



เครื่องวัดทางจิตเวช: ขอควรพิจารณาสำหรับผู้ใช้

มานิต ศรีสุรภานนท์ พ.บ.*

พริ้มเพรา ดิษยวณิช ปร.ด.*

จำลอง ดิษยวณิช พ.บ.*

บทคัดย่อ

ความหลากหลายของเครื่องวัดทางจิตเวชมักทำให้ผู้ใช้เครื่องวัดเกิดความไม่แน่ใจในการเลือกใช้เครื่องวัดที่เหมาะสมกับเวชปฏิบัติหรือการวิจัยของตน บทความนี้จึงขอแนะนำเสนอข้อควรพิจารณาพื้นฐานที่ผู้ใช้เครื่องวัดควรให้ความสนใจในการประเมินเครื่องวัดทางจิตเวช โดยสรุปแล้ว ผู้ใช้เครื่องวัดควรเริ่มพิจารณาจากภาพรวมของเครื่องวัด (overview of a measure) ก่อนว่าเครื่องวัดนั้น ๆ เหมาะสำหรับงานที่ผู้ใช้จะนำไปใช้หรือไม่ หากพิจารณาแล้วว่าเหมาะสมก็ควรพิจารณาคุณสมบัติอื่น ๆ ของเครื่องวัดต่อไป ซึ่งประกอบด้วย หัวข้อของเครื่องวัด (items of the measure), ความเชื่อมั่น (reliability), ความแม่นยำ (validity), กลุ่มตัวอย่าง (sample) ที่ใช้ในการหาความเชื่อมั่นและความแม่นยำ และการตอบสนองต่อการเปลี่ยนแปลง (response to change) หลังจากที่พิจารณาโดยถี่ถ้วนแล้ว ผู้ใช้ย่อมสามารถเลือกใช้เครื่องวัดที่เหมาะสมกับเวชปฏิบัติหรืองานวิจัยของตนเองได้ นอกจากนี้ ผู้ใช้เครื่องวัดยังสามารถที่จะประเมินเครื่องวัดทางจิตเวชที่มีผู้พัฒนาขึ้นใหม่ว่าเครื่องวัดใหม่มีข้อดีกว่าเครื่องวัดเดิมที่มีอยู่แล้วหรือไม่อย่างไร วารสารสมาคมจิตแพทย์แห่งประเทศไทย 2544; 46(2):127-136.

คำสำคัญ เครื่องวัดทางจิตเวช ความเชื่อมั่น ความแม่นยำ

บทนำ

ในขณะที่ความผิดปกติทางกายสามารถตรวจวัดได้ง่าย ความผิดปกติทางจิตใจและพฤติกรรม เป็นสิ่งที่ตรวจวัดได้ยาก ความยากลำบากในการตรวจวัดดังกล่าวทำให้มีการพัฒนาเครื่องวัด (measures) ทางจิตเวชขึ้นเป็นจำนวนมากในช่วงหลายทศวรรษที่ผ่านมา ยิ่งกว่านั้น ความผิดปกติทางจิตใจและทางพฤติกรรมบางชนิดก็มีเครื่องวัดหลายชุดให้เลือกใช้

ความหลากหลายดังกล่าวมักทำให้ผู้ใช้เครื่องวัดเกิดความไม่แน่ใจในการเลือกใช้เครื่องวัดที่เหมาะสมกับเวชปฏิบัติหรือการวิจัยของตน ประกอบกับการที่เครื่องวัดที่มีคุณภาพสูงมักต้องใช้เวลาในการประเมินและบันทึกข้อมูล ในขณะที่เครื่องวัดที่ใช้เวลาในการประเมินและบันทึกข้อมูลน้อยมักมีคุณภาพต่ำ ดังนั้นหากผู้ใช้เครื่องวัดมีแนวทางที่เหมาะสมในการประเมินเครื่องวัดก็ย่อมทำให้สามารถเลือกใช้เครื่องวัดที่เหมาะสมกับงานของตน และทราบข้อควรระวังในการใช้เครื่องวัดแต่ละชุดได้

บทความนี้มีวัตถุประสงค์ที่จะนำเสนอข้อควรพิจารณาพื้นฐานที่ผู้ใช้เครื่องวัดควรให้ความสนใจ โดยไม่กล่าวถึงรายละเอียดบางประการที่เข้าใจยาก นอกจากนี้ บทความนี้จะไม่ลงสูตรต่างๆ ที่ใช้ในการคำนวณทางสถิติ เนื่องจากโปรแกรมคอมพิวเตอร์บางชุด เช่น SPSS สามารถคำนวณค่าทางสถิติที่กล่าวถึงในบทความนี้ได้อยู่แล้ว โดยผู้ใช้ไม่จำเป็นต้องทราบถึงสูตรการคำนวณซึ่งเป็นเรื่องที่ซับซ้อนและเข้าใจได้ยาก

โดยทั่วไปแล้ว ผู้ใช้เครื่องวัดควรเริ่มพิจารณาจากภาพรวมของเครื่องวัด (overview of a measure) ก่อนว่าเครื่องวัดนั้นๆ เหมาะสำหรับงานที่ผู้ใช้จะนำไปใช้หรือไม่ หากพิจารณาแล้วว่าเหมาะสมก็ควรพิจารณาคูณสมบัติอื่นๆ ของเครื่องวัดต่อไป ซึ่งประกอบด้วย หัวข้อของเครื่องวัด (items of the measure), ความเชื่อมั่น (reliability), ความแม่นยำ (validity), กลุ่มตัวอย่าง (sample) ที่ใช้ในการหา

ความเชื่อมั่นและความแม่นยำ และการตอบสนองต่อการเปลี่ยนแปลง (response to change)

ภาพรวมของเครื่องวัด

ก่อนใช้เครื่องวัดใด ผู้ใช้ควรพิจารณาภาพรวมของเครื่องวัดในแง่มุมเหล่านี้ คือ ประเภทของข้อมูล (types of data), วัตถุประสงค์ของการวัด (purposes of measurement), ผู้ทำการวัด (measure administrators) และเวลาที่ใช้ในการประเมินและบันทึกข้อมูล (completion time)

ประเภทของข้อมูล

ข้อมูลที่ได้จากเครื่องวัดสามารถเรียงลำดับจากข้อมูลที่มีรายละเอียดน้อยไปมาก ดังนี้ คือ

1. ข้อมูลแยกจากกัน (dichotomous data) เช่น ป่วยเป็นโรค/ไม่ป่วยเป็นโรค เครื่องวัดที่ให้ข้อมูลชนิดนี้มักเป็นเครื่องวัดที่ใช้ในการวินิจฉัย (diagnostic measures)
2. ข้อมูลเรียงลำดับ (ordinal data) เช่น คะแนนความรุนแรงของอาการซึมเศร้า เครื่องวัดความรุนแรงของอาการทางจิตส่วนใหญ่ให้ข้อมูลประเภทนี้
3. ข้อมูลช่วงระยะ (interval data) เช่น visual analog scale เครื่องวัดในทางจิตเวชที่เป็นประเภทนี้มีน้อยมาก

อย่างไรก็ตาม เครื่องวัดหนึ่งอาจให้ข้อมูลมากกว่าหนึ่งประเภทได้ เช่น เครื่องวัดที่ให้ทั้งข้อมูลแยกจากกันและข้อมูลเรียงลำดับ เป็นต้น วิธีการหนึ่งที่ทำให้เครื่องวัดสามารถให้ข้อมูลได้มากกว่าหนึ่งประเภท คือ การที่นักวิจัยบางท่านได้กำหนดจุดตัดของเครื่องวัดชนิดข้อมูลเรียงลำดับหรือช่วงระยะขึ้น แล้วนำข้อมูลของผู้ป่วยมาระบุเป็นข้อมูลแยกจากกัน เช่น ผู้ป่วยที่มีคะแนนของ Hamilton Rating Scale for Depression (HRSD)¹ ลดลงมากกว่าร้อยละ 50 ถือว่าเป็นผู้ที่ตอบสนองต่อการรักษา ส่วนผู้ที่มีคะแนนลดลงน้อยกว่าร้อยละ 50 ถือว่า

เป็นผู้ที่ไม่ตอบสนองต่อการรักษา เป็นต้น

ก่อนใช้เครื่องวัด ผู้ใช้ควรทราบก่อนว่าข้อมูลที่จะได้จากเครื่องวัดเป็นข้อมูลประเภทใด หากเป็นไปได้ ผู้ใช้ควรเลือกเครื่องวัดที่ให้ข้อมูลที่มีรายละเอียดมากที่สุดเท่าที่จะทำได้ เนื่องจากข้อมูลที่มีรายละเอียดมากจะทำให้สามารถเลือกใช้การทดสอบทางสถิติบางประเภทที่มีความไวสูงต่อการพบความแตกต่างหรือความสัมพันธ์อย่างมีนัยสำคัญทางสถิติของข้อมูล² หรืออาจกล่าวอีกนัยหนึ่งได้ว่า จำนวนตัวอย่างที่ใช้ในการหาความแตกต่างหรือความสัมพันธ์อย่างมีนัยสำคัญทางสถิติจะน้อยลงเมื่อใช้ข้อมูลที่มีรายละเอียดมากขึ้น นอกจากนี้ การทราบถึงประเภทของข้อมูลยังนำไปสู่การเลือกใช้การทดสอบทางสถิติที่เหมาะสมสำหรับข้อมูลนั้น ๆ อีกด้วย

วัตถุประสงค์ของการวัด

ผู้ใช้เครื่องวัดควรทราบวัตถุประสงค์ของการวัดด้วยว่า เครื่องวัดนั้น ๆ ถูกพัฒนาขึ้นมาเพื่อใช้วัดอะไร เครื่องวัดอาจแบ่งได้เป็นหลายประเภท โดยประเภทที่มีการใช้บ่อย คือ เครื่องวัดเพื่อการวินิจฉัย (diagnostic measures) และเครื่องวัดความรุนแรงของอาการ (symptom-severity measures)

เครื่องวัดเพื่อการวินิจฉัยที่มีการกล่าวถึงบ่อยและจัดว่าเป็นมาตรฐานที่ดี (gold standard) มีอยู่ไม่มากนัก ตัวอย่างของเครื่องวัดประเภทนี้ คือ CIDI 2.1³ ซึ่งต้องใช้เวลาประมาณ 2-3 ชั่วโมงในการประเมินและบันทึกข้อมูลของผู้ป่วยแต่ละราย ด้วยเหตุนี้จึงมีผู้พัฒนาเครื่องวัดเพื่อการวินิจฉัยอื่นๆ ขึ้น โดยมีวัตถุประสงค์ที่จะลดเวลาในการประเมินและบันทึกข้อมูลของผู้ป่วยแต่ละรายลง แต่ยังคงความแม่นยำไว้ให้ได้มากที่สุด เครื่องวัดเหล่านี้อาจพัฒนาขึ้นเพื่อการวินิจฉัยโรคหลายโรค เช่น Mini-International Neuropsychiatric Interview (MINI)⁴ หรือเพื่อการวินิจฉัยโรคเพียงโรคเดียว เช่น Confusion Assessment Method (CAM)⁵ ก็ได้

ดังนั้น ขั้นตอนหนึ่งที่สำคัญมากในการพัฒนาเครื่องวัดเหล่านี้ คือ การหาความแม่นยำ (validity) ของเครื่องวัดที่พัฒนาขึ้นใหม่ (ซึ่งใช้เวลาประเมินและบันทึกข้อมูลผู้ป่วยสั้นลง) โดยเปรียบเทียบเครื่องวัดซึ่งเป็นมาตรฐานที่ดี (ซึ่งใช้เวลามากในการประเมินและบันทึกข้อมูลผู้ป่วย)

เครื่องวัดความรุนแรงของอาการมักประกอบด้วยหลายหัวข้อ และในแต่ละหัวข้อก็มีหลายระดับความรุนแรง เมื่อนำคะแนนจากหัวข้อทั้งหมดมารวมกันก็จะสามารถบอกระดับความรุนแรงของอาการได้ ผู้ใช้เครื่องวัดความรุนแรงของอาการไม่ควรใช้เครื่องวัดชนิดนี้ในการวินิจฉัย ยกเว้นว่าจะมีการพิสูจน์ให้เห็นถึงความแม่นยำของเครื่องวัดดังกล่าวในการวินิจฉัยไว้แล้ว เช่น คะแนนของ HRSD ที่มากกว่าหรือเท่ากับ 20 มักบ่งชี้ว่าผู้ป่วยมีอาการซึมเศร้าในระดับปานกลางขึ้นไป แต่ผู้ใช้เครื่องมือไม่สามารถใช้คะแนนดังกล่าวในการวินิจฉัยว่าผู้ป่วยป่วยเป็นโรคซึมเศร้า เพราะผู้ป่วยโรคจิตเภทที่มีอาการซึมเศร้ามากก็อาจมีคะแนน HRSD มากกว่าหรือเท่ากับ 20 ได้เช่นกัน

ผู้ทำการวัด

ในแง่ของผู้ทำการวัด เครื่องวัดอาจแบ่งได้เป็น 2 ประเภท คือ

1. เครื่องวัดที่ทำโดยผู้สัมภาษณ์ (interviewer-administered measures) โดยผู้สัมภาษณ์อาศัยข้อมูลจากผู้ป่วย ผู้ดูแลผู้ป่วย และ/หรือการสังเกตของผู้สัมภาษณ์เอง
2. เครื่องวัดที่ทำด้วยตนเอง (self-administered measures) ซึ่งส่วนใหญ่จะมีลักษณะเป็นแบบสอบถามให้ผู้ป่วยตอบ

เวลาที่ใช้ในการประเมินและบันทึกข้อมูล

เวลาที่ใช้ในการประเมินและบันทึกข้อมูลของแต่ละเครื่องวัดก็เป็นเรื่องที่ต้องพิจารณา การใช้

เครื่องวัดชุดเดียวที่ใช้เวลาประเมินและบันทึกข้อมูลราว 45-60 นาทีอาจเป็นไปได้ แต่การใช้เครื่องวัดหลายชุดที่แต่ละชุดใช้เวลาประเมินและบันทึกข้อมูลราว 45-60 นาที คงทำให้เกิดความยุ่งยากอย่างมากแก่ผู้ประเมินและผู้ถูกประเมิน ดังนั้นผู้พัฒนาเครื่องวัดจึงมักระบุถึงเวลาที่ใช้ในการประเมินและบันทึกข้อมูลของแต่ละเครื่องวัดไว้

หัวข้อของเครื่องวัด

แหล่งที่มาของหัวข้อ (Sources of items)

ผู้พัฒนาเครื่องวัดมักระบุถึงที่มาของหัวข้อในเครื่องวัดไว้ แหล่งที่มาส่วนใหญ่ คือ เครื่องวัดที่มีอยู่แล้ว การสังเกตทางคลินิก ความเห็นของผู้เชี่ยวชาญ การบอกเล่าประสบการณ์ของผู้ป่วย สิ่งที่ได้ค้นพบจากการวิจัย และทฤษฎี แต่ละแหล่งที่มาทั้งข้อดีและข้อเสีย หลังจากที่ใช้ทราบหัวข้อมาจากแหล่งใด ผู้ใช้ก็จะสามารถทราบได้ว่าหัวข้อที่ปรากฏในเครื่องวัดเกิดจากอคติ (bias) ของผู้พัฒนาหรือไม่และครอบคลุมสิ่งที่ต้องการวัดเพียงใด เช่น หัวข้อที่มาจากผลการวิจัยย่อมมีอคติน้อยกว่าและน่าจะครอบคลุมสิ่งที่ต้องการวัดได้ดีกว่าหัวข้อที่มาจากการสังเกตทางคลินิกหรือความเห็นของผู้เชี่ยวชาญ

การให้คะแนน (Scoring)

คะแนนที่ได้จากแต่ละหัวข้อไม่ควรเป็นไปในทิศทางใดทิศทางหนึ่งมากเกินไป (restriction in direction) และไม่ควรจำกัดอยู่เพียงช่วงใดช่วงหนึ่งของคะแนน (restriction in range) ถ้าคำตอบของหัวข้อใดเป็นไปในทิศทางใดทิศทางหนึ่งมากกว่า 90% หรือ 95% หัวข้อนั้นควรตัดออก เพราะนอกจากหัวข้อดังกล่าวจะไม่มีประโยชน์แล้วยังอาจเป็นโทษอีกด้วย⁶

ความเชื่อมั่น (Reliability)

ความเชื่อมั่น หมายถึง ความแน่นอน

(consistency) หรือความสามารถในการทำซ้ำ (reproducibility) ของการวัดเมื่อใช้เครื่องวัดใดเครื่องวัดหนึ่งในสถานการณ์ที่บุคคลและสภาวะต่างๆ ไม่มีการเปลี่ยนแปลง ความเชื่อมั่นของเครื่องวัดที่มีการกล่าวถึงบ่อยๆ คือ ความแน่นอนภายใน (internal consistency) ความเชื่อมั่นระหว่างผู้ประเมิน (inter-rater reliability) และความเชื่อมั่นชนิดทดสอบและทดสอบซ้ำ (test-retest reliability)

ความแน่นอนภายใน

ความแน่นอนภายในบ่งชี้ว่า คะแนนในแต่ละหัวข้อของเครื่องวัดเป็นไปในทิศทางเดียวกันมากน้อยเพียงใด เช่น ในการวัดความวิตกกังวล คะแนนของหัวข้อที่บ่งชี้ถึงความวิตกกังวล เช่น ความรู้สึกเครียด ความไม่สบายใจ อาการทางกายของความวิตกกังวล ควรเป็นไปในทิศทางเดียวกัน หากคะแนนในหัวข้อดังกล่าวไม่ไปด้วยกัน อาจบ่งชี้ว่าเครื่องวัดกำลังประเมินสภาพจิตใจหรือพฤติกรรมอย่างน้อย 2 สภาพที่แตกต่างกัน

สถิติที่มีการใช้บ่อยในการระบุความแน่นอนภายในของเครื่องวัด คือ Cronbach's alpha⁷ ซึ่งค่าที่ได้จะอยู่ระหว่าง 0.00 ถึง 1.00 ค่าที่ใกล้ 1.00 บ่งชี้ว่าเครื่องวัดมีความแน่นอนภายในสูง โดยทั่วไปแล้ว ค่า Cronbach's alpha จะสูงขึ้นเมื่อจำนวนหัวข้อในเครื่องวัดเพิ่มขึ้น ดังนั้นเครื่องวัดที่มีมากกว่า 20 หัวข้อจึงมักมีค่า Cronbach's alpha ไม่น้อยกว่า 0.90⁸ เครื่องวัดที่ดีควรมีค่า Cronbach's alpha อย่างน้อย 0.70⁸

ความเชื่อมั่นระหว่างผู้ประเมิน

ความเชื่อมั่นระหว่างผู้ประเมินบ่งชี้ว่า การที่ผู้วัดสองคนหรือมากกว่าใช้เครื่องวัดหนึ่งประเมินบุคคลคนเดียวกันแล้วได้คะแนนใกล้เคียงกันเพียงใด ผู้พัฒนาเครื่องวัดที่ใช้โดยผู้สัมภาษณ์มักแสดงความเชื่อมั่นในลักษณะนี้ให้ผู้ใช้ได้ทราบทั้งค่าความเชื่อมั่นของคะแนนรวมของเครื่องวัดและค่าความเชื่อมั่นของคะแนนในแต่ละหัวข้อของเครื่องวัด

หากค่าความเชื่อมั่นของคะแนนรวมของเครื่องวัดมีค่าต่ำมาก ผู้ใช้ควรหลีกเลี่ยงการใช้เครื่องวัดดังกล่าว แต่หากค่าความเชื่อมั่นของคะแนนในหัวข้อใดมีค่าต่ำมาก ผู้ใช้ควรระมัดระวังในการประเมินและการบันทึกข้อมูลในหัวข้อดังกล่าวให้มาก เนื่องจากค่าที่ต่ำมากดังกล่าวบ่งชี้ว่าผู้ประเมินและบันทึกข้อมูลมักมีความเห็นไม่ตรงกันในหัวข้อดังกล่าว

การทดสอบทางสถิติที่ใช้ในการหาค่าความเชื่อมั่นชนิดนี้มีดังนี้ คือ

1. ข้อมูลแยกจากกันใช้ Cohen's Kappa
2. ข้อมูลเรียงลำดับใช้ Spearman rank order correlation coefficient (r_s) หรือ intraclass correlation coefficient (ICC)
3. ข้อมูลช่วงระยะใช้ Pearson's correlation coefficient (r) หรือ ICC

ค่าทางสถิติที่บ่งบอกถึงความเชื่อมั่นชนิดนี้จะมีค่าอยู่ระหว่าง -1.00 ถึง 1.00 โดยค่า 1.00 จัดว่าเป็นค่าที่แสดงถึงความเชื่อมั่นอย่างสมบูรณ์ (perfect)

เนื่องจากเครื่องวัดที่ให้ข้อมูลเรียงลำดับหรือช่วงระยะมีการทดสอบทางสถิติ 2 วิธีให้เลือกใช้ บทความนี้จะจึงขอกกล่าวถึงรายละเอียดของการเลือกการทดสอบทางสถิติที่ใช้กับเครื่องวัด 2 ประเภทนี้เท่านั้น

ในอดีต ค่า r และค่า r_s เป็นค่าที่ใช้บ่อยในการระบุความเชื่อมั่นชนิดนี้ แต่ในปัจจุบัน ผู้พัฒนาเครื่องวัดไม่ค่อยนิยมใช้ค่า r และค่า r_s แล้ว เนื่องจากค่า r และค่า r_s บ่งชี้ถึงแต่เพียงความสัมพันธ์กัน (association) แต่ไม่ได้บ่งชี้ถึงความเห็นที่ตรงกัน (agreement) เช่น ผู้ประเมินท่านหนึ่งให้คะแนนมากกว่าผู้ประเมินอีกท่านหนึ่งประมาณ 1 คะแนนทุกครั้ง เมื่อดำเนินหาค่า r หรือค่า r_s จะพบว่าเท่ากับ 1.00 แต่แท้จริงแล้ว ผู้ประเมินทั้งสองยังมีความเห็นแตกต่างกันในการให้คะแนนอยู่ นอกจากนี้ ค่า r และค่า r_s ยังคำนวณได้เฉพาะกรณีที่มีข้อมูล 2 กลุ่ม

เท่านั้น หมายความว่า การทดสอบนี้จะไม่สามารถทำได้หากมีผู้ประเมินมากกว่า 2 คน

ในปัจจุบัน ค่า ICC เป็นค่าที่ใช้บ่อยในการบ่งชี้ความเชื่อมั่นชนิดนี้ เนื่องจากค่า ICC เป็นค่าที่บ่งชี้ถึงความเห็นที่ตรงกัน ซึ่งสะท้อนให้เห็นถึงความเชื่อมั่นระหว่างผู้ประเมินอย่างแท้จริง ค่า ICC นี้สามารถคำนวณได้โดยใช้การทดสอบทางสถิติที่เรียกว่า random-effect, two-way ANOVA การหาค่า ICC ด้วยการทดสอบทางสถิติชนิดนี้ทำให้สามารถคำนวณค่า ICC ได้ในกรณีที่ข้อมูลมากกว่า 2 กลุ่ม เช่น กรณีที่มีผู้ประเมินมากกว่า 2 คน เป็นต้น ความเชื่อมั่นระหว่างผู้ประเมินของเครื่องวัดที่ดีควรมีค่า r ค่า r_s หรือค่า ICC อย่างน้อย 0.70⁸

การหาค่าความเชื่อมั่นชนิดนี้ทำได้ 2 แบบ คือ การที่ผู้ประเมินอยู่พร้อมหน้ากันในขณะที่ประเมินผู้ป่วย (simultaneous assessment) กับการที่ผู้ประเมินแยกกันประเมินผู้ป่วย (independent assessment) ซึ่งการประเมินในลักษณะแรกมักให้ความเชื่อมั่นระหว่างผู้ประเมินสูงกว่าการประเมินในลักษณะหลังอยู่บ้าง

เนื่องจากความเชื่อมั่นระหว่างผู้ประเมินของแต่ละเครื่องวัดเป็นประเด็นที่มีความสำคัญมาก เช่น การใช้เครื่องวัดที่มีความเชื่อมั่นชนิดนี้สูงจะทำให้สามารถตรวจพบความแตกต่างหรือความสัมพันธ์อย่างมีนัยสำคัญทางสถิติโดยใช้จำนวนผู้ป่วยในการวิจัยน้อย เป็นต้น ในระยะหลัง ผู้เชี่ยวชาญบางท่านจึงนำวิธีการบางอย่างมาใช้ในการตัดแปลงเครื่องวัดให้มีค่าความเชื่อมั่นระหว่างผู้ประเมินสูงขึ้น วิธีการที่นิยมใช้บ่อยๆ มี 2 วิธี คือ

1. การระบุรายละเอียดของการให้คะแนน: ต้นฉบับของ Brief Psychiatric Rating Scale (BPRS)⁹ และ HRSD¹ เป็นเครื่องวัดที่ระบุเพียงว่าแต่ละหัวข้อสามารถให้คะแนนในช่วงใด เช่น 1-7, 0-4 โดยไม่ได้บอกรายละเอียดว่าผู้ป่วยควรมีความรุนแรงของอาการในหัวข้อนั้นเพียงใดจึงจะให้

คะแนนเท่าใด ซึ่งการไม่ระบุรายละเอียดในลักษณะนี้ มักส่งผลให้ผู้ประเมินแต่ละท่านให้คะแนนแตกต่างกัน แม้ว่าผู้ประเมินจะมีความเห็นตรงกันในเรื่อง ความรุนแรงของอาการในหัวข้อนั้น ในระยะหลังจึง มีผู้ปรับปรุงเครื่องวัดดังกล่าวโดยระบุรายละเอียด ของการให้คะแนนในแต่ละหัวข้อลงไป เช่น BPRS-Anchored¹⁰ และ NIMH-HRSD¹¹

2. การระบุคำถามที่ใช้ในการสัมภาษณ์: จะ เห็นได้ว่า หากผู้ประเมินใช้คำถามที่แตกต่างกันในการประเมินผู้ป่วยรายเดียวกัน ผลของการประเมิน และการให้คะแนนย่อมแตกต่างกัน ดังนั้นนอกจาก การระบุรายละเอียดของการให้คะแนนแล้ว ยังมี ผู้เชี่ยวชาญบางท่านเพิ่มคำถามที่ใช้ในการสัมภาษณ์ เข้าไปในเครื่องวัดอีกด้วย ซึ่งเครื่องวัดชนิดนี้อาจ เรียกได้ว่าเป็นมาตรวัดการสัมภาษณ์ทางคลินิกอย่าง มีแบบแผน (structured clinical interview scale) ตัวอย่างของเครื่องวัดชนิดนี้ คือ Structured Interview Guide for the Hamilton Depression Rating Scale (SIGH-D)¹²

ความเชื่อมั่นชนิดทดสอบและทดสอบซ้ำ

ความเชื่อมั่นชนิดทดสอบและทดสอบซ้ำ มีความสำคัญเช่นเดียวกับความเชื่อมั่นระหว่าง ผู้ประเมิน และแนวคิดของความเชื่อมั่นชนิดนี้ก็ไม่ แตกต่างจากแนวคิดของความเชื่อมั่นระหว่าง ผู้ประเมินมากนัก ความเชื่อมั่นชนิดทดสอบและ ทดสอบซ้ำนี้บ่งชี้ว่า ในกรณีที่สภาวะของบุคคลที่ถูก วัดไม่มีการเปลี่ยนแปลงไปตามกาลเวลา การวัดใน เวลาที่แตกต่างกันจะได้ผลที่ใกล้เคียงกันเพียงใด โดยทั่วไปแล้ว ผู้พัฒนาเครื่องวัดที่ใช้ด้วยตนเอง โดยเฉพาะแบบสอบถาม มักแสดงความเชื่อมั่นใน ลักษณะนี้ให้ผู้ใช้ได้ทราบ

การทดสอบทางสถิติและค่าทางสถิติที่ยอมรับ ได้ของความเชื่อมั่นชนิดนี้จะเป็นเช่นเดียวกับการ ทดสอบทางสถิติและค่าทางสถิติที่ยอมรับได้ของ ความเชื่อมั่นระหว่างผู้ใชซึ่งได้กล่าวมาแล้ว แต่เนื่อง

จากข้อมูลที่ใช้ในการหาค่าความเชื่อมั่นชนิดนี้มักได้ มาจากการประเมิน 2 ครั้งในเวลาห่างกันพอควรจึง ทำให้ข้อมูลที่จะนำไปหาค่าทางสถิติมีเพียง 2 กลุ่ม และไม่มีปัญหาในเรื่องของความเห็นที่ตรงกัน (agreement) ค่า r และค่า r จึงมักถูกนำมาใช้ในการ บ่งชี้ความเชื่อมั่นชนิดทดสอบและทดสอบซ้ำ

เช่นเดียวกับความเชื่อมั่นระหว่างผู้ประเมิน ความเชื่อมั่นชนิดทดสอบและทดสอบซ้ำของ เครื่องวัดที่ดีควรมีค่า r และ r อย่างน้อย 0.70⁷

ความแม่นยำ

ความแม่นยำ หมายถึง ความสามารถของ เครื่องวัดในการวัดสิ่งที่ต้องการ เช่น เครื่องวัดที่ใช้ สำหรับการวัดการประคับประคองทางสังคม (social support) หากใช้กับบุคคลหนึ่งแล้วมีคะแนนสูง (หรือต่ำแล้วแต่ชนิดของเครื่องวัด) ก็ต้องบ่งชี้ว่า บุคคลนั้นมีการประคับประคองทางสังคมที่ดีจริง ๆ ความแม่นยำของเครื่องวัดมีหลายประเภทและการ จำแนกประเภทก็ทำได้หลายแบบ สำหรับบทความนี้ ผู้เขียนได้จำแนกความแม่นยำเป็น 4 ประเภท คือ ความแม่นยำด้านหน้าตา (face validity), ความแม่นยำ ด้านเนื้อหา (content validity), ความแม่นยำ ตามเกณฑ์ (criterion validity) และความแม่นยำ ด้านการสร้าง (construct validity) โดยความแม่นยำ 2 ประเภทแรกไม่ต้องพิสูจน์ด้วยการทดสอบ ทางสถิติ แต่ความแม่นยำ 2 ประเภทหลังต้อง พิสูจน์ด้วยการทดสอบทางสถิติ

ความแม่นยำด้านหน้าตา

ความแม่นยำด้านหน้าตาบ่งชี้ว่าองค์ประกอบ ที่ปรากฏอยู่ในเครื่องวัดจะสามารถวัดในสิ่งที่ต้องการ วัดได้จริงหรือไม่ เครื่องวัดที่ดี (โดยเฉพาะแบบ สอบถาม) ควรมีการเรียงลำดับหัวข้อให้เข้าใจง่าย ไม่ซับซ้อน และดูเป็นทางการ การมีความ แม่นยำ ด้านนี้จะทำให้ผู้ตอบแบบสอบถามไม่สับสนและ มีความตั้งใจสูงที่จะตอบแบบสอบถาม ส่วนผู้ใช้

เครื่องวัดก็สามารถใช้เครื่องวัดได้อย่างไม่ผิดพลาดหรือผิดพลาดน้อย โดยทั่วไปแล้ว ผู้พัฒนาเครื่องวัดมักสรุปด้วยตนเองว่า ความแม่นยำตรงด้าน หน้าตาของเครื่องวัดดูดีแล้วก่อนถูกนำไปทดสอบหาความเชื่อมั่นและความแม่นยำ

ความแม่นยำตรงด้านเนื้อหา

ความแม่นยำตรงด้านเนื้อหาบางข้อหวัข้อที่ปรากฏอยู่ในเครื่องวัดนั้นครอบคลุมทุกแง่มุมที่จะทำการวัดแล้วหรือไม่ โดยทั่วไปแล้ว ผู้พัฒนามักใช้ทฤษฎี, องค์ความรู้ หรือความเห็นของตนในแง่ นั้นมาสนับสนุนว่าข้อที่ปรากฏอยู่ในเครื่องวัดสามารถวัดสิ่งที่ต้องการได้อย่างครอบคลุมแล้ว

ความแม่นยำตรงกับเกณฑ์

ความแม่นยำตรงกับเกณฑ์ บ่งชี้ว่า เครื่องวัดมีความแม่นยำตรงเพียงใดเมื่อเทียบกับเกณฑ์อื่น ๆ ซึ่งเกณฑ์อื่น ๆ มักหมายถึงเครื่องวัดที่มีอยู่แล้ว โดยเครื่องวัดที่มีอยู่แล้วอาจแบ่งได้เป็น 2 ประเภท คือ เครื่องวัดซึ่งเป็นมาตรฐานที่ดี (gold standard measure) และเครื่องวัดที่พอยอมรับได้ (acceptable measure)

ในกรณีที่มีเครื่องวัดซึ่งเป็นมาตรฐานที่ดีอยู่แล้ว ผู้พัฒนาเครื่องวัดใหม่จะต้องแสดงให้เห็นว่าเครื่องวัดใหม่มีความแม่นยำตรงใกล้เคียงกับเครื่องวัดเดิมซึ่งเป็นมาตรฐานที่ดี แต่เครื่องวัดใหม่มีข้อดีกว่าในแง่อื่น ๆ เช่น ใช้เวลาในการประเมินน้อยกว่า, สะดวกกว่า เป็นต้น เนื่องจากเครื่องวัดซึ่งเป็นมาตรฐานที่ดีมักเป็นเครื่องวัดที่ใช้ในการวินิจฉัย เช่น CIDI ดังนั้นเครื่องวัดใหม่ที่พัฒนาขึ้นโดยเปรียบเทียบับเครื่องวัดซึ่งเป็นมาตรฐานที่ดีจึงมักเป็นเครื่องวัดที่ใช้ในการวินิจฉัย เช่น MINI³ ซึ่งใช้เวลาในการประเมินสั้นลงมากแต่ยังมีความแม่นยำที่ใกล้เคียงกับ CIDI เป็นต้น สำหรับเครื่องวัดที่วัดความรุนแรงของอาการ ยังไม่มีเครื่องวัดใดที่ได้รับการยอมรับอย่างกว้างขวางว่าเป็นมาตรฐานที่ดี

เนื่องจากเครื่องวัดความรุนแรงของอาการที่จัดว่าเป็นมาตรฐานที่ดีมีอยู่น้อยมาก ดังนั้นการหา

ความแม่นยำตรงตามเกณฑ์ของเครื่องวัดความรุนแรงของอาการที่พัฒนาขึ้นใหม่จึงมักเปรียบเทียบกับเครื่องวัดที่พอยอมรับได้ เช่น ในการพัฒนา Amphetamine Withdrawal Questionnaire¹³ ผู้พัฒนาได้หาความแม่นยำตรงตามเกณฑ์โดยเปรียบเทียบกับ Clinical Global Impression¹¹ ซึ่งได้ค่า $r = 0.62$ เป็นต้น

การทดสอบทางสถิติที่ใช้ในการหาความแม่นยำตรงตามเกณฑ์มีการใช้ดังนี้ คือ

1. ข้อมูลชนิดแยกจากกันใช้ความไว (sensitivity) และความจำเพาะ (specificity) อย่างไรก็ดีตามค่าต่าง ๆ ที่ใช้คำนวณความไวและความจำเพาะก็อาจนำมาคำนวณหาค่าการคาดหมายที่เป็นบวก (positive predictive value), ค่าการคาดหมายที่เป็นลบ (negative predictive value) และอัตราความเป็นไปได้ (likelihood ratio) ได้ด้วยเช่นกัน

2. ข้อมูลชนิดเรียงลำดับใช้ Spearman rank order correlation coefficient (r_s)

3. ข้อมูลชนิดต่อเนื่องใช้ Pearson's correlation (r)

ค่าความไวและค่าความจำเพาะจะมีค่าอยู่ระหว่างร้อยละ 0 ถึง 100 ส่วนค่า r และ r_s จะมีค่าอยู่ระหว่าง -1.00 ถึง 1.00 โดยค่าร้อยละ 100 และ 1.00 จัดว่าเป็นค่าที่แสดงถึงความแม่นยำอย่างสมบูรณ์ (perfect) ตามลำดับ

สำหรับการหาความแม่นยำตรงตามเกณฑ์ของเครื่องวัดใหม่ชนิดที่ให้ข้อมูลเรียงลำดับหรือต่อเนื่อง หากเครื่องวัดซึ่งเป็นมาตรฐานที่ดีถูกนำมาใช้ในการหาค่าความแม่นยำตรงดังกล่าว ค่า r หรือค่า r_s ที่ได้ควรมีค่าอย่างน้อย 0.8 ซึ่งหมายความว่าเครื่องวัดใหม่วัดได้ใกล้เคียงกับเครื่องวัดซึ่งเป็นมาตรฐานที่ดี⁵

แต่หากเครื่องวัดใหม่หาค่าความแม่นยำตรงตามเกณฑ์โดยเทียบกับเครื่องวัดที่พอยอมรับได้ ค่า r หรือ r_s ควรอยู่ระหว่าง 0.3 ถึง 0.7 ซึ่งหมายความว่า

เครื่องวัดใหม่มีความสามารถในการวัดใกล้เคียงกับเครื่องวัดเดิมแต่ไม่เหมือนกันทั้งหมด⁵ โดยทั่วไปแล้วผู้พัฒนาเครื่องวัดใหม่มักนำเอาทฤษฎี, องค์ความรู้หรือความเห็นของตนที่ใช้ในการพัฒนาเครื่องวัดมาสรุปว่าความแตกต่างดังกล่าวเกิดจากการที่เครื่องวัดใหม่วัดได้แม่นยำตรงกว่าเครื่องวัดเดิมที่มีอยู่ในกรณีค่า r หรือ r น้อยกว่า 0.3 แสดงว่าเครื่องวัดใหม่วัดในสิ่งที่แตกต่างอย่างมากจากเครื่องวัดเดิมที่มีอยู่ ดังนั้น เครื่องวัดใหม่จึงไม่ควรได้รับการยอมรับและในกรณีที่ค่า r หรือ r มากกว่า 0.7 แสดงว่าเครื่องวัดใหม่วัดได้ไม่ต่างไปจากเครื่องวัดเดิมที่มีอยู่ ดังนั้นเครื่องวัดใหม่ที่พัฒนาขึ้นจึงเป็นสิ่งที่ไม่จำเป็นเพราะพัฒนาขึ้นมาแล้วแทบไม่มีอะไรที่แตกต่างจากเครื่องวัดเดิมที่มีอยู่

ความแม่นยำด้านการสร้าง (Construct validity)

ความแม่นยำด้านการสร้างบ่งชี้ว่า เครื่องวัดสามารถแยกผู้ที่ควรได้คะแนนสูงออกจากผู้ที่ได้คะแนนต่ำได้ดีเพียงใด ความแม่นยำชนิดนี้มักนำมาใช้ในกรณีที่ไม่มีเครื่องวัดซึ่งเป็นมาตรฐานที่ดีและไม่มีเครื่องวัดที่ยอมรับได้ (ซึ่งทำให้ไม่สามารถหาความแม่นยำด้านเกณฑ์ได้) ตัวอย่างของเครื่องวัดที่ใช้วิธีการนี้ คือ Mini-Mental State Examination¹⁴ ซึ่งผู้พัฒนาแสดงให้เห็นว่า ผู้ป่วยสมองเสื่อม (dementia) จะได้คะแนนจากเครื่องวัดนี้ต่ำกว่าผู้ป่วยโรคซึมเศร้า ผู้ป่วยโรคจิตเภท และคนปกติอย่างมีนัยสำคัญทางสถิติ ดังนั้นการทดสอบทางสถิติที่ใช้ในการหาความแตกต่างอย่างมีนัยสำคัญทางสถิติ เช่น Mann-Whitney U test, Student-t test จึงมักถูกนำมาใช้ในการหาความแม่นยำชนิดนี้

แม้ว่าการใช้คะแนนสูงต่ำจะเป็นวิธีการที่ใช้บ่อยในการทดสอบความแม่นยำด้านการสร้าง แต่บางครั้งการวิเคราะห์ปัจจัย (factor analysis) ก็อาจถูกนำมาใช้ทดสอบความแม่นยำชนิดนี้ได้ โดยเฉพาะการทดสอบความครอบคลุมของอาการต่างๆ

ที่ต้องการวัด ตัวอย่างเช่น การใช้การวิเคราะห์ปัจจัยทดสอบให้เห็นว่า Positive and Negative Syndrome Scale (PANSS) ประกอบด้วยหลายปัจจัย โดยในหลายปัจจัยนั้นมีปัจจัยชนิดบวก (positive factor) และปัจจัยชนิดลบ (negative factor) อยู่ด้วย¹⁵ การทดสอบนี้แสดงให้เห็นว่า PANSS เหมาะสำหรับนำมาใช้ประเมินผู้ป่วยโรคจิตเภท เนื่องจากอาการของผู้ป่วยโรคจิตมีทั้งอาการชนิดบวกและชนิดลบ

กลุ่มตัวอย่างที่ใช้ในการหาความเชื่อมั่นและความแม่นยำ

ผู้ใช้ควรพิจารณาด้วยว่า ผู้พัฒนาเครื่องวัดใช้กลุ่มตัวอย่างใดในการหาความเชื่อมั่นและความแม่นยำ หากกลุ่มตัวอย่างที่ใช้ในการพัฒนาเครื่องมือแตกต่างจากประชากรที่ใช้เครื่องวัดต้องการนำไปใช้แล้ว ผู้ใช้เครื่องวัดควรทำการทดสอบความเชื่อมั่นและความแม่นยำใหม่ในกลุ่มประชากรที่จะนำเครื่องวัดไปใช้

การตอบสนองต่อการเปลี่ยนแปลง

โดยทั่วไปแล้ว ผู้ใช้ควรให้ความสำคัญอย่างมากกับความเชื่อมั่นและความแม่นยำของเครื่องวัดอย่างไรก็ตาม ผู้เชี่ยวชาญบางท่านให้ทัศนะว่าเครื่องวัดที่ดีควรมีการตอบสนองที่ดีต่อการเปลี่ยนแปลงด้วย¹⁶ ตัวอย่างเช่น ผู้พัฒนา Montgomery-Asberg Rating Scale for Depression (MARS) ได้แสดงให้เห็นว่า MARS มีการตอบสนองต่อการเปลี่ยนแปลงของอาการซึมเศร้าได้ดีกว่า HRSD¹⁷

สรุป

จากบทความนี้จะเห็นได้ว่า ก่อนนำเครื่องวัดทางจิตเวชไปใช้ ผู้ใช้ควรพิจารณาเครื่องวัดในหลายแง่มุม โดยเฉพาะความเชื่อมั่นและความแม่นยำหลังจากที่พิจารณาโดยถี่ถ้วนแล้ว ผู้ใช้ย่อมสามารถเลือกใช้เครื่องวัดที่เหมาะสมกับเวชปฏิบัติหรืองาน

วิจัยของตนเองได้ นอกจากนี้ ผู้ใช้ยังสามารถที่จะประเมินเครื่องวัดทางจิตเวชที่มีผู้พัฒนาขึ้นใหม่ว่า เครื่องวัดใหม่มีข้อดีกว่าเครื่องวัดเดิมที่มีอยู่แล้วหรือไม่อย่างไร

เอกสารอ้างอิง

1. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960; 23: 56-62.
2. Kraemer HC. To increase power in randomized clinical trials without increasing sample size. *Psychopharmacol Bull* 1991; 27:217-24.
3. World Health Organization. Composite International Diagnostic Interview (CIDI), Version 2.1. Geneva, Switzerland: World Health Organization, 1998.
4. Sheehan DV, Lecrubier Y, Sheehan H, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998; 59(Suppl 20): 22-33.
5. Inouye S, van Dyck C, Alessi C, Balkin S, Siegal AP, Horwitz RI. Clarifying confusion: the Confusion Assessment Method. *Ann Intern Med* 1990; 113:941-8.
6. Streiner DL. A checklist for evaluating the usefulness of rating scales. *Can J Psychiatry* 1993; 38:140-8.
7. Cronbach LJ. Coefficient alpha and the internal structure of a test. *Psychometrika* 1951; 16:297-334.
8. Leary MR. Introduction to behavioral research methods. 2nd ed. Pacific Grove: Brooks/Cole, 1995.
9. Overall JE, Gorham DR. The brief psychiatric rating scale. *Psychol Rep* 1962;10:799-812.
10. Woerner MG, Mannuzza S, Kane JM. Anchoring the BPRS: an aid to improve reliability. *Psychopharmacol Bull* 1988; 24:112-7.
11. Guy W. ECDEU assessment manual for psychopharmacology, revised 1976. DHEW Publication No. (ADM) 76-338, 1976.
12. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry* 1988; 45:742-7.
13. Srisurapanont M, Jarusuraisin N, Jittiwutikan J. Amphetamine withdrawal: I. Reliability, validity and factor structure of a measure. *Aust N J Psychiatry* 1999; 33:89-93.
14. Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *J Psychiat Res* 1975; 12:189-98.
15. Kay SR, Sevy S. Pyramidal model of schizophrenia. *Schizophr Bull* 1990; 16:537-45.
16. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chron Dis* 1987; 40:171-8.
17. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979;134:382-9.



Measures in Psychiatry: Users' Guide

*Manit Srisurapanont, M.D.**

*Primprao Disayavanish, Ph.D.**

*Chamlong Disayavanish, M.D.**

Abstract

Because of the variety of psychiatric measures, a user may be unconfident in choosing a measure appropriate for s/he clinical practice or research. The authors, therefore, present general issues of consideration in evaluating a psychiatric measure. In summary, a user should initially evaluate the overview of a measure to see its appropriateness to be used. If appropriate, the issues of items, reliability, validity, the sample for reliability and validity tests, and the response to change should be considered. After a rigorous evaluation, the user should be able to choose a measure appropriate for s/he clinical practice or research. In addition, s/he should be able to evaluate a new psychiatric measure to see whether it is better than the existing ones.

J Psychiatr Assoc Thailand 2001; 46(2):127-136.

Key words: psychiatric measure, reliability, validity

* Department of Psychiatry, Faculty of Medicine, Chiang Mai University, Amphur Muang, Chiang Mai 50200.

E-mail: msrisura@mail.med.cmu.ac.th