การวัดทางจิต

ความหมาย

การวัดทางจิต (Psychometrics) หรืออีกนัยหนึ่งใช้ชื่อว่า จิตมาตร หมายถึง การวัดลักษณะที่เกี่ยวกับจิตใจ เช่น การวัดความรู้ ความสามารถ เจตคติ ลักษณะบุคลิกภาพ และอื่นๆ ที่เกี่ยวข้อง โดยใช้การพรรณนาในเชิงของตัวเลข และเชิงปริมาณ ประกอบด้วยการวัดตัวแปรทางจิต และรูปแบบเชิงคณิตศาสตร์ของการทดสอบ และการวัด

ความเป็นมา

ประวัติช่วงต้นๆของการวัดทางจิตเดินตามเส้นทาง 2 สาย (Guilford, 1954, p. 2) คือ สายแรกเป็น psychophysical tradition ซึ่งเป็นขั้นเริ่มต้นและนำไปสู่จิตวิทยาเชิงทดลองที่แท้จริง สายที่ 2 เป็น mental test tradition ซึ่งมีความแตกต่างระหว่างบุคคลเป็นจุดศูนย์กลางของความสนใจ ดังมีเรื่องราวโดยย่อดังนี้

กลุ่มแรก พัฒนามาจาก สรีรวิทยาเชิงทดลอง (experimental physiology) และวิธีการเชิงปริมาณ (quantitative methods) ที่เติบโตขึ้นโดยเชื่อมโยงกับวิทยาศาสตร์ธรรมชาติ งานพื้นฐานของ psychophysics คือ ความรู้เกี่ยวกับการรับสัมผัสความรู้สึก(sensory) กล่าวคือ เฮอร์บาร์ท (Herbart) (1776-1841) (อ้างอิงจาก Guilford, 1954, p. 3) ได้เสนอ concept of an absolute threshold หรือ lower limit of sensation ส่วน เว็บเบอร์ (Weber) (1795-1878) (อ้างอิงจาก Guilford, 1954, p. 3) ได้เสนอหลักการเกี่ยวกับ just noticeable difference (jnd) จากกฎของเว็บเบอร์นี้ เฟคเนอร์ (Fechner) ได้สร้าง psychophysical relationship ขึ้นมา โดยเขานิยาม psychophysics ว่า "ศาสตร์ของความสัมพันธ์ระหว่างกายกับจิต"(functional relations of dependency between body and mind) (Guilford, 1954, p.3) แนวคิดนี้นอกเหนือจากการวัดขนาดของความรู้สึกสัมผัสยังได้มีการนำไปใช้ กับการวัดขนาดของการรับรู้ (perception) ความรู้สึก (feeling) การกระทำ (action) และความใส่ใจ(attention) ซึ่งสิ่งเหล่านี้ก็เป็นกระบวนการทางจิตที่สามารถสัมพันธ์กับสิ่งเร้าได้นั่นเอง เฟคเนอร์ ได้พัฒนาการศึกษาเกี่ยวกับความรู้สึกสัมผัสไปจนกลายเป็นพื้นฐานของ psychophysical methods นั่นคือ the method of average error, the method of minimal changes, และ the constant method ส่วน Thurstone ก็ได้อาศัยแนวคิดนี้มาพัฒนา Thurstone's law of comparative judgment

กลุ่มที่สอง กลุ่มนี้ได้รับแรงดลใจจากวิวัฒนาการในชีววิทยาเกี่ยวกับลักษณะของพันธุกรรมซึ่งยืมโดยตรงมาจากนักคณิตศาสตร์ผู้ที่อาศัยปัญหาของความน่าจะเป็นและสร้างวิธีการทางสถิติขึ้นมา เซอร์ ฟรานซิส แกลตัน (Sir Francis Galton) (1822-1911)(อ้างอิงจาก Guilford, 1954, p.3) ได้สร้างห้องทดลองทาง anthropometric ซึ่งมีอุปกรณ์การทดสอบระหว่างการรับความรู้สึก (sensory) กับกลไกการเคลื่อนไหว (motor) เขาได้สร้างเครื่องมือทางสถิติมากมายด้วยความ

ช่วยเหลือของ คาร์ล เพียร์สัน (Karl Pearson) เช่น วิธีหาความสัมพันธ์(correlation) การใช้คะแนนมาตรฐาน(standard scores) และวิธีการสร้างมาตราทางจิตวิทยา(psychological scaling method) ดังเช่น rating scale method บุคคลที่มีชื่อเสียงเด่นในด้านของสถิตินอกจาก คาร์ล เพียร์สัน ยังมี อาร์. เอ. ฟิชเชอร์ (R. A. Fisher)

ส่วนด้านของการพัฒนาแบบทดสอบ มีบุคคลที่เด่นๆดังนี้

ชาร์ลส์ เอ็ดเวอร์ด สเปียร์แมน (Charles Edward Spearman) (1863-1945) นักจิตวิทยาชาวอังกฤษ เป็นที่รู้จักกันในเรื่องสถิติ และ human intelligence (g-factor)

เอ็ดเวิร์ด ลี ธอร์นไดค์ (Edward Lee Thorndike (1874-1949) นักจิตวิทยาชาวอเมริกันมีผลงานเกี่ยวกับกระบวนการเรียนรู้มากมาย เป็นประธานคนที่ 2 ของ Psychometric Society คนแรก คือ หลุยส์ ลีออน เธอร์สโตน (Louis Leon Thurstone)

เลวิส แมดิสัน เทอร์แมน (Lewis Madison Terman) (1877-1956) นักจิดวิทยาชาวอเมริกันผู้สร้าง แบบทดสอบวัดสติปัญญา(Stanford-Binet IQ test) โดยพัฒนา (แปล ปรับปรุง และสร้างข้อคำถามใหม่) มาจาก Binet-Simon Intelligence Scale ซึ่งเป็นภาษาฝรั่งเศส

หลุยส์ ลีออน เธอร์สโตน (Louis Leon Thurstone) (1887-1955) นักจิดวิทยาชาวอเมริกันมีผลงานเกี่ยวกับการวิเคราะห์องค์ประกอบ(factor analysis) และสร้างรูปแบบสติปัญญาที่เรียกว่าPrimary Mental Abilities (PMAs)ซึ่งประกอบด้วย 7 ตัวแปร เขาเป็นผู้ก่อตั้ง Psychometric Society และวารสารวิชาการของสมาคมนี้คือ Psychometrika ซึ่งยังคงมีการพิมพ์อยู่ในปัจจุบัน

เรมอน เบอร์นาร์ด แคทเทล (Raymond Bernard Cattell) (1905-1998) นักจิตวิทยาชาวอังกฤษและอเมริกันผู้ค้นพบว่าบุคลิกภาพของบุคคลมี 16 องค์ประกอบโดยการใช้การวิเคราะห์องค์ประกอบ เรียกเครื่องมือวัดบุคลิกภาพนี้ว่า 16 personality factor model หรือ 16PF Questionnaire

อัลเฟรด บิเนท์ (Alfred Binet) (1857-1911) นักจิตวิทยาชาวฝรั่งเศส เขาได้ร่วมมือกับธีโอดอร์ ไซมอน (Theodore Simon) สร้างเครื่องมือที่เป็นปรนัยวัดความบกพร่องทางสมอง (mental retardation) ที่เรียกว่า Binet – Simon Intelligence Test ซึ่งพิมพ์ในปี 1905 และฉบับปรับปรุงพิมพ์ในปี 1908 และ 1911 ฉบับใหม่นี้ขยายรวมไปถึงคนปกติด้วย ฉบับแปลเป็นภาษาอังกฤษพิมพ์ในปี ค.ศ.1916 (พ.ศ.2459) หรือชื่อเต็มคือ Stanford Binet Intelligence Scales ได้มีการใช้และปรับปรุงมาเป็นลำดับ ฉบับปรับปรุงสุดท้ายคือครั้งที่ 5 พิมพ์ปี ค.ศ.2003 (พ.ศ.2546) ผู้ที่สนใจความเป็นมาและสิ่งที่ปรับปรุงสามารถอ่านได้จาก Becker (2003)

ความเชื่อมโยงระหว่างการทดสอบทางจิตและจิตวิทยาเชิงทดลอง (mental testing and experimental psychology)

จากเส้นทางสองสายดังกล่าวแล้วจะเห็นได้ว่างานทางสาขาสรีรวิทยาและการทดสอบทางจิต มีความเชื่อมโยงกันโดยตรงอย่างใกล้ชิด นั่นคือจิตวิทยาเชิงทดลองซึ่งเชื่อมโยงอย่างใกล้ชิดกับการทดลองในสรีรวิทยา นักจิตวิทยาเชิงทดลองเองก็ได้ใช้การทดสอบทางจิตเป็นเครื่องมือวัดอยู่แล้ว ซึ่งก็เป็นการวัดความแตกต่าง (ของตัวแปรหนึ่ง ๆ) ในบุคคลคนเดียวกัน แม้ว่านักจิตวิทยาเชิงทดลองทำการวัดในเทอมของตัวแปรทางกายภาพเป็นจำนวนมาก (เช่น ลักษณะกายภาพของทั้งสิ่งเร้าและการตอบสนองในระบบ เซนติเมตร กรัม วินาที) อย่างไรก็ดีการวัดของนักจิตวิทยาเชิงทดลองส่วนมากออกมาในรูปของคะแนนจากการทดสอบ เช่น การทดลองเกี่ยวกับการเรียนรู้ ความจำ การจูงใจ และการคิด มักให้คะแนนในเทอมของจำนวนคำตอบถูก จำนวนความคลาดเคลื่อน จำนวนข้อในรายการ เป็นต้น และใช้สถิติในการทดสอบนัยสำคัญเพื่อสรุปผลจากข้อมูล ซึ่งเป็นการใช้เหตุผลเชิงสถิติและคณิตศาสตร์เป็นฐานสำหรับการวัดทางจิต

วิธีการสร้างมาตราวัดทางจิต (The psychological scaling methods)

มาตราวัดทางจิตที่ได้รับการสร้างและพัฒนามีหลายชนิด ดังเช่น การเปรียบเทียบเป็นคู่ ๆ (pair comparison) การจัดอันดับ (ranking method – order of merit) มาตราจัดอันดับ (rating scales) มาตราช่วงเท่ากัน (equal – appearing intervals) และมาตราอื่น ๆ ที่ปรับเปลี่ยนรูปแบบไปจากที่ได้กล่าวแล้ว ได้ช่วยให้พบจุดร่วมกันของ psychophysics และการทดสอบทางจิต เป้าหมายหลักของทั้งสองฝ่ายคือการประเมินวัตถุที่เป็นสิ่งเร้าบนมาตราเชิงเส้นตรง (linear scale) ดังเช่น 1) ค่านิยมเชิงความรู้สึก (affective values)หรือความเชื่อ 2) คุณภาพของลายมือที่เขียน การวาดภาพ การเขียนเรียงความ 3) ลักษณะบุคลิกภาพดังเช่น ความเป็นผู้นำ tactfulness หรือ การเข้าสังคม (sociability) เป็นต้น ในกรณีเหล่านี้ไม่มีการประเมินเชิงกายภาพ (physical) ของสิ่งเร้า แต่วิธีการสร้างมาตราจำนวนมากก็มีพื้นฐาน จุดตั้งต้นมาจาก psychophysics

ประโยชน์ของวิธีการสร้างมาตราในปัญหาทางการศึกษา โดยเฉพาะอย่างยิ่งในลักษณะที่เป็นระบบและเป็นปรนัย ที่ช่วยให้การตัดสินใจเกี่ยวกับแต่ละบุคคลมีความถูกต้อง ทำให้นักทดสอบทางจิตพอใจ เนื่องจากวิธีนี้เป็นประโยชน์ในการตรวจสอบความเที่ยงตรง (validity) ของแบบทดสอบ และใช้ในการประมาณลักษณะ (traits) เมื่อยังไม่มีการทดสอบที่เป็นที่ยอมรับใช้กันอยู่ ดังนั้นวิธีการสร้างมาตราจึงกล่าวได้ว่าเป็นวิธีการร่วมของ psychophysics และการทดสอบทางจิต โดยฝ่ายแรกเกื้อหนุนในรูปแบบของหลักเหตุผล และหลักคณิตศาสตร์ และกระตุ้นให้มีการใช้ในจิตวิทยาเชิงทดลอง ฝ่ายหลังเกื้อหนุน

ในข้อมูลเชิงประจักษ์ (empirical information) จากการใช้ในการศึกษาและในปัญหาของความแตกต่างระหว่างบุคคล

มาตราวัดทางจิตที่สำคัญมีดังต่อไปนี้

Method of average error เป็นวิธีการทางจิตวิทยาที่เก่าและเป็นพื้นฐานที่สุด เป้าหมายของวิธีการนี้คือเพื่อหาสิ่งเร้าที่เท่ากันโดยการให้ผู้สังเกตเป็นผู้ปรับ

Method of pair comparisons ในวิธีนี้ สิ่งเร้าทั้งหมดถูกประเมินบนมาตราทางจิต (psychological scale) สิ่งเร้าเหล่านี้ถูกนำเสนอให้กับผู้ตอบ (O) ในลักษณะเป็นคู่ ๆ ทุกคู่ที่เป็นไปได้ (all possible pairs) ผู้ตอบตัดสิน (judge) ว่า 1 ใน 2 อย่างนั้น อันไหนมากกว่าในมิติที่กำหนด

Method of rank order เป็นวิธีที่ได้รับความนิยมมากและใช้ประโยชน์ได้มากที่สุด เนื่องจากความง่ายที่สิ่งเร้าจำนวนมากสามารถถูกประเมินโดยอ้างอิงซึ่งกันและกัน และใช้ได้กว้างขวาง ให้ผู้ตอบเรียงลำดับสิ่งเร้าทั้งหลาย ค่าที่ได้ (scale values) อยู่ในมาตราช่วงเท่ากัน

Equal – appearing intervals (ช่วงเท่ากัน) ในวิธีนี้มีการกำหนดสิ่งเร้า เช่น ลายมือหรือการวาดรูปจำนวนหนึ่ง (โดยปกติมีจำนวนมาก) ให้กับผู้สังเกต (O) แล้วให้ O แบ่งสิ่งเร้าออกเป็นกลุ่ม ๆ (piles) โดยให้แต่ละกลุ่มที่อยู่ติดกันดูเหมือนห่างเท่า ๆ กัน เธอร์สโตน (Thurstone) เป็นผู้เสนอให้ใช้มาตรานี้เป็นมาตราวัดเจตคติ

Rating scales ในจำนวนวิธีการวัดทางจิตวิทยาที่ขึ้นอยู่กับการตัดสินของบุคคล (หรือผู้ตอบ) นั้น กระบวนการ rating scale เป็นที่นิยมใช้กันมากที่สุด รูปแบบของ rating scales ที่ใช้กันเป็นปกติมี 5 ประเภท คือ ตัวเลข (numerical) กราฟ (graphic) มาตรฐาน (standard) คะแนนสะสม (cumulated points) และตัวเลือกที่บังคับ (forced choice)

ทฤษฎีทั่วไปของการวัด (general theory of measurement)

นิยามของการวัดที่อ้างอิงกันบ่อยมากและใช้กันอยู่จนถึงปัจจุบันเป็นความคิดของแคมป์เบลล์) (Campbell, 1940) (Guilford, 1954, p. 5) ที่ว่าเป็นการกำหนดตัวเลขให้กับวัตถุหรือเหตุการณ์ตามกฎ (the assignment of numerals to objects or events according to rules) คำว่า "numerals" นี้เป็นเพียงสัญลักษณ์ที่ใช้บอกประเภทหรือกลุ่มเท่านั้น เช่น กลุ่ม 1 หรือ กลุ่ม 2 เป็นต้น นักวิชาการบางคนใช้คำว่า "numbers" แทน "numerals" ซึ่งมีความหมายครอบคลุมตัวเลขที่มีความหมายดังที่ใช้กันในวิชาคณิตศาสตร์

การวัด (measurement) มีความสัมพันธ์ใกล้ชิดกับคณิตศาสตร์ (mathematics) มาก เราจะเข้าใจธรรมชาติของการวัดไม่ได้ถ้าไม่รู้คุณสมบัติของคณิตศาสตร์ จากกฎเกณฑ์ในวิชาคณิตศาสตร์โมเดลทางคณิตศาสตร์ (mathematical models) สามารถให้โมเดลที่สะดวกและมีประโยชน์ต่อการบรรยายธรรมชาติ แม้ว่าธรรมชาติจะไม่

สามารถบรรยายได้อย่างถูกต้องได้ด้วยโมเดลทางคณิตศาสตร์ มันเป็นเพียงการประมาณซึ่งบางครั้งก็ดีบางครั้งก็ไม่ค่อยดี ความพอดีหรือพอเหมาะนี้สามารถทดสอบได้และถ้าพบว่ายอมรับได้ การหาข้อสรุปและการทำนายธรรมชาติโดยโมเดลเชิงคณิตศาสตร์จึงเป็นประโยชน์อย่างยิ่ง ซึ่งการหาข้อสรุปและการทำนายจะมีความคลาดเคลื่อนเล็กน้อยมาก โมเดลทางคณิตศาสตร์ที่ใช้กันบ่อยที่สุดคือ โมเดลโค้งปกติ ซึ่งเป็นรูปแบบการกระจายหรือการแจกแจงของข้อมูลที่มีลักษณะคล้ายระฆังคว่ำ

ระดับของการวัดโดยทั่วไปมี 4 ระดับ (four general levels of measurement)

ตัวเลขที่ถูกกำหนดให้กับวัตถุหรือเหตุการณ์มีหลายระดับ เรียกว่าระดับของการวัด

ระดับของการวัดโดยทั่วไปแบ่งออกเป็น 4 ระดับตามสตีเวนส์ (Stevens) (อ้างอิงจาก Guilford, 1954 : 11) จากระดับต่ำสุดไปถึงสูงสุดคือ 1. นามบัญญัติ (nominal) 2. อันดับ (ordinal) 3. ช่วงเท่ากัน (interval) และ 4. อัตราส่วน (ratio) ระดับทั้งสี่ นี้จำแนกออกตามเกณฑ์ของ "กฎ" ในนิยามของคำว่า "การวัด" ตัวเลขในระดับการวัดที่สูงขึ้น จะมีเกณฑ์ที่มีข้อจำกัดมากขึ้น เราสามารถจัดกระทำกับตัวเลขในเชิงคณิตศาสตร์/สถิติได้มากขึ้น สำหรับรายละเอียดเรื่องนี้ผู้อ่านสามารถอ่านได้จากตำราเกี่ยวกับการวัดผลทั่ว ๆ ไปหรือดูจากกิลฟอร์ด (Guilford) (1954 : 11 – 17) ในที่นี้ขอยกตัวอย่างการวัดในระดับต่าง ๆ โดยย่อดังนี้

1. มาตรานามบัญญัติ (nominal scale) ตัวเลขในระดับนี้สามารถบอกกลุ่มหรือการจำแนกประเภทเท่านั้นเช่น กลุ่ม 1 กลุ่ม 2 เป็นต้น การจัดกระทำทางสถิติที่สามารถใช้ได้กับตัวเลขระดับนี้มีน้อยมาก เช่น การแจกแจงความถี่ ฐานนิยม (mode) และสามารถคำนวณ coefficient of contingency เพื่อบอกความเกี่ยวข้องกันของการจำแนก 2 แบบได้

2. มาตราอันดับ (ordinal scale) ตัวเลขในระดับนี้สามารถบอกอันดับ เช่น จัดอันดับตามน้ำหนัก ความดัง เป็นต้น สถิติที่ใช้กับตัวเลขในระดับนามบัญญัติสามารถนำมาใช้ในระดับนี้ได้ นอกจากนั้นยังสามารถใช้สถิติต่อไปนี้ได้ด้วย ได้แก่ มัธยฐาน (median) เปอร์เซ็นไทล์ และ rank – order coefficient of correlation

3. มาตราช่วงเท่ากัน (interval scale or equal – unit scale) ตัวเลขในระดับนี้บอกช่วงเท่ากัน นั่นคือ ตัวเลขที่แสดงระยะทางที่เท่ากันแทนระยะทางเท่ากันในมิติใดมิติหนึ่งของวัตถุ แต่มาตรานี้มีจุดตั้งต้น (ศูนย์) ไม่แน่นอน ดังนั้นการบวกตัวเลขจึงไม่มีความหมายมากนัก แต่ระยะทางสามารถนำมาบวกกันได้ ตัวอย่างมาตราวัดในระดับนี้คือ องศาเซนติเกรต (เซลเซียส) หรือฟาเรนไฮต์ เป็นต้น สถิติแทบทั้งหมดใช้ได้กับตัวเลขในมาตรานี้ เช่น ค่าเฉลี่ย (mean) ส่วนเบี่ยงเบนมาตรฐาน (standard deviation) สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson product – moment r)เป็นต้น ยกเว้นสัมประสิทธิ์ของความแปรปรวน (coefficient of variation)

4. มาตราอัตราส่วน (ratio scale) ตัวเลขในมาตรานี้มีศูนย์แท้ (absolute zeros, true zero) ซึ่งหมายถึงไม่มีปริมาณที่วัดนั้น ๆ โดยแท้จริง สัดส่วนของตัวเลขสามารถเทียบเคียงกันได้ เช่น $\frac{1}{2} = \frac{10}{20}$ สถิติทั้งหมดสามารถใช้ได้กับตัวเลขบนมาตรานี้ รวมทั้งสัมประสิทธิ์ของความแปรปรวน

ผลการวัดที่ได้จากการนับจำนวนของสิ่งต่างๆ เป็นตัวเลขบนมาตราอัตราส่วน เนื่องจากมีศูนย์แท้และสัดส่วนของความถี่เป็นสิ่งที่มีความหมาย เช่น คะแนนจากแบบทดสอบที่นับมาจากจำนวนคำตอบถูก ย่อมเป็นตัวเลขบนมาตราอัตราส่วนถ้าเรายังคงอยู่ในความหมาย "จำนวนของข้อที่ตอบถูก" แต่ถ้าตัวเลขนี้ถูกใช้เพื่อแทนตำแหน่งของคนในเรื่องความสามารถหรือคุณลักษณะใด ๆ (trait) ตัวเลขนี้ย่อมสูญเสียความเป็นมาตราอัตราส่วนหรือแม้แต่มาตราช่วงเท่ากัน อย่างไรก็ดีมีหลักฐานแสดงว่าคะแนนนี้เข้าใกล้มาตราช่วงเท่ากันได้เมื่อแบบทดสอบมีความยาวเพียงพอและข้อต่างๆ (items) มีความยาก (difficulty) ที่มีการกระจายที่ดี (Guilford, 1954 : 86)

ชนิดของมาตราวัดในการทดสอบ (types of test scales)

การวัดตัวแปรเชิงจิตวิทยาด้วยคะแนนด้านกายภาพ (physical measures of psychological variables)

แบบทดสอบที่ให้ผลการวัดบนมาตราด้านกายภาพ เช่น การวัดเวลา ตัวอย่าง เด็กคนหนึ่งทำข้อสอบฉบับหนึ่งเสร็จในเวลา 20 นาที อ่านข้อความหนึ่งเสร็จในเวลา 5 นาที เป็นต้น มาตราที่วัดเป็นหน่วยด้านกายภาพ (เวลา) และผลการวัดอยู่ในมาตราอัตราส่วน (ratio scale) แต่ผลการวัดเช่นนี้ยังเป็นปัญหาเนื่องจากคะแนนที่ได้นี้ (เป็นคำตอบกายภาพ) ถูกใช้เป็นปริมาณที่แสดงผลงานในเชิงจิตวิทยา (ความสามารถในเชิงจิตวิทยา) คะแนนนี้อาจไม่ได้แทนหน่วยเชิงจิตวิทยาที่เท่ากันหรือไม่มีศูนย์แท้ที่มีความหมาย ไม่มีหลักฐานที่แสดงว่าการเปลี่ยนแปลงของหน่วยทางกายภาพจะมีความหมายเหมือนการเปลี่ยนแปลงของหน่วยในทางจิตวิทยา

อีกตัวอย่างหนึ่งของการใช้มาตราเวลาในทางจิตวิทยาคือ มาตราอายุสมอง (mental – age scale) ซึ่งได้รับการวิพากษ์วิจารณ์มากในเรื่องความเหมาะสมในการวัด แม้ว่าจะมีการใช้กันมาเป็นเวลานาน

คะแนนในทางจิตวิทยาที่นำมารวมกัน (summational psychological scores)

ชนิดของคะแนนทดสอบที่ใช้กันทั่วไปคือ การรวมคะแนนรายข้อ วิธีการคือการรวมจำนวนการตอบสนองตามเกณฑ์ที่ได้กำหนดไว้ เช่น การตอบถูก หรือจำนวนการตอบผิด น้ำหนักรายข้อจะเท่ากันหรือไม่ก็ไม่สำคัญ หลักฐานยังไม่เป็นอันยุติว่าคะแนนจากการรวมคะแนนรายข้อนี้ให้คะแนนในมาตราช่วงเท่ากัน อย่างไรก็ดีเมื่อกลุ่มตัวอย่างได้มาอย่างสุ่มและคะแนนของลักษณะที่

วัดในกลุ่มประชากรมีการกระจายเป็นแบบปกติ (normal distribution) เราสามารถจัดกระทำกับข้อมูลประหนึ่งว่าอยู่ในมาตราช่วงเท่ากันได้ โดยเฉพาะเมื่อความยาวของแบบทดสอบมีมากพอ คืออย่างน้อย 30 – 40 ข้อ

เมื่อคะแนนที่ได้มีการแจกแจงไม่เป็นปกติอย่างที่น่าจะเป็น คะแนนอาจมีช่วงไม่เท่ากัน เราสามารถดำเนินตามกระบวนการปรับมาตราเพื่อให้คะแนนจากแบบทดสอบเข้าใกล้หรือกลายเป็นคะแนนที่มีช่วงเท่ากัน มาตราวัดที่ใช้กันทั่วไป คือ T – scale ซึ่งมีการแจกแจงของคะแนนเป็นแบบปกติ โดยมีค่าเฉลี่ย 50 และส่วนเบี่ยงเบนมาตรฐาน 10

ทฤษฎีของคะแนนแบบทดสอบ (theory of test score)

● **ทฤษฎีดั้งเดิมที่มีคุณค่า (classical test theory)**

ทฤษฎีของคะแนนแบบทดสอบที่ใช้กันอยู่เดิม (Classical Test Theory – CTT) มีหลักการตรวจสอบคุณสมบัติของคะแนน 2 ประการที่สำคัญคือ ความเชื่อมั่น (reliability) และความเที่ยงตรง (validity) ความเชื่อมั่นเป็นความจำเป็นแต่ไม่เพียงพอสำหรับความเที่ยงตรง

เหตุผลของความเชื่อมั่นของคะแนนแบบทดสอบ (the rationale of test reliability)

ภายใต้สังกัปของความเชื่อมั่น เราสนใจความแม่นยำ (accuracy) ที่คะแนนตัวหนึ่งจะแทนสถานะของบุคคลหนึ่งไม่ว่าจะทดสอบเขาในด้านใด นั่นคือการวัดตัวแปรอย่างคงเส้นคงวา เมื่อเวลา บุคคล และสถานการณ์เปลี่ยนแปลงไป เป็นที่ยอมรับกันว่าคะแนนย่อมมีความคลาดเคลื่อน (error) นั่นคือ คะแนนที่วัดได้ (X) เป็นผลบวกของส่วนที่เป็นคะแนนจริง (true component) (T) และส่วนที่เป็นคะแนนความคลาดเคลื่อน (error component) (E) หรือเขียนเป็นรูปสมการได้ดังนี้

$$X = T + E$$

เมื่อ T หมายถึงคะแนนที่บุคคลหนึ่งจะทำได้ภายใต้สถานการณ์ที่เป็นอุดมคติหรือการใช้เครื่องมือวัดที่สมบูรณ์แบบ หรือ T คือค่าเฉลี่ยของคะแนนที่วัดได้จากการวัดที่เป็นอิสระกันหลายครั้งมาก ๆ โดยใช้แบบทดสอบฉบับเดียวกันวัดบุคคลเดียวกัน

E คือส่วนที่เพิ่ม (เปลี่ยน) (increment) (บวกหรือลบ) ที่ขึ้นอยู่กับสภาพการณ์ในโอกาสหนึ่งของการทดสอบบุคคลเดียวกัน สิ่งที่มีอิทธิพลต่อ E มีมากมาย ซึ่งอาจจะระบุได้หรือระบุไม่ได้ ในประชากรกลุ่มใหญ่ ค่าเฉลี่ยของ E เป็นศูนย์ E ไม่สัมพันธ์กับ T และ E ของคะแนนจากฉบับหนึ่งไม่สัมพันธ์กับ E ในอีกฉบับหนึ่งที่เป็นคู่ขนานกันด้วยข้อตกลงเบื้องต้นทั้งสามที่กล่าวแล้ว ความเชื่อมั่นสามารถนิยามได้ว่า เป็นสัดส่วนของความแปรปรวนของคะแนนจริง (σ_T^2) ในความแปรปรวนของคะแนนที่วัดได้ σ_x^2 หรือเขียนเป็นสมการดังนี้

$$r_{xx} = \frac{\sigma_x^2}{\sigma_x^2}$$

เมื่อ r_{xx} คือ สัมประสิทธิ์ของความเชื่อมั่น ซึ่งอยู่ในรูปของความสัมพันธ์กับตัวเองของคะแนนที่วัดได้ ซึ่งไม่สามารถกระทำได้โดยตรง วิธีประมาณค่าความเชื่อมั่นมีหลายวิธีดังนี้

1. วิธีแบ่งครึ่งข้อสอบ (split – half) และใช้ Spearman – Brown formula เพื่อประมาณค่าความเชื่อมั่นของข้อสอบทั้งฉบับ การแบ่งข้อสอบออกเป็นส่วน ๆ นี้สามารถแบ่งย่อยลงไปอีก 3 ส่วน 4 ส่วน ฯลฯ ตลอดจนถึงแบ่งย่อยเป็นข้อ ๆ ซึ่งเป็นการตรวจสอบความเท่าเทียมกันของส่วนต่าง ๆ ของข้อสอบหรือเรียกว่าความคงเส้นคงวาภายใน (internal consistency)

2. วิธีใช้ข้อสอบคู่ขนาน (alternate – form) เพื่อดูความเท่าเทียมกันของเนื้อหาในฉบับหนึ่งกับอีกฉบับหนึ่งของข้อสอบเดียวกัน

3. วิธีสอบซ้ำ (retest) เพื่อดูความคงที่หรือเสถียรภาพ เมื่อสอบในโอกาสต่างกัน

จากสมการที่แสดงค่าความเชื่อมั่น เราสามารถจัดกระทำต่อไปจนได้สมการที่แสดงความคลาดเคลื่อนมาตรฐานของการวัด (standard error of measurement) ดังนี้

$$\sigma_E = \sigma_x\sqrt{1-r_{xx}}$$

นี่เป็นความคลาดเคลื่อนของคะแนนที่วัดได้ตัวใต ๆ และเป็นประโยชน์อย่างยิ่งในการตีความคะแนนสำหรับการนำไปใช้ในโอกาสต่าง ๆ

เหตุผลของความเที่ยงตรงของคะแนนแบบทดสอบ (The rationale of test validity)

ความเที่ยงตรงเกี่ยวกับคำถามที่ว่าคะแนนแบบทดสอบนั้นวัดอะไร หรือคะแนนนั้นสามารถทำนายเกณฑ์ของผลงานได้ดีเพียงใด เมื่อแบบทดสอบถูกใช้เพื่อทำนายผลงาน ความเที่ยงตรงจึงอธิบายในเทอมของสหสัมพันธ์ระหว่างคะแนนแบบทดสอบและตัววัดผลงานนั้นหรือที่เรียกว่าเกณฑ์ (ซึ่งอาจเป็นคะแนนจากแบบทดสอบฉบับอื่น) ความสัมพันธ์ระหว่างคะแนนจากแบบทดสอบ 2 ฉบับก็คือผลบวกของผลคูณ (sum of cross products) ของ factor loadings ในผลการวิเคราะห์องค์ประกอบ (factor analysis) เมื่อ factor loadings ระหว่างแบบทดสอบนั้นกับเกณฑ์มีขนาดใหญ่ แสดงถึงความสัมพันธ์สูงระหว่างคะแนนจากแบบทดสอบและจากเกณฑ์ ความสัมพันธ์นี้เรียกว่าสัมประสิทธิ์ความเที่ยงตรง (validity coefficient)

การที่จะให้ได้ค่าสัมประสิทธิ์ความเที่ยงตรงที่ดี ผู้วิจัยต้องสามารถหาเกณฑ์ที่ดี (เที่ยงตรง) เป็นจุดเริ่มต้น มิฉะนั้นการศึกษาความเที่ยงตรงนั้นจะไม่เป็นประโยชน์ในการอ้างอิง (Guilford, 1954 : 402)

วิธีการหาความสัมพันธ์ระหว่างคะแนนจากแบบทดสอบกับคะแนนจากเกณฑ์เรียกกันว่าความเที่ยงตรงชนิดเกณฑ์ที่เกี่ยวข้อง (criterion

related validity) นอกจากนี้ในทางปฏิบัติยังมีวิธีการอื่นซึ่งเป็นไปตามความมุ่งหมายในการอ้างอิงคะแนน นั่นคือ การศึกษาความเที่ยงตรงเชิงโครงสร้าง (construct validity) ซึ่งสามารถอธิบายโครงสร้างเชิงทฤษฎีของตัวแปรที่วัดนั้น การศึกษาความเที่ยงตรงในลักษณะนี้มีแนวโน้มเป็นการทดสอบสมมติฐาน ซึ่งอาจมีหลายสมมติฐาน หรืออาจมีการศึกษาหลายครั้งก็ได้ เพื่อให้ได้หลักฐานที่สามารถอธิบายโครงสร้างของตัวแปรได้ครอบคลุมตามทฤษฎี

ส่วนการตรวจสอบความเที่ยงตรงเชิงเนื้อหา (content validity) นั้นเป็นเพียงการบรรยายความเกี่ยวข้องของข้อความแต่ละข้อกับเนื้อหาจากนิยาม

● **ทฤษฎีการตอบข้อคำถาม** (Item Response Theory – IRT) ในช่วงต้น ๆ ทฤษฎีใหม่นี้ได้รับการศึกษาและพัฒนามาอย่างช้า ๆ และต่อมาปรากฏชัดเจนด้วยผลงานของ A. Birnbaum ในหนังสือชื่อ *Statistical theories of mental test scores* ของ Frederic Lord and Melvin Novick (Lord and Novick, 1968) หลังจากนั้นจึงมีการศึกษาวิจัยทั้งในเชิงทฤษฎีและนำไปปฏิบัติอย่างกว้างขวางรวดเร็ว แทนที่จะใช้คะแนนรวมทั้งฉบับ IRT วิเคราะห์ข้อมูลจากผลการตอบต่อข้อคำถามแต่ละข้อ โมเดลใน IRT แสดงความสัมพันธ์ระหว่างตัวแปรที่วัด (latent trait) กับผลการตอบต่อข้อคำถามแต่ละข้อ โมเดลเชิงคณิตศาสตร์ที่ใช้แสดงความสัมพันธ์ดังกล่าวคือ โมเดลโอไจฟ์ (Ogive model) และโมเดลโลจิสติก (logistic model) แต่โมเดลโลจิสติกมีความเหมาะสมมากกว่าจึงนิยมใช้กันทั่วไป โมเดลนี้บอกความน่าจะเป็นของผลการตอบข้อคำถามในลักษณะที่เป็นฟังก์ชันของตัวแปรแฝงนั้น (ดูเพิ่มเติม เช่น ผจงจิต อินทสุวรรณ (2525, หน้า 51 – 69 และ2534 หน้า 79 - 90); Lord, 1980; Hambleton & Swaminathan, 1984) ฟังก์ชันนี้กำหนดด้วยพารามิเตอร์ 2 กลุ่มคือ พารามิเตอร์ของผู้ตอบ (examinee parameter) และพารามิเตอร์ของข้อ (item parameter) พารามิเตอร์ของผู้ตอบอาจเรียกว่า ability parameter เมื่อใช้กับแบบทดสอบวัดผลสัมฤทธิ์หรือความสามารถใด ๆ ส่วนพารามิเตอร์ของข้อมี 3 ตัว นั่นคือ อำนาจจำแนก (discrimination index) ดัชนีความยาก (difficulty index) และการเดา (guessing) ผู้วิจัยสามารถเลือกใช้โมเดลที่มีพารามิเตอร์ 3 หรือ 2 หรือ 1 ตัวตามความเหมาะสม โมเดลที่มีพารามิเตอร์ 1 ตัวได้รับการพัฒนามาอย่างเป็นอิสระกันและจากฐานที่แตกต่างกันโดย จอร์ช รัช (George Rasch) (1960 / 1980) ผู้สนใจสามารถศึกษาในหนังสือจำนวนมาก ดังเช่น ไรท์ และ สโตน (Wright & Stone)(1979) แอนดริช (Andrich) (1988) ในปัจจุบัน IRT ได้รับความสนใจอย่างกว้างขวางในการเรียนการสอนระดับบัณฑิตศึกษาและมีการทำวิจัยขยายขอบข่ายที่ซับซ้อนยิ่งขึ้น ทั้งในแง่ของทฤษฎีและการประยุกต์ใช้ และได้มีการนำไปใช้ในหน่วยงานทดสอบที่สำคัญ ดังเช่น ETS (Educational Testing Service)

ประเทศสหรัฐอเมริกา

IRT มีข้อได้เปรียบ CTT หลายประการ ที่สำคัญคือ เราสามารถคำนวณความคลาดเคลื่อนในการวัดที่ตำแหน่งนั้น ๆ ของผลการตอบ (ไม่ใช่คะแนนทุกตำแหน่ง ใช้ค่าความคลาดเคลื่อนเดียวกันดังใน CTT) อีกประการหนึ่งคะแนนของผู้สอบไม่ขึ้นอยู่กับชุดของข้อคำถามที่ตอบ และไม่ขึ้นอยู่กับกลุ่มผู้สอบ (CTT ขาดคุณสมบัตินี้เช่นกัน) และด้วยเหตุที่มีโปรแกรมคอมพิวเตอร์ (software) ที่สามารถใช้วิเคราะห์ข้อมูลในทฤษฎีนี้ได้มากมาย นักวิจัยจึงตรวจสอบการประยุกต์ใช้ในปัญหาการวัดต่าง ๆ ที่ CTT ไม่สามารถทำได้ เช่น การสร้างแบบทดสอบ การเทียบคะแนน (test equating) การตรวจสอบความลำเอียง (bias) ของข้อสอบ และการทดสอบโดยใช้คอมพิวเตอร์ (computerized adaptive testing)

การวัดทางจิตกับการศึกษา

ความรู้ทางการวัดทางจิตเป็นประโยชน์อย่างยิ่งต่อวงการศึกษา และมีการนำไปใช้อย่างกว้างขวางมาตลอดเวลา ในด้านต่าง ๆ ดังนี้

1. การสอบคัดเลือกผู้สมัครเข้าเรียน เพื่อตรวจสอบความพร้อมในพื้นฐานการศึกษา เพื่อจัดกลุ่มการเรียน หรือเพื่อการเลือกวิชา / สาขาในการเรียน

2. การประเมินผลการเรียน ในระดับต่าง ๆ เช่น ประเมินความสามารถในโดเมนต่าง ๆ เช่น การอ่าน การเขียน คณิตศาสตร์ เป็นต้น

ในข้อ 1 และ 2 นั้น เป็นการวัดผลงานสูงสุด (maximum performance) และใช้แบบทดสอบที่มีการตรวจให้คะแนนถูก-ผิด เช่น การสอบวัดความถนัด วัดสติปัญญาหรือผลสัมฤทธิ์ทางการเรียน

3. การวัดบุคลิกภาพ ซึ่งรวมทั้งเจตคติ ความเชื่อ ค่านิยม ความสนใจ พฤติกรรมของบุคคล เพื่อประโยชน์ในการปรับตัวของนักเรียน และเพื่อการเลือกสาขาวิชาที่เรียน โดยใช้ประกอบกับคะแนนด้านความสามารถของบุคคลนั้น ๆ ด้วย

การวัดในข้อนี้เป็นการวัดพฤติกรรมตามปกติ (typical performance) รูปแบบมักเป็นมาตราจัดอันดับ ตัวอย่างของเครื่องมือวัดดังเช่น Minnesota Multiphasic Personality Inventory (MMPI)(e.g., Friedman, et.al., 2001), The 16 Personality Factor Questionnaire (16 PF) (e.g., Cattell & Mead, 2007), the Five – Factor Model (Big 5)(e.g., Caprara, et.al., 1993)

บรรณานุกรม

ผจงจิต อินทสุวรรณ. (เมษายน 2525). "Latent Trait Theory," **วารสารการวัดผลการศึกษา.** 3(3) : 51 - 69.

-------. (พฤษภาคม - สิงหาคม 2534). "การวิเคราะห์แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนด้วยโมเดลโลจิสติก," **วารสารการวัดผลการศึกษา.** 13 (37) : 79 - 90.

Andrich, D. (1988). **Rasch models for measurement.** Newbury Park, Ca : SAGE Publications, Inc.

Becker, K.A. (2003). **History of the Stanford-Binet intelligence scales : Content and psychometrics.** (Stanford – Binet Intelligence Scales. Fifth Edition Assessment Service Bulletin No.1). Itasca., IIS Riverside Publishing.

Caprara, G.V. ; Barbaranelli, C. ; Borgogni, L. & Perugini, M. (1993). The "big five questionnaire : A new questionnaire to assess the five factor model," **Personality and Individual Differences.** 15(37) : 281 -288.

Cattell, H.E.P. and Mead, A.D. (2007). "The 16 Personality Factor Questionnaire (16 PF)." in G.J. Boyle, G., Matthews, and D.H. Sakeofske (Eds.) **Handbook of Personality theory and testing Vol.2 Personality measurement and assessment.** London : Sage.

Friedman, A.F., Lewak, R., Nichols, D.S., & Webb, J.T. (2001). **Psychological assessment with the MMPI-2.** Mahwah, NJ: Lawrence Erlbaum Associates.

Guilford, J.P. (1954). **Psychometric Methods.** New York : McGraw Hill.

Hambleton, R.K. and Swaminathan, H. (1984). **Item response theory: Principles and applications.** Hingham, MA: Kluwer, Nijhaff.

Lord, F.M. (1980). **Applications of item response theory to practical testing problems.** New Jersey: Lawrence Erlbaum Associates.

Morris, W. (Editor). (1978). **American Heritage Dictionary of the English Language.** Boston: Houghton Mifflin.

Rasch, G. (1960 / 1980). **Probabilistic models for some intelligence and attainment tests.** (Copenhagen, Danish Institute for Educational Research), expaned edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.