ทบทวนวรรณกรรม

การเลือกใช้มาตรวัดสุขภาพ: ข้อควรพิจารณาเบื้องต้น

มานิต ศรีสุรภานนท์

ภาควิชาจิตเวชศาสตร์ คณะแพทยศาสตร์ มหาวิทยาลัยเชียงใหม่

ในปัจจุบัน มนุษย์มีความต้องการด้านสุขภาพมากขึ้น ทำให้บุคลากรทางการแพทย์ต้องอาศัยการวัดหรือการประเมินสุขภาพที่ละเอียดอ่อนขึ้น การใช้มาตรวัดสุขภาพเพื่อประเมินผลลัพธ์ทางสุขภาพที่เป็นนามธรรม เช่น ความเชื่อ, ความรู้สึก, เจตคติ และพฤติกรรมต่าง ๆ ได้เข้ามาเกี่ยวข้องกับการประเมินสุขภาพมากขึ้น บุคลากรทางการแพทย์จึงจำเป็นต้องเลือกใช้มาตรวัดสุขภาพที่สามารถประเมินสิ่งต่าง ๆ เหล่านี้ได้อย่างถูกต้องแม่นยำ รวมทั้งคำนึงถึงความเป็นไปได้และความคุ้มค่าของการประเมินด้วย ซึ่งส่งผลให้ผลลัพธ์ที่เป็นนามธรรมเหล่านี้ก็สามารถแปรเปลี่ยนเป็นคะแนนที่สามารถนำไปคำนวณและสื่อสารได้ บทความนี้ได้กล่าวถึงประเด็นสำคัญที่ควรพิจารณาก่อนนำมาตรวัดสุขภาพไปใช้ในเวชปฏิบัติหรือการวิจัย ผู้ใช้ควรให้ความสำคัญกับภาพรวมของมาตรวัดก่อน หลังจากนั้นจึงใช้ทฤษฎีการทดสอบแบบดั้งเดิมในการพิจารณาความเชื่อถือได้และความถูกต้องของมาตรวัด ความรู้เกี่ยวกับทฤษฎีการตอบสนองรายข้อสามารถนำมาใช้เสริมทฤษฎีการทดสอบแบบดั้งเดิมได้และจะช่วยให้เห็นถึงคุณสมบัติของแต่ละรายข้อของมาตรวัดได้ดีขึ้น เมื่อทราบคุณสมบัติดังกล่าวมาแล้ว ผู้ใช้ย่อมสามารถตัดสินใจเลือกมาตรวัดที่เหมาะกับงานของตนเองได้
เชียงใหม่เวชสาร 2560;56(1):49-61.

คำสำคัญ: มาตรวัด การประเมิน ความเชื่อถือได้ ความถูกต้อง

บทนำ

ในอดีตการประเมินสุขภาพทำได้ง่าย ผลลัพธ์ที่นิยมใช้ คือ การตายหรือรอดชีวิต ซึ่งผลลัพธ์นี้เป็นค่าที่ตรงไปตรงมาและเข้าใจได้ง่าย ด้วยเทคโนโลยีทางการแพทย์ที่ก้าวหน้าขึ้น ความต้องการของมนุษย์ในด้านสุขภาพก็เพิ่มมากขึ้นไปด้วย ส่งผลให้ผลลัพธ์ด้านสุขภาพมีความละเอียดอ่อนมากขึ้น เช่น จากอดีตที่เราให้ความสำคัญกับอัตราการตายในเด็กแรกเกิด ในปัจจุบัน ทุกภาคส่วนคำนึงถึงผลลัพธ์อื่นร่วมด้วย เช่น เด็กแรกเกิดที่รอดตายมาได้ มีความเจ็บป่วยหรือพิการหรือไม่ และหากป่วยหรือพิการเด็กนั้นป่วยหรือพิการรุนแรงเพียงใด ในยุคปัจจุบันผลลัพธ์ด้านสุขภาพไม่เพียงแต่เป็นรูปธรรมที่นับได้ง่าย (เช่น การตาย การป่วย) เท่านั้น แต่ยังมีลักษณะ

ติดต่อเกี่ยวกับบทความ: ผู้รับผิดชอบบทความ, มานิต ศรีสุรภานนท์, พ.บ., ภาควิชาจิตเวชศาสตร์ คณะแพทยศาสตร์ มหาวิทยาลัยเชียงใหม่, เชียงใหม่ 50200, ประเทศไทย อีเมล์: manit.s@cmu.ac.th
วันรับเรื่อง 14 มีนาคม 2560, **วันยอมรับการตีพิมพ์** 3 เมษายน 2560

เป็นนามธรรมที่มีความซับซ้อนเพิ่มขึ้นมาด้วย เช่น ความผาสุก (well-being), คุณภาพชีวิต (quality of life)

ความต้องการด้านสุขภาพของมนุษย์ที่เพิ่มขึ้นดังกล่าว ทำให้บุคลากรทางการแพทย์ต้องอาศัยการวัดหรือการประเมินสุขภาพที่ละเอียดอ่อนขึ้น การใช้มาตรวัดสุขภาพเพื่อประเมินผลลัพธ์ทางสุขภาพที่เป็นนามธรรมเช่น ความเชื่อ ความรู้สึกเจตคติ และพฤติกรรมต่าง ๆ ได้เข้ามาเกี่ยวข้องกับการประเมินสุขภาพมากขึ้น บุคลากรทางการแพทย์จึงจำเป็นต้องเลือกใช้มาตรวัดสุขภาพที่สามารถประเมินสิ่งต่าง ๆ เหล่านี้ได้อย่างถูกต้องแม่นยำ รวมทั้งคำนึงถึงความเป็นไปได้และความคุ้มค่าของการประเมินด้วย ซึ่งจะส่งผลให้ผลลัพธ์ที่เป็นนามธรรมเหล่านี้ก็สามารถแปรเปลี่ยนเป็นคะแนนที่สามารถนำไปคำนวณและสื่อสารได้

จากเหตุผลดังกล่าว มาตรวัดสุขภาพจึงถูกนำมาใช้บ่อยในเวชปฏิบัติหรือการวิจัย ไม่ว่าจะเป็นมาตรวัดที่บุคลากรทางการแพทย์เป็นผู้ประเมิน (rating scale) หรือแบบสอบถามที่แต่ละบุคคลประเมินตนเอง (questionnaire) ส่งผลให้มีการพัฒนามาตรวัดสุขภาพเพิ่มขึ้นมาก ซึ่งในจำนวนมาตรวัดที่มากมายมีทั้งที่มีคุณภาพสูงและคุณภาพต่ำ ผู้ใช้มาตรวัดจึงจำเป็นต้องมีความรู้เบื้องต้นในการเลือกมาตรวัดดังกล่าว การทราบถึงคุณสมบัติของมาตรวัดสุขภาพจะช่วยให้ผู้ใช้มาตรวัดสามารถเข้าใจจุดอ่อนจุดแข็งของแต่ละมาตรวัด และสามารถเลือกใช้มาตรวัดที่มีคุณภาพและเหมาะสมกับงานของตนได้

บทความนี้มีวัตถุประสงค์ที่จะนำเสนอข้อควรพิจารณาในการเลือกใช้มาตรวัดสุขภาพ แต่เนื่องจากศาสตร์ในด้านนี้มีรายละเอียดและความซับซ้อนมาก ผู้นิพนธ์จึงขอกล่าวถึงข้อควรพิจารณาเบื้องต้นเท่านั้น ในประเด็นที่ผู้เชี่ยวชาญมีความเห็นหลากหลาย บทความนี้จะกล่าวถึงเฉพาะแง่มุมที่ได้รับการยอมรับอย่างกว้างขวางเท่านั้น ซึ่งความรู้เบื้องต้นในบทความนี้น่าจะเหมาะกับผู้ใช้มาตรวัดสุขภาพ แต่สำหรับผู้ที่ต้องการพัฒนามาตรวัดสุขภาพขึ้นมาเอง ผู้พัฒนาควรศึกษาจากตำราอื่นเพิ่มเติม (1,2)

นอกจากการคำนึงถึงภาพรวมของมาตรวัดแล้ว ประเด็นที่ควรคำนึงถึงในการประเมินคุณภาพมาตรวัดสุขภาพ คือ (3)

1. ความเชื่อถือได้ (reliability)

   a. ความกลมกลืนภายใน (internal consistency)

   b. ความเชื่อถือได้ระหว่างผู้ประเมิน (interrater reliability) และความเชื่อถือได้ด้านการทดสอบซ้ำ (repeat reliability) ซึ่งอาจเรียกอีกชื่อหนึ่งว่าการทดสอบ-การทดสอบใหม่ (test-retest)

2. ความถูกต้อง (validity)

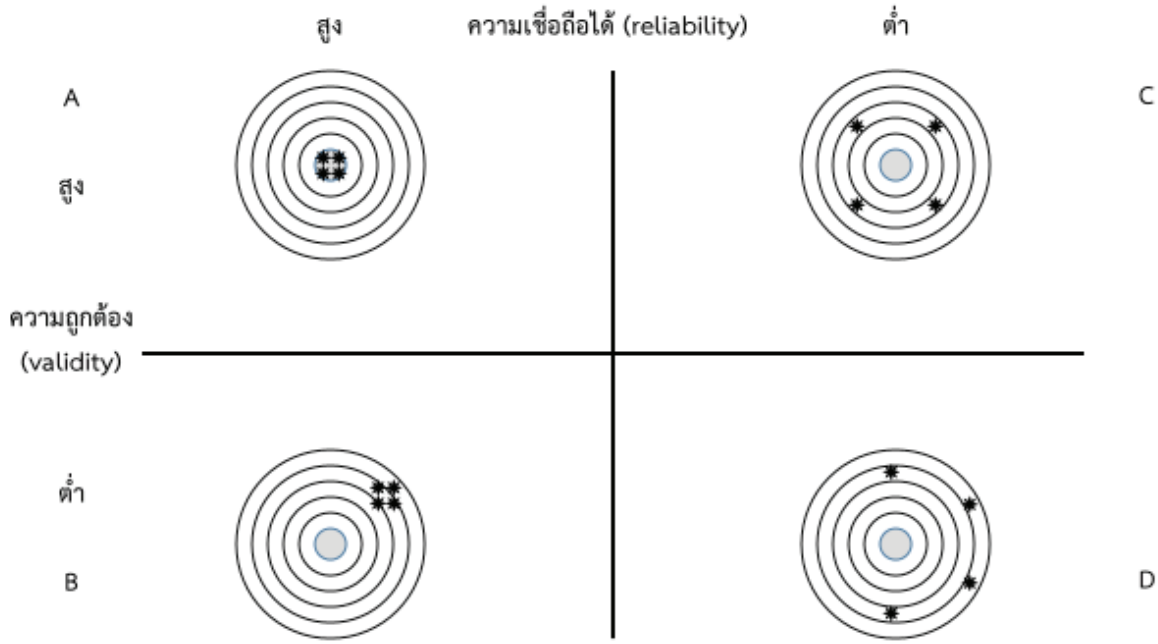
   a. ความถูกต้องด้านเนื้อหา (content validity)

   b. ความถูกต้องตามเกณฑ์ (criterion validity)

   c. ความถูกต้องด้านโครงสร้าง (structural validity) หรือความถูกต้องด้านการสร้าง (construct validity)

ในการทำความเข้าใจเบื้องต้นเกี่ยวกับความเชื่อถือได้และความถูกต้อง ผู้ใช้อาจมองว่าภาวะทางสุขภาพที่ต้องการวัดเป็นเป้าสำหรับการยิงปืนหรือยิงธนู ความเชื่อถือได้ดูได้จากการเกาะกลุ่มกันของจุดที่ยิงบนเป้า ส่วนความถูกต้องสามารถดูได้จากความใกล้เคียงของจุดที่ยิงกับเป้าที่ต้องการยิง (รูปที่ 1)

ข้อควรพิจารณาที่กล่าวมาแล้วจัดอยู่ในทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory, CTT) ซึ่งเป็นทฤษฎีที่ถูกใช้มาหลายสิบปีแล้ว แต่ระยะหลังได้มีการนำทฤษฎีการตอบสนองของรายข้อ (Item Response Theory, IRT) เข้ามาใช้ในการ



รูปที่ 1. แสดงให้เห็นถึงความเชื่อถือได้และความถูกต้องของมาตรวัดสุขภาพ

A มาตรวัดที่ดีควรมีความเชื่อถือได้และความถูกต้องสูง เปรียบเสมือนจุดที่ยิงมีการเกาะกลุ่มกันอย่างใกล้ชิดและจุดส่วนใหญ่อยู่ใกล้จุดศูนย์กลางของเป้า

B แสดงให้เห็นถึงปัญหาของการยิงที่ไม่เข้าเป้าแต่ยังเกาะกลุ่มกันดีอยู่ (ความเชื่อถือได้สูงแต่ความถูกต้องต่ำ)

C แสดงให้เห็นว่ามีการยิงเข้าเป้าบ้าง แต่จุดที่ยิงไม่เกาะกลุ่มกัน (ความเชื่อถือได้ต่ำ แต่ความถูกต้องยังสูงอยู่)

D แสดงให้เห็นว่าการยิงทั้งไม่เกาะกลุ่มและไม่เข้าเป้า (ความเชื่อถือได้ต่ำและความถูกต้องต่ำ)

ประเมินมาตรวัดสุขภาพด้วย ซึ่งบทความนี้จะกล่าวถึงทฤษฎีการทดสอบแบบดั้งเดิม เป็นส่วนใหญ่และกล่าวถึงทฤษฎีการตอบสนองรายข้อพอสังเขป

ภาพรวมของมาตรวัดสุขภาพ

ก่อนใช้มาตรวัดสุขภาพ ผู้ใช้ควรทำความเข้าใจกับภาพรวมของมาตรวัดก่อน ซึ่งประกอบด้วย วัตถุประสงค์ของการวัด (purpose of measurement), ประเภทของข้อมูล (types of data) ผู้ทำการวัด (measure administers) ผู้ถูกประเมิน และเวลาที่ใช้ในการประเมิน (completion time)

วัตถุประสงค์ของการวัด

วัตถุประสงค์ของการใช้มาตรวัดสุขภาพอาจแบ่งได้เป็น 3 ประเภท (1) คือ

1. เพื่อการประเมิน (evaluation) โดยเฉพาะความรุนแรงของโรคหรือภาวะทางสุขภาพ เช่น แบบประเมินความรุนแรงของความพิการที่เรียกว่า World Health Organization Disability Assessment Scale, Version 2.0 (WHODAS 2.0) (4)

2. เพื่อการวินิจฉัยโรค (diagnosis) เช่น แบบประเมินเพื่อการวินิจฉัยโรคทางจิตเวชที่เรียกว่า Mini-International Neuropsychiatric Interview (5)

3. เพื่อการทำนาย (prediction) ซึ่งใช้ประเมินเพื่อทำนายภาวะทางสุขภาพในอนาคต เช่น Apgar score ที่ใช้ประเมินทารกแรกเกิดเพื่อทำนายภาวะทางสุขภาพของทารกในอนาคต (6) มาตรวัดสุขภาพประเภทนี้มีจำนวนไม่มากนัก เนื่องจากพัฒนาได้ยากและใช้เวลานานในการติดตามผลลัพธ์ทางสุขภาพ

แม้ว่ามาตรวัดสุขภาพส่วนใหญ่จะพัฒนาขึ้นเพื่อวัตถุประสงค์ใดวัตถุประสงค์หนึ่ง แต่บางมาตรวัดอาจใช้ได้หลายวัตถุประสงค์ เช่น 9-Item Patient Health Questionnaire (PHQ-9) สามารถใช้ประเมินความรุนแรงของโรคซึมเศร้าได้ ในขณะที่คะแนนจุดตัดที่ 9/10 ก็ใช้คัดกรองโรคซึมเศร้าได้ (7)

ประเภทของข้อมูล

เช่นเดียวกับข้อมูลทางสถิติ ข้อมูลที่ได้จากมาตรวัดสุขภาพอาจแบ่งได้เป็น 4 ประเภท (1) คือ

1. ข้อมูลประเภทชื่อหรือประเภท (nominal or categorical data) แม้ว่าในการบันทึกข้อมูลอาจมีการใส่ตัวเลขแทนค่าต่าง ๆ เช่น female = 0, male = 1 แต่ข้อมูลประเภทนี้ไม่สามารถเรียงลำดับจากมากไปน้อยได้

2. ข้อมูลประเภทลำดับ (ordinal data) มีการใช้ตัวเลขที่แสดงถึงขนาดของการวัดเช่น ผู้ที่มีคะแนนของ PHQ-9 เท่ากับ 20 มีความเศร้ามากกว่าผู้ที่มีคะแนน PHQ-9 15 คะแนน, และคนที่มีคะแนนของ PHQ-9 เท่ากับ 15 มีความเศร้ามากกว่าผู้ที่มีคะแนน PHQ-9 10 คะแนน อย่างไรก็ตามความแตกต่างกัน 5 คะแนนของความเศร้าในสองสภาวะดังกล่าวอาจไม่เท่ากันก็ได้

3. ข้อมูลประเภทระยะห่าง (interval data) ข้อมูลที่เป็นตัวเลขแต่ละตัว เช่น 3 กับ 2 และ 2 กับ 1 มีระยะห่างเท่ากับ 1 เหมือนกัน แต่ค่าที่ได้จากการวัดอาจต่ำกว่า 0 ได้ ตัวอย่างเช่น อุณหภูมิ ซึ่งมีค่าเป็นบวกหรือลบก็ได้

4. ข้อมูลประเภทอัตราส่วน (ratio data) คล้ายคลึงกับข้อมูลประเภทระยะห่าง แต่ค่าต่ำสุดที่ได้จากการวัด คือ 0 (ไม่มีค่าติดลบ) มาตรวัดสุขภาพประเภทนี้มีน้อยมาก และอาจต้องมีการศึกษาเป็นอย่างดีก่อนจึงจะเชื่อได้ว่าเป็นข้อมูลประเภทอัตราส่วนจริง เช่น Myles และคณะได้ทำการศึกษา visual analog scale (VAS) ที่ใช้วัดความปวดในผู้ป่วยหลังผ่าตัด และพบว่าคะแนน VAS ที่ใช้วัดความปวดมีคุณสมบัติแบบข้อมูลประเภทอัตราส่วนจริง (8)

ผู้ทำการวัด (measure administers)

แต่ละมาตรวัดสุขภาพจะระบุผู้ทำการวัดไว้ชัดเจน ซึ่งอาจแบ่งออกได้เป็น

1. มาตรวัดที่ทำโดยผู้สัมภาษณ์ (interview-administered measures) ผู้สัมภาษณ์ให้คะแนนในมาตรวัดโดยอาศัยข้อมูลที่ได้จากผู้ป่วย ญาติผู้ป่วย การตรวจร่างกายและการสังเกต การตรวจทางห้องปฏิบัติ มาตรวัดประเภทนี้อาจเรียกอีกชื่อหนึ่งว่า rating scale

2. มาตรวัดที่ทำด้วยตนเอง (self-administered measures) ซึ่งก็คือมาตรวัดที่ให้บุคคลที่ต้องถูกประเมินเป็นผู้ตอบ มาตรวัดประเภทนี้มักเรียกว่า แบบสอบถาม (questionnaire)

3. มาตรวัดที่ทำโดยญาติหรือผู้ดูแลผู้ป่วย (proxy-administered measures) ซึ่งก็คือแบบสอบถามที่ให้ญาติหรือผู้ดูแลของบุคคลที่ต้องได้รับการประเมินเป็นผู้ตอบ

มาตรวัดสุขภาพบางฉบับอาจทำขึ้นหลายรูปแบบเพื่อให้ใช้ได้กับผู้ทำการวัดหลายประเภท เช่น World Health Organization Disability Assessment Schedule 2.0 ที่จัดทำขึ้นให้สามารถใช้ได้กับผู้ทำการวัดทั้งสามรูปแบบ (4)

ผู้ถูกประเมิน

ในการศึกษาคุณสมบัติต่าง ๆ ของมาตรวัด ผู้ศึกษาควรให้ความสำคัญกับกลุ่มประชากรที่ใช้ในการศึกษาด้วยว่าคล้ายคลึงกับผู้ถูกประเมินที่กำลังจะนำมาตรวัดไปใช้หรือไม่ แม้ว่าจะเป็นมาตรวัดเดียวกัน หากใช้ในประชากรคนละกลุ่ม คุณสมบัติของมาตรวัดก็อาจเปลี่ยนไปได้ เช่น การศึกษาความกลมกลืนภายในของ Montreal Cognitive Assessment (MoCA) ในผู้สูงอายุที่มีการศึกษาสูง (ค่าเฉลี่ยของ MoCA เท่ากับ 26.5), ผู้สูงอายุทั่วไป (ค่าเฉลี่ยของ MoCA เท่ากับ 23.8) และผู้สูงอายุที่มีความผิดปกติของสมอง (ค่าเฉลี่ยของ MoCA เท่ากับ 21.0) Bernstein และคณะ (9) พบว่าค่า Cronbach's alpha ของทั้งสามกลุ่มมีความแตกต่างกัน คือ 0.5, 0.6 และ 0.8 ตามลำดับ

เวลาที่ใช้ในการประเมิน (completion time)

ผู้ใช้มาตรวัดควรคำนึงถึงเวลาที่ต้องใช้ในการประเมินและบันทึกข้อมูลเสมอ การใช้เวลาในการประเมินนานเกินไปอาจทำให้ผู้ถูกประเมินไม่ร่วมมือหรือไม่ตั้งใจที่จะให้คำตอบตามความเป็นจริงได้

ความเชื่อถือได้ (reliability)

ความเชื่อถือได้ หมายถึง ระดับความคลาดเคลื่อนของการวัด (measurement error) ที่เกิดขึ้นในมาตรวัด (3) มาตรวัดที่ดีควรมีความคลาดเคลื่อนน้อย หรือกล่าวอีกนัยหนึ่ง คือ มีค่าใกล้เคียงเดิมเมื่อทำการวัดซ้ำในผู้ถูกประเมินที่มีภาวะทางสุขภาพเหมือนเดิม (reproducibility) ความเชื่อถือได้อาจแบ่งออกได้เป็น 2 ประเภทใหญ่ ๆ คือ ความกลมกลืนภายใน และความเชื่อถือได้ระหว่างผู้ประเมินและด้านการทำซ้ำ

ความกลมกลืนภายใน

ความกลมกลืนภายใน คือ ระดับความสัมพันธ์ของแต่ละรายข้อ (item) ในมาตรวัด การประเมินความสัมพันธ์นี้จะกระทำกับมาตราวัดหรือมาตรวัดย่อย (subscale) ที่มีความเป็นมิติเดียว (unidimensionality) เท่านั้น วิธีที่ใช้บ่อยในการแสดงถึงความเป็นมิติเดียวกันคือ การวิเคราะห์ปัจจัย (factor analysis) โดยอาจเลือกใช้วิธีสำรวจ (exploratory) หรือวิธียืนยัน (confirmatory) ตามความเหมาะสม

สำหรับสถิติที่นิยมใช้บ่อยที่สุดในการประเมินความกลมกลืนภายใน คือ Cronbach's alpha (10) ซึ่งแสดงค่าระหว่าง 0.00 ถึง 1.00 ค่าที่ใกล้ 1.00 แสดงว่ามาตรวัดมีความกลมกลืนภายในสูง ค่าที่ยอมรับได้ คือ 0.70-0.90 (11) เนื่องจากค่านี้สัมพันธ์อย่างยิ่งกับจำนวนรายข้อของมาตรวัด มาตรวัดที่มีจำนวนรายข้อมากจึงมักมีค่า Cronbach's alpha สูงไปด้วย เช่น มาตรวัดสุขภาพที่มีมากกว่า 15 รายข้อส่วนใหญ่จะมีค่า Cronbach's alpha สูง แม้ว่าจะมีความกลมกลืนภายในต่ำ (12) ดังนั้นมาตรวัดที่มีจำนวนรายข้อน้อยและมีค่า Cronbach's alpha ต่ำกว่า 0.70 ก็อาจยอมรับได้ สำหรับค่า Cronbach's alpha ที่มากกว่า 0.90 แม้จะดูดี แต่ก็บ่งชี้ว่า บางรายข้อของมาตรวัดนั้นกำลังวัดในสิ่งที่คล้ายกันมากเกินไป

ค่าเฉลี่ยความสัมพันธ์ระหว่างรายข้อ (average หรือ mean item intercorrelation) เป็นอีกดัชนีหนึ่งที่ใช้ประเมินความกลมกลืนภายในของมาตรวัดสุขภาพ ข้อดีของดัชนีนี้ คือ ค่าที่ได้ไม่ขึ้นกับจำนวนรายข้อของมาตรวัด (13)

ความเชื่อถือได้ด้านการทำซ้ำและระหว่างผู้ประเมิน

หลักการสำคัญในการประเมินความเชื่อถือได้ คือ การประเมินว่าคะแนนที่ได้จากการประเมินมากกว่าหนึ่งครั้งในผู้ถูกประเมินคนเดิมซึ่งยังคงมีภาวะทางสุขภาพเหมือนเดิม จะยังคงมีคะแนนเท่าเดิมหรือไม่ โดยการประเมินแต่ละครั้งต้องมีความเป็นอิสระและไม่มีอิทธิพลต่อกัน ในการประเมินซ้ำ ผู้ประเมินต้องไม่ทราบหรือไม่สามารถจำคำตอบของการประเมินครั้งก่อนได้

1. ความเชื่อถือได้ระหว่างผู้ประเมิน (inter-rater reliability) มาตรวัดที่ทำโดยผู้สัมภาษณ์หรือ rating scale ควรต้องมีความเชื่อถือได้ระหว่างผู้ประเมินในระดับสูง หากผู้ถูกประเมินมีภาวะสุขภาพเช่นเดิม ผลลัพธ์ (คะแนน) ของมาตรวัดที่ได้จากผู้ประเมินแต่ละคนควรใกล้เคียงกันมาก สถิติที่ใช้บ่อยในการศึกษาประเด็นนี้ คือ ค่าความสัมพันธ์ระหว่างคะแนน ซึ่งแบ่งได้ดังนี้

- ข้อมูลประเภทอัตราส่วนหรือระยะห่าง มักใช้ Intraclass correlation coefficient (ICC) (14) อย่างไรก็ตาม Pearson correlation coefficient (r) ก็ถูกนำมาใช้บ้าง ข้อจำกัดของ r ในกรณีนี้ คือ หากผู้ประเมินรายหนึ่งให้คะแนนสูงกว่าผู้ประเมินอีกราย (เช่น มากกว่า 1 คะแนน) อยู่ตลอดเวลา ค่า r ที่ได้จะเท่ากับ 1 ทั้ง ๆ ที่ผู้ประเมินทั้งสองรายมีความเห็นที่ไม่ตรงกันเลย (disagreement) แต่หากใช้ ICC ปัญหานี้จะไม่เกิดขึ้น (15)

- ข้อมูลประเภทลำดับ มักใช้ Weighted Kappa Coefficient (16), (Unweighted) Kappa Coefficient (17) อย่างไรก็ตาม Spearman correlation coefficient (rs), ก็ถูกนำมาใช้บ้าง แต่ก็จะพบปัญหาเดียวกันกับการใช้ r ดังกล่าวมาแล้ว คือ rs อาจมีค่าสูงโดยที่ผู้ประเมินทั้งสองรายมีความเห็นแทบที่ไม่ตรงกันเลย

- ข้อมูลชื่อหรือประเภท เช่น Kappa Coefficient (17)

การศึกษาความเชื่อถือได้ประเภทนี้ควรทำการประเมินแบบเป็นอิสระต่อกัน (independent assessment) โดยผู้ประเมินแต่ละคนควรแยกกันสัมภาษณ์และให้คะแนน ในกรณีที่เป็นการประเมินร่วม (joint assessment) เช่น ผู้ประเมินคนที่ 1 สัมภาษณ์ผู้ป่วยและผู้ประเมินคนที่ 1 และ 2 (หรือมากกว่า) แยกกันให้คะแนนก็ยังถือว่าไม่เป็นอิสระต่อกันอย่างแท้จริง การประเมินร่วมในลักษณะนี้อาจทำให้ได้ค่าความสัมพันธ์สูงกว่าความเป็นจริงได้

นอกจากข้อจำกัดของ Pearson r และ Spearman rs ดังกล่าวแล้วการคำนวณในลักษณะนี้ยังมีข้อจำกัดเพิ่มเติมอีก คือ สามารถใช้ได้กับข้อมูลที่ได้จากผู้ประเมินเพียง 2 คนเท่านั้น

ค่าสัมประสิทธิ์ของความเชื่อถือได้ (reliability coefficient) สามารถแปลผลได้ดังนี้ น้อยกว่า 0.40 = ต่ำ, 0.40-0.59 = พอใช้, 0.60-0.74 = ดี และ 0.75-1.00 = ดีมาก (18)

2. ความเชื่อถือได้ประเภททำซ้ำ (repeat reliability) มาตรวัดสุขภาพที่ดีควรมีความคงที่ หากผู้ถูกประเมินมีสุขภาวะ (ที่ต้องการวัด) คงที่ คะแนนที่ได้จากผู้ถูกประเมินควรใกล้เคียงกันทุกครั้ง ความเชื่อถือได้ประเภทนี้สำคัญต่อมาตรวัดประเภทแบบสอบถามมาก วิธีที่ใช้บ่อย คือ การให้ผู้ถูกประเมินตอบแบบสอบถามมากกว่า 1 ครั้ง แล้วนำผลลัพธ์ (คะแนน) ที่ได้มาหาความสัมพันธ์กัน

วิธีการทางสถิติที่ใช้ในการคำนวนหาค่าสัมประสิทธิ์ของความเชื่อถือได้ประเภททำซ้ำ (repeat reliability coefficient) และการแปรผลจะคล้ายคลึงกับวิธีการหาค่าสัมประสิทธิ์ของความเชื่อถือได้ระหว่างผู้ประเทิน อย่างไรก็ตามเนื่องจากการ

ศึกษาความเชื่อถือได้ประเภททำซ้ำมักทำซ้ำไม่เกิน 1 ครั้ง ผู้ประเมินแต่ละรายจึงมักมีข้อมูลไม่เกิน 2 ชุด สถิติที่ใช้ในการหาความสัมพันธ์แบบเป็นคู่จึงอาจนำมาใช้ได้ เช่น Pearson r สำหรับข้อมูลประเภทอัตราส่วนหรือระยะห่าง, Spearman rs สำหรับข้อมูลประเภทลำดับ, Cohen's Kappa สำหรับข้อมูลชื่อหรือประเภท ผู้เชี่ยวชาญบางท่านมองว่ามาตรวัดที่จะนำมาใช้ทางคลินิกควรมีค่าความเชื่อถือได้ประเภททำซ้ำไม่น้อยกว่า 0.90 และมาตรวัดที่จะนำมาใช้ในการวิจัยควรมีค่าความเชื่อถือได้ประเภททำซ้ำไม่น้อยกว่า 0.70 (19)

ปัญหาที่มักเกิดขึ้นในการหาความเชื่อถือได้ประเภทนี้ คือ การทำอย่างไรให้เชื่อถือได้ว่า ผู้ถูกประเมินไม่สามารถจำคำตอบที่ให้ในครั้งก่อนได้ แม้ว่าจะไม่มีการศึกษาที่ชัดเจน แต่ผู้เชี่ยวชาญส่วนใหญ่ยอมรับว่า การตอบแบบทดสอบซ้ำที่ห่างกันอย่างน้อย 2 สัปดาห์เป็นสิ่งที่ยอมรับได้ในการศึกษาความเชื่อถือได้ประเภทนี้ (20)

ความถูกต้อง (validity)

ความถูกต้องเป็นหัวใจสำคัญของมาตรวัดสุขภาพ ไม่ว่ามาตรวัดจะมีความน่าเชื่อถือเพียงใด หากมาตรวัดนั้นไม่สามารถวัดภาวะทางสุขภาพที่ต้องการวัดได้อย่างถูกต้อง มาตรวัดนั้นไม่มีประโยชน์

ความถูกต้องด้านเนื้อหา (content validity)

ความถูกต้องด้านเนื้อหาเป็นการประเมินว่ารายข้อของมาตรวัดเกี่ยวข้องและครอบคลุมทุกด้านของภาวะทางสุขภาพที่ต้องการประเมิน ในการประเมินความถูกต้องประเภทนี้ ผู้พัฒนามาตรวัดมักแสดงให้เห็นว่ารายข้อที่ปรากฏได้มาอย่างไร เช่น มาตรวัดอื่นที่มีมาก่อน, การสัมภาษณ์ผู้ประเมินหรือผู้ถูกประเมิน, การสังเกตทางคลินิก, ความเห็นของผู้เชี่ยวชาญ, ผลการศึกษาวิจัยเดิม, แนวคิด, ทฤษฎีต่าง ๆ (19) ตัวอย่างที่เห็นได้จากการพัฒนา Medical Outcomes Study (MOS) 36-item short-form health survey หรือ SF-36 จะพบว่ารายข้อในมาตรวัดนี้มีพื้นฐานมาจากแนวคิดที่ว่าสุขภาพประกอบด้วย 8 ด้านที่สำคัญ (21)

แม้ว่าการประเมินความถูกต้องด้านเนื้อหาไม่จำเป็นต้องใช้สถิติที่ซับซ้อน แต่ถ้ารายข้อใดถูกตอบไปในทิศทางใดทิศทางหนึ่งมากเกินไป เช่น มากกว่าร้อยละ 90 รายข้อนั้นสมควรถูกตัดออก เนื่องจากเป็นรายข้อที่ไม่มีประโยชน์ในการแยกระหว่างผู้ที่มีและไม่มีภาวะทางสุขภาพที่ต้องการประเมิน (22)

ประเด็นย่อยประการหนึ่งของความถูกต้องด้านเนื้อหา คือ ความถูกต้องด้านหน้าตา (face validity) ซึ่งมาตรวัดที่ดีควรใช้ภาษาที่ชัดเจนและเข้าใจง่าย หลีกเลี่ยงการใช้คำเฉพาะสำหรับวิชาชีพ (jargon) ไม่ใช้คำถามกำกวม (เช่น การถามสองประเด็นในเวลาเดียวกัน) (19)

ความถูกต้องตามเกณฑ์ (criterion validity)

ผู้เชี่ยวชาญยังมีความเห็นเกี่ยวกับความถูกต้องตามการเกณฑ์แตกต่างกันพอควร Mokkink และคณะ มีมุมมองที่เฉพาะมาก โดยมองว่า ความถูกต้องตามเกณฑ์ของมาตรวัดที่สนใจควรถูกเปรียบเทียบกับมาตราวัดที่มีมาตรฐานสูงหรือมาตรฐานอ้างอิง (gold standard or reference standard) เท่านั้น (3)

อีกมุมมองที่ได้รับการยอมรับมากกว่า คือ การแบ่งความถูกต้องตามเกณฑ์ออกเป็น ความถูกต้องแบบเข้ากันได้ (concurrent validity) และความถูกต้องในการทำนาย (predictive validity) การศึกษาความถูกต้องทั้งสองแบบทำได้โดยการประเมินผู้ถูกประเมินด้วย 2 มาตรวัด แล้วหาค่าความ

สัมพันธ์หรือค่าเปรียบเทียบการทำนายของทั้งสองมาตราวัด ความถูกต้องทั้งสองประเภทนี้มีความแตกต่างที่สำคัญ คือ เวลาที่ใช้ในการประเมิน หากการประเมินโดยใช้มากกกว่า 1 มาตรวัดเกิดขึ้นในเวลาเดียวกัน จะเป็นการศึกษาความถูกต้องแบบเข้ากันได้ แต่หากมีการใช้มาตรวัดที่สนใจประเมินก่อน หลังจากนั้นเมื่อเวลาผ่านไปช่วงหนึ่งจึงใช้อีกผลลัพธ์หนึ่งมาประเมินผู้ถูกประเมินอีกครั้ง จะเป็นการศึกษาเพื่อหาความถูกต้องในการทำนาย

1. ความถูกต้องแบบเข้ากันได้ (concurrent validity) การศึกษาความถูกต้องแบบเข้ากันได้มักทำโดยการเปรียบเทียบคะแนนที่ได้จากมาตรวัดที่สนใจกับมาตรวัดที่มีอยู่เดิม โดยสองมาตรวัดใช้ประเมินภาวะสุขภาพเดียวกันในเวลาเดียวกัน มาตรวัดเดิมที่ถูกนำมาเปรียบเทียบต้องมีมาตรฐานสูง (gold standard) ในกรณีที่มาตรวัดที่นำมาใช้เปรียบเทียบมีมาตรฐานสูงมาก ค่าสัมประสิทธิ์ของความสัมพันธ์ (เช่น ค่า Pearson r หรือ Spearman rs) ระหว่างคะแนนที่ได้จากทั้งสองมาตรวัดก็ควรมีค่าสูงมาก (เช่น มากกว่า 0.80) แต่หากมาตรวัดที่นำมาใช้เปรียบเทียบมีมาตรฐานที่ไม่สูงมาก ค่าสัมประสิทธิ์ของความสัมพันธ์ระหว่างสองมาตรวัดก็ไม่จำเป็นต้องมีค่าสูงมาก (เช่น 0.30-0.80) แต่ค่าสัมประสิทธิ์ที่ต่ำกว่า 0.30 แสดงว่าคะแนนที่ได้จากมาตรวัดทั้งสองไม่สัมพันธ์กันเลย (19)

ตัวอย่างของการศึกษาความถูกต้องแบบเข้ากันได้ คือ การประเมินการซึมเศร้าในผู้ป่วย traumatic brain injury ของ Fann และคณะ (23) ค่า Pearson r ของคะแนนที่ได้จากของแบบสอบถาม PHQ-9 (7) กับคะแนนของ Hamilton Rating Scale for Depression (HRSD) (24) เท่ากับ 0.78 เนื่องจากการประเมินความรุนแรงของการซึมเศร้าไม่มีมาตรวัดที่มีมาตรฐานสูงมาก ค่า Pearson r = 0.78 จึงน่าพึงพอใจแล้ว ประกอบกับ PHQ-9 เป็นแบบสอบถามที่สะดวกในการใช้มากกว่า HRSD (ซึ่งต้องใช้บุคลากรแพทย์ในการประเมิน) แบบสอบถาม PHQ-9 จึงน่าจะยอมรับให้นำมาใช้ประเมินการซึมเศร้าในผู้ป่วย traumatic brain injury ได้

2. ความถูกต้องในการทำนาย การศึกษาความถูกต้องแบบทำนายสามารถทำได้โดยใช้มาตรวัดที่สนใจประเมินก่อน หลังจากผ่านไปช่วงหนึ่งแล้วจึงประเมินประชากรกลุ่มเดิมด้วยมาตรวัดที่สอง การศึกษาความถูกต้องประเภทนี้มักใช้กับมาตรวัดสุขภาพที่ให้ข้อมูลชื่อหรือประเภท การเปรียบเทียบค่าที่ได้ระหว่างมาตรวัดทั้งสองทำได้หลายรูปแบบดังตารางที่ 1

ตารางที่ 1. ความสัมพันธ์ระหว่างผลการทดสอบกับการมีหรือไม่มีโรค

ผลการทดสอบ	โรค		รวมทั้งหมด
	มี	ไม่มี	
ผลการทดสอบเป็นบวก	a (ผลบวกถูกต้อง)	b (ผลบวกเท็จ)	a + b
ผลการทดสอบเป็นลบ	c (ผลลบเท็จ)	d (ผลลบถูกต้อง)	c + d
รวมทั้งหมด	a + c	b + d	N

- ความไว แสดงให้เห็นว่า มาตรวัดสามารถระบุถึงการมีโรคได้ดีเพียงใด ผู้ที่ป่วยเป็นโรคมีโอกาสมากน้อยเพียงใดที่จะถูกระบุว่ามีโรค สูตรในการคำนวณ คือ a / (a + c)
- ความจำเพาะเจาะจง แสดงให้เห็นว่ามาตรวัดสามารถระบุถึงการไม่มีโรคได้ดีเพียงใด สูตรในการคำนวณ คือ d / (b + d)

ดัดแปลงจาก Gordis L (2016)(25)

- ตัวอย่างของการศึกษาความถูกต้องประเภทนี้ คือ การหาค่าความเสี่ยงสัมพัทธ์ที่ใช้ทำนายการตายของทารกแรกเกิดที่คลอดแบบครบเทอมซึ่งมีคะแนน Apgar ที่ 5 นาทีระหว่าง 0 ถึง 3 ซึ่ง Casey และคณะ พบว่าทารกแรกเกิดเหล่านี้มีโอกาสเสียชีวิตมากขึ้น 1,460 เท่า ผลการศึกษานี้แสดงให้เห็นว่า Apgar score เป็นมาตรวัดสุขภาพที่มีความถูกต้องในการทำนายสูง (25) (26)

ความถูกต้องด้านโครงสร้าง (structural validity)

ความถูกต้องด้านโครงสร้างมักถูกศึกษาในกรณีที่ไม่มีมาตรวัดเดิมที่มีมาตรฐานสูงและคล้ายคลึงกับมาตรวัดที่สนใจ (ทำให้ไม่สามารถศึกษาความถูกต้องตามเกณฑ์ได้) องค์ประกอบของความถูกต้องด้านการโครงสร้างอาจแบ่งได้เป็นความถูกต้องแบบแยกจาก (discriminant validity) และความถูกต้องแบบลู่เข้า (convergentce validity) รูปแบบของการศึกษาความถูกต้องทั้งสองประการนี้ คือ การแสดงให้เห็นถึงความต่างกันหรือความสัมพันธ์กันอย่างมีนัยสำคัญของมาตรวัดที่สนใจกับอีกมาตรวัดหนึ่งที่มีทฤษฎีหรือผลการวิจัยรองรับความต่างกันหรือความสัมพันธ์ดังกล่าวมาก่อนแล้ว ตามลำดับ

ตัวอย่างของการศึกษาความถูกต้องด้านการโครงสร้างคือ การศึกษาความถูกต้องแบบแยกจากที่พบว่า ผู้ที่มีคะแนน PHQ-9 (มาตรวัดการซึมเศร้า) สูง จะมีคะแนน SF-20 (มาตรวัดสุขภาวะ) ต่ำกว่าผู้ที่มีคะแนน PHQ-9 ต่ำ อย่างมีนัยสำคัญทางสถิติ นอกจากนี้ การศึกษาความถูกต้องแบบลู่เข้ายังพบอีกด้วยว่า คะแนนของ PHQ-9 สัมพันธ์กับสุขภาวะ 6 ด้านของ SF-20 อย่างมีนัยสำคัญทางสถิติ (7) การศึกษานี้ทำได้เพราะมีผลการวิจัยในอดีตมากมายรองรับว่าผู้ป่วยซึมเศร้ามีสุขภาวะแย่กว่าผู้ที่ไม่ซึมเศร้า

ความถูกต้องตามปัจจัย (factor validity) ก็ถือว่าเป็นประเด็นหนึ่งของความถูกต้องด้านการโครงสร้าง หลังพัฒนามาตรวัดขึ้นแล้ว ผู้พัฒนามักทำการศึกษาความถูกต้องตามปัจจัยเพิ่มเติมเพื่อสนับสนุนทฤษฎีที่สนับสนุนว่าภาวะสุขภาพดังกล่าวมีหลายมิติ (dimension) อย่างไรก็ตาม การวิเคราะห์ปัจจัย (factor analysis) ที่ใช้ในการศึกษาความถูกต้องประเภทนี้มีความแปรปรวนได้มาก เช่น โครงสร้างของปัจจัย (factor structure) ของ Childhood Trauma Questionnaire ในชายและหญิงมีความแตกต่างกัน (26, 27) นอกจากนี้ การวิเคราะห์ปัจจัยยังอาจพบว่ามาตรวัดไม่ได้มีมิติตรงตามทฤษฎีที่ใช้ในการพัฒนามาตรวัดก็ได้ เช่น Positive and Negative Syndrome Scale (PANSS) ที่พัฒนาขึ้นโดยใช้ทฤษฎีว่าอาการโรคจิต (psychotic symptoms) ของผู้ป่วยโรคจิตเภท (schizophrenia) มี 3 มิติ (27, 28) แต่ผลการวิเคราะห์ปัจจัยในภายหลังพบว่ามาตรวัดนี้มี 5 มิติ (28, 29)

การวิเคราะห์ปัจจัยนอกจากใช้พิสูจน์ทฤษฎีที่ใช้พัฒนามาตรวัดแล้วยังอาจนำมาใช้ศึกษาเพื่อตั้งสมมติฐานเกี่ยวกับการจัดกลุ่มรายข้อที่อยู่ในมาตรวัดก็ได้ เช่น Amphetamine Withdrawal Scale ถูกพัฒนาขึ้นโดยไม่ได้คำนึงถึงมิติของอาการถอนยาแอมเฟตามีน แต่เมื่อทำการวิเคราะห์ปัจจัยก็พบว่ามาตราวัดนี้อาจประกอบด้วย 3 มิติหรือมาตรวัดย่อย (29, 30)

ความไวต่อการเปลี่ยนแปลง (sensitivity to change)

ผู้เชี่ยวชาญบางท่านจัดให้ความไวต่อการเปลี่ยนแปลงเป็นอีกประเด็นหนึ่งของความถูกต้องด้านโครงสร้าง แสดงให้เห็นว่ามาตรวัดที่มีโครงสร้างดี ควรมีการเปลี่ยนแปลงของคะแนนอย่างรวดเร็ว

เมื่อภาวะทางสุขภาพที่ประเมินมีการเปลี่ยนแปลงไป เช่น ผลการศึกษาของ Wlodyka-Demaille และคณะ ที่พบว่า คะแนน Neck Pain and Disability Scale มีการเปลี่ยนแปลงเร็วที่สุดเมื่อผู้ป่วยรู้สึกว่าอาการปวดคอของตนเปลี่ยนไป (30, 31)

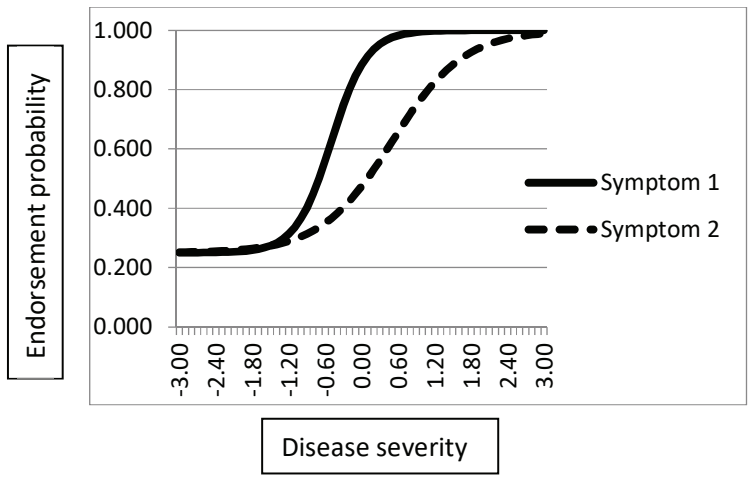
ทฤษฎีการตอบสนองรายข้อ (Item response theory)

ในขณะที่ทฤษฎีการทดสอบแบบดั้งเดิม (classical test theory) เน้นการศึกษาภาพรวมของมาตราวัดหรือมาตราวัดย่อยที่เป็นองค์ประกอบของมาตรวัดรวม ทฤษฎีการตอบสนองของรายข้อเน้นที่การศึกษารายข้อที่ประกอบขึ้นเป็นมาตรวัดจะเห็นได้ว่าทั้งสองทฤษฎีมีส่วนเสริมกันและกัน ทำให้บางมาตรวัดที่พัฒนาขึ้นในช่วงหลังมักใช้ทั้งสองทฤษฎีประกอบกันในการศึกษาคุณสมบัติของมาตรวัด

ทฤษฎีการตอบสนองรายข้อเริ่มถูกนำมาใช้ในงานวิจัยด้านการศึกษาตั้งแต่ช่วงปีค.ศ. 1970-1979 แนวคิดหลักสองประการของของทฤษฎีนี้ คือ คะแนนที่ได้จากข้อสอบแต่ละข้อควรสัมพันธ์กับคะแนนรวมที่ได้จากข้อสอบทั้งฉบับ และข้อสอบแต่ละข้อไม่ควรยากหรือง่ายเกินไป (เช่น ร้อยละ 90 ของนักเรียนตอบข้อสอบข้อนั้นถูกหรือผิด) หรือกล่าวอีกนัยหนึ่งว่า คะแนนที่ได้จากข้อสอบแต่ละข้อควรสามารถแยกนักเรียนที่ทำคะแนนรวมได้สูงออกจากนักเรียนที่ทำคะแนนรวมได้ต่ำ

หากนำทฤษฎีดังกล่าวมาใช้กับรายข้อในมาตรวัดย่อมหมายความว่า คะแนนของแต่ละรายข้อในมาตรวัดควรสัมพันธ์กับคะแนนรวมของมาตรวัดนั้น และในแต่ละข้อ ผู้ประเมินส่วนใหญ่ไม่ควรให้คะแนนสูงมากหรือต่ำมากเกินไปเช่น (เช่น ร้อยละ 90 ของผู้ประเมินให้คะแนนสูงมากหรือต่ำมากในรายข้อใดรายข้อหนึ่งของมาตรวัด)

นอกจากการนำทฤษฎีการตอบสนองรายข้อมาใช้ศึกษามาตรวัดดังกล่าวแล้ว ทฤษฎีนี้ยังถูกนำมาใช้ในการประเมินว่าผู้ถูกประเมินแต่ละกลุ่มมีแนว



รูปที่ 2. คุณสมบัติตามทฤษฎีการตอบสนองรายข้อ ทั้ง symptom 1 และ 2 แสดงคุณสมบัติเป็นรูปตัว S
Symptom 1 มีลักษณะ shift-to-the-left แสดงให้เห็นว่าผู้ป่วยที่มีอาการไม่รุนแรงก็มีแนวโน้มที่จะตอบว่าตนเองมีอาการนี้ นอกจากนี้ Symptom 1 ยังมีความชันมาก แสดงให้เห็นว่าอาการนี้ใช้แยกระหว่างผู้ที่ป่วยน้อยและผู้ที่ป่วยมากออกจากกันได้
Symptom 2 มีลักษณะ shift-to-the-right แสดงให้เห็นว่าผู้ป่วยที่ป่วยรุนแรงเท่านั้นจึงจะตอบว่าตนเองมีอาการนี้ นอกจากนี้ Symptom 2 ยังมีความชันต่ำ แสดงให้เห็นว่าอาการนี้ใช้แยกระหว่างผู้ที่ป่วยน้อยและผู้ที่ป่วยมากได้ไม่ดี

โน้มที่จะให้คำตอบที่แตกต่างกันหรือไม่อีกด้วย การประเมินในลักษณะนี้มีชื่อเรียกเฉพาะว่า Differential Item Functioning (DIF) ซึ่งสามารถคำนวณได้หลายวิธี (31) (32) หลักการสำคัญของ DIF ที่ใช้ในการวิเคราะห์ข้อสอบ คือ นักเรียนที่มีความสามารถใกล้เคียงกัน (ได้คะแนนรวมของข้อสอบชุดนั้นใกล้เคียงกัน) ไม่ควรตอบข้อคำถามใดคำถามหนึ่งแตกต่างกันอันเนื่องมาจากลักษณะอื่นเช่น เพศ เชื้อชาติ ศาสนา เมื่อนำหลักการนี้มาใช้ในการประเมินแต่ละรายข้อของมาตรวัด จะเห็นได้ว่าผู้ที่มีคะแนนรวมใกล้เคียงกันไม่ควรตอบรายข้อใดรายข้อหนึ่งในมาตรวัดแตกต่างกันอันเนื่องจากมากลักษณะอื่น เช่น (เพศ เชื้อชาติ ศาสนา)

ตัวอย่างของการใช้ DIF ในการศึกษาคุณสมบัติของ PHQ-9 ในผู้ป่วยโรคซึมเศร้าชาวอเมริกัน พบว่าผู้ป่วยเชื้อสายจีนมีแนวโน้มที่จะมีปัญหาการนอนและการเคลื่อนไหวมากกว่าผู้ป่วยผิวขาว (32, 33) ผลการศึกษานี้อาจแสดงให้เห็นว่าการมี 2 รายข้อนี้ใน PHQ-9 จะทำให้ผู้ป่วยเชื้อชาติจีนที่ถูกประเมินด้วย PHQ-9 มีแนวโน้มที่จะมีอาการรุนแรงกว่าผู้ป่วยผิวขาวได้

ในปัจจุบัน มีบทความจำนวนไม่น้อยที่กล่าวถึงทฤษฎีการตอบสนองของรายข้อ ซึ่งผู้ที่สนใจสามารถศึกษาเพิ่มเติมได้ (34–36)

สรุป

บทความนี้ได้กล่าวถึงประเด็นสำคัญที่ควรพิจารณาก่อนนำมาตรวัดสุขภาพไปใช้ในเวชปฏิบัติหรือการวิจัย ผู้ใช้ควรให้ความสำคัญกับภาพรวมของมาตรวัดก่อน หลังจากนั้นจึงใช้ทฤษฎีการทดสอบแบบดั้งเดิมในการพิจารณาความเชื่อถือได้และความถูกต้องของมาตรวัด ความรู้เกี่ยวกับทฤษฎีการตอบสนองรายข้อสามารถนำมาใช้เสริมทฤษฎีการทดสอบแบบดั้งเดิมได้และจะช่วยให้เห็นถึงคุณสมบัติของแต่ละรายข้อของมาตรวัดได้ดีขึ้น เมื่อทราบคุณสมบัติดังกล่าวมาแล้ว ผู้ใช้ย่อมสามารถตัดสินใจเลือกมาตรวัดที่เหมาะกับงานของตนเองได้

การขัดกันของผลประโยชน์

ไม่มี

เอกสารอ้างอิง

1. McDowell I. Measuring Health: A Guide to Rating Scales and Questionnaires. 3rd ed. Oxford, UK: Oxford University Press; 2006.
2. Streiner DL, Norman G R, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, UK: Oxford University Press; 2015.
3. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN Ckecklist Manual [Internet]. [cited 2017 Feb 26]. Available from: www.cosmin.nl
4. Üstün TB, Chatterji S, Kostanjsek N, Rehm J, Kennedy C, Epping-Jordan J, et al. Developing the World Health Organization Disability Assessment Schedule 2.0. Bull World Health Organ. 2010;88:815–23.
5. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, 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.
6. Apgar V. A proposal for a new method of evaluation of the newborn infant. Curr Res Anesth Analg. 1953;32:260–7.
7. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity meas-

ure. J Gen Intern Med. 2001;16:606–13.
8. Myles PS, Troedel S, Boquest M, Reeves M. The pain visual analog scale: is it linear or nonlinear? Anesth Analg. 1999;89:1517–20.
9. Bernstein IH, Lacritz L, Barlow CE, Weiner MF, DeFina LF. Psychometric evaluation of the Montreal Cognitive Assessment (MoCA) in three diverse samples. Clin Neuropsychol. 2011;25:119–26.
10. Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. Psychometrika. 2009;74:107–20.
11. Loewenthal K, Eysenck MW. An Introduction to Psychological Tests and Scales. 2nd ed. Philadelphia: Psychology Press; 2001. 184 p.
12. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78:98-104.
13. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80:99–103.
14. Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. J Clin Epidemiol. 1991;44:381–90.
15. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6: 284–90.
16. Jakobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci. 2005;19:427–31.
17. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85:257-68.
18. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic. 1981;86:127–37.
19. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res. 2010;68:319-23.
20. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63:737–45.
21. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30:473–83.
22. Streiner DL. A checklist for evaluating the usefulness of rating scales. Can J Psychiatry Rev Can Psychiatr. 1993;38:140–8.
23. Fann JR, Bombardier CH, Dikmen S, Esselman P, Warms CA, Pelzer E, et al. Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. J Head Trauma Rehabil. 2005;20:501–11.
24. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56–62.
25. Gordis L. Epidemiology. 5th ed. Philadelphia, PA: Elsevier Saunders; 2016.
26. Casey BM, McIntire DD, Leveno KJ. The Continuing Value of the Apgar Score for the Assessment of Newborn Infants. N Engl J Med. 2001;344:467–71.
27. Wright KD, Asmundson GJ, McCreary DR, Scher C, Hami S, Stein MB. Factorial validity of the Childhood Trauma Questionnaire in men and women. Depress Anxiety. 2001;13:179–83.
28. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1987;13:261–76.
29. Marder SR, Davis JM, Chouinard G. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: Combined results of the North American trials. J Clin Psychiatry. 1997;58:538–46.
30. Srisurapanont M, Jarusuraisin N, Jittiwutikan J. Amphetamine withdrawal: I. Reliability, valid-

ity and factor structure of a measure. Aust N Z J Psychiatry. 1999;33:89–93.
31. Wlodyka-Demaille S, Poiraudeau S, Catanzariti J-F, Rannou F, Fermanian J, Revel M. The ability to change of three questionnaires for neck pain. Jt Bone Spine Rev Rhum. 2004;71: 317–26.
32. Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. Med Care. 2006;44:S152-170.
33. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J Gen Intern Med. 2006; 21:547–52.
34. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38:II28-42.
35. Hays RD, Morales LS, Reise SP. Item Response Theory and Health Outcomes Measurement in the 21st Century. Med Care. 2000;38:II28-II42.
36. Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol. 2009;5:27–48.

Selecting health measurement scales: basic issues for considerations

Manit Srisurapanont

Department of Psychiatry, Faculty of Medicine, Chiang Mai University

At present, people have more needs on health. It is, therefore, necessary for medical professionals to elaborately measure or evaluate health. Health measurement scales for evaluating health abstract outcomes, eg, beliefs, feeling, attitudes, and behavior, has been more and more involved in health assessment. Health professionals, therefore, need competency in selecting health measurement scales, including the feasibility and cost-effectiveness of assessment. In doing so, they will be able to convert such abstract outcomes into scores, which can be further computed and communicated. This article describes key issues for consideration in applying health measurement scales in clinical practice or research. Scale users should, firstly, consider the scale overview and, then, apply the classical test theory to determine the scale reliability and validity. Knowledge on item response theory can be additionally used with classical test theory and will be helpful to determine item properties. After all properties are taken into account, scale users would be able to choose the scales appropriate for their work. **Chiang Mai Medical Journal 2017;56(1):49-31.**

Keywords: measurement, assessment, reliability, validity