | 36 – 64% |
| > 0.81 | เกือบสมบูรณ์ (Almost Perfect) | 64 – 100% |

12. วิธีหาค่าสัมประสิทธิ์ความเที่ยงระหว่างผู้ประเมินแบบ Gwet's gamma Coefficient ($\gamma$) ซึ่งมีอยู่ 2 วิธีย่อย คือ Gwet's AC1 และ Gwet's AC2 ซึ่งนำเสนอโดย K.L. Gwet ในปี 2008 (Gwet, 2008) เพื่อแก้ปัญหาของค่าที่ขัดแย้งกับความเป็นจริงของวิธี Cohen's kappa โดยวิธีแรกใช้เพื่อหาความสอดคล้องของผู้เชี่ยวชาญ 2 คนต่อความเห็นของข้อสอบหรือข้อคำถามของเครื่องมือวัดอย่างอื่นที่เป็นมาตรวัดนามบัญญัติที่ผู้เชี่ยวชาญทั้งสองคนจะต้องประเมินครบทุกข้อ ส่วนวิธีที่ 2 ใช้ได้กว้างขวางมากกว่าวิธีแรก กล่าวคือ

1. สำหรับผู้เชี่ยวชาญในเนื้อหาจำนวน 2 คน หรือมากกว่า
2. สำหรับมาตรวัดนามบัญญัติ มาตรวัดลำดับที่ มาตรวัดอันตรภาค หรือมาตรวัดอัตราส่วนก็ได้
3. ผู้เชี่ยวชาญไม่จำเป็นต้องทำการประเมินข้อสอบหรือข้อคำถามครบทุกข้อ

สูตรของ Gwet's AC1 คือ (Girard, 2017)

$$\text{Gwet's } \gamma = \frac{p_o - p_c}{1 - p_c}$$

เมื่อ

$$p_o = \frac{n_{11} + n_{22}}{n}$$

$$p_c = (\frac{m_1}{n} * \frac{m_2}{n}) + (\frac{m_1}{n} * \frac{m_2}{n})$$

$$m_1 = \frac{n_{+1} + n_{1+}}{2}$$

$$m_1 = \frac{n_{+2} + n_{2+}}{2}$$

$n_{11}$= จำนวนข้อคำถามที่ผู้เชี่ยวชาญทั้ง 2มีความเห็นตรงกันว่าสอดคล้องกับเนื้อหา

$n_{22}$= จำนวนข้อคำถามที่ผู้เชี่ยวชาญทั้ง 2 มีความเห็นตรงกันว่าไม่สอดคล้องกับเนื้อหา

n = จำนวนข้อคำถามทั้งหมด

$n_{1+}$= จำนวนข้อคำถามที่ผู้เชี่ยวชาญคนที่ 1 มีความเห็นว่าสอดคล้องกับเนื้อหา

$n_{2+}$= จำนวนข้อคำถามที่ผู้เชี่ยวชาญคนที่ 1 มีความเห็นว่าไม่สอดคล้องกับเนื้อหา

$n_{+1}$= จำนวนข้อคำถามที่ผู้เชี่ยวชาญคนที่ 2 มีความเห็นว่าสอดคล้องกับเนื้อหา

$n_{+2}$= จำนวนข้อคำถามที่ผู้เชี่ยวชาญคนที่ 2 มีความเห็นว่าไม่สอดคล้องกับเนื้อหา

สูตรของ Gwet's AC2คือ (Girard, 2017)

$$\text{Gwet's } \gamma = \frac{p_o - p_c}{1 - p_c}$$

เมื่อ

$$p_o = \frac{1}{n'}\sum_{i=1}^{n'}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}^{*}-1)}{r_i(r_i-1)}$$

$$r_{ik}^{*} = \sum_{l=1}^{q} w_{kl} r_{il}$$

$$p_c = \frac{T_w}{q(q-1)}\sum_{k,l}\pi_k(1-\pi_k)$$

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r_i}$$

$$T_w = \sum_{k,l} w_{kl}$$

เนื่องจาก Gwet's gamma Coefficient ($\gamma$ เป็นวิธีหาค่าสัมประสิทธิ์ความสอดคล้องระหว่างผู้ประเมินที่เกิดขึ้นใหม่มาก และสามารถใช้แก้ปัญหาเรื่องความขัดแย้งระหว่างค่าที่ได้จากการคำนวณได้กับความเป็นจริงที่เกิดจากความเห็นของผู้เชี่ยวชาญรวมกันด้านใดด้านหนึ่งแตกต่างจากอีกด้านหนึ่งมาก (prevalence problem) และค่าที่คำนวณได้มีความคงเส้นคงวา รวมทั้งใช้ได้หลายกรณี จึงเป็นวิธีที่ปัจจุบันนี้กำลังได้รับความนิยม นอกจากนี้ ปัจจุบันนี้มีงานวิจัยจำนวนมากที่ทำการศึกษาเปรียบเทียบค่าสัมประสิทธิ์ที่ได้จากวิธี Gwet's gamma Coefficient ($\gamma$) กับวิธี Cohen's kappa ซึ่งได้รับความนิยมมานาน ปรากฏว่าวิธี Gwet's gammaให้ค่าที่ดีกว่าวิธี Cohen's kappa เพราะสามารถแก้ปัญหาดังที่ได้กล่าวมาแล้วได้ (Haley et al., 2008; Wongpakaran et al., 2013)

13. วิธีหาค่าสัมประสิทธิ์ความเที่ยงระหว่างผู้ประเมินแบบ Krippendorff's alpha ใช้วัดความสอดคล้องของความเห็นของผู้เชี่ยวชาญในเนื้อหาจำนวนมากต่อสิ่งที่วัดหรือประเมินซึ่งอาจเป็นมาตรวัดได้หลายชนิด

วิธีนี้คิดขึ้นใช้โดย Klaus Krippendorff เมื่อไม่นานมานี้ (Krippendorff, 2011) เพื่อหาค่าความตรงเชิงเนื้อหาของแบบทดสอบ และเครื่องมือวิจัยอื่นๆ วิธีนี้พัฒนาขึ้นเพื่อมุ่งแก้ปัญหาของค่า Cohen's kappa เมื่อการกระจายของความคิดเห็นของผู้เชี่ยวชาญมีลักษณะเบ้มากเช่นเดียวกับวิธีที่ 11 และ 12 และใช้ในกรณีต่อไปนี้

1. สิ่งที่ผู้เชี่ยวชาญประเมินเป็นมาตรวัดชนิดใดก็ได้ เช่น นามบัญญัติ (nominal scale) ลำดับที่ (ordinal scale) อันตรภาค (interval scale) หรืออัตราส่วน (ratio scale)
2. ผู้เชี่ยวชาญในเนื้อหามีจำนวน 2 คน หรือมากกว่า 2 คนขึ้นไป
3. ผู้เชี่ยวชาญแต่ละคนไม่จำเป็นต้องประเมินข้อสอบหรือข้อคำถามทุกข้อก็ได้
4. จำนวนค่าของมาตรวัด (n of categories) ค่าของมาตรวัด (scale values) และจำนวนมาตรวัด (measures) มีเท่าใดก็ได้
5. ไม่มีการกำหนดขนาดกลุ่มตัวอย่างที่เล็กที่สุด

อนึ่ง เนื่องจากวิธีนี้สามารถใช้ได้กับมาตรวัดหลายชนิดในกรณีเฉพาะที่แตกต่างกันหลายอย่าง เพื่อคำนวณหาค่าความเที่ยงระหว่างผู้เชี่ยวชาญ (inter-rater reliability) สำหรับใช้วัดความตรงเชิง

เนื้อหาของแบบสอบ หรือเครื่องมือเพื่อการวิจัยอื่นๆ เช่น ข้อมูลทวิภาค (binary or dichotomous data) มีผู้เชี่ยวชาญ 2 คน ทำการประเมินข้อสอบทุกข้อ หรือมาตรวัดนามบัญญัติ มีผู้เชี่ยวชาญ 2 คน ทำการประเมินข้อสอบบางข้อ หรือมาตรวัดนามบัญญัติ มีผู้เชี่ยวชาญจำนวนมาก ทำการประเมินข้อสอบบางข้อ เป็นต้น จึงทำให้มีหลายสูตรที่มีลักษณะซับซ้อนมาก (Osborne, 2008) แต่มีสูตรพื้นฐานดังนี้ (Krippendorff, 2011)

$$\alpha = 1 - \frac{D_0}{D_e}$$

เมื่อ $\alpha$ = ค่าสัมประสิทธิ์ความเที่ยงระหว่างผู้ประเมินแบบ Krippendorff's alpha

$D_0$ = คะแนนความไม่สอดคล้องของความเห็นของผู้เชี่ยวชาญที่ปรากฏ คือ

$$D_0 = \frac{1}{n}\sum_c \sum_k o_{ck\ metric}\delta^2_{ck}$$

$D_e$ = คะแนนความไม่สอดคล้องของความเห็นของผู้เชี่ยวชาญที่คาดหวัง คือ

$$D_e = \frac{1}{n(n-1)}\sum_c \sum_k n_c * n_{k\ metric}\delta^2_{ck}$$

และ n, $n_c$, $n_k$และ $\delta_{ck}$ คือความถี่ของค่าต่างๆ ที่เกิดจากการประเมินของผู้เชี่ยวชาญที่สอดคล้องกันในกรณีต่างๆ

ค่า Krippendorff's alpha สูงสุดคือ 1.00 เมื่อผู้เชี่ยวชาญทุกคนมีความคิดเห็นสอดคล้องกัน และมีค่า -1.00 เมื่อมีความคิดเห็นขัดแย้งกันมากแต่ว่าเกณฑ์ทั่วไปของวิธีนี้ เหมือนกับเกณฑ์ของ Cohen's kappa และ Scott's pi (Krippendorff, 2004 อ้างถึงใน Hollingshead et al., 2012) คือ

<0.67 = สอดคล้องกันต่ำและยอมรับไม่ได้

0.67-0.80 = สอดคล้องกันปานกลางและพอยอมรับได้

>0.80 = สอดคล้องกันสูง และยอมรับได้

14. วิธีหาค่าสัมประสิทธิ์ความสอดคล้องแบบ Martin & Femia's delta (Δ) เป็นวิธีการคำนวณหาค่าความสอดคล้องของผู้เชี่ยวชาญต่อข้อคำถามแบบเลือกตอบ (Multiple-choice Test Item) หรือข้อคำถามของเครื่องมือวิจัยใหม่ล่าสุด เพราะว่านำเสนอแนวคิดโดย A. A. Martín และM. P. Femiaเมื่อปี 2004 (Martín & Femia, 2004) เพื่อที่จะแก้ไขปัญหาต่างๆ ที่เกิดจากการใช้ Cohen's kappa ( $\kappa$ ) สูตรที่ 1 เมื่อมีผู้เชี่ยวชาญในเนื้อหาเพียง 2 คนเพื่อประเมินข้อคำถามหรือข้อคำถามที่เป็นมาตรวัดแบบนามบัญญัติเพียง 2 ระดับเท่านั้น โดยเฉพาะอย่างยิ่งเมื่อการกระจายของคะแนนรวมของความเห็นของผู้เชี่ยวชาญมีความแตกต่างกันมาก (prevalence) แล้วทำให้เกิดปัญหาเรื่องค่าที่ขัดแย้งกับความเป็นจริง (paradox) ดังได้กล่าวมาแล้ว ผลการคำนวณของวิธี Martin และ Femia's delta (Δ) จะใกล้เคียงกับผลที่ได้จากวิธี Cohen's kappa ( $\kappa$ ) แต่จะไม่เกิดปัญหาเรื่องค่าที่ขัดแย้งกับความเป็นจริง (paradox) วิธีนี้ใช้ในกรณีต่อไปนี้

1. ใช้ผู้เชี่ยวชาญในการประเมินความสอดคล้อง 2 คน
2. มาตรเป็นแบบนามบัญญัติที่มี 2 ระดับหรือมากกว่าก็ได้
3. เป็นวิธีที่คิดขึ้นเพื่อใช้ประเมินความสอดคล้องของข้อคำถามแบบเลือกตอบโดยเฉพาะ แต่ก็สามารถใช้เพื่อการประเมินเครื่องมืออย่างอื่นที่มีลักษณะที่คล้ายกันได้
4. ค่าที่คำนวณได้เป็นค่าโดยประมาณ (estimator)

สูตรในการคำนวณเพื่อหาค่า Martin และ Femia's delta (Δ) มีลักษณะซับซ้อน ใช้สถิติชั้นสูงและมีหลายรูปแบบขึ้นอยู่กับจำนวนระดับของมาตรวัด ถ้าคำนวณด้วยมือต้องอาศัยเครื่องคิดเลขชนิดที่สามารถเขียนโปรแกรมได้ (Martin & Femia, 2004) ต่อมาในปี 2019 ได้มีการพัฒนา Martin และÁlvarez's delta (Δ) ขึ้นโดย A. A. Martín และ H. M. Álvarez เพื่อให้ใช้ได้กว้างขวางยิ่งขึ้น กล่าวคือ ใช้ผู้เชี่ยวชาญในการประเมินความสอดคล้อง 2 คนหรือมากกว่าก็ได้ ซึ่งนับว่าเป็นวิธีหาสัมประสิทธิ์ของความสอดคล้องชนิดใหม่ล่าสุด แต่ว่าสูตรในการคำนวณยังคงมีลักษณะซับซ้อนมาก เช่นเดียวกันกับวิธีแรก และค่าที่คำนวณได้จะใกล้เคียงกับผลที่ได้จากวิธี Weighted Cohen's kappa ( $\kappa_w$ ) แต่จะไม่เกิดปัญหาเรื่องค่าที่ขัดแย้งกับความเป็นจริง (paradox)เพราะการกระจายของคะแนนรวมของผู้เชี่ยวชาญแตกต่างกันมาก หรือเบ้มาก (Martín & Álvarez, 2019)

15. วิธีหาค่าสัมประสิทธิ์สรุปอ้างอิง (Generalizability Coefficient) เป็นค่าสัมประสิทธิ์ที่พัฒนามาจากทฤษฎีสรุปอ้างอิง (Generalizability Theory) ซึ่งเป็นทฤษฎีใหม่ทางการวัดและประเมินผลที่พิจารณานำแหล่งของความผิดพลาดทั้งหลายที่เกี่ยวข้องกับการวัดและประเมินผลในหลายๆ ด้านมาไว้ในรูปแบบ (model) ของสูตรการคำนวณ (Cronbach et al., 1972 อ้างถึงใน Shavelson & Webb, 2005) ยกตัวอย่าง ในกรณีของการประเมินผลของผู้เชี่ยวชาญต่อข้อคำถามของเครื่องมืออื่นๆ อาจต้องคำนึงถึงปัจจัยต่างๆ ที่เกี่ยวข้องแล้วนำมาเป็นส่วนหนึ่งของสูตรการคำนวณ เช่น จำนวนผู้เชี่ยวชาญ เพศของผู้เชี่ยวชาญ คุณวุฒิทางวิชาการ ประสบการณ์การประเมิน ช่วงเวลาการประเมิน จำนวนข้อคำถามหรือข้อคำถามในการประเมินแต่ละครั้ง จำนวนครั้งในการประเมิน และจำนวนข้อคำถามทั้งหมด เป็นต้น รวมทั้งพิจารณาว่าปัจจัยของแต่ละด้านได้มาจากการสุ่มเลือก (random effects) หรือการเลือกแบบเจาะจง (fixed effects) หรือว่าเป็นแบบผสม (mixed effects) ดังนั้นสูตรในการคำนวณหาค่าสัมประสิทธิ์ความสอดคล้องแบบนี้จะมีความซับซ้อนมากที่สุด แต่จะได้ค่าที่ดีที่สุด เพราะว่านำปัจจัยที่เกี่ยวข้องหลายๆ ด้าน รวมทั้งใช้ค่าคะแนนจริง (true scores) ในสูตรการคำนวณ จึงเป็นค่าสัมประสิทธิ์สรุปอ้างอิง(Generalizability Coefficient) และค่านี้คือค่าสัมประสิทธิ์ความเที่ยงระหว่างผู้ประเมินชนิดหนึ่ง (O'brien, 1995)

**คำแนะนำในการเลือกใช้วิธีการหาค่าความตรงเชิงเนื้อหา**

1. ศึกษาลักษณะเฉพาะของข้อมูลของข้อคำถามหรือเครื่องมือวิจัยที่ต้องการหาค่าความตรงเชิงเนื้อหาว่าเป็นมาตรวัดชนิดใด (เช่น นามบัญญัติลำดับที่ อันตรภาคหรืออัตราส่วนเป็นต้น) และจำนวน

ผู้เชี่ยวชาญในเนื้อหาที่จะประเมินมีจำนวนกี่คน (2 คน หรือมากกว่า) เช่นวิธี Scott's pi ($\pi$) ใช้สำหรับข้อมูลที่วัดด้วยมาตรนามบัญญัติและใช้ผู้เชี่ยวชาญในเนื้อหาเพียง 2 คน แต่วิธี Krippendorff's alpha($\alpha$) ใช้สำหรับข้อมูลที่วัดด้วยมาตรวัดชนิดใดก็ได้และใช้ผู้เชี่ยวชาญในเนื้อหาเพียง 2 คน หรือมากกว่าก็ได้ เป็นต้น

2. ตัดสินใจว่าต้องการหาค่าความตรงเชิงเนื้อหารายข้ออย่างเดียวหรือทั้งฉบับ หรือทั้ง 2 อย่าง เช่น วิธี I-CVI (Item-Content Validity Index) วิธี IOC (Item-Objective Congruence Index) และวิธี IOCI (Item-Objective Consistency Index) รวมทั้งวิธีอื่นอีก 2-3 วิธีใช้หาค่าความตรงรายข้อ ส่วนวิธีอื่นๆ ใช้เพื่อหาค่าความตรงของแบบทดสอบ หรือเครื่องมือการวิจัยอย่างอื่นทั้งฉบับ
3. เลือกหาวิธีที่มีข้อบกพร่องนี้ในการแปลความหมายน้อย เช่น ไม่ควรใช้วิธี Cohen's kappa ($\kappa$) แบบทั่วไป เพราะว่ามีข้อบกพร่องที่สำคัญมากโดยเฉพาะในบางกรณีที่ให้ค่าที่ขัดแย้งกับความเป็นจริง และไม่ควรใช้วิธีหาค่า IOC (ล้วน สายยศ และอังคณา สายยศ, 2539 อ้างถึงใน จักรกฤษณ์ สำราญใจ, 2554) เพราะที่มาของสูตรไม่มีแนวคิดที่ชัดเจนรองรับ และไม่ได้คำนึงถึงความคลาดเคลื่อนจากการประเมินของผู้เชี่ยวชาญเลย เป็นต้น
4. เลือกวิธีที่มีเกณฑ์ในการแปลความที่เป็นที่ยอมรับกันทั่วไป เช่น วิธี Fleiss' kappa($\kappa$) วิธี Krippendorff's alpha($\alpha$) วิธีหาค่า I-CVI (Item-Content Validity Index) และวิธี IOC (Item-Objective Congruence Index) เป็นต้น
5. เลือกวิธีที่มีวิธีการคำนวณที่ง่าย และสะดวกจากแหล่งที่เชื่อถือได้ ปัจจุบันนี้การคำนวณหาค่าความตรงโดยวิธีต่างๆ ดังกล่าวแล้วทำได้ง่าย เพราะบางวิธีมีอยู่ในโปรแกรมสำเร็จรูปเพื่อการวิจัย เช่น วิธี Cohen's kappa ($\kappa$), Scott's pi ($\pi$) และ Fleiss' kappa($\kappa$) มีอยู่ในโปรแกรม SPSS (Statistical Package for the Social Sciences) ตั้งแต่ Version 25 ขึ้นไป และหลายวิธีสามารถ download โปรแกรมฟรีมาใช้ได้จาก http://www.real-statistics.com/free-download/real-statistics-resource-pack/หรือสามารถใช้โปรแกรมย่อยสำเร็จรูป (package) สำหรับใช้กับโปรแกรม R ได้ฟรี ซึ่งมีครบทุกวิธี แต่สำหรับการหาค่าความตรงที่มีสูตรยุ่งมาก คือ วิธี Generalizability Coefficient จำเป็นจะต้องใช้โปรแกรมสำเร็จรูปขั้นสูง เช่น GENOVA, SAS, EduG, G-String และ LISREL ซึ่งโปรแกรมเหล่านี้สามารถคำนวณได้ค่าดังกล่าวไม่แตกต่างกัน (Yelboga, 2011; Clauser, 2008)
6. เพื่อเพิ่มความน่าเชื่อถือเรื่องคุณภาพของข้อคำถาม แบบทดสอบ หรือเครื่องมือวิจัยอย่างอื่น ๆ นักทดสอบหรือผู้วิจัยควรคำนวณหาค่าความตรงเชิงเนื้อหาหลายๆ วิธีแล้วเสนอผลไว้ในรายงานด้วย (Syed & Nelson, 2015) ทั้งนี้เพราะว่าวิธีการหาค่าความตรงเชิงเนื้อหาต่างวิธีได้ค่าที่แตกต่างกัน และยังไม่มีเกณฑ์กลางในการเทียบค่าจากวิธีต่างๆ เช่น

ตารางที่ 6 การเปรียบเทียบค่าสัมประสิทธิ์ของความตรงเชิงเนื้อหาที่ได้จากวิธีต่างๆ

| Methods | Coefficients | Standard Error of Estimate |
|---|---|---|
| Cohen's kappa | 0.676 | 0.088 |
| Gwet's AC1 | 0.868 | 0.039 |
| Scott's pi | 0.675 | 0.089 |
| Krippendorff's alpha | 0.677 | 0.088 |
| % Agreement | 0.890 | 0.031 |

7. โปรดจำไว้เสมอว่า ค่าสัมประสิทธิ์ความเที่ยง หรือค่าความสอดคล้องของความเห็นของผู้เชี่ยวชาญในเนื้อหาสูง ไม่ได้หมายความโดยอัตโนมัติว่าข้อคำถามของแบบทดสอบ หรือเครื่องมือวิจัยอย่างอื่นมีความตรงเชิงเนื้อหาสูง เพราะว่านักทดสอบและประเมินผล หรือนักวิจัยจะต้องคำนึงถึงปัจจัยอื่นที่มีผลโดยตรงต่อค่าความตรงเชิงเนื้อหาในมิติอื่นด้วย คือ 1) จำนวนเนื้อหามีมากเพียงพอหรือไม่ 2) เนื้อหาเป็นตัวแทนที่ดีของเนื้อหาทั้งหมดที่ต้องการทดสอบ หรือวัด/ประเมินหรือไม่ และ 3) คุณภาพของข้อคำถามแต่ละข้อดีหรือไม่นอกจากความสอดคล้องของข้อคำถาม หรือข้อคำถาม ที่ได้จากความคิดเห็นของผู้เชี่ยวชาญในเนื้อหาจากวิธีต่างดังกล่าวแล้ว (Syed & Nelson, 2015)
8. นักทดสอบและนักวิจัยควรรายงานขั้นตอนต่างๆ ในการคำนวณหาค่าความตรงเชิงเนื้อหาที่ตนเองใช้อย่างละเอียดให้มากที่สุดในรายงาน เพื่อให้เกิดความคงเส้นคงวา และความน่าเชื่อถือของรายงาน (Syed & Nelson, 2015)
9. จากการศึกษาเปรียบเทียบค่าสัมประสิทธิ์ความสอดคล้องของความเห็นระหว่างผู้เชี่ยวชาญ 6 วิธี ของ Ato และคณะ(Ato, et al., 2011)) คือ วิธี Bennet's $\sigma$ (1954) Scott's $\pi$ (1955) Cohen's $\kappa$ (1960) Aickin's $\alpha$ (1990) และ Martin's & Femia's $\Delta$ (2004) พบว่า
   ก. Scott's $\pi$ และ Cohen's $\kappa$ มีปัญหาเรื่องค่าที่ขัดแย้งกับความเป็นจริงเมื่อคะแนนรวมของความเห็นด้านต่างๆ แตกต่างกันมาก (prevalence problem) และความมีอคติของผู้ประเมิน (rater bias problem) ยกเว้นเมื่อค่าที่คำนวณได้อยู่ในระดับปานกลาง (moderate level)
   ข. วิธีที่ดีที่สุดคือ Bennet's $\sigma$ และ Martin's & Femia's $\Delta$ ซึ่งเป็นวิธีที่ใช้สำหรับการประเมินข้อคำถามแบบเลือกตอบ (Multiple-choice Test Item) โดยเฉพาะ เพราะว่าค่าที่คำนวณได้มีความคงเส้นคงว่าในทุกระดับมากกว่าวิธีอื่นๆ และไม่มีปัญหาเรื่องการเกิดค่าที่ขัดแย้งกับความเป็นจริงเมื่อคะแนนรวมของความเห็นด้านต่างๆ แตกต่างกันมาก (prevalence problem)
10. ในกรณีที่ผู้วิจัยต้องการหาค่าความตรงเชิงเนื้อหาโดยการหาค่า IOC โดยสูตรของ Rovinelli และ Hambleton (1977) สามารถ download โปรแกรมมาใช้ได้ฟรีพร้อมคู่มือการใช้ได้จาก https://drive.google.com/open?id=1KgFvdtg-2qGaKQuG0gsxFf8OHfGSlsW9

**รายการอ้างอิง**


จักรกฤษณ์ สำราญใจ. (2554). *IOC = ความตรง?. วารสารหลักสูตรและการเรียนการสอน*. *4*(1–2), มหาวิทยาลัยขอนแก่น. https://www.scribd.com/doc/86608731/IOC.

พิศิษฐ ตัณฑวณิช และพนา จินดาศรี. (2561). ความหมายที่แท้จริงของค่า **IOC.** *วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม***,** *24*(2), 3–12. https://so02.tci-thaijo.org/index.php/jemmsu/article/view/174521/124950

ล้วน สายยศ และอังคณา สายยศ. (2539). *หลักการสร้างแบบทดสอบความถนัดทางการเรียน*. วัฒนาพานิช.

เยาวดี รางชัยกุล วิบูลย์ศรี. (2556). *การวัดผลและการสร้างแบบสอบผลสัมฤทธิ์*. สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

Abbott, R. D. & Perkin, D. (1982). Reliability and validity evidence for scale measuring dimensions of student ratings of instruction, *Journal of Educational and Psychological Measurement*, *42*(2), 563–569. https://doi.org/10.1177/001316448204200220

Andres, A. M. & Marzo, P. F. (2004). Delta: A new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology, 57*, 1–19. https://doi.org/10.1348/000711004849268

Ato, M., López, J. J., & Benavente, A. M. (2011). A simulation study of rater agreement measures with 2x2 contingency tables. *Psicológica, 32*(2), 385–402. https://www.uv.es/psicologica/articulos2.11/12ATO.pdf

Ayre, C. & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Journal of Measurement and Evaluation in Counseling and Development*, *47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Cherry, K. (2017). Why Validity is Important to Psychological Tests. Verywellmind. Retrieved from https://www.verywellmind.com/what-is-validity-2795788

Cherry, K. (2023). *Validity in psychological tests: why measures like validity and reliability are important*. https://www.verywellmind.com/what-is-validity-2795788

Choudhury, A. (2018). *Top 4 characteristics of a good test.* http://www.yourarticlelibrary.com/education/test/top-4-characteristics-of-a-good-test/64804

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Journal of Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Disha, M. (2018). *Validity of a test: 6 types | statistics*. http://www.yourarticlelibrary.com/statistics-2/validity-of-a-test-6-types-statistics/92597

Ebel, R. L. (1972). *Essential educational measurements*. Prentice Hall.

Fitzpatrick, A. R. (1983). The Meaning of content validity. *Applied Psychological Measurement*, *7*(1), 3–13. https://doi.org/10.1177/014662168300700102

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Garrett, H. E. (1964). *Testing for teachers*. American Book Company.

Girard, J. M. (2022). *Scott's pi coefficient*. https://github.com/jmgirard/mReliability/wiki/Scott%27s-pi-coefficient.

Goodwin, L. D. (2001). Interrater Agreement and Reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13–34. https://doi.org/10.1207/S15327841MPEE0501_2

Gwet, K. (2002). Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, *2*, 1–9. http://www.agreestat.com/research_papers/inter_rater_reliability_dependency.pdf

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1). https://doi: 10.1348/000711006X126600. PMID: 18482474

Haley, D. T., Thomas, P., Petre, M., & Roeck, A. D. (2008). *Using a new inter-rater reliability statistic*. Technical Report No. 2008/15. https://pdfs.semanticscholar.org/765d/f9d90295d5ca2b59e5092c4a5f7a09668d23.pdf

Haynes. S. N., Richard, D. C. S. & Kubany, E. S. (1995). Content validity in psychological assessment: A Functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hughes, A. (1995). *Testing for Language Teachers*. Bell & Bain, Ltd.

Kleeman, J. (2018). *Six tips to increase content validity in competence tests and exams*. https://www.questionmark.com/resources/blog/six-tips-to-increase-reliability-in-competence-tests-and-exams/

Krippendorff, K. (2011). *Computing Krippendorff 's Alpha-Reliability*. https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/07/fulltext.pdf

Lado, R. (1975). *Language testing*. Wing Tasi Cheung Printing.

Laerd Statistics. (2019). *Fleiss' kappa using SPSS Statistics. Statistical tutorials and software guides.* https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php

Laerd Research. (2018). *Content validity*. http://dissertation.laerd.com/content-validity.php

Lawshe, C. H. (1975). A Quantitative approach to content validity. *Journal of Personnel Psychology*, *28*, 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Martín A. A. & Álvarez, H. M. (2019). *Multi-rater delta: extension to many raters of the measure delta of nominal agreement*. https://arxiv.org/ftp/arxiv/papers/1909/1909.05575.pdf

Martín A. A. & Femia, P. (2004). Delta: A new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology*, 57(Pt 1), 1–19. https://doi.org/10.1348/000711004849268

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Medica, 22*(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

O'brien, R. M. (1995). Generalizability coefficients are reliability coefficients. *Quality & Quantity*, *29*, 421–428. https://doi.org/10.1007/BF01106066

Osborn, J. W. (Ed.). (2008). *Best Practices in Qualitative Methods*. Sage. https://books.google.co.th/books?id=M5_FCgCuwFgC&pg=PA35&lpg=PA35&dq=Krippendorff's+Alpha,+advantages,+disadvantages&source=bl&ots=SwoektNeeE&sig=ACfU3U3zRzvAtSWStHpCF9ypi82q2F7tkw&hl=th&sa=X&ved=2ahUKEwinw8qQ7_XlAhXhzTgGHV3gBP4Q6AEwEXoECAkQAQ#v=onepage&q=Krippendorff's%20Alpha%2C%20advantages%2C%20disadvantages&f=false

Polit, D. F., & Beck, C. T. (2006). The Content validity index: are you sure you know what's being reported? Critique and recommendations. *Journal of Research Nurse Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Polit, D.F., Beck, C.T. & Owen, S.V. (2007). Is the CVI an acceptable indicator of content validity? appraisal and recommendations. *Journal of Research Nurse Health, 30*(4). https://doi.org/10.1002/nur.20147

Professional Testing. (2006). Test validity. http://www.proftesting.com/test_topics/pdfs/test_quality_validity.pdf.

PTI. (2006). *Test Validity. Professional.* https://proftesting.com/test_topics/pdfs/test_quality_validity.pdf

Rovinelli, R. J., & Hambleton, R. K. (1976). *On the use of content specialists in the assessment of criterion-referenced test item validity*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. https://files.eric.ed.gov/fulltext/ED121845.pdf

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Tijdschrift Voor Onderwijs Research*, 2, 49-60.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325. https://doi.org/10.1086/266577

Shavelson, R. J., & Webb, N.M. (2005). *Generalizability theory*. https://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20AERA.pdf

Shuttleworth, M. (2009). *Content validity, explorable*. https://explorable.com/content-validity

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299–321. https://doi.org/10.1207/s15326977ea0504_2

Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, *3*(6). https://doi.org/10.1177/2167696815587648

Tang, W., Hu, J., Zhange, H., Wu, P., & He, H. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai Archives of Psychiatry*, *27*(1), 62–67. https://www.researchgate.net/publication/274727961_Kappa_coefficient_a_popular_measure_of_rater_agreement

Turner, R. C. & Carlson, L. (2003). Indexes of iItem-objective congruence for multidimensional items. *International Journal of Testing*, *3*(2), 163-17. https://www.tandfonline.com/doi/abs/10.1207/S15327574IJT0302_5

Turner, R. C., Mulvenon, S. W., Thomas, S. P., & Balkin, R. S. (2002). *Computing indices of item congruence for test development validity assessments*. https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p255-27.pdf

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). *A Comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples*. *BMC Medical Research Methodology*, *13*(1), https://doi.org/10.1186/1471-2288-13-61

Xie, Q. (2013). *Agree or disagree? A demonstration of an alternative statistic to Cohen's Kappa for measuring the extent and reliability of agreement between observers*. https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/J4_Xie_2013FCSM.pdf

Yelboga, A. (2011). *Investigation of generalizability theory analysis results with different statistical programs*. Poster presented at the XII. European Congress of Psychology, Istanbul, Turkey.

Zaiontz, C. (2019). *Real statistics using excel*. http://www.real-statistics.com

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: development of an instrument for measuring patient-centered communication. *Journal of Caring Science*, *4*(2), 165–167. https://doi.org/10.15171/jcs.2015.017