12 SHEE Research: Inter-rater reliability

ในสถานการณ์ตัดสินการนำเสนอผลงานวิจัยที่มีคณะกรรมการ จำนวน 3 ท่าน ถ้าพบว่าคะแนนที่กรรมการแต่ละท่านให้นั้นมีความแตกต่างกันมาก เช่น กรรมการท่านที่หนึ่งให้ 9 คะแนน ท่านที่สองให้ 3 คะแนนในการนำเสนอผลงานวิจัยชิ้นเดียวกัน ท่านคิดว่าการตัดสินนี้อาจขาดความน่าเชื่อถือหรือไม่ ถ้าผู้นำเสนอการวิจัยได้รับข้อมูลป้อนกลับจากผู้ประเมินที่มีความเห็นไปในทิศทางที่ไม่สอดคล้องกัน ผลการประเมินจากผู้ประเมินท่านใดจะมีความน่าเชื่อถือมากกว่ากัน จากสถานการณ์ดังกล่าวจึงเป็นที่มาของการตรวจสอบความสอดคล้องหรือความน่าเชื่อถือระหว่างผู้ประเมินซึ่งเป็นเรื่องที่สำคัญ ผู้เขียนบทความจึงได้กล่าวถึงใน SHEE Research ครั้งนี้

ความสอดคล้องระหว่างผู้ประเมิน (Inter-rater reliability) หมายถึงระดับของความเห็นพ้องต้องกันระหว่างผู้ประเมินหรือผู้ตัดสินหลายคนเมื่อทำการประเมินปรากฏการณ์เดียวกัน ในการวิจัยทางการศึกษา การประเมินมักเกี่ยวข้องกับการตัดสินผลแบบอัตวิสัย เช่น การประเมินผลงานของนักเรียน ประสิทธิภาพของครู หรือพลวัตในห้องเรียน ดังนั้นความสอดคล้องระหว่างผู้ประเมินจึงเป็นมาตรการเชิงป้องกันที่สำคัญเพื่อให้มั่นใจว่าการวิจัยนั้นมีคุณภาพและมีความน่าเชื่อถือ

นักวิจัยด้านการศึกษานำแนวคิดเรื่องความสอดคล้องระหว่างผู้ประเมินมาใช้เมื่อมีบุคคลหลายคนมีส่วนร่วมในการรวบรวม หรือวิเคราะห์ข้อมูลที่ต้องใช้การตัดสิน เช่น การสังเกตในห้องเรียน การให้คะแนนการประเมินแบบจดบันทึก เป็นต้น หากพบว่าความสอดคล้องระหว่างผู้ประเมินไม่เพียงพอ ผลการวิจัยอาจได้รับผลกระทบจากการให้คะแนนที่ไม่สอดคล้องกัน หรือมีความลำเอียงจากผู้ประเมิน ทำให้ความถูกต้องของข้อสรุปนั้นลดลงได้

อ. ดร.ปาริชาต อภิเดชากุล
ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ

หากนักวิจัยต้องการให้การประเมินผลการปฏิบัติมีความสอดคล้องกัน เมื่อผู้ประเมินทำการประเมินในสถานการณ์เดียวกัน ความคงเส้นคงวาของการให้คะแนนจึงเป็นปัจจัยสำคัญ อย่างไรก็ตาม อาจพบความคลาดเคลื่อนระหว่างผู้ประเมินได้ ซึ่งอาจเกิดจากการให้คะแนนในช่วงที่แคบเกินไป หรือมีอคติที่เกิดจากความประทับใจแรกพบของผู้ประเมิน ส่งผลให้คะแนนที่ผู้รับการประเมินได้รับนั้นอาจมีความแตกต่างกัน

เพื่อค้นหาหลักฐานที่แสดงถึงความคงเส้นคงวาของผลการประเมินภาคปฏิบัติจากผู้ประเมินตั้งแต่ 2 ท่านขึ้นไป เราสามารถใช้วิธีการกำกับติดตามคุณภาพการตรวจให้คะแนน โดยอาศัยการวิเคราะห์ค่าสถิติของผลการประเมิน (Rater statistics) หนึ่งในวิธีที่ใช้ คือการตรวจสอบค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของคะแนน เมื่อมีการให้คะแนนชิ้นงานเดียวกันจากผู้ประเมินหลายคน การเปรียบเทียบค่าเฉลี่ยเลขคณิตของผลการประเมินกับค่าที่ได้จากผู้ประเมินรายบุคคล อาจช่วยให้มองเห็นแนวโน้มการให้คะแนนที่เข้มงวดหรือลดหย่อนเกินไป นอกจากนี้ การจดบันทึก (tally) เพื่อตรวจสอบความถี่ของคะแนนที่ให้ ยังช่วยให้เห็นลักษณะการกระจายตัวของคะแนนว่ามีการกระจุกตัวหรือกระจายตัวออกไปในรูปแบบใด ซึ่งสามารถนำไปใช้ในการปรับปรุงมาตรฐานการให้คะแนนให้มีความสม่ำเสมอมากขึ้น

ประเด็นสำคัญในครั้งนี้ คือการตรวจสอบความสอดคล้องระหว่างผู้ประเมิน (Rater agreement) โดยวิธีหนึ่งที่ใช้ คือการตรวจสอบร้อยละของความสอดคล้องระหว่างผู้ประเมิน (Percent agreement) ซึ่งคำนวณจากสัดส่วนของการประเมินทั้งหมดที่ผู้ประเมินได้ให้คะแนนที่ตรงกัน

	Rater 1	Rater 2	Rater 3	Rater 4
นักศึกษา A	3	2	1	1
นักศึกษา B	2	3	2	2
นักศึกษา C	1	4	3	3
นักศึกษา D	4	5	4	4

ตารางที่ 1 แสดงคะแนนที่นักศึกษาได้รับจากผู้ประเมินแต่ละท่าน

จากตารางที่ 1 เราอาจประเมินความสอดคล้องระหว่างผู้ประเมินสำหรับท่านที่ 3 และ 4 ว่ามีความสอดคล้องของผลประเมินกันอย่างสมบูรณ์ (Absolute agreement) แต่ในการปฏิบัติจริงเราสามารถใช้หลักการคำนวณร้อยละความสอดคล้องระหว่างผู้ประเมิน (Percent agreement) คือการนำจำนวนเหตุการณ์ที่ผู้ประเมินให้คะแนนที่ตรงกันหารด้วยจำนวนเหตุการณ์ทั้งหมดที่ต้องสังเกตหรือให้คะแนน

ตัวอย่างที่ 1

การตรวจสอบความสอดคล้องระหว่างผู้ประเมิน ในการทดสอบ 2 ชุด นักเรียน 25 คน ผ่านคือต้องได้ร้อยละ 80 พบว่าผลการประเมินด้วยความเห็นชอบของผู้สอน 2 คนตรงกัน โดยพบว่าผ่าน 11 คน ไม่ผ่าน 9 คน คำนวณร้อยละความสอดคล้องระหว่างผู้ประเมิน

Percent absolute agreement = (สัดส่วนต่อร้อยละของการตัดสินว่าผ่าน) + (สัดส่วนต่อร้อยของการตัดสินว่าไม่ผ่าน)

Percent absolute agreement = (11/25) + (9/25) = 20/25 = 0.80

ตัวอย่างที่ 2

การตรวจสอบความสอดคล้องระหว่างผู้ประเมินในการประเมินค่าร้อยละความเห็นพ้องของผู้ประเมินทักษะการอ่านพบว่า

ผลการประเมินของผู้ประเมินคนที่ 2	คะแนน	ผลการประเมินของผู้ประเมินคนที่ 1				รวมแถว
		1	2	3	4	
	1	6	2			8
	2	3	4			10
	3		2	5	2	9
	4				3	3
รวมหลัก		9	8	8	5	30

ผลการประเมินทักษะการอ่านของนักศึกษาจำนวน 30 คน ซึ่งประเมินโดยอาจารย์จำนวน 2 คน อาจารย์ทั้งสองคนใช้เกณฑ์รูบริก 4 ระดับ ในการประเมินทักษะร้อยละความเห็นพ้องของผู้ประเมิน

$$= \frac{\textit{จำนวนผู้รับการประเมินที่ได้คะแนนเท่ากันจากผู้ประเมิน 2 คน}}{\textit{จำนวนผู้รับการประเมินทั้งหมด}} \times 100$$

= (18/30) X 100 = 60

การวัดความสอดคล้องหรือสัมประสิทธิ์สหสัมพันธ์ความสอดคล้อง เป็นการประเมินค่าความเที่ยง (Reliability) ของผู้ประเมิน หรือเครื่องมือวัดหลายชนิดที่ใช้วัดสิ่งเดียวกัน เป้าหมายคือตรวจสอบว่าการวัดสิ่งเดียวกันโดยผู้วัดหลายคน หรือโดยเครื่องมือวัดหลายชนิดนั้นให้ผลลัพธ์ที่สอดคล้องกันหรือไม่ โดยทั่วไปสถิติที่ใช้ในการวัดความสอดคล้องสามารถแบ่งออกเป็น 2 ประเภท ตามชนิดของข้อมูล หรือระดับการวัดของข้อมูล

1. ข้อมูลชนิดไม่ต่อเนื่อง (Discrete Data) เป็นการวัดความสอดคล้องสำหรับข้อมูลประเภทไม่ต่อเนื่อง มักใช้ค่าสถิติโคเฮนแคปปา (Kappa Statistic) ซึ่งเป็นวิธีที่คำนึงถึงโอกาสที่ผู้ประเมินจะให้คะแนน หรือประเมินตรงกันโดยบังเอิญด้วย

ตัวอย่าง

ผู้ประเมิน 2 คน ใช้แบบประเมินทักษะการซักประวัติผู้ป่วย เพื่อประเมินผู้เรียนจำนวน 20 คน โดยผลการประเมินแบ่งเป็น 2 ระดับ ได้แก่ ไม่ผ่าน (0) และ ผ่าน (1) หากทำการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินโดยใช้ Cohen's Kappa พบว่า แบบประเมินนี้มีความเที่ยงระหว่างผู้ประเมินอยู่ในระดับสูง (Cohen's kappa = .68, p = .000) โดยทั่วไป หากค่า Kappa ≥ 0.7 จะถือว่ามีความสอดคล้องที่ดี

2. ข้อมูลชนิดต่อเนื่อง (Continuous Data) สามารถทำการวิเคราะห์ได้หลายวิธี เช่น สัมประสิทธิ์สหสัมพันธ์ความสอดคล้อง (Concordance Correlation Coefficient) และสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient) เป็นต้น

ตัวอย่าง

การประเมินทักษะการเขียนรายงานของนักเรียนมัธยมศึกษาปีที่ 4 จำนวน 120 คน โดยใช้ผู้ประเมิน 3 คน ซึ่งเป็นครูผู้สอน ใช้เกณฑ์การให้คะแนน (Rubric) ที่มีองค์ประกอบ 5 ด้าน ประกอบด้วย ด้านเนื้อหา ด้านการจัดลำดับความคิด ด้านการใช้ภาษา ด้านความคิดสร้างสรรค์ และด้านการเขียนสะกดคำ โดยแต่ละด้านให้คะแนนได้ในช่วง 1-5 คะแนน ซึ่งจัดเป็นข้อมูลชนิดต่อเนื่อง การตรวจสอบความสอดคล้องของผู้ประเมินสามารถใช้การคำนวณสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient, ICC) เนื่องจาก ICC ใช้ได้กับผู้ประเมินที่มีตั้งแต่สองคนขึ้นไป ในบริบทนี้ มีผู้ประเมิน 3 คน และข้อมูลที่ได้จากการประเมินเป็นคะแนนซึ่งเป็นข้อมูลเชิงปริมาณ ถ้าผลการวิเคราะห์พบว่า ICC มีค่าสูงแสดงให้เห็นว่ามีความสอดคล้องระหว่างผู้ประเมินมาก เกณฑ์การให้คะแนนมีความชัดเจน และผู้ประเมินมีความเข้าใจตรงกันในการให้คะแนน โดยค่าสหสัมพันธ์ภายในชั้น (ICC) ≥ 0.75 ถือว่าอยู่ในระดับที่ดี

การนำการตรวจสอบความสอดคล้องระหว่างผู้ประเมิน มาใช้ในงานวิจัยควรมีการวางแผนตั้งแต่ขั้นตอนการออกแบบการศึกษา โดยมีแนวทางดังนี้

1. กำหนดจำนวนผู้ประเมินที่เหมาะสม สำหรับข้อมูลที่ต้องใช้การตัดสินใจเชิงคุณภาพควรมีผู้ประเมินอย่างน้อย 2 คน หากเป็นข้อมูลที่ซับซ้อนอาจเพิ่มจำนวนผู้ประเมินให้มากขึ้นเพื่อความแม่นยำในการประเมิน
2. เตรียมความพร้อมของผู้ประเมิน จัดอบรมเพื่อให้ผู้ประเมินมีความเข้าใจตรงกันเกี่ยวกับเกณฑ์การประเมิน ใช้ Rubric ที่มีเกณฑ์ชัดเจน พร้อมคำอธิบายของแต่ละระดับคะแนนอย่างละเอียด
3. การพัฒนาเครื่องมือวิจัย มีการออกแบบและทดสอบเครื่องมือโดยให้ผู้ประเมินทดลองใช้กับข้อมูลชุดเล็กก่อน หากพบว่าความสอดคล้องต่ำควรปรับปรุงเกณฑ์ให้ชัดเจนยิ่งขึ้นก่อนนำไปใช้จริง
4. การเก็บข้อมูลอย่างอิสระ ควรให้ผู้ประเมินนั้นประเมินแยกกันโดยไม่มีการปรึกษากันระหว่างการให้คะแนน บันทึกผลการประเมินแยกกัน เพื่อลดอิทธิพลของผู้ประเมินต่อกัน
5. การคำนวณและวิเคราะห์ความสอดคล้อง โดยหลังจากเก็บข้อมูลครบถ้วน นำข้อมูลมาวิเคราะห์ค่าสถิติความสอดคล้อง เช่น Cohen’s Kappa สำหรับผู้ประเมิน 2 คน, Fleiss’ Kappa สำหรับผู้ประเมินที่มีมากกว่า 2 คนโดยข้อมูลที่ได้เป็นชนิดไม่ต่อเนื่อง หรือ Intraclass Correlation Coefficient (ICC) สำหรับผลการประเมินที่จัดเป็นข้อมูลชนิดต่อเนื่อง
6. การแปลผลและรายงานผลการวิจัย ควรมีการรายงานค่าความสอดคล้องที่คำนวณได้ พร้อมอธิบายกระบวนการตรวจสอบวิเคราะห์ความแตกต่างในกรณีที่พบความไม่สอดคล้องกันเพื่อหาสาเหตุ หากพบปัญหาอาจต้องปรับปรุงเกณฑ์ หรือปรับแนวทางการฝึกอบรมผู้ประเมิน

การตรวจสอบความสอดคล้องระหว่างผู้ประเมินเป็นขั้นตอนสำคัญที่ช่วยยืนยันคุณภาพของข้อมูลวิจัยและเพิ่มความน่าเชื่อถือให้กับผลการวิจัย ความสอดคล้องระหว่างผู้ประเมินเปรียบเสมือนการมีเครื่องมือวัดที่ให้ผลลัพธ์เหมือนกัน ไม่ว่าใครจะเป็นผู้ใช้เครื่องมือ ซึ่งช่วยให้มั่นใจได้ว่าผลการวิจัยไม่ได้ขึ้นอยู่กับความคิดเห็นส่วนตัวของผู้ประเมินคนใดคนหนึ่ง แต่เป็นผลที่สะท้อนความเป็นจริงอย่างแท้จริง เหตุผลที่ต้องมีการตรวจสอบความสอดคล้องระหว่างผู้ประเมินคือ การประเมินที่ดีต้องไม่ขึ้นอยู่กับตัวผู้ประเมินแต่ละคน เช่นเดียวกันกับ เครื่องชั่งน้ำหนักที่ดี ต้องให้ค่าที่แม่นยำและสม่ำเสมอไม่ว่าใครจะเป็นผู้ใช้งานก็ตาม ถ้าทำได้จะช่วยให้ผลการวิจัยมีความถูกต้องและเป็นที่ยอมรับมากยิ่งขึ้น

ท้ายที่สุดความสอดคล้องของผู้ประเมินยังคงเป็นรากฐานสำคัญของวิธีการวิจัยทางการศึกษา โดยเฉพาะอย่างยิ่งในกรณีที่การตัดสินของผู้ประเมินนั้นมีบทบาทสำคัญในการรวบรวม และวิเคราะห์ข้อมูล เพื่อให้การประเมินมีความน่าเชื่อถือ นักวิจัยควรที่จะมีการฝึกอบรมผู้ประเมินอย่างเป็นระบบและครอบคลุม มีการกำหนดเกณฑ์การประเมินที่ชัดเจนและเหมาะสม อาจมีการบูรณาการระบบการประเมินอัตโนมัติร่วมกับผู้ประเมิน พัฒนาแนวทางการสร้างความสอดคล้องของผู้ประเมินในบริบทเฉพาะของแต่ละสาขาวิชารวมถึงสำรวจและวิเคราะห์ความสอดคล้องของผู้ประเมินในสภาพแวดล้อมที่ใช้เทคโนโลยีเสริมการประเมิน เมื่อดำเนินการตามแนวทางข้างต้น นักวิจัยทางการศึกษาจะสามารถเพิ่มความน่าเชื่อถือของผลการศึกษา สร้างหลักฐานเชิงประจักษ์ที่แข็งแกร่ง และสนับสนุนการพัฒนาองค์ความรู้ในสาขานั้นๆ ได้อย่างมั่นคงและมีประสิทธิภาพ

References

1. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-63.

2. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.

3. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Pract Assess Res Eval. 2004;9(1).

4. กมลวรรณ ตังธนกานนท์. การวัดและประเมินทักษะการปฏิบัติ. พิมพ์ครั้งที่ 3. กรุงเทพฯ: สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย; 2563.