บทที่ 2

วรรณคดีที่เกี่ยวข้อง

แนวคิดและงานวิจัยที่เกี่ยวข้องกับงานวิจัยเรื่องนี้แบ่งออกเป็น 6 ตอน คือ

ตอนที่ 1. มโนทัศน์เกี่ยวกับการรับรู้ผลประทบของการสอบ

ตอนที่ 2. มโนทัศน์เกี่ยวกับทฤษฎีการตอบสนองข้อสอบ (IRT)

ตอนที่ 3. มโนทัศน์เกี่ยวกับการวิเคราะห์ความสอดคล้องระหว่างความสามารถกับความยากของข้อสอบ

ตอนที่ 4. มโนทัศน์เกี่ยวกับการตรวจให้คะแนน

ตอนที่ 5. มโนทัศน์เกี่ยวกับความคลาดเคลื่อนมาตรฐานของการวัด

ตอนที่ 6. งานวิจัยที่เกี่ยวข้องกับการเพิ่มคุณภาพของแบบสอบ

ตอนที่ 1 มโนทัศน์เกี่ยวกับการรับรู้ผลกระทบของการสอบ

การรับรู้ผลกระทบของการสอบของนักเรียนว่าผลการสอบจะนำไปทำอะไรนั้นถือว่าเป็นสิ่งจูงใจให้นักเรียนเกิดความตั้งใจ และใช้ความสามารถในการทำแบบสอบมากน้อยแตกต่างกัน นั่นคือถ้าผลการสอบนั้นกระทบกับผลการเรียนของนักเรียนโดยตรงก็น่าจะมีความตั้งใจมาก แต่ถ้าผลการสอบไม่กระทบกับผลการเรียนของนักเรียนแล้ว ผู้สอบก็น่าจะมีความตั้งใจน้อย สิ่งจูงใจนี้ต่อมานักจิตวิทยาใช้คำว่าแรงเสริม (ชัยพร วิชชาวุธ 2519 : 140) นอกจากนี้ ชัยพร วิชชาวุธ ยังกล่าวว่าทฤษฎีการเสริมแรงได้พัฒนามาจากกฎแห่งผลกรรม (Law of Effect) ซึ่งเป็นกฎของการเรียนรู้ที่เสนอโดยนักจิตวิทยาชาวอเมริกัน คือธอร์นไดค์ (Thorndike) ซึ่งตามแนวคิดของธอร์นไดค์นั้นพฤติกรรมของมนุษย์เกิดขึ้นตามกฎของผลกรรม พฤติกรรมที่ยังผลเป็นที่พึงพอใจแก่ผู้กระทำพฤติกรรมนั้นย่อมเกิดขึ้นอีกและพฤติกรรมใดที่ไม่เป็นที่พึงพอใจแก่ผู้กระทำ พฤติกรรมนั้นย่อมหดหายไป

ความหมายของการเสริมแรงมีผู้ให้ความหมายไว้หลายคน เช่น สมโภชน์ เอี่ยมสุภาษิต (2524 : 35) ได้ให้ความหมายของการเสริมแรงไว้ คือ การที่ทำให้พฤติกรรมหนึ่งของอินทรีย์เพิ่มขึ้นอันเป็นผลเนื่องมาจากการได้รับผลกรรม (Consequence) ที่พึงพอใจหลังจากการแสดงพฤติกรรมนั้นหรืออันเนื่องมาจากความสำเร็จในการหลีก (Avoidance) หรือการหนี (Escape) จากสิ่งเร้าที่อินทรีย์ไม่พึงพอใจ (Aversive Stimulus) ซึ่งผลกรรมที่พึงพอใจนั้นเรียกว่าตัวเสริมแรงทางบวก (Positive Reinforcer) ส่วนสิ่งเร้าที่ไม่พึงพอใจนั้นเรียกว่าตัวเสริมแรงทางลบ (Negative Reinforcer) พัฟลอฟ (Pavlov) นักจิตวิทยาชาวรัสเซียได้ให้ความหมายของการเสริมแรงว่าเป็นสิ่งเร้าที่ไม่มีเงื่อนไข (Unconditioned stimulus) ซึ่งเกิดขึ้นภายหลังสิ่งเร้าที่มีเงื่อนไข (Conditioned stimulus) ในการทดลองเกี่ยวกับการเรียนรู้แบบวางเงื่อนไข (Conditioned stimulus) ในการทดลองเกี่ยวกับการเรียนรู้แบบวางเงื่อนไข (Munn 1961 : 377) นอกจากนี้การเสริมแรงยังหมายถึงสิ่งที่เกิดเป็นผลตามมาภายหลังการตอบสนอง แล้วมีผลให้โอกาสที่จะมีการตอบสนองเกิดเพิ่มขึ้น (Kuhlen 1968 : 260)

การเสริมแรงมีหลายชนิด อาจเป็นการเสริมแรงทางบวก (Positive reinforcement) เช่นการยอมรับของสังคม การเสริมแรงทางลบ (Negative reinforcement) เช่นการลงโทษ การเสริมแรงโดยตรง (Direct reinforcement) ได้แก่การชมเชยเด็กตรง ๆ การเสริมแรงทางอ้อม (Vacarious reinforcement) ได้แก่ความรู้สึกเหมือนกับตนได้รับคำชมเชยด้วยของผู้ที่ร่วมอยู่ในขณะที่ผู้อื่นได้รับคำชมเชย (Kuhlen 1968 : 260)

การรับรู้ผลการสอบเป็นการเสริมแรงทางบวก (Positive reinforcement) ไม่ว่าผลการสอบนั้นจะมีผลโดยตรงหรือไม่โดยตรงก็ตาม เป็นผลก่อให้เกิดบุคลิกภาพขึ้นอันเป็นตัวกำหนดให้พฤติกรรมในการตอบแบบสอบเกี่ยวกับแรงเสริมบุคลิกภาพและพฤติกรรมนี้ แอลล์พอร์ต (Allport quoted in Wolman 1965 : 418-419) ได้กล่าวถึงความสัมพันธ์ระหว่างแรงจูงใจกับบุคลิกภาพโดยมีความเห็นตรงกับสเทอร์น (Stern) ว่าบุคลิกภาพเป็นหน่วยย่อยที่สลับซับซ้อน และได้พยายามที่จะประเมินองค์ประกอบพื้นฐานของหน่วย

เหล่านี้ โดยเน้นถึงความสำคัญของพฤติกรรมที่นำไปสู่เป้าหมาย และชี้ให้เห็นว่าลักษณะที่กำหนดพฤติกรรมเหล่านี้คือลักษณะทางบุคลิกภาพซึ่งเป็นลักษณะอารมณ์ในสภาพการปรับตัวของบุคคลในยามปกติ เป็นผลรวมของแรงเสริมหรือแรงจูงใจและอุปนิสัย กล่าวได้ว่าเป็นระบบจิตประสาท อันเป็นจุดรวมที่เด่นเป็นตัวกำหนดขอบเขตของการรับรู้และชนิดของการตอบสนองของบุคคล บุคคลแต่ละคนต่างมีโครงสร้างของจิตที่เป็นเอกลักษณ์ จึงทำให้มีความสามารถที่จะแสดงปฏิกริยาต่อสิ่งเร้าทั้งหลายแตกต่างกัน

การเสริมแรงทางบวกมีด้วยกัน 5 ชนิด คือ (สมโภชน์ เอี่ยมสุภาษิต 2524 : 42)

1. แรงเสริมที่เป็นสิ่งที่เสพได้ เช่น อาหาร ขนม น้ำ บุหรี่ เหล้า เป็นต้น

2. แรงเสริมทางสังคม (Social reinforcer) เช่น การยอมรับ การยิ้ม การยกย่อง การชมเชย เป็นต้น

3. หลักการของพรีแม็ค (Premack's Principle) ได้แก่การใช้พฤติกรรมที่อินทรีย์ชอบทำมาก มาเสริมแรงพฤติกรรมที่อินทรีย์ชอบทำน้อยกว่า เช่น การใช้พฤติกรรมการเล่นฟุตบอล (พฤติกรรมที่ชอบมาก) มาเสริมแรงพฤติกรรมการทำการบ้าน (พฤติกรรมที่ชอบน้อยกว่า) ทำให้พฤติกรรมการทำการบ้านเพิ่มขึ้น เป็นต้น

4. การให้ข้อมูลย้อนกลับ (Informative Feedback) เป็นการให้ข้อมูลเกี่ยวกับการกระทำของอินทรีย์รู้ว่าตนได้ทำพฤติกรรมที่เหมาะสมหรือไม่ เช่น ในการสอบเก็บคะแนนครั้งนี้แดงสอบได้ 80% เป็นต้น

5. เบี้ยอรรถกร (Token Economy) คือ การใช้การเสริมแรงด้วยเบี้ยคะแนนหรือแต้ม เป็นต้น ซึ่งเบี้ยคะแนนหรือแต้มสามารถนำไปแลกเป็นตัวเสริมแรงอื่นได้ เช่นนำไปแลกเป็นขนม เวลาว่าง สิทธิพิเศษต่าง ๆ ฯลฯ

เพื่อความเข้าใจเกี่ยวกับทฤษฎีการเสริมแรงจะขอยกตัวอย่างซึ่ง ชัยพร วิชชาวุธ (2519 : 142) ได้ให้ไว้ดังนี้ สำหรับนิสิตมหาวิทยาลัยการสำเร็จการศึกษาได้รับปริญญาบัตรเป็นสิ่งที่ทุกคนปรารถนา การให้สำเร็จการศึกษาจึงเป็นการให้แรงเสริมอย่างหนึ่ง สถาบันการศึกษาต่าง ๆ จึงได้กำหนดเงื่อนไขของการให้สำเร็จการศึกษาเอาไว้ เช่น ต้อง

สอบให้ได้หน่วยกิตสะสมเท่านั้นเท่านี้หน่วยกิต ชำระเงินเท่านั้นเท่านี้ ได้แต้มเฉลี่ยไม่ต่ำกว่าเท่านั้นเท่านี้ จะต้องเรียนวิชานั้นวิชานี้ ถ้าจะเรียนวิชานี้จะต้องมาเรียนในวันนี้เวลานี้ห้องนี้กับอาจารย์คนนี้ และต้องมาสอบในวันนั้นวันนี้ ต้องตอบข้อสอบอย่างนั้นอย่างนี้ ฯลฯ เงื่อนไขต่าง ๆ เหล่านี้มากมาย และที่ไม่ได้เขียนเป็นระเบียบชัดเจนแต่อาจารย์แต่ละคนกำหนดขึ้นเองก็มีอีกมากมาย นิสิตที่ประสงค์จะสำเร็จการศึกษาก็ต้องกระทำตามเงื่อนไขเหล่านี้ มิฉะนั้นก็จะไม่ได้รับปริญญา ซึ่งเป็นแรงเสริม ถ้าหากมหาวิทยาลัยเปลี่ยนแปลงเงื่อนไขเหล่านี้ พฤติกรรมของนิสิตก็จะเปลี่ยนแปลงตามเงื่อนไขที่เปลี่ยนไป

การรับนิสิตนักศึกษาเข้าศึกษาในมหาวิทยาลัยก็เช่นกัน การให้เข้าศึกษาเป็นสิ่งที่บุคคลจำนวนมากต้องการ ปัจจุบันเงื่อนไขในการเข้าคือต้องจบการศึกษาชั้นนั้นชั้นนี้ ต้องสอบคัดเลือกวิชานั้นวิชานี้ ผู้ประสงค์จะเข้ามหาวิทยาลัยก็จะปรับพฤติกรรมของตนเองให้เข้ากับเงื่อนไขต่าง ๆ เหล่านี้ ลองคิดเอาเองก็ได้ว่าพฤติกรรมของเยาวชน ผู้ประสงค์จะเข้าศึกษาในมหาวิทยาลัยจะเป็นอย่างไร หากเรากำหนดว่าผู้ที่จะสมัครได้นั้นต้องมีหลักฐานมาแสดงว่า (ก) เคยเป็นลูกเสือหรือเนตรนารีมาก่อน (ข) เคยร่วมกิจกรรมบำเพ็ญประโยชน์ต่อสังคมมาก่อน เช่น ทำความสะอาดสถานสาธารณะ สอนหนังสือแก่เด็กสลัม ร่วมกิจกรรมค่ายอาสาพัฒนา ฯลฯ นอกเหนือจากการเรียนจบประโยคมัธยมศึกษาตอนปลายและสอบคัดเลือกวิชาเฉพาะได้ผลเป็นที่พอใจ

นอกจากนี้ ชัยพร วิชชาวุธ (2519 : 144) ยังให้ทรรศนะเกี่ยวกับทฤษฎีการเสริมแรงไว้ว่า พฤติกรรมของมนุษย์ถูกควบคุมโดยเงื่อนไขผลกรรมทั้งที่เป็นผลกรรมทางบวกและผลกรรมทางลบ ตามทฤษฎีนี้ผลกรรมเป็นตัวกำหนดพฤติกรรม ผลกรรมทางบวกเป็นแรงเสริมให้เกิดพฤติกรรม ผลกรรมทางลบเป็นแรงปรามไม่ให้เกิดพฤติกรรม อย่างไรก็ดีทฤษฎีการเสริมแรงก็ประสบปัญหาเช่นเดียวกันกับความคิดเกี่ยวกับสิ่งจูงใจ นั่นคือผลกรรมเกิดที่หลังพฤติกรรม ผลกรรมจะเป็นเหตุของพฤติกรรมได้อย่างไร ทางออกในปัญหานี้ก็เช่นเดียวกันกับทางออกของความคิดเกี่ยวกับสิ่งจูงใจ กล่าวคือเราสามารถอธิบายว่าคนเราเรียนรู้ความสัมพันธ์ระหว่างพฤติกรรมกับแรงเสริมต่าง ๆ ทั้งทางบวกและทางลบ ผลของการเรียนรู้ทำให้คนเราสามารถคาดหวังว่าพฤติกรรมอะไร จะทำให้ได้รับแรงเสริมอะไร ความ

คาดหวังว่าพฤติกรรมอะไร จะทำให้ได้รับแรงเสริมอะไร ความคาดหวังที่จะได้รับแรงเสริมหนึ่ง ๆ เกิดก่อนที่จะกระทำพฤติกรรมจึงเป็นแรงผลักดันให้เกิดพฤติกรรมที่นำไปสู่แรงเสริมนั้น ดังนั้นจึงกล่าวโดยสรุปได้ว่า ตามทฤษฎีการเสริมแรง ตัวการที่ทำให้เกิดพฤติกรรมคือแรงเสริม และถ้าจะพูดในแง่ของการผลักดันพฤติกรรมก็สามารถกล่าวได้ว่า ความคาดหวังที่จะได้รับแรงเสริมเป็นแรงผลักดันพฤติกรรม

จากที่กล่าวมาข้างต้นเกี่ยวกับแรงเสริมจะเห็นว่าแรงเสริมมีทั้งแรงเสริมที่สนองความต้องการของร่างกายและแรงเสริมที่สนองความต้องการทางด้านจิตใจ การรับรู้ผลกระทบของการสอบ เพื่อใช้เป็นส่วนหนึ่งของการตัดสินผลการเรียน หรือไปทำวิจัยเป็นตัวแปรหนึ่งที่ผู้วิจัยศึกษา ซึ่งถือว่าเป็นแรงเสริมทางด้านจิตใจอย่างหนึ่งในอันที่จะมีผลกระทบต่อพฤติกรรมในการตอบแบบสอบและก็กระทบต่อผลการสอบตามมา

ตอนที่ 2 มโนทัศน์เกี่ยวกับทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) เรียกย่อ ๆ ว่า ทฤษฎี IRT เป็นทฤษฎีลักษณะแอบแฝง (Latent trait theory) ผู้ที่ริเริ่มขึ้นคือ เฟอร์กูสัน (Fergusen) ในปี ค.ศ. 1942 ลอเลย์ (Lawley) ในปี ค.ศ. 1943 (Warm 1979 : 19; Lord and Novick 1968 : 369) ได้เสนอหลักการว่า ผลการสอบของผู้สอบจากแบบสอบใดนั้นขึ้นอยู่กับความสามารถ (Latent trait or ability or skill) ของผู้สอบซึ่งสามารถจะเขียนอยู่ในรูปฟังชั่น (Function) ทางคณิตศาสตร์ได้ ซึ่งฟังชั่นนี้จะประกอบด้วยค่าพารามิเตอร์ต่าง ๆ ได้แก่ค่าความยาก ค่าอำนาจจำแนกและค่าการเดา ซึ่งแล้วแต่ว่าฟังชั่นนั้นจะใช้ค่าพารามิเตอร์กี่ตัว

ต่อมาในปี ค.ศ. 1952 ลอร์ด (Lord) ได้เสนอ IRT ขึ้นใหม่ชื่อทฤษฎีโค้งลักษณะข้อสอบ (Item Characteristic Curve Theory) โดยเสนอว่าลักษณะของโค้งข้อสอบ (Item Characteristic Curve) แต่ละข้อสอบมีลักษณะเป็น Normal Ogive ซึ่งต่อมาเรียกว่า Normal Ogive Model โมเดลนี้มีการคำนวณที่ยุ่งยากมาก

ดังนั้นทำให้ลอร์ดหยุดความสนใจใน IRT ไปพักหนึ่งในปี ค.ศ. 1960 ราส์ช (Rasch) ได้เสนอ IRT ในรูปของพารามิเตอร์เดียวคือความยาก และในปี 1965 ลอร์ดก็กลับมาพัฒนาทฤษฎี IRT ใหม่ (Warm 1979 : 19)

ปี ค.ศ. 1960 เบิร์นบอม (Birmbaum) ได้เสนอ Logistic Model แบบ 2 พารามิเตอร์ขึ้น ซึ่งเป็นโมเดลที่ง่ายกว่าของลอร์ด (Lord) จึงทำให้ IRT นิยมใช้กันแพร่หลายต่อมา และได้มีการพัฒนาใช้กับพารามิเตอร์ 1 ตัว 2 ตัว และ 3 ตัว (Warm 1979: 19 - 21)

ทฤษฎีลักษณะแอบแฝง (Latent trait theory) มีข้อสมมติฐานว่าในสถานการณ์การสอบผลการปฏิบัติของผู้สอบต่อแบบสอบชุดหนึ่ง สามารถอธิบายได้ด้วยการนิยามลักษณะของข้อสอบโยงไปยังลักษณะ (Trait) ของผู้สอบได้ หรือสามารถประมาณคะแนนของผู้สอบบนลักษณะที่วัดได้ เพราะว่าลักษณะเหล่านี้ไม่สามารถจะวัดได้และสังเกตได้โดยตรง นักจิตวิทยาถือว่าลักษณะแอบแฝง (Latent trait) ของบุคคลเป็นสิ่งที่เราสามารถสังเกตได้หรือวัดได้ แต่ผลการปฏิบัติของเขาเราสามารถวัดความสามารถหรือลักษณะ (Trait) ได้โดยดูว่าความสามารถของเขาสัมพันธ์กับการสนองต่อข้อสอบอย่างไร ซึ่งทางทฤษฎีลักษณะแอบแฝงจะอธิบายได้ด้วยฟังชั่นทางคณิตศาสตร์ (Lord and Novick 1968: 359)

ข้อตกลงเบื้องต้นของ IRT มี 3 ประการ ดังนี้ (Hambleton and Cook 1977 : 77-80)

1. เกี่ยวกับมิติของลักษณะที่วัด (Dimension of the Latent space) มิติของลักษณะที่วัดนี้ขึ้นอยู่กับจำนวน Traits ที่กำหนดไว้ในแบบสอบที่จะวัด โดยมีข้อตกลงว่าต้องเป็นมิติเดียวนั่นคือ ข้อสอบในแบบสอบเป็นเอกพันธ์ (Homogeneous) หรือวัดความสามารถด้านเดียว ตัวอย่างเช่น วัดคำศัพท์ วัดการอ่าน วัดการสะกดคำ เป็นต้น ซึ่งลอร์ด (Lord) ก็กล่าวว่าข้อตกลงนี้ไม่เป็นจริงในทุกแบบสอบ แฮมเบิลตันและทรอบ (Hambleton

and Traub) ก็ได้แสดงให้เห็นวิธีตรวจสอบความเป็นมิติเดียวโดยใช้เทคนิคการวิเคราะห์องค์ประกอบ (Factor Analysis) และกล่าวว่าถ้าแบบสอบไม่เป็นมิติเดียว ข้อสอบในแบบสอบจะแบ่งออกเป็นกลุ่มตามองค์ประกอบ (Factor) แบบสอบชุดใดมีลักษณะอย่างนี้เราเรียกว่ามีหลายมิติ (Multidimensional) เบจา (Bejar 1980 : 283) ได้ทำการศึกษาความเป็นมิติเดียวของแบบสอบวัดผลสัมฤทธิ์ในวิชาชีววิทยา โดยใช้กระบวนการศึกษา 2 แบบ คือวิธีแรกเป็นการศึกษาจากการเขียนกราฟเปรียบเทียบ 2 เส้นระหว่าง Theoretical axis กับ Principal axis และวิธีที่สองเป็นการวิเคราะห์องค์ประกอบผลปรากฏว่าแบบสอบวิชาชีววิทยาไม่เป็นมิติเดียว

2. ความเป็นอิสระที่ความสามารถระดับเดียวกัน (Local Independence) ซึ่งมีข้อตกลงว่า การตอบสนองข้อสอบหนึ่ง ๆ ของแบบสอบของผู้ตอบคนใดเป็นอิสระหรือไม่ขึ้นอยู่กับผลการปฏิบัติหรือตอบสนองข้อสอบอื่น ๆ ในแบบสอบชุดเดียวกัน หรือการตอบข้อสอบข้อหนึ่งเป็นอิสระจากการตอบข้ออื่น ๆ สำหรับผู้สอบที่มีความสามารถระดับหนึ่ง แต่มิได้หมายความว่าข้อสอบเหล่านี้ไม่มีความสัมพันธ์กันในกลุ่มผู้สอบที่มีความสามารถต่างกัน ผู้สอบที่มีความสามารถสูงกว่าก็จะมีแนวโน้มที่จะตอบได้ถูกต้องกว่าผู้สอบที่มีความสามารถต่ำกว่า

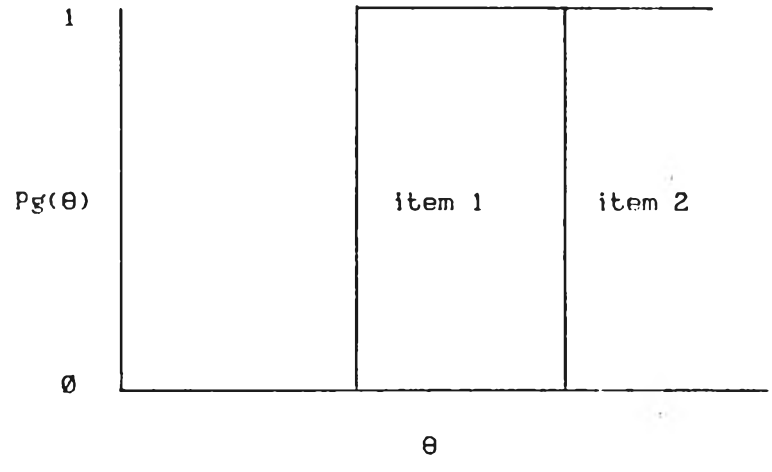
ในความหมายทางคณิตศาสตร์ของความเป็นอิสระที่ความสามารถระดับเดียวกัน (Local Independence) ลาซาร์สเฟลด์ (Lazarsfeld quoted in Lord 1980 : 19) กล่าวว่าความน่าจะเป็น (Probability) ของความสำเร็จในการตอบข้อ i และ j ได้ถูกต้องย่อมเท่ากับผลคูณของความน่าจะเป็นของการตอบถูกข้อ i และความน่าจะเป็นของการตอบ j ถูก เมื่อกำหนดความสามารถ (θ) ระดับหนึ่งดังสมการ

$$P(U_i = 1, U_j = 1/\theta) = P(U_i = 1/\theta))\ P(U_j = 1/\theta)$$

ในเมื่อ U_i และ U_j เท่ากับ ๐ หรือ 1 คะแนนสำหรับข้อสอบข้อ i และ j

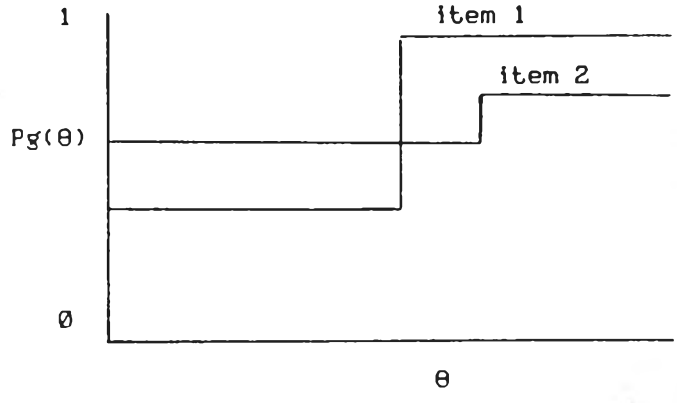
3. เกี่ยวกับโค้งลักษณะของข้อสอบ (Item Characteristic Curve)

เขียนย่อว่า i.c.c ซึ่งเป็นฟังชั่นทางคณิตศาสตร์ที่แสดงความสัมพันธ์ของโอกาสที่ผู้สอบจะประสบความสำเร็จในการทำข้อสอบ กับความสามารถของเขา i.c.c จะมีลักษณะเพิ่มขึ้นทางเดียว (Monotonic increasing) ตามระดับความสามารถที่เพิ่มขึ้น i.c.c มีรูปแบบแตกต่างกัน 7 ลักษณะ ดังภาพต่อไปนี้



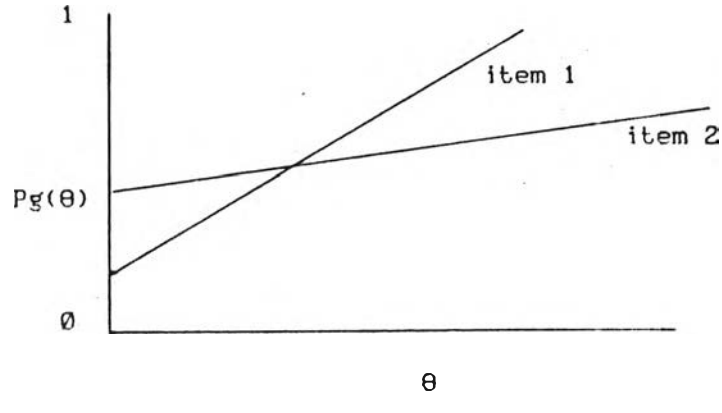
ภาพที่ 1 perfect scale curve

กราฟรูป a นี้เรียกว่าฟังชั่นสเตฟ (Step function) โอกาสที่จะตอบข้อสอบถูกเป็น Ø หรือ 1 เท่านั้น

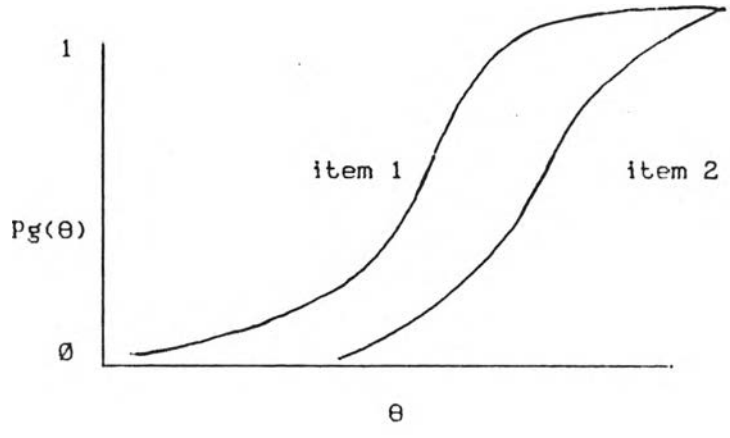


ภาพที่ 2 Latent distance curves

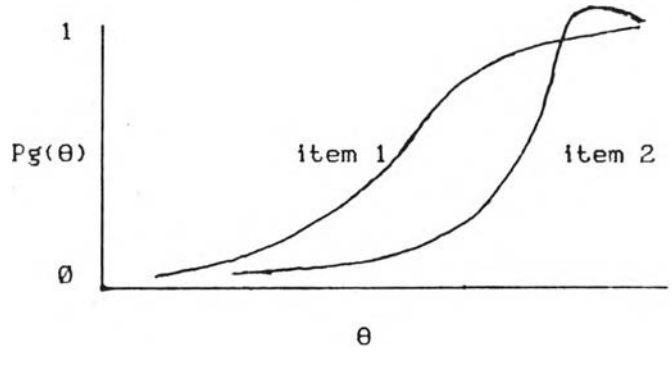
โค้งนี้อธิบายการเปลี่ยนตามรูปแบบเสกลที่สมบูรณ์ของ Guttman โค้งลักษณะของข้อสอบเป็นรูปแบบฟังชั่นเสตฟ (Step function) แต่โอกาสตอบผิดและตอบถูกแปรเปลี่ยนระหว่าง Ø และ 1



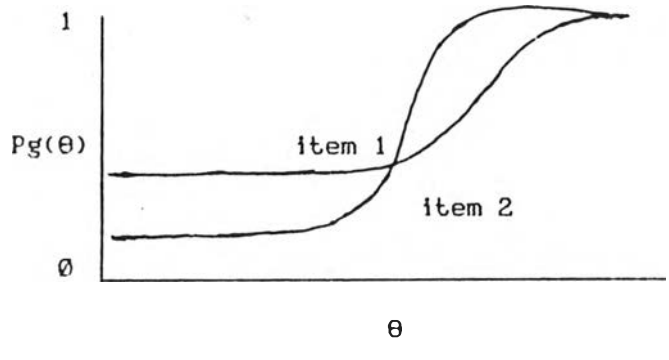
ภาพที่ 3 Latent linear curves



ภาพที่ 4 one - parameter logistic curves



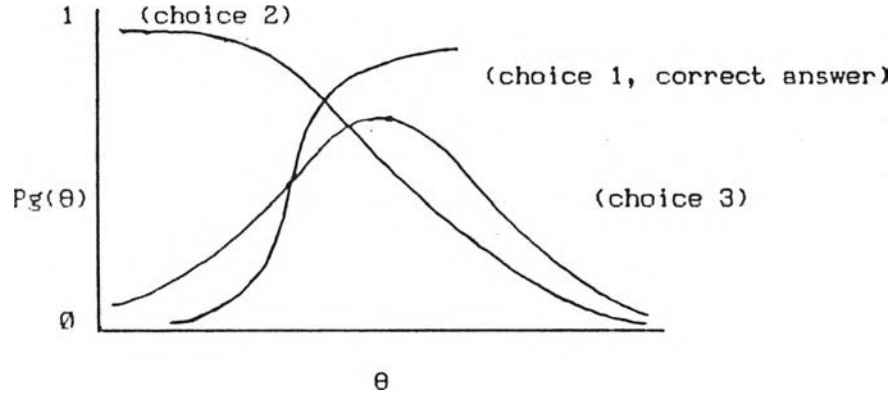
ภาพที่ 5 two - parameter logistic curves



ภาพที่ 6 three - parameter logistic curves

โค้งตามรูปภาพที่ 3,4,5 และ 6 สามารถอธิบายได้ด้วยค่าพารามิเตอร์ (Parameter) 3 ตัว คือค่าความยาก (Difficulty) ค่าอำนาจจำแนก (Discrimination) และค่าการเดา (Guessing) โมเดล (Model) ที่มีพารามิเตอร์เดียวโค้งจะแปรเปลี่ยนไปตามระดับค่าความยาก (θ) ดังภาพที่ 4 ส่วนโมเดลที่มีพารามิเตอร์ 2 ค่า โค้งจะแปรเปลี่ยนไปตามค่าอำนาจจำแนกและค่าความยากดังภาพที่ 5 และโมเดลที่มีค่าพารามิเตอร์ 3 ค่า โค้งจะแปรเปลี่ยนไปตามค่าอำนาจจำแนก ค่าความยากและค่าการเดา ดังภาพที่ 3 และ 6

ยังมีโมเดลอีกแบบหนึ่งที่แสดงให้เห็นโค้งลักษณะของตัวเลือกของข้อสอบ (Item option characteristic curves) ดังนี้



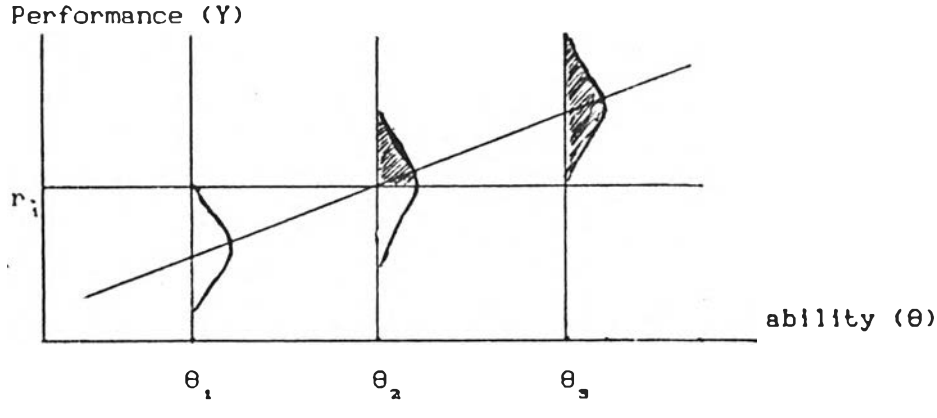
ภาพที่ 7 Item - option characteristic curve

i.c.c (Item characteristic curve) นี้จะแสดงโอกาสของการตอบถูกข้อสอบข้อหนึ่งของผู้สอบ โอกาสของผู้สอบที่จะตอบถูกแต่ละคนเป็นอิสระจากกันไม่ขึ้นอยู่กับการกระจายของกลุ่มหรือความสามารถของกลุ่ม รูปร่างโค้งของ i.c.c ไม่ขึ้นอยู่กับการกระจายของความสามารถในประชากรผู้สอบหรือรูปร่างของ i.c.c จะไม่แปรปรวน (Invariant) เมื่อข้อสอบข้อนี้ไปใช้กับตัวอย่างกลุ่มอื่น ๆ ในประชากรเดียวกัน

โมเดลต่าง ๆ ใน IRT

โมเดลต่าง ๆ ใน IRT ได้มาจากมโนทัศน์เรื่องความสัมพันธ์ระหว่างความสามารถ (Ability) กับการกระทำ (Performance) ถ้ามีความสามารถ (θ) สูงก็น่าจะมีการกระทำ (Performance) สูงด้วย ถ้ามีเกณฑ์ (r_i) ตัวหนึ่งเป็นตัวบอกว่า Y (performance) สูงแค่ไหนจึงจะทำข้อสอบ i ถูก ดังนั้น $Y > r_i$ แสดงว่าทำข้อสอบ r_i ถูก ถ้า $Y < r_i$ แสดงว่าทำข้อสอบ i ผิด

ภาพที่ 8 แสดงความสัมพันธ์ระหว่างความสามารถ (θ) กับ Performance (Y)



ดังนั้น ถ้านำโอกาสที่จะตอบถูก (พ.ท.ส่วนที่แรเงา) ในระดับความสามารถต่าง ๆ มาเขียนกราฟใหม่ จะได้เป็นรูป i.c.c ต่าง ๆ ซึ่ง i.c.c เหล่านี้จะแตกต่างกันไปตามฟังชั่นทางคณิตศาสตร์ และจำนวนพารามิเตอร์ที่ใช้อธิบาย i.c.c ของแต่ละโมเดลซึ่งเขียนเป็นรูปฟังชั่นทั่วไปคือ

$P(\theta) = f(\theta)$

$P(\theta)$ = ความน่าจะเป็นในการตอบถูก

θ = ความสามารถที่แท้จริง

f = ความสัมพันธ์

ซึ่ง $f(\theta)$ จะเปลี่ยนไปตามโมเดลต่าง ๆ ดังนี้

1. The latent linear model (Hambleton and Cook 1977 : 78)

$P_g(\theta) = b_g + a_g$

$p_g(\theta)$ คือความน่าจะเป็นของคนในระดับความสามารถ θ จะตอบข้อสอบ ย ถูก

a_g = ค่าอำนาจจำแนก

b_g = ค่าความยาก

2. The latent distance model (Torgerson 1958 : 374)

$$P_g(\theta) = \begin{cases} b_g - a_g ; & \theta \leq \theta_g \\ b_g + a_g ; & \theta > \theta_g \end{cases}$$

โมเดลนี้ใช้กับ Guttman scale และ scale อื่น ๆ

3. Normal ogive model ; (Lord and Novick 1968 : 366)

$$P_g(\theta) = \int_{-\infty}^{a_g(\theta - b_g)} \emptyset(t)\, dt \; ; \; g = 1,2,3.....n$$

$P_g(O)$, a_g,b_g มีความหมายเช่นเดียวกับในข้อ 1

O = ระดับความสามารถของผู้สอบ

$O(t)$ = Normal density function

4. Logistic Model (Hambleton and Cook 1977 : 81-84) ซึ่งมี i.c.c เหมือนกับ Normal Ogive Model ถ้าเพิ่มค่าคงที่ 1.7 เข้าไปใน Model แบ่งย่อยออกเป็น 3 โมเดล ดังนี้

4.1 Two parameter Logistic model Birnbaum (1968) ได้เสนอโมเดลลักษณะแอบแฝง (Latent trait model) ซึ่งเป็นโค้ง i.c.c มี 2 พารามิเตอร์ เป็นฟังชั่นการกระจายแบบ Logistic (Two-parameter logistic distribution function)

$$P_g(\theta) = e^{Da_g(\theta-b_g)}/(1 + e^{Da_g(\theta-b_g)}) \quad (g = 1,2.....n)$$

$P_g(\theta)$ = โอกาสที่ผู้สอบจะตอบข้อสอบได้ถูกต้องด้วยความสามารถระดับหนึ่ง

a_g = ค่าอำนาจจำแนกของข้อสอบ

b_g = ค่าดรรชนีความยากของข้อสอบ ณ จุดเปลี่ยนโค้ง

D = ค่าคงที่เรียกว่า Scaling Factor มีค่า 1.7

4.2 Three - parameter Logistic Model เป็นโมเดลที่มี 3 พารามิเตอร์เพิ่มขึ้นจากรูปแบบที่ 1. คือเพิ่มค่าการเดา (C_g) ของข้อสอบเข้ามาในโมเดล ด้วยมีโมเดลทางคณิตศาสตร์ดังนี้

$$P_g(\theta) = C_g + (1 - C_g)e^{Da_g(\theta-b_g)}/(1 + e^{Da_g(\theta-b_g)}) \quad (g = 1,2...n)$$

4.3 One - parameter Logistic Model (Rasch Model) Rasch ได้พัฒนาโมเดลของเขาขึ้นมาโดยไม่ได้ศึกษาโมเดลทาง Latent trait มาก่อนเลย เรียกว่า Rasch Model เมื่อได้ตีพิมพ์เสนอผลการศึกษาของเขาขึ้นมาก็ปรากฏว่าสอดคล้องกับโมเดลทาง Latent trait ในกรณีพารามิเตอร์เดียว Rasch Model เป็นกรณีพิเศษของ Two - parameter Logistic Model ของ Birnbaum ซึ่งทุกข้อมีค่า a_g เท่ากัน ดังสมการ

$$P_g(\theta) = e^{Da_g(\theta-b_g)}/(1 + e^{Da_g(\theta-b_g)}) \quad (g = 1,2.....n)$$

$$= e^{(\theta-b_g)}/(1 + e^{(\theta-b_g)}) \quad (\text{Da มีค่าคงที่รวมอยู่ใน Scale})$$

ข้อตกลงในเรื่องค่า a_g ของ Rasch นี้จะถูกละเมิด (Violeted) ถึงแม้จะบอกว่าโมเดลนี้มีความแกร่ง (Model Robustness) ก็ตามแต่ก็ยังไม่เป็นที่ตกลงกันได้

นอกจาก Logistic Model แล้วแฮมเบิลตันและคุก ยังได้เสนอโมเดลอื่น ๆ อีก 2 โมเดล คือ

1. Nominal Response Model จากโมเดลที่กล่าวมาในข้อ 1,2 และ 3 นั้นใช้กับข้อสอบที่ให้คะแนนเป็น 2 แบบ คือตอบถูกให้ 1 ผิดให้ 0 โมเดลแบบ Nominal Response Model เสนอโดย Bock (1972) และ Samejima (1972) ใช้ในกรณีที่ตัวเลือกแต่ละตัวมีคะแนนหรือน้ำหนักแตกต่างกัน แต่ละข้ออธิบายได้ด้วย Characteristic Curve ที่เสนอมาแล้ว วิธีการให้น้ำหนักคะแนนตัวเลือกแต่ละตัวแตกต่างกันนี้ช่วยทำให้ค่าความเที่ยง (Reliability) และค่าความตรง (Validity) ของการให้คะแนนดีขึ้น (Wang and Stanley : 1970 : 663)

จุดมุ่งหมายของโมเดลนี้ก็เพื่อจะให้มีความถูกต้องมากที่สุดในการประมาณค่าความสามารถของผู้สอบมีโมเดลทางคณิตศาสตร์ ดังนี้

$$P_{gk}(\theta) = e^{b_{gk}+a_{gk}\theta} / \sum_{h=1}^{m} e^{b_{gh}+a_{gh}\theta} \qquad (g = 1,2.....n)$$

$$k = 1,2.....,.m)$$

b_{gk} = ค่าความยากของข้อ g ที่ตัวเลือก k

a_{gk} = ค่าอำนาจจำแนกของข้อ g ที่ตัวเลือก k

ผลรวมของโอกาสของการเลือกตัวเลือกทุกตัวในแต่ละข้อมีค่าเท่ากับ 1 ถ้า m = 2 โมเดลนี้ก็เหมือนกับ Two - parameter Logistic Model

2. Other Models เป็นกรณีพิเศษของ Nominal Response Model พัฒนาโดย Samejima (1969) ใช้ในกรณีที่ตัวเลือกในข้อสอบแต่ละข้อให้คะแนนเป็นลำดับที่ (Order) เช่นกรณีแบบสอบวัดด้านทัศนคติ เป็นต้น

ความแกร่งของ IRT(Robustness of IRT) จากข้อตกลงเบื้องต้นของ IRT ข้างต้นจะเห็นว่าข้อตกลงทางด้าน Dimension และ Local independence เป็นการยากที่จะสร้างให้แบบสอบมีคุณสมบัติดังกล่าว จากการศึกษาของ Hambleton และนักวิจัยคนอื่น ๆ ในเรื่องนี้ ถึงแม้การศึกษาจะใช้วิธีการแตกต่างกัน ผลที่สรุปได้เหมือนกันคือ ตามข้อตกลงในโมเดลของ IRT นั้นไม่มีข้อมูลใด ๆ ที่จะมีคุณสมบัติครบสมบูรณ์ตามโมเดลของ IRT ที่กำหนดไว้ แต่โมเดลของ IRT ก็ยังสามารถใช้ได้กับข้อมูลประเภทนี้ เพราะผลการวิเคราะห์ลักษณะนี้ก็ยังสอดคล้อง (Fit) กับโมเดล สรุปแล้วโมเดลของ IRT มีความแกร่ง

เกี่ยวกับด้านคุณภาพของแบบสอบในการวัดแบบประเพณีนิยมที่ทำกันอยู่เราพิจารณาคุณภาพแบบสอบโดยดูค่าความเที่ยง (Reliability) ความตรง (Validity) และความคลาดเคลื่อนมาตรฐานของการวัด (Standard error of measurement) แต่ในการวัดผลตามแนวทางทฤษฎี IRT นี้เราพิจารณาค่า Test information function และ Item information function

ลอร์ด (Lord 1980 : 65) ได้นิยาม Information function หมายถึงปริมาณหนึ่งซึ่งมีค่าเป็นส่วนกลับกับความยาวยกกำลังสองของความกว้างของช่วงความเชื่อมั่นได้ของความคลาดเคลื่อนมาตรฐานรอบ ๆ ค่าที่ประมาณได้ของความสามารถของผู้สอบ (O) ที่หาได้จากคะแนน Y

Item information function ($I\{\theta,U_g\}$) ของข้อสอบแต่ละข้อกำหนดได้ดังนี้ (Birnbaum 1968 : 449)

$$I\{\theta,U_g\} = P_g(\theta)^2/P(\theta)Q_g(\theta)$$

เมื่อ $P_g(\theta)$ เป็น derivative ของ P_i with respect to θ หรือ slope ของ i.c.c ที่ระดับความสามารถ θ $P'(\theta) = \partial P(\theta) / \partial\theta$

เบรนบอม (Birnbaum 1968) (อ้างใน ผจงจิต อินทรสุวรรณ 2525 : 65-66) ได้แสดงให้เห็นว่า Test information function (I(θ)) คือผลรวมของ Item information function นั่นเอง แสดงในรูปสมการได้ดังนี้

$$I(\theta) = I(\theta, U_i) = (\Sigma_i P'_i)^2 / \Sigma P_i Q_i$$

เมื่อใช้น้ำหนักคะแนนที่เหมาะสม (Optimal Weight) กับข้อสอบแต่ละข้อเป็น $W_i(\theta) = P'_i / P_i Q_i$ แทนค่าลงในสมการข้างบนจะได้ค่า Test information function สูงสุดดังสมการข้างล่างนี้ (Lord 1980 : 74)

$$I(\theta) = (\Sigma_i P'^2_i / P_i Q_i)^2 / \Sigma_i P_i Q_i (P'_i / P_i Q_i)^2$$

$$= \Sigma_i P'^2_i / P_i Q_i$$

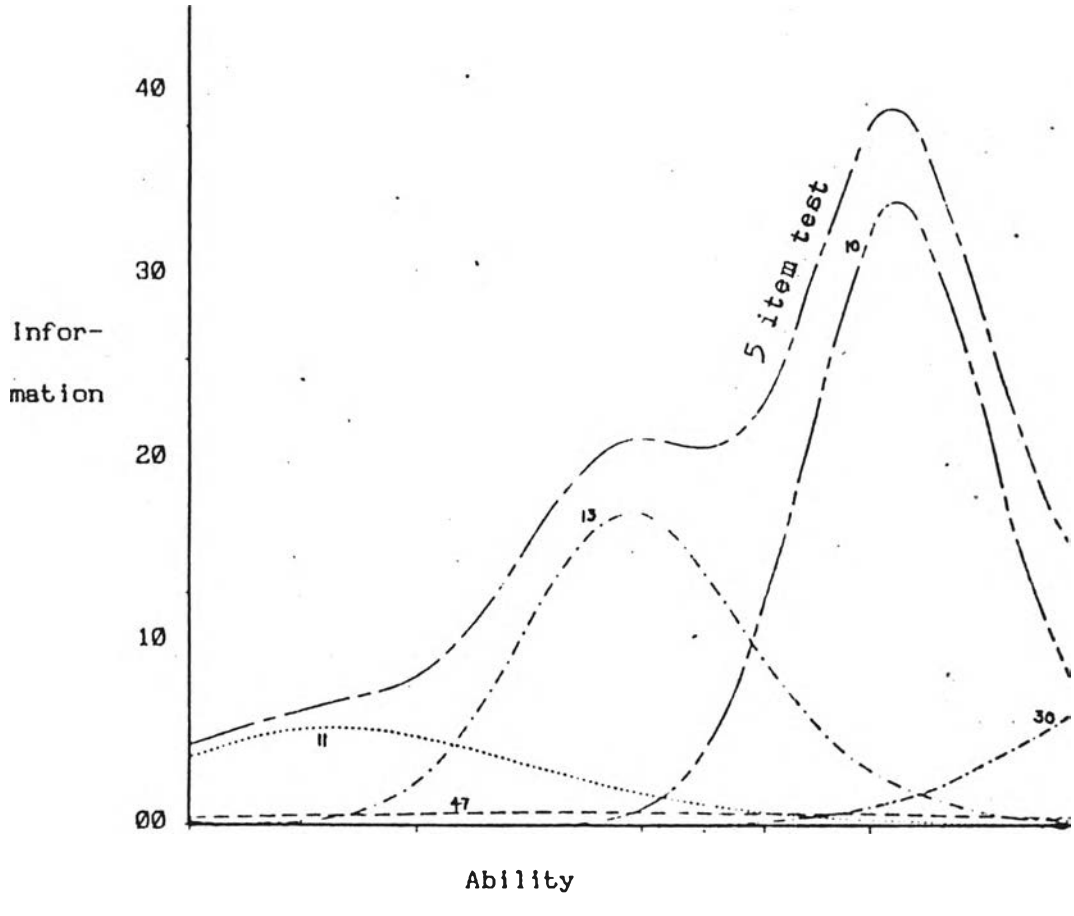
Test information function เป็นประมาณสูงสุดของ information ที่เราสามารถได้รับจากการตอบข้อสอบแต่ละข้อ โดยการตรวจให้คะแนนที่ใช้ Optimal score Y คะแนน Y นี้เป็นตัวประกอบเชิงเส้นตรง (Linear composite) ของคะแนนรายข้อ u_i (u_i = 0 หรือ 1) กับ Optimal weight(W_i) รูปสมการเชิงเส้นตรงดังกล่าวคือ $Y = \sum_{i=1}^{n} w_i u_i$ น้ำหนักนี้แตกต่างกันไปตามโมเดลที่แตกต่างกัน สำหรับ Logistic model นั้น Optimal weight ขึ้นอยู่กับระดับความสามารถ θ นั่นคือที่ระดับความสามารถสูง ๆ น้ำหนัก w_i นี้จะเป็นอิสระจากระดับความสามารถ แต่จะเป็นสัดส่วนโดยตรงกับ a_i เนื่องจากไม่มีการเดา ส่วนที่ระดับความสามารถต่ำ ๆ น้ำหนักของคำถามข้อยาก ๆ เริ่มลดลง (มีการเดาเกิดขึ้น) และยิ่งความสามารถลดลงอีกน้ำหนักจะเข้าใกล้ศูนย์ ข้อคำถามยาก ๆ เหล่านี้จะไม่ถูกนำมาใช้กับผู้สอบซึ่งมีความสามารถต่ำเช่นนี้ ประโยชน์โดยตรงของความรู้เรื่อง Information นี้คือ ถ้าเรามีกลุ่มของข้อคำถามที่เราทราบ Information curves เราก็สามารถสร้างข้อสอบฉบับหนึ่งให้มี Test information curve ณ ระดับหนึ่งของความสามารถตามที่เราต้องการได้ เนื่อ

จุดประสงค์ในการใช้ข้อสอบฉบับนี้นั้นโดยเฉพาะ เช่น เราต้องการสร้างข้อสอบฉบับหนึ่งเพื่อคัดเลือกนักเรียนให้ได้รับทุนเราก็ต้องการข้อสอบที่มีประสิทธิภาพสูงสุดที่ระดับสูง ๆ นั่นคือต้องการ Information curve สูงสุดที่ระดับความสามารถสูง ๆ เป็นต้น

ข้อสอบ 2 ฉบับที่วัดความสามารถอย่างเดียวกัน สามารถนำมาเปรียบเทียบกันได้ โดยใช้ค่าอัตราส่วนของ Information function ของข้อสอบ Y ต่อ Information function ของข้อสอบ X เราเรียกว่า Relative efficiency ของข้อสอบ Y (เมื่อเราเปรียบเทียบกับ X) ถ้า Relative efficiency ของข้อสอบ Y = 1.50 ที่ระดับหนึ่งของความสามารถ หมายความว่าเราต้องใช้ข้อคำถามจำนวน 1½ เท่าของข้อสอบฉบับ X เพื่อให้ได้อำนาจจำแนกที่ระดับนั้นเท่ากันกับที่ข้อสอบ Y ให้ Information curves ของ Item information curves และ Test information curves แสดงได้ดังนี้

ภาพที่ 9 Item information function ของข้อสอบ 5 ข้อ ที่วัดความสามารถในทางภาษา และค่า Test information function ของแบบสอบชุดนี้ โดยมีค่า Parameter ตามโมเดล Logistic ของ 5 ข้อ ดังนี้ (Lord 1980 : 22)

ข้อ	b_g	a_g	c_g
10	1.1	2.0	.05
11	-1.5	.9	.20
13	-0.1	1.6	.16
30	2.4	1.1	.09
47	-0.4	.4	.20



ประโยชน์ที่ได้จากการรู้ค่า Information เราสามารถสร้างแบบสอบหรือข้อสอบให้มีค่า Information ณ ระดับหนึ่งของความสามารถตามที่เราต้องการได้เพื่อจุดประสงค์ของการใช้ข้อสอบหรือแบบสอบฉบับนั้น เช่น ต้องการสร้างแบบสอบเพื่อคัดเลือก เราก็ต้องการแบบสอบที่ให้ค่า Information curve สูง ๆ ที่ระดับความสามารถสูง ๆ เป็นต้น

ตอนที่ 3 มโนทัศน์เกี่ยวกับการวิเคราะห์ความสอดคล้องระหว่างความสามารถกับความยากของข้อสอบ

ไรท์และสโตน (Wright and Stone 1979 : 69-79) กล่าวว่าโมเดลของราส์ช (Rasch) เกี่ยวกับการคำนวณและการวัดตั้งอยู่บนแนวคิดพื้นฐาน 2 ข้อ คือ

1. บุคคลที่มีความสามารถมากกว่าควรจะมีโอกาสประสบความสำเร็จในการทำข้อสอบมากกว่าคนที่มีความสามารถน้อยกว่า

2. คนแต่ละคนจะประสบความสำเร็จในข้อสอบที่ง่าย ๆ หรือข้อยากน้อยกว่าความสามารถของตนเอง

การวิเคราะห์ความสอดคล้องด้วยมือ (Analysis of Fit by Hand) ขั้นแรกของการวิเคราะห์เราจะดูค่าความแตกต่างระหว่างความสามารถของผู้สอบ (b_v) และความยากของข้อสอบ (d_i) หรือค่าของ $b_v - d_i$ ถ้าความแตกต่างนี้เป็นค่าบวกแสดงว่าข้อสอบข้อนี้ง่ายสำหรับผู้สอบ และผู้สอบก็มีโอกาสที่จะตอบถูกได้มากขึ้น ในทางตรงกันข้าม ถ้าค่า $b_v - d_i$ มีค่าเป็นลบก็แสดงว่าข้อสอบข้อนี้ยากสำหรับผู้สอบ และผู้สอบก็มีโอกาสที่จะตอบผิดมากขึ้นถ้าค่าเป็นลบมากขึ้น เพื่อให้เห็นชัดเจนตามแนวคิดนี้จะยกตัวอย่างผลการสอบของผู้สอบมา 6 คน ทำข้อสอบ 7 ข้อ ที่แบบแผนการตอบไม่สมเหตุสมผลตามแนวคิดของราส์ช (Rasch) ดังนี้

I 10300983

ตารางที่ 1 การตอบข้อสอบ 7 ข้อของผู้สอบ 6 คน

ผู้สอบคนที่	ข้อสอบที่							จำนวนข้อของการตอบที่ไม่คาดหวัง	ความสามารถของผู้สอบ (b_v)
	4	5	7	6	8	12	14		
11	0	1	1	1	1	0	0	1	-1.2
12	1	1	1	0	0	0	0	2	-1.2
17	1	0	1	1	1	0	0	1	-0.6
3	1	1	1	1	1	1	0	1	0.0
13	1	1	0	0	1	1	0	3	0.0
29	1	1	0	1	0	0	1	3	0.0
จำนวนข้อของการตอบที่ไม่คาดหวัง	1	1	2	2	2	2	1	11	
ความยากของข้อ (d_i)	-3.9	-3.3	-3.3	-2.9	-2.0	1.7	2.8		
	"1" คือคะแนนของการตอบที่คาดหวัง (Expected) "0" คือคะแนนของการตอบที่ไม่คาดหวัง (Unexpected)					"1" คือ Un Expected "0" คือ Expected			

จากตารางจะเห็นว่าคนที่ 11 ทำข้อสอบข้อ 4 ผิดทั้งที่มีความสามารถ (b_v) มากกว่าความยากของข้อ (d_i) ซึ่งตามแนวคิดของราส์ชแล้วจะต้องทำข้อ 4 ถูก คนที่ 12 ทำข้อ 7 และ 8 ผิดทั้งที่ความยากของข้อ 7 และ 8 น้อยกว่าความสามารถของเขา (b_v)

ตามแนวคิดของราส์ชก็น่าจะตอบถูก ส่วนคนที่ 3,13 และ 29 ที่ทำข้อ 12 และข้อ 14 ถูกนั้น ทั้ง ๆ ที่ความยากของข้อสอบ ทั้ง 2 ข้อนี้สูงกว่าความสามารถของผู้สอบซึ่งตามแนวคิดของราส์ชแล้ว ผู้สอบจะต้องตอบผิดจากตัวอย่างที่ยกมาในลักษณะที่ผิดไปจากโมเดลการตอบตามแนวคิดของราส์ช เราสามารถวิเคราะห์ดูความสอดคล้องของค่า b_v และ d_i ของแต่ละคนแต่ละข้อได้

โมเดลของราส์ชสำหรับการกระจายแบบไบนารี (Binary) ความน่าจะเป็นในการได้คะแนน (X_{vi}) ของผู้ตอบ v ในข้อ 1 เป็นดังนี้ (Wright and stone 1977 : 62 - 64)

$$P\{X_{vi}|\beta_v,\delta_i\} = \exp[X_{vi}(B_v-\delta_i)]/[1+\exp(\beta_v-\delta_i)] \qquad (1)$$

ในเมื่อ $X_{vi} = \begin{cases} 1 & \text{ถ้าตอบถูก} \\ 0 & \text{ถ้าตอบผิด} \end{cases}$

B_v = ค่าพารามิเตอร์ความสามารถของผู้ตอบ

S_i = ค่าพารามิเตอร์ความยากของข้อสอบ i

ค่า Likelihood Λ ของข้อมูลที่เป็นเมตริก ((X_{vi})) จากสมการ (1) เขียนเป็นค่า Λ ได้ดังนี้

$$\Lambda = \exp[\sum_v^N \sum_i^L X_{vi}(\beta_v-\delta_i)] \Big/ \prod_v^N \prod_i^L [1+\exp(\beta_v-\delta_i)] \qquad (2)$$

กำหนดให้ $\sum_i^L X_{vi} = r_v$ เป็นคะแนนของ คนที่ v

$\sum_v^N X_{vi} = s_i$ เป็นคะแนนของข้อสอบ i

ดังนั้น Take Log Likelihood (λ) สมการ (2) จะได้

$$\lambda = \ln \Lambda = \sum_{v}^{N} r_v \beta_v - \sum_{i}^{L} \delta_i \delta_i - \sum_{v}^{N} \ln \frac{1+\exp(B_v - \delta_i)}{} \qquad (3)$$

ผลรวมของ δ_i คือ $\sum \delta_i = 0$ จากสมการ (3) ทำ First และ Second Partial Derivative ของ with respect to B_v และ S_i ได้สมการ Derivative ดังนี้

$$\frac{\partial \lambda}{\partial \beta} = r_v - \sum_{i}^{L} \pi_{vi} \qquad v = 1,N \qquad (4)$$

$$\frac{\partial^2 \lambda}{\partial B_v^2} = - \sum_{i}^{L} \pi_{vi}(1-\pi_{vi}) \qquad (5)$$

และ

$$\frac{\partial \lambda}{\partial \delta_i} = -\delta_i + \sum_{v}^{N} \pi_{vi} \qquad i = 1,L \qquad (6)$$

$$\frac{\partial^2 \lambda}{\partial \delta_i^2} = - \sum_{v}^{N} \pi_{vi}(1-\pi_{vi}) \qquad (7)$$

ในเมื่อค่า $\pi_{vi} = \exp(\beta_v - \delta_i) / [1+\exp(\beta_v - \delta_i)]$ (8)

ในการประมาณค่าความน่าจะเป็นของแต่ละคนที่มีคะแนน r จะประสบความสำเร็จในการทำข้อสอบ i เรากำหนดให้

b_r = ค่าประมาณความสามารถ (Ability estimate) สำหรับแต่ละคนที่ได้คะแนน r

d_i = ค่าประมาณความยาก (Difficulty estimate) ของข้อสอบ i

n_r = เป็นจำนวนของผู้สอบที่ได้คะแนน r

ดังนั้นสูตรที่ใช้ในการประมาณค่าความน่าจะเป็นของแต่ละคนที่จะประสบความสำเร็จในข้อ i จากสมการ (8) เขียนเป็นค่าประมาณของความน่าจะเป็นที่จะประสบความสำเร็จในข้อ i ของแต่ละคนได้ดังสมการ

$$P_{ri} = \exp(b_r - d_i)/[1+\exp(b_r - d_i)] \qquad (9)$$

ค่า P_{ri} เป็นค่าประมาณความน่าจะเป็นของ π_{vi}

จากตารางที่แสดงผลการตอบของผู้สอบ 6 คนที่ไม่สมเหตุสมผลตามโมเดลของราส์ชชั้นแรกเราสามารถหาค่าความน่าจะเป็นของแต่ละคนที่ได้คะแนน X_{vi} จากสมการ (9) ได้ดังนี้

$$P_{vi} = \exp(b_v - d_i)/[1+\exp(b_v - d_i)] \qquad (10)$$

ในเมื่อ b_v = ความสามารถของผู้สอบ v

d_i = ความยากของข้อสอบที่ i

เราสามารถใช้ค่า P_{vi} เป็นค่าที่คาดหวัง (Expected value) ของ X_{vi} ค่าความแปรปรวนที่ประมาณได้ (Expected variance) ของคะแนน X_{vi} คือ $\pi_{vi}(1-v_i)$ ซึ่งเราประมาณได้ด้วย $P_{vi}(1 - P_{vi})$ เนื่องจาก π_{vi} เป็นค่าพารามิเตอร์ของความน่าจะเป็นที่จะประสบความสำเร็จ ดังนั้นสูตรที่ใช้ในการตรวจสอบความสอดคล้องคือ

$$Z_{vi} = (X_i - P_i)/[P_i(1-P_i)]^{1/2} \qquad (11)$$

Z_{vi} = ค่าส่วนที่เหลือมาตรฐาน (Standard Residual)

ค่า Z_{vi} จะมีการประจายเป็นโค้งปกติ (Normal Distribution) ซึ่งมีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 1

$Z_{vi} \simeq N(0,1)$

เมื่อยกกำลังสองค่า Z_{vi} ก็จะมีการกระจายเป็น Chi-Square ที่ Degree of freedom เท่ากับ 1

$$Z_{vi} \approx \chi^2_1$$

จากสมการ (11) $Z_x = (X - P) / [P(1 - P)]^{1/2}$

ถ้า X = 0 แล้ว $Z_o = (-P) / [P(1 - P)]^{1/2}$

ถ้า X = 1 แล้ว $Z_1 = (1 - P) / [P(1 - P)]^{1/2}$

เนื่องจากค่า $P = \exp(b - d) / [1 + \exp(b - d)]$

แล้ว $P/(1-P) = \exp(b - d)$

และ $(1-P) / P = \exp(d - b)$

ดังนั้น $Z_o = -\exp[(b - d)]^{1/2}$

$$Z_o^2 = \exp(b - d) \qquad (12)$$

และ $Z_1 = +\exp[(d - b)]^{1/2}$

$$Z_1^2 = \exp(d - b) \qquad (13)$$

จากสมการ 12 และ 13 จะเห็นว่า exp(b-d) ชี้ให้เห็นถึงค่าที่ไม่คาดหวังของคำตอบที่ผิดทั้ง ๆ ที่เป็นข้อสอบที่ง่าย ขณะที่ exp(d-b) ชี้ให้เห็นถึงค่าที่ไม่คาดหวังของคำตอบที่ถูกทั้ง ๆ ที่เป็นข้อสอบที่ยาก ค่า Z_o^2 และ Z_1^2 สามารถหาได้สำหรับค่า X_{vi} ที่มีค่าเป็น 0 หรือ 1 เพื่อดูความสมเหตุสมผลในแต่ละข้อแต่ละคนที่ตอบไม่เป็นไปตามแนวคิดของราส์ช

จากตัวอย่างการตอบข้อสอบของผู้สอบ 6 คน ที่ยกมาเป็นตัวอย่างในตอนต้นจะแสดงตัวอย่างในการคำนวณหาค่า Z_o^2 และ Z_1^2 ของแต่ละคนแต่ละข้อที่ผลการตอบไม่เป็นไปตามความหวัง ดังนี้

ขั้นแรกหาค่า (b-d) ของค่า X_{vi} ที่ไม่คาดหวังว่าจะตอบผิดในข้อที่ง่าย และค่า (d-b) ของค่า X_{vi} ที่ไม่คาดหวังว่าจะตอบถูกในข้อที่ยาก ดังตารางต่อไปนี้

ตารางที่ 2 ค่า (b-d) ของข้อง่ายและ (d-b) ของข้อยากของข้อสอบที่ไม่คาดหวัง

ผู้สอบที่	ข้อสอบที่							ค่าความสามารถของผู้สอบ (b_v)
	4	5	7	6	8	12	14	
11	2.7							-1.2
12				1.7	0.8			-1.2
17		2.7						-0.6
3						1.7		0.0
13			3.3	2.9		1.7		0.0
29			3.3		2.0		2.8	0.0
ค่าความยากของข้อ d_i	-3.9	-3.3	-3.3	-2.9	-2.0	1.7	2.8	
	ข้อง่ายค่า X_{vi} ควรเป็นดังนี้					ข้อยากค่า X_{vi} ควรเป็นดังนี้		
	1 ค่าที่คาดหวัง (Expected)					Ø คือที่คาดหวัง (Expected)		
	Ø ค่าที่ไม่คาดหวัง (Unexpected)					1 ค่าที่ไม่คาดหวัง (Unexpected)		

ขั้นที่สองหาค่า exp(b-d) ของข้อง่ายและ exp(d-b) ของข้อที่ยากในแต่ละค่าของ (b-d) และ (d-b) เป็นค่า Z^2 ดังตาราง

ตารางที่ 3 ค่า Z^2 ของข้อสอบ 7 ข้อเฉพาะข้อที่ไม่คาดหวัง

ผู้สอบคนที่	ค่า Z^2 ของข้อสอบที่							รวม
	4	5	7	6	8	12	14	
11	15							15
12				6	2			8
17		17						15
3						6		6
13			27	18		6		51
29			27		7		17	51
รวม	15	15	54	24	9	12	17	146

จากตารางแสดงว่า Z^2 ซึ่งค่านี้จะมี Distribution ใกล้เคียงกับ χ^2 ไรท์และสโตน (Wright and Stone) ได้เสนอวิธีหาค่าระดับนัยสำคัญในแต่ละค่าของ χ^2 ไว้ดังนี้ $P = 1/(1 + Z^2)$ เมื่อได้ค่า P แล้วสามารถไปเปิดตาราง χ^2 ที่ Degree of freedon = 1 ในแต่ละค่าของ P แล้วนำมาเปรียบกับค่า Z^2 ในตาราง ถ้าค่า Z^2 มากกว่า แสดงว่าผู้สอบใช้ความสามารถไม่สอดคล้องกับความยากของข้อ

การวิเคราะห์ความสอดคล้องของผู้สอบ (Person Fit) ค่า Z^2 จากที่กล่าวมานั้นใช้สำหรับการวิเคราะห์ข้อสอบ แต่ละข้อแต่ละคน ถ้าจะวิเคราะห์ผู้สอบคนใดคนหนึ่งที่ตอบข้อสอบทั้งหมดในแบบสอบเราต้องหาค่ารวมของ Z^2 (Standard mean squares residual) ของผู้สอบคนนั้นก่อนดังนี้

$$V_v = \sum_i^L Z^2_{vi}/(L-1) \sim F_{L-1}$$ (มีการกระจายใกล้เคียงกับ F - Distribution ที่ df = L-1)

แล้วใช้สูตรวิเคราะห์ความสอดคล้องของผู้สอบ ดังนี้

$$t_v = [\ln(V_v) + V_v - 1]\,[(L-1)/8]^{1/2} \sim N(0,1)$$

ในเมื่อ i = 1, L = ข้อสอบทั้งหมดหรือความยากของแบบสอบ ถ้าวิเคราะห์ข้อสอบ (Item) แต่ละข้อจากผู้สอบทั้งหมดจะได้สูตรคล้าย ๆ กับสูตรวิเคราะห์ความสอดคล้องของผู้สอบดังนี้

$$V_i = \sum_v^N Z_i / (N - 1) \sim F_{N-1,\infty}$$

แล้วใช้สูตรวิเคราะห์ความสอดคล้องของข้อสอบแต่ละข้อของผู้สอบทั้งหมด ดังนี้

$$t_1 = [\ln(V_i) + V_i - 1]\,[(N-1)/8]^{1/2} \sim N(0;1)$$

เลข 8 ซึ่งเป็นตัวหารในสูตร t_v และ t_1 มาจากการเฉลี่ยค่าของ Standardization ค่าตรงข้ามกัน 2 ค่า คือ

$$t_1 = (V - 1)\,[(L - 1)/2]^{1/2} \sim N(0,1)$$

$$t_2 = [\ln(V)]\,[(L - 1)/2]^{1/2} \sim N(0,1)$$

$$t = (t_1 + t_2)/2 = [\ln(V) + (V - 1)]\,[(L - 1)/8]^{1/2} \sim N(0,1)$$

ถ้าวิเคราะห์ความสามารถของผู้ตอบตามสถิติที่เสนอมาข้างต้น จะได้ผลการวิเคราะห์ดังตัวอย่างในตารางต่อไปนี้ (Wright and Stone 1979 : 167-169)

ตารางที่ 4 ค่า d,(d-b) และค่า Z^2 ของผู้สอบ 3 คน

ผู้สอบ	ความสามารถ	ค่าสถิติ	รูปแบบการตอบสนองข้อสอบแต่ละข้อ														
	b	d	-6.2	-4.3	-4.1	-2.7	-2.6	-2.6	-2.1	-2.1	-1.5						
12M	-2.8	x	0	1	1	1	1	1	0	0	0						
		(d-b)	-3.4	-1.5	-1.3	0.1	0.2	0.2	0.7	0.7	1.3						
		z^2	30.0	0.2	0.3	1.1	1.2	1.2	0.5	0.5	0.3						
		d	-4.1	-2.7	-2.6	-2.6	-2.1	-2.1	-1.5	-1.0	-0.9	-0.8	-0.5	-0.1	1.4	1.9	2.0
35F	-0.3	x	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0
		(d-b)	-3.8	-2.4	-2.3	-2.3	-1.8	-1.8	-1.2	-0.7	-0.6	-0.5	-0.2	0.2	1.7	2.2	2.3
		z^2	0.0	0.1	0.1	10.0	0.2	0.2	0.3	2.0	0.5	0.6	1.2	1.2	0.2	0.1	0.1
		d	-0.5	-0.1	1.4	1.9	2.0	2.9	3.3	3.3	4.5	5.8	6.3				
88M	3.2	x	1	1	1	0	1	1	0	1	0	0	0				
		(d-b)	-3.7	-3.3	-1.8	-1.3	-1.2	-0.3	0.1	0.1	1.3	2.6	3.1				
		z^2	0.0	0.0	0.2	3.7	0.3	0.7	0.9	1.1	0.3	0.1	0.0				

จากตาราง จะได้ค่า Z_i^2 ของข้อสอบแต่ละข้อของผู้สอบ 3 คน ที่ยกมาเป็นตัวอย่างขั้นต่อไปจะตรวจสอบเพื่อความสอดคล้องข้อผู้สอบแต่ละคนน่าจะสอดคล้องกับ Pattern การตอบของตัวเองหรือไม่ ดังแสดงค่าให้เห็นในตารางที่ 5

ตารางที่ 5 ความสอดคล้องระหว่างความสามารถกับความยากง่ายของข้อสอบ

ผู้สอบ	ความสามารถ	ค่าสถิติ			
		$SS(Z^2)$	df(L-1)	MS(V)	t
12M	-2.8	35.0	8	4.4	4.9*
35F	-0.3	16.8	14	1.2	0.5
88M	3.2	7.3	10	0.7	-0.7

*ไม่สอดคล้อง

จากตารางจะเห็นว่าผู้สอบ 12M ผลการทำแบบสอบหรือแบบแผนการตอบไม่สอดคล้องกับโมเดลการตอบที่ควรจะเป็น ส่วนผู้ตอบ 35F และ 88M ไม่พบความแตกต่างระหว่างแบบแผนการตอบกับโมเดลที่ควรจะเป็น

นอกจากนี้ไรท์และสโตน (Wright and Stone) ยังได้เสนอเกณฑ์ในการพิจารณาค่า t_v ดังนี้ ถ้าค่า $t_v<3$ แสดงว่าแบบแผนการตอบตรงตามความสามารถ ถ้าค่า $3<t_v<5$ แสดงว่าการตอบไม่ตรงตามความสามารถ ซึ่งเกิดจากการตอบถูกในข้อที่ยากกว่าความสามารถมาก ๆ กับตอบผิดในข้อที่ง่ายกว่าความสามารถมาก ๆ แต่ส่วนอื่น ๆ เป็นแบบปกติ ส่วนค่า $t_v>5$ แสดงว่าแบบแผนการตอบไม่ตรงตามความสามารถ

สรุปแนวทางการวิเคราะห์ความสอดคล้องระหว่างความสามารถของผู้สอบกับความยากของข้อตามแนวคิดราส์ช ทำได้ดังนี้

1. วิเคราะห์ความสอดคล้องของข้อสอบแต่ละข้อของแต่ละคนโดยหาค่า Z^2 ถ้าค่า Z^2 ออกมา Misfit ก็แสดงว่าผู้สอบทำข้อสอบนั้นผิด ในกรณีที่ค่าความยากของข้อน้อยกว่าความสามารถของผู้สอบ หรือถูกในกรณีที่ค่าความยากของข้อมากกว่าความสามารถของผู้สอบ

2. วิเคราะห์ความสอดคล้องของข้อสอบหรือแบบสอบ ทั้งฉบับของแต่ละคนโดยหาค่า t_v ถ้าค่า t_v ออกมา Misfit ก็แสดงว่าผู้สอบทำข้อสอบทั้งหมดไม่สอดคล้องกับความสามารถของตน

จากแนวคิดของราส์ชในเรื่องของการวิเคราะห์ความสอดคล้องนี้ผู้วิจัยได้นำมาใช้ในการจำแนกผู้สอบออกตามแบบแผนการตอบ 2 แบบแผน คือ แบบแผนการตอบสอดคล้องกับความสามารถ และแบบแผนการตอบไม่สอดคล้องกับความสามารถ

<u>ตอนที่ 4</u> มโนทัศน์เกี่ยวกับการตรวจให้คะแนน

วิธีการตรวจให้คะแนนที่ใช้ในการวิจัยมี 3 วิธีคือ 1. วิธีตามประเพณีนิยมให้คะแนน 0,1 2. วิธีประยุกต์ใช้ IRT (P(θ)) 3. วิธีให้น้ำหนักคะแนนที่เหมาะสม (W_i) วิธีให้คะแนนตามประเพณีนิยมนั้นเป็นการให้คะแนนแบบสอบรายข้อโดยให้ 1 สำหรับข้อที่ผู้สอบตอบถูกหรือกาเครื่องหมายตรงคำเฉลยและให้ 0 สำหรับข้อที่ผู้สอบตอบผิดหรือกาเครื่องหมายไม่ตรงกับคำเฉลย ส่วนวิธีให้คะแนนโดยการประยุกต์ใช้ IRT (P(θ)) กับวิธีให้น้ำหนักคะแนนที่เหมาะสมมีรายละเอียดดังนี้

วิธีตรวจให้คะแนนโดยการประยุกต์ใช้ IRT (P(θ) จะใช้โมเดลที่มี 3 พารามิเตอร์ (Three-parameter Logistic Model) ซึ่งมีโมเดลทางคณิตศาสตร์สำหรับหาค่า P (θ) ที่กล่าวมาแล้วในตอนที่ 2 ดังนี้

$$P_g(\theta) \text{ หรือ } P_i(\theta) = C_g + (1-C_g)\{e^{Da_g(\theta-b_g)}\}/\{1 + e^{Da_g(\theta-b_g)}\}$$

(g หรือ i = 1,2,.....n)

เมื่อ $P_g(\theta)$ หรือ $P_i(\theta)$ = โอกาสที่ผู้สอบจะตอบข้อสอบข้อ g ได้ถูกต้องด้วยความสามารถ θ

a_g = ค่าอำนาจจำแนกของข้อสอบข้อ g

b_g = ค่าความยากของข้อสอบข้อ g

D = ค่าคงที่ที่เรียกว่า Scaling Facter มีค่า 1.7

การให้คะแนนโดยประยุกต์ใช้ IRT หาได้โดยเอาค่า $P_i(\theta)$ แต่ละข้อคูณกับค่า U_i ในเมื่อ U_i มีค่า 0 เมื่อผู้สอบตอบข้อ i ผิดและมีค่า 1 เมื่อผู้สอบตอบข้อ i ถูกเขียนเป็นสูตรได้ดังนี้

$$X = \sum_{i=1}^{n} P_i(\theta)U_i$$

ลอร์ดและโนวิค (Lord and Novick 1968 charpter 2, quoted in Lord 1980 : 45-46) ได้แสดงวิธีหาค่าคะแนนจริง (True score) ของผู้สอบที่ทำข้อสอบถูกจำนวนหนึ่งจากแบบสอบ โดยคิดจากผลบวกของค่า $P_i(\theta)$ เขียนเป็นสูตรได้ดังนี้

$$\xi = \sum_{i=1}^{n} P_i(\theta)$$

ในเมื่อ ξ = คะแนนจริงของผู้สอบ

$P_i(\theta)$ = ค่าความสามารถของผู้สอบที่มีความสามารถ θ ที่จะตอบข้อ i ถูก

ค่า $P_i(\theta)$ แต่ละค่าจะมากขึ้นอยู่กับค่า θ และคะแนนจริงที่หาได้จากผลรวมของค่า $P_i(\theta)$ ของแต่ละค่าก็จะมากขึ้นด้วย

วิธีตรวจให้คะแนนโดยวิธีให้น้ำหนักคะแนนที่เหมาะสม (W_i)

ลอร์ด (Lord 1980 : 73-76) ได้เสนอวิธีคิดคะแนนตามน้ำหนักคะแนน กำหนดให้ $Y = \sum_i W_i U_i$ เมื่อ W_i เป็นน้ำหนักของข้อสอบและ U_i เป็นตัวแปรแบบไบโนเมียลที่เป็นอิสระหรือคะแนน 0,1 จากการตอบข้อสอบ ดังนั้นค่า Expectation ซึ่งเป็นคะแนนของผู้สอบทั้งหมดที่ทำแบบสอบจะได้

$$Y = \sum_{i=1} W_i U_i$$ เมื่อค่า U_i คือ 0 คะแนนเมื่อตอบผิด
1 คะแนนเมื่อตอบถูก

จากค่า Information Function

$$I\{\theta, \textstyle\sum_i W_i U_i\} = (\sum_i W_i P_i)^2 / \sum_i W_i^2 P_i Q_i$$

ถ้าค่าน้ำหนักของข้อสอบเป็น 1 เท่ากันหมด คะแนน Y ก็คือคะแนนที่ผู้สอบตอบถูกทั้งหมด (X) ค่า Test Information function ของคะแนนที่ตอบถูก (X) นี้คือ

$$I\{\theta, X\} = (\sum_i P'_i)^2 / \sum_i P_i Q_i$$

ลอร์ดได้เสนอวิธีหาค่าน้ำหนักคะแนนประจำข้อไว้คือ $W_i(\theta) = P'_i / P_i Q_i$ เมื่อแทนค่านี้ในสมการของ Test information function จะได้ดังนี้

$$I\{0, \textstyle\sum_i P_i U_i / P_i Q_i\} = (\sum_i P'^2_i / P_i Q_i)^2 / \{\sum_i P_i Q_i (P'_i / P_i Q_i)^2$$

$$= \sum_i P'^2_i / P_i Q_i$$

เราได้ค่า Information function ของคะแนนรวมที่คิดน้ำหนักเป็น $\Sigma_i(P_i'/P_iQ_i)U_i$ เหมือนกับ Test information function ซึ่งเป็น Information ที่มีค่าสูงสุดด้วยวิธีการให้น้ำหนักที่เหมาะสมของข้อสอบแต่ละข้อด้วยสูตร

$$W_i(\theta) = P_i'(\theta)/P_i(\theta)Q_i(\theta)$$

ในทางปฏิบัติเราไม่รู้ค่า θ ของผู้สอบแต่ละคนและค่า $W_i(\theta)$ แต่เราสามารถใช้วิธีการประมาณค่าของ θ และ $W_i(\theta)$ ได้ในกรณี Three - parameter Logistic Model เราจะประมาณค่าได้ดังนี้

$$P_i = Da_iQ_i(P_i - C_i)/(1 - C_i)$$

จากสมการวิธีการให้น้ำหนักคะแนนที่เหมาะสม $W_i(\theta)$ และสมการหาค่า P_i จะได้สมการวิธีการให้น้ำหนักคะแนนที่เหมาะสมของข้อสอบแต่ละข้อใหม่ ดังนี้

$$W_i(\theta) = Da_i(P_i - C_i)/\{P_i(1 - C_i)\} = Da_i/(1 - C_i^{-DL_i})$$

เมื่อ $L_i = a_i(\theta - b_i)$

การให้น้ำหนักคะแนนที่เหมาะสมสำหรับแต่ละคน $(W_i(\theta))$ ดังกล่าวนั้นเราไม่สามารถจะหาได้อย่างแน่นอน เพราะเราไม่สามารถรู้ค่าความสามารถของผู้สอบ (θ) ดังนั้นวิธีการคร่าว ๆ สำหรับการให้น้ำหนักคะแนนข้อสอบแต่ละข้อเราใช้ค่าความยากของข้อสอบแต่ละข้อ (P_i) แทนค่า $P_i(\theta)$ ในสมการของ $W_i(\theta)$ สมการที่ใช้หาค่าน้ำหนักคะแนนที่เหมาะสมคร่าว ๆ สำหรับข้อสอบแต่ละข้อจึงเปลี่ยนไป

$$W_i = \{a_i/(1 - C_i)\}\{(P_i - C_i)/P_i\}$$

เมื่อค่า D = 1.7 เป็นค่าคงที่ของข้อสอบทุกข้อที่เราสามารถเอาออกจากสมการได้ ดังนั้นค่า w_i จึงขึ้นอยู่กับค่าอำนาจจำแนกของข้อสอบ (a_i) ค่าความยากของข้อสอบ (P_i) และการเดาของข้อสอบ (C_i)

ลอร์ด (Lord 1977 a : 123) ได้วิเคราะห์ผลการสอบและพบว่า

1. เมื่อความสามารถของผู้สอบมีมาก โค้งแสดงค่าน้ำหนักคะแนนจะขนานกับแกนนอน

2. เมื่อความสามารถของผู้สอบต่ำ โค้งแสดงค่าน้ำหนักคะแนนจะขึ้นอยู่กับความยากของข้อสอบ

นอกจากนี้ลอร์ดยังสรุปอีกว่าในกรณีที่ผู้สอบมีความสามารถสูงมาก การให้คะแนนตามประเพณีนิยมจะดีกว่าการให้คะแนนข้อสอบต่างกัน เพราะค่าน้ำหนักคะแนนระดับสูง ๆ จะคงที่แต่ตัวผู้สอบมีความสามารถต่ำหรือปานกลางการให้คะแนนข้อสอบต่างกันจะถูกต้องมากกว่า

ในปี ค.ศ. 1975 ลอร์ด (Lord 1975 : 10) ได้สรุปความคิดเห็นเกี่ยวกับการให้คะแนนอีกดังนี้

1. การให้คะแนนตามประเพณีนิยมสำหรับผู้สอบที่ทำไม่ครบทุกข้อ ผลการสอบจะเป็นตัวทำนายไม่ดีเท่ากับการให้คะแนนโดยวิธีให้คะแนนข้อสอบต่างกัน

2. การให้คะแนนข้อสอบต่างกันทำให้แบบสอบมีค่าความเที่ยงและความตรงเพิ่มขึ้นแม้จะเพิ่มขึ้นเพียงเล็กน้อยแต่ก็เป็นการพัฒนาที่คุ้มค่า เพราะสามารถลดจำนวนข้อสอบลงได้อย่างน้อยประมาณ 1/9

ในปี ค.ศ. 1982 ซูและชูศักดิ์ (Hsu and C.Khampalikit 1982 : 7) ได้เปรียบเทียบผลการให้คะแนนโดยวิธีให้ค่าน้ำหนักคะแนนที่เหมาะสม (w_i) กับการให้คะแนนตามวิธีประเพณีนิยมแก่ข้อสอบแต่ละข้อ ในวิชาภาษาจีนและวิชาประวัติศาสตร์ พบว่าวิธีให้ค่าน้ำหนักคะแนนที่เหมาะสม (w_i) ทำให้แบบสอบมีค่าอำนาจจำแนกและค่าความเที่ยงสูงกว่าวิธีการให้คะแนนตามประเพณีนิยม

จากที่กล่าวมาข้างต้นจะเห็นว่า การประยุกต์ใช้ทฤษฎี IRT มาให้คะแนนและการให้น้ำหนักคะแนนที่เหมาะสมรายข้อ (w_i) เป็นสิ่งที่น่าสนใจที่จะนำมาศึกษาเพื่อแก้ปัญหาความคลาดเคลื่อนของการวัดที่เกิดจากการไม่ได้รับความร่วมมือ ในการสอบกรณีที่ผู้สอบรับรู้ว่าผลการสอบไม่มีผลโดยตรงต่อตัวเขาแต่นำผลไปวิจัย

ตอนที่ 5 มโนทัศน์เกี่ยวกับความคลาดเคลื่อนมาตรฐานของการวัด (Standard Error of Measurement)

กิลลิคเซน (Gulliksen 1950 : 42) ได้ให้ความหมายของความคลาดเคลื่อนของการวัดว่าเกิดจากการใช้คะแนนสอบ (Observed score) แทนคะแนนจริง (True score) ซึ่งเหมือนกับ ลอร์ด (Lord 1980 : 46) ได้ให้ความหมายไว้ว่า ความคลาดเคลื่อนของการวัด คือความแตกต่างระหว่างคะแนนสอบ (Observed score) กับคะแนนจริง (True score) ซึ่งมีสมการที่เหมือนกันดังนี้

$$E = X - T \quad \text{or} \quad e = X - \xi$$

เมื่อ E หรือ e = ความคลาดเคลื่อน

X = คะแนนสอบ

T หรือ ξ = คะแนนจริง หรือ ลอร์ด เรียกว่า Equivalent True Score (ξ)

ในการวัดผลจะเกิดความคลาดเคลื่อน 2 แบบ คือความคลาดเคลื่อนคงที่ (Constant Error) และความคลาดเคลื่อนแบบสุ่ม (Randon Error) แต่นักวัดผลให้ความสนใจค่าความคลาดเคลื่อนแบบสุ่มมากกว่า ซึ่งตามทฤษฎีการวัดแบบเดิม (Classical Test Theory) และทฤษฎีการวัดแบบ IRT (Item Response Theory) ได้ให้นิยามเกี่ยวกับความคลาดเคลื่อนแบบสุ่ม ซึ่งถือว่าเป็นความคิดเบื้องต้นอันเป็นพื้นฐานของการวัดเหมือนกันดังนี้ (Gullilsen 1950 : 6-9, Lord 1980 : 3-4)

นิยามที่ 1. ถ้าทำการวัดคนจำนวนมากแล้วค่าเฉลี่ยหรือค่าที่คาดหวัง (Expectation) ของ E เท่ากับ Ø

$$M_E = Ø$$

หรือ $$U_E/T = Ø$$

นิยามที่ 2. สหสัมพันธ์ระหว่าง T กับ E มีค่าเป็น Ø

$$r_{TE} = Ø$$

หรือ $$e_{ET} = Ø$$

นอกจากนี้กิลลิคเซน (Gulliksen) ยังให้นิยามอีกว่า ค่าสหสัมพันธ์ระหว่าง E_1 ในแบบสอบหนึ่งกับ E_2 ในอีกแบบสอบหนึ่งที่คู่ขนานกันมีค่าเป็น Ø

$$r_{E_1E_2} = Ø$$

เกี่ยวกับค่าความแปรปรวน (Variance) ของคะแนน T,X และ E ได้แสดงความสัมพันธ์ของความแปรปรวนเหมือนกันดังสมการ

$$\sigma_x^2 = \sigma_T^2 + \sigma_E^2 \qquad (1)$$

$$1 = \sigma_T^2/\sigma_x^2 + \sigma_E^2/\sigma_x^2 \qquad (2)$$

เกี่ยวกับเครื่องหมายกิลลิคเซนใช้เครื่องหมายเท่ากับ (=) ส่วนลอร์ดใช้เครื่องหมาย Equivatent (≡)

ค่าสัดส่วนระหว่างความแปรปรวนของคะแนนจริง (σ_T^2) กับความแปรปรวนของคะแนนสอบ (σ_x^2) คือค่าความเที่ยงของแบบสอบ (Test Reliability : r_{xx})

(Ebel 1965 : 331) หรือสหสัมพันธ์ยกกำลังสองระหว่างคะแนนที่ปรากฏกับคะแนนจริง ρ^2_{xT} (Lord 1980 : 5) ดังนั้นจากสมการ 2 ก็จะได้

$$1 = \rho^2_{xT} + \sigma_E^2/\sigma_x^2 \quad (3)$$

$$\sigma_E^2 = \sigma_x^2(1 + \rho^2_{xT}) \quad (4)$$

$$\sigma_E = \sigma_x\sqrt{1 - \rho^2_{xT}} \quad (5)$$

หรือตามทฤษฎีการวัดแบบเดิมจะได้สูตรที่มีสัญลักษณ์ดังนี้

$$S_e = S_x\sqrt{1 - r_{xx}} \quad (6)$$

จะเห็นว่าสมการ 5 และสมการ 6 เป็นสมการเดียวกันที่ใช้สำหรับหาค่าความคลาดเคลื่อนมาตรฐานของการวัด (Standard error of measurement)

ความหมายของความคลาดเคลื่อนมาตรฐานของการวัดและการคำนวณค่า

ความหมายของความคลาดเคลื่อนมาตรฐานของการวัดทั้งทฤษฎีแบบเดิมและทฤษฎี IRT ให้ความหมายไว้เหมือนกัน คือ ความแปรปรวนของคะแนนคลาดเคลื่อนที่คิดจากผู้สอบแต่ละคน $\sigma_E^2 = \sigma^2(E_{ij})$ เพื่ออ้างอิงความแปรปรวนของความคลาดเคลื่อนของการวัด ถอดรากที่สองของ σ_E^2 จะได้ σ_E เราเรียกว่าความคลาดเคลื่อนมาตรฐานของการวัด (Lord and Novick 1968 : 60) กิลลิคเซน (Gulliksen 1950 : 33) ได้เสนอสูตรเพื่อหาค่าความคลาดเคลื่อนมาตรฐานของการวัดซึ่งสอดคล้องกับความหมายดังกล่าวไว้ดังนี้

$$S_e^2 = \Sigma(X-T)^2/N$$

$$S_e = \sqrt{\Sigma(X-T)^2/N}$$

สูตรนี้เป็นสูตรพื้นฐานของการหาค่าความคลาดเคลื่อนมาตรฐานของการวัดเป็นสูตรในการหาค่าความคลาดเคลื่อนมาตรฐานของการวัดทางตรง ตามทฤษฎีการวัดแบบเดิม ถ้าเรารู้ค่าคะแนนจริง (True Score) ของผู้สอบแต่ละคนเราก็สามารถหาค่าความคลาดเคลื่อนแต่ละคนได้โดยสูตร E = X-T ดังตัวอย่าง (Ebel 1965 : 332)

นักเรียน	คะแนนจริง	ความคลาดเคลื่อนของการวัด	คะแนนที่ปรากฏ
อาร์ลี	18	-2	16
ดัน	9	+1	10
ยีน	15	+2	17
จอห์น	21	+1	22
วิคเตอร์	12	-2	10
ค่าเฉลี่ย	15	0	15
ค่าความแปรปรวน	18	2.8	20.8

ค่าความเที่ยง = σ_T^2 / σ_x^2 = 18/20 = 0.865

ความคลาดเคลื่อนมาตรฐานของการวัดที่คำนวณทางตรงตามสูตรได้ดังนี้

$$S_e = \sqrt{\Sigma(X-T)^2/N}$$

$$= \quad 2.8 \quad = 1.67$$

แต่ในทางปฏิบัติตามทฤษฎีแบบเดิมเราไม่สามารถหาคะแนนจริงเป็นรายบุคคลได้ (Thorndike 1964 quoted in Lord and Novick 1960 : 27) ดังนั้นการหาค่า S_e จึงไม่สามารถจะหาได้ทางตรง แต่หาได้ทางอ้อมโดยใช้สูตร

$$S_e = S_x\sqrt{1-r_{XX}}$$

$$S_x = 20.8 = 4.56$$

จากตัวอย่างข้างต้นหาค่า S_e โดยทางอ้อมดังนี้

$$S_e = 4.56\sqrt{1-0.865}$$
$$= 4.56 \times 0.367$$
$$= 1.67$$

การคำนวณหาค่า S_e ทั้งทางตรงและทางอ้อมได้ค่าเท่ากัน

ตามทฤษฎีแบบเดิมคะแนนจริง (T) ในทางปฏิบัติเราไม่สามารถจะหาค่าได้ เพราะตามนิยามกิลลิคเซน (Gulliksen 1950 : 30) คะแนนจริงของแต่ละคนหาได้จากค่าเขตจำกัด (Limit) ของคะแนนเฉลี่ยจากการสอบแต่ละคนที่สอบแบบสอบคู่ขนาน K ฉบับ ดังสูตร

$$T_i = \lim_{K\to\infty} \sum_{g=1}^{k} X_{ig}/K$$

ในเมื่อ g = แบบสอบชุดที่ g (g = 1,2,3.....k)

K = จำนวนแบบสอบที่มีถึง ∞ ฉบับ

X_{ig} = คะแนนสอบของคนที่ 1 ที่สอบแบบสอบฉบับ g

แอลเลนและเยน (Allen and Yen 1979 : 57) ได้ให้นิยามของคะแนนจริงไว้ว่าเป็นค่า Expected ของคะแนนสอบ (X) ($\mathcal{E}(X) = T$) ซึ่งตามนิยามนี้ค่า T หาได้จากค่าเฉลี่ยของ Theoretical Distribution ของคะแนน X ที่ได้จากการสอบด้วยแบบสอบชุดเดียวกันหลาย ๆ ครั้ง ซึ่งแต่ละครั้งเป็นอิสระจากกัน

แต่ตามทฤษฎี IRT ค่าคะแนนจริงของผู้สอบแต่ละคนหาได้จากนิยามของลอร์ดและโนวิค (Lord and Novick 1968 : Chapter 2 quoted in Lord 1980 : 45) ที่ว่าคะแนนจริงของผู้สอบที่สอบแบบสอบฉบับหนึ่งนั้นคือ $U_{x/\theta} = \sum_{i=1}^{n} P_i(\theta)$ เมื่อค่า $U_{x/\theta}$ เป็นค่าเฉลี่ยเราหาค่าคะแนนจริงของผู้สอบแต่ละคน ได้ดังนี้

$$\xi = \sum_{i=1}^{n} P_i(\theta)$$

ξ = คะแนนจริง (Equivalent true scores)

$P_i(\theta)$ = โอกาสที่ผู้สอบจะตอบข้อสอบข้อ i ถูกเมื่อมีความสามารถ θ

ในแต่ละค่าของ $P_i(\theta)$ นี้จะเพิ่มมากขึ้นเป็นฟังชั่นกับความสามารถของผู้สอบ (θ) และคะแนนจริงที่ได้ก็จะเพิ่มขึ้นเป็นฟังชั่นกับความสามารถ ดังนั้นความคลาดเคลื่อนของการวัดของผู้สอบตามนิยามข้างต้นเขียนเป็นสูตรได้ดังนี้ $e = X - \xi$ ถ้ายกกำลังสอง ค่า e ของผู้สอบแต่ละคนแล้วนำมารวมกันหารด้วยจำนวนผู้สอบทั้งหมดก็จะเป็นค่าเฉลี่ยของความแตกต่าง หรือค่าความแปรปรวนาของความคลาดเคลื่อน (Error Variance) เขียนเป็นสูตรได้ดังนี้

ตามวิธีการให้คะแนนโดยการประยุกต์ใช้ IRT

$$X = \sum_{i=1}^{n} P_i(\theta)U_i$$

จาก $e = X - \xi$

$$= \sum_{i=1}^{n} P_i(\theta)U_i - \sum_{i=1}^{n} P_i(\theta)$$

$$S^2_{e \cdot \xi} = \sum^{N}\left(\sum_{i=1}^{n} P_i(\theta)U_i - \sum_{i=1}^{n} P_i(\theta)\right)^2/N$$

หรือ $S^2_{e.\xi} = 1/N \sum^{N} \sigma^2_{e.\xi}$

ถอดรากที่สองของค่า $S^2_{e.\xi}$ จะได้ $S_{e.\xi}$ เรียกว่าค่าความคลาดเคลื่อนมาตรฐานของการวัด ซึ่งเหมือนกับวิธีหาตามทฤษฎีการวัดแบบเดิมที่กล่าวไว้ข้างต้น

อีกวิธีหนึ่งในการหาค่าความคลาดเคลื่อนมาตรฐานของการวัดโดยอาศัยทฤษฎี IRT ที่แตกต่างจากที่กล่าวมาแล้ว คือการตรวจให้คะแนนใช้ค่าน้ำหนักคะแนนที่เหมาะสมเพื่อให้ได้คะแนนสอบ (X) เพราะตามสูตร $S_e = S_x\sqrt{1 - r_{xx}}$ และ $S^2_{e.\xi} = 1/N\sum^{n} \sigma^2_{e/\xi}$ นั้น คะแนนสอบ (X) ได้มาจากการให้คะแนน 0 ข้อที่ตอบผิด และคะแนน 1 ข้อที่ตอบถูก แต่มีการตรวจอีกวิธีหนึ่งที่ผู้วิจัยนำมาศึกษาด้วย คือการให้น้ำหนักคะแนนที่เหมาะสม (W_i) ซึ่งมีรายละเอียดในตอน 4 ซึ่งมีสูตรที่ใช้ในการให้คะแนนสอบเป็น $U_{wu/e} = \Sigma_i W_i U_i$

ถ้าแทนค่า $U_{wu/e}$ ด้วยคะแนน X จะได้สูตรหาคะแนนสอบดังนี้

$$X = \Sigma_i W_i U_i$$

ตามสูตรหาคะแนนจริง $\xi = \Sigma_i W_i(\theta)$ ดังนั้นหาค่าคะแนนคลาดเคลื่อนเป็นรายคน ได้คือ

$$e = X - \xi$$

$$e = \Sigma_i^n W_i U_i - \Sigma_i^n W_i(\theta)$$

เมื่อได้ค่าคะแนนคลาดเคลื่อน (e) รายคน ถ้ายกกำลังสองของค่า e แล้วนำมารวมกันหารด้วยจำนวนผู้สอบทั้งหมด เราจะได้ความแปรปรวนของความคลาดเคลื่อน (Error Variance) ถอดรากที่สองก็จะเป็นค่าความคลาดเคลื่อนมาตรฐานของการวัด

สรุป จากวิธีหาค่าความคลาดเคลื่อนมาตรฐานของการวัด ที่จะนำมาใช้ในการวิจัยมี 3 วิธี คือ

วิธีที่ 1 เป็นการหาค่าความคลาดเคลื่อนมาตรฐานของการวัดแบบรวม ๆ หรือทางอ้อม เพราะค่าคะแนนจริง (T) ตามนิยามของทฤษฎีแบบเดิมหาไม่ได้ในทางปฏิบัติ สูตรจะเป็นดังนี้

$$S_e = S_x\sqrt{1 - r_{xx}}$$

แต่ถ้าเรารู้ค่าคะแนนจริงดังตัวอย่างของ Ebel เราก็สามารถหาค่าความคลาดเคลื่อนมาตรฐานของการวัดแบบตรงได้ โดยหาค่าความคลาดเคลื่อนเป็นรายคนก่อนแล้วใช้สูตรดังนี้

$$S_e = \sqrt{\Sigma(X-T)^2/N}$$

วิธีที่ 2 เป็นการหาค่าความคลาดเคลื่อนมาตรฐานของการวัดแบบทางตรงจากวิธีการตรวจให้คะแนนโดยการประยุกต์ใช้ IRT เพราะเราสามารถหาค่าคะแนนจริงตามแนวทางที่ ลอร์ด และโนวิค (Lord and Novick) ให้นิยามไว้ S_e หาได้ดังนี้

$$S_{e\cdot\hat{\theta}} = \sqrt{\sum^{N}\left(\sum_{i=1}^{n} P_i(\theta)U_i - \sum_{i=1}^{n} P_i(\theta)\right)^2/N}$$

หรือ $$S_{e\cdot\hat{\theta}} = \sqrt{1/N\sum_{i=1}^{N} \sigma_e^2/\hat{\theta}}$$

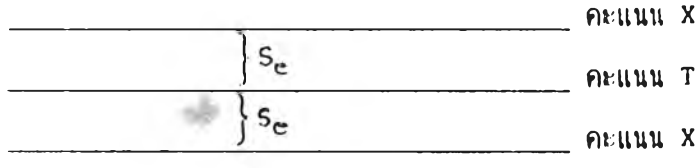
วิธีที่ 3 เป็นการหาค่าความคลาดเคลื่อนมาตรฐานของการวัดแบบทางตรงเหมือนวิธีที่ 2 แต่คะแนนสอบ (X) หาได้จากการตรวจให้คะแนน โดยวิธีใช้น้ำหนักคะแนนที่เหมาะสม (W_i) และคะแนนจริงหาได้ตามนิยามของลอร์ด และโนวิค S_e หาได้ดังนี้

$$S_{e\cdot\hat{\theta}} = \sqrt{\sum^{N}\left(\sum^{n} W_iU_i - W_i(\theta)\right)^2/N}$$

หรือ $$S_{e\cdot\hat{\theta}} = \sqrt{1/N\sum_{i=1}^{N}\sigma^2_{e/\hat{\theta}}}$$

สูตรการหาค่า S_e ของทั้ง 3 วิธีที่กล่าวมานี้มีแหล่งที่มาจากนิยามและโมเดลเดียวกัน คือเริ่มต้นด้วยโมเดล $E = X - T$ ดังที่กล่าวมาตอนต้น

ค่าความคลาดเคลื่อนมาตรฐาน (S_e) ที่ได้ทั้ง 3 วิธีเป็นค่าเฉลี่ยของความคลาดเคลื่อนที่คิดจากผู้สอบทุกคนไม่ว่าจะคิดทางตรงหรือทางอ้อม ค่า S_e เป็นค่าที่บอกให้ทราบว่าคะแนนสอบ (X) ห่างจากคะแนนจริง (T) ซึ่งถือว่าเป็นจุดหลักเฉลี่ยแล้วเท่าไร ซึ่งทั้ง 3 วิธีที่สามารถแปลความหมายได้เหมือนกัน ดังรูป



ความสัมพันธ์กันระหว่างค่าความเที่ยงและความคลาดเคลื่อนมาตรฐานของการวัดแบบสอบใดที่มีค่าความเที่ยงมาก ค่าความคลาดเคลื่อนมาตรฐานของการวัดก็จะน้อย ซึ่งพิจารณาได้จากสูตรการคำนวณหาค่า S_e

$$S_e = S_x\sqrt{1 - r_{xx}}$$

ดังนั้นปัจจัยต่าง ๆ ที่ส่งผลกระทบต่อค่าความเที่ยงของแบบสอบก็ย่อมจะส่งผลกระทบโดยตรงต่อค่าความคลาดเคลื่อนมาตรฐานของการวัดด้วย ซึ่งเรื่องนี้ได้มีนักวัดผลกล่าวไว้หลายท่าน แม้ว่าจะเป็นทฤษฎีการวัดแบบดั้งเดิม (Classical Theory) หรือทฤษฎีการวัดแบบ IRT การให้ได้มาซึ่งข้อมูลทางด้านผลสัมฤทธิ์ทางการเรียนก็ต้องอาศัยแบบสอบเหมือนกัน ปัจจัยต่าง ๆ ที่ก่อให้เกิดความคลาดเคลื่อนของการวัดก็สามารถใช้อธิบายแหล่งความคลาดเคลื่อนของทั้ง 2 ทฤษฎีได้

ครอนบัคและเมอร์วิน (Cronbach and Merwin 1955 : 337) กล่าวว่าค่าความเที่ยงของแบบสอบชนิดเลือกตอบนั้นขึ้นอยู่กับความใกล้เคียงของตัวเลือกถ้ามีความ

ใกล้เคียงกันมาก ๆ จะทำให้ค่าความเที่ยงสูง ส่วนอีเบล (Ebel 1965 : 20) ก็กล่าวว่าความเที่ยงขึ้นอยู่กับองค์ประกอบของการสอบหลายประการ คือลักษณะของข้อคำถามการดำเนินการสอบ เวลา เงื่อนไขในการสอบและการเรียงลำดับข้อสอบ อาดัมส์ (Adams 1964 : 340) ก็ได้กล่าวว่าคุณภาพของแบบสอบชนิดเลือกตอบจะมีมากหรือน้อยเพียงใดขึ้นอยู่กับการเขียนตัวเลือกของข้อคำถามในแบบสอบฉบับนั้นเป็นสำคัญ ดังนั้นการเขียนตัวลวงที่ดีมีความยากใกล้เคียงกันจะทำให้แบบสอบมีคุณภาพสูงขึ้น ซึ่งคำกล่าวของ อาดัมส์ (Adams) ก็สอดคล้องกับคำกล่าวของครอนบัค เมอร์วิน (Cronbach and Merwin) และอีเบล (Ebel) ซึ่งกล่าวถึงความใกล้เคียงกันของตัวเลือกในแบบสอบปรนัยชนิดเลือกตอบ นอกจากนี้เมเรนส์และเลแมนน์ (Mehrens and Lahmann 1975 : 100-103) ได้กล่าวถึงปัจจัยที่มีอิทธิพลต่อค่าความเที่ยงมี 5 ปัจจัย คือ

1. ความยาวของแบบสอบ (Test Lenght) เมื่อแบบสอบมีความยาวเพิ่มขึ้นค่าความเที่ยงจะมากขึ้นตามสูตรการหาค่าความเที่ยงแบบแบ่งครึ่ง (Split - half) ของสเปียร์แมน - บราวน์ (Spearman - Brown) ดังนี้

$$r_{xx} = Kr/(1 + (K - 1)r)$$

เมื่อ r_{xx} = ค่าความเที่ยงที่คำนวณได้เมื่อแบบสอบยาวเพิ่มขึ้นเป็น K เท่า

r = ค่าความเที่ยงของแบบสอบชุดเดิมหรือชุดเริ่มต้น

K = ค่าอัตราส่วนของข้อสอบชุดใหม่ที่เพิ่มขึ้น

สมมติว่ามีแบบสอบชุดหนึ่งมีค่าความเที่ยง .60 ถ้าเพิ่มความยาวของแบบสอบชุดนี้เป็น 3 เท่า เราจะคำนวณหาค่าความเที่ยงของแบบสอบชุดใหม่ได้ดังนี้

$$r_{xx} = 3x.60/\ 1+(3-1)x.60$$

$$= .818$$

2. เวลาในการทำแบบสอบ (Speed) ถ้ากำหนดเวลาในการทำแบบสอบไม่เหมาะสมจะมีผลต่อค่าความเที่ยงของแบบสอบ เช่น ถ้ากำหนดเวลาน้อยเกินไป ผู้สอบทำไม่ทันก็อาจจะเดาข้อสอบที่เหลือ ซึ่งจะทำให้การกระจายของคะแนนแตกต่างจากแบบสอบที่ให้เวลาเพียงพอ แต่ถ้าหากให้เวลามากเกินไปในกรณีแบบสอบ Speed Test จะทำให้คนเก่งและคนอ่อนได้คะแนนเท่า ๆ กัน คะแนนมีการกระจายน้อยก็จะมีผลต่อค่าความเที่ยงเหมือนกัน

3. ความเป็นเอกพันธ์ของกลุ่มผู้สอบ (Group Homogeneity) ถ้าปัจจัยอย่างอื่น ๆ มีผลต่อแบบสอบ 2 ฉบับพอ ๆ กันแล้ว แบบสอบที่ใช้สอบกับผู้สอบที่เป็นกลุ่มวิวิธพันธ์กัน (Heterogeneous) จะให้ค่าความเที่ยงมากขึ้น จากสูตรดังนี้

$$r_{xx} = 1 - S_e^2/S_x^2$$

เมื่อค่า S_e^2 คือความแปรปรวนของความแตกต่างระหว่างคะแนนสอบ (Observed Score) กับคะแนนจริง (True Score) ค่า S_e^2 จะมีค่าคงที่ แม้ว่ากลุ่มผู้สอบจะเปลี่ยนไป แต่ค่า S_x^2 จะมีค่าเพิ่มมากขึ้นเมื่อกลุ่มผู้สอบมีความแตกต่างกันมากขึ้น ดังนั้นค่าความเที่ยงของแบบสอบมีค่ามากขึ้นถ้าใช้สอบกับกลุ่มวิวิธพันธ์ (Heterogeneous) และแบบสอบจะมีค่าความเที่ยงลดลงถ้าสอบกับกลุ่มที่เป็นเอกพันธ์กัน เช่น กลุ่มผู้สอบที่ผ่านการคัดสรรแล้วอย่างกลุ่มที่ผ่านการคัดเลือกเข้ามหาวิทยาลัยมาแล้ว

4. ความยากง่ายของข้อสอบ (Item Difficulty) ข้อสอบที่ง่ายเกินไปทุกคนทำถูกหมด ในทางตรงกันข้ามข้อสอบที่ยากเกินไปทุกคนทำผิดหมด จะมีการกระจายของคะแนนน้อยจะมีผลให้ค่าความเที่ยงต่ำ ดังนั้นข้อสอบควรมีค่าความยากง่ายพอดีจะทำให้แบบสอบมีค่าความเที่ยงสูง

5. ความเป็นปรนัย (Objectivity) แบบสอบที่มีการให้คะแนนที่เป็นปรนัยจะทำให้ค่าความเที่ยงสูงกว่าแบบสอบที่มีการให้คะแนนไม่เป็นปรนัย

กรอนลันด์ (Gronlund 1971 : 112 - 123) ได้กล่าวไว้ที่นอกเหนือจาก เมเรนส์และเลแมนน์ (Mehrens and Lahmann) ถึงปัจจัยที่มีอิทธิพลต่อค่าความเที่ยงของแบบสอบอีก คือวิธีการประมาณค่า การคิดค่าความเที่ยงมีหลายวิธีแต่ละวิธีให้ผลต่างกัน และแต่ละวิธีก็เหมาะสำหรับจุดมุ่งหมายเฉพาะในการตีความของค่าความเที่ยง ค่าความเที่ยงที่คิดแบบแบ่งครึ่ง (Split - half method) จะให้ค่าความเที่ยงสูง คิดแบบการสอบซ้ำ (Test - retest method) จะให้ค่าความเที่ยงปานกลางคิดแบบใช้ฟอร์มคู่ขนานโดยไม่ได้กำหนดช่วงเวลาการสอบ (Equivalent - forms without time interval) จะให้ค่าความเที่ยงปานกลางถึงค่าสูงสุดและคิดแบบใช้ฟอร์มคู่ขนาน โดยกำหนดช่วงเวลาการสอบ (Equivalent - forms method with time interval) จะให้ค่าความเที่ยงน้อยที่สุด

เกี่ยวกับปัจจัยที่มีผลต่อความเที่ยงตามที่ฮอฟคินส์ (Hopkins 1964 : 271 - 281) ได้ทำการศึกษาหาปัจจัยต่าง ๆ ที่มีผลต่อความเที่ยงของแบบสอบ และได้เสนอปัจจัยไว้ดังนี้

1. ความเร็วในการทำแบบสอบ
2. การเดา
3. รูปแบบ (Format) ของแบบสอบ
4. กิจนิสัยในการทำแบบสอบ (Response style)
5. คำสั่งในการทำแบบสอบ
6. เวลาที่กำหนดในการทำแบบสอบ
7. การทำข้อสอบไม่ครบทุกข้อ
8. ระยะเวลาในการสอบซ้ำ
9. เวลาหรือโอกาสทำแบบสอบ
10. ความเหน็ดเหนื่อยและความเบื่อหน่ายในการทำแบบสอบ

ตอนที่ 6 งานวิจัยที่เกี่ยวข้องกับการเพิ่มคุณภาพของแบบสอบ

การตรวจให้คะแนนที่ทำกันอยู่ในปัจจุบันนี้ ในคำสั่งชี้แจงจะให้นักเรียนเลือกเฉพาะคำตอบที่ถูกที่สุดเพียงคำตอบเดียว ถ้าตอบถูกจะได้ 1 คะแนน ผิดได้ 0 ก็มีการวิจารณ์กันว่าไม่ยุติธรรมเลยสำหรับผู้ที่มีความรู้จริง ๆ ในเรื่องนั้นตอบถูกได้คะแนน 1 กับผู้ที่ไม่มีความรู้เลย หรือมีความรู้บางส่วนแต่ตอบถูกก็ได้ 1 คะแนนเหมือนกัน ซึ่งการตอบนี้อาจจะตอบแบบเดาก็ได้บังเอิญกาตัวเลือกที่ถูก เกี่ยวกับปัญหาอย่างนี้จึงได้มีผู้เสนอแนวทางในการแก้ไขในการตอบขึ้นดังที่ เพรสซี่ (Pressey 1950 : 417-447 quoted in Whetton and Childs 1981 : 336) ได้พัฒนาแบบสอบเลือกตอบให้สามารถบอกคะแนนและตัวถูกให้ทราบทันทีในขณะทำแบบสอบวิธีตอบแบบใหม่ต่อมาได้เรียกว่าวิธีตอบจนถูก (Answer - until - correct หรือ AUC) วิธีการนี้อาจทำให้พิสัยของคะแนนของแต่ละข้อคำถามเพิ่มขึ้น และทำให้ส่วนของความรู้ต่อเนื่องกันไป แต่วิธีตอบจนถูกนี้มีปัญหาตรงกระดาษคำตอบที่ต้องทำขึ้นมาเป็นพิเศษ ซึ่งเป็นการยุ่งยากและสิ้นเปลือง ต่อมา คูมบ์ส (Coombs 1953 : 308-313) ได้เสนอให้มีวิธีตอบโดยให้เลือกคำตอบที่ผิดแทนการเลือกคำตอบถูก คูมบ์ส ให้ความเห็นว่าการเลือกคำตอบที่ผิดซึ่งมีอยู่หลายตัว เลือกในข้อคำถามหนึ่ง ๆ นั้น ความรู้ที่วัดได้จะสมบูรณ์มากกว่า เพราะผู้ตอบรู้ว่าสิ่งใดผิดสิ่งใดถูก วิธีนี้สามารถวัดส่วนของความรู้ได้ละเอียดกว่าวิธีธรรมดา สำหรับการตรวจให้คะแนนนั้นคูมบ์สให้คะแนนดังนี้ ถ้านักเรียนทำเครื่องหมายผิดแก่ตัวเลือกที่ผิดจริง ให้คะแนนตัวเลือกละ 1 คะแนน แต่ถ้านักเรียนทำเครื่องหมายผิดแก่ตัวเลือกที่ถูก ตัวเลือกนั้นจะได้คะแนน 1-k (k คือจำนวนตัวเลือกทั้งหมดในแต่ละข้อ) เพราะฉะนั้นถ้ามี 4 ตัวเลือก คะแนนที่ได้จะอยู่ระหว่าง -3 ถึง +3 ซึ่งดีกว่าวิธีธรรมดาที่มี 2 ช่วงคะแนน

วิธีที่คูมบ์สเสนอแนะนับว่ามีข้อดีที่ทำให้พิสัยของคะแนนกว้างขึ้น สามารถจำแนกความรู้ของนักเรียนได้ดีกว่าวิธีให้คะแนน 0 - 1 ที่ทำกันอยู่ แต่วิธีการนี้ก็ยังมีข้อบกพร่องที่ผู้ตอบจะมีโอกาสได้คะแนนจากตัวเลือกที่ถูกและผิดต่างกัน อย่างเช่น แบบสอบที่มี 4 ตัวเลือก นักเรียนคนหนึ่งรู้ว่าตัวเลือกใดถูก ตัวเลือกใดผิด เขาสามารถทำเครื่องหมายลงบนตัวเลือกทุกตัวได้ถูกต้องได้คะแนน 3 ส่วนอีกคนหนึ่งรู้เฉพาะตัวเลือกที่ถูก แต่ไม่รู้ว่าตัว

เลือกใดผิด เขาทำเครื่องหมายลงบนตัวเลือกที่ผิดและให้คะแนน 3 เหมือนกันทั้ง ๆ ที่นักเรียนทั้ง 2 คนนี้มีความรู้ไม่เท่ากัน

นอกจากนี้ อนันต์ ศรีโสภา (2516 : 13-19) ได้เสนอวิธีการตอบวิธีใหม่โดยให้ตอบทุกตัวเลือก การตอบวิธีนี้นักเรียนจะต้องพิจารณาตอบทุกตัวเลือกว่าตัวเลือกใดถูกตัวเลือกใดผิด ซึ่งจะทำให้เป็นการวัดความรู้ในข้อคำถามนั้น ๆ ได้สมบูรณ์ยิ่งกว่าการเลือกคำตอบถูกเพียงคำตอบเดียว และยังสามารถทราบความรู้ที่ผิดของนักเรียนได้ดีกว่าวิธีธรรมชาติอีกด้วย การให้คะแนนก็ให้เป็นรายตัวเลือก คือถ้านักเรียนทำเครื่องหมายได้ตรงกับสภาพเป็นจริงว่าตัวเลือกนั้นถูกหรือผิด จะได้ตัวเลือกละ 1 คะแนน ถ้าทำเครื่องหมายตรงข้ามกับสภาพความเป็นจริง จะได้ตัวเลือกละ -1 คะแนน ซึ่งวิธีการนี้จะทำให้คะแนนที่ได้จากการวัดถูกต้อง และเหมาะสมกับความรู้ของนักเรียนมากขึ้น

วิธีการอีกวิธีหนึ่งที่ใช้เพื่อลดผลที่เกิดจากการเดาก็คือ วิธีการที่เรียกว่าวิธีบอกความมั่นใจในการตอบ (Confidence Weighting Procedures) หรือแบบสอบความมั่นใจ (Confidence Testing) วิธีนี้ผู้ตอบจะต้องบอกระดับความมั่นใจของข้อที่เขาเลือกว่าจะถูกต้องเพียงไร แบบบอกความมั่นใจในการตอบนี้มีแตกต่างกันหลายวิธี ในครั้งแรกเริ่มใช้ในระหว่างปี ค.ศ. 1930 ถึง 1939 ในฐานะเป็นวิธีในการเพิ่มข้อมูลให้กับการสอบและลดผลที่จะเกิดจากการเดาของแบบสอบประเภทถูก-ผิด (Echternacht 1972 : 218)

ในเรื่องวิธีตอบโดยให้บอกความมั่นใจในการตอบด้วยนี้ นนทนา เผือกผ่อง (2510 : 30-33) ได้ให้ความเห็นสนับสนุนวิธีตอบแบบนี้ โดยกล่าวว่าการที่นักเรียนตอบคำถามแบบสอบชนิดเลือกตอบได้ถูกต้องนั้น ผู้ตรวจไม่ทราบเลยว่านักเรียนตอบถูกเพราะว่ามีความรู้จริงหรือตอบถูกเพราะเดา ฉะนั้นเพื่อที่จะวัดดูความรู้ของนักเรียนให้แน่นอน และเพื่อป้องกันการเดา จึงเสนอให้มีวิธีการตอบโดยบอกความมั่นใจในการตอบให้แตกต่างไปจากวิธีตอบแบบธรรมดาที่ใช้กันอยู่ในปัจจุบันนี้ กล่าวคือนอกจากผู้ตอบจะต้องเลือกตัวเลือกที่คิดว่าเป็นคำตอบที่ถูกต้องแล้ว ผู้ตอบยังต้องตัดสินใจอีกว่า ตัวเลือกที่เป็นคำตอบที่ถูกที่สุดนั้น

ตอบด้วยความมั่นใจว่าถูกต้องจริงหรือไม่มั่นใจในคำตอบเท่าใดนัก กระดาษคำนอบก็ต้องเปลี่ยนเสียใหม่เพื่อให้การตอบแต่ละตัวเลือกมีช่องให้ผู้ตอบบอกความมั่นใจในการตอบด้วย การตรวจให้คะแนนก็ต้องให้แตกต่างจากธรรมดา โดยให้น้ำหนักความมั่นใจมาเกี่ยวข้องด้วย เช่น ให้คะแนนเป็น 5,4,3,2 และ 1 เมื่อตอบถูกและบอกว่ามั่นใจมากที่สุด มั่นใจมาก ค่อนข้างมั่นใจ ไม่ค่อยมั่นใจ และไม่มั่นใจเลย ตามลำดับ ถ้านักเรียนตอบผิดก็จะได้คะแนนเป็น -5,-4,-3,-2 และ -1 เมื่อบอกว่ามั่นใจมากที่สุด มั่นใจมาก ค่อนข้างมั่นใจ ไม่ค่อยมั่นใจ และไม่มั่นใจเลย ตามลำดับ หรือจะให้คะแนนเป็นอย่างอื่นที่แตกต่างจากนี้ไปก็ได้ แต่ต้องให้น้ำหนักของความมั่นใจในการตอบมาเกี่ยวข้องด้วย

การทำแบบสอบโดยบอกความมั่นใจนี้มีลักษณะต่าง ๆ กัน หรือมีเทคนิคแตกต่างกันขึ้นอยู่กับการออกแบบของผู้วิจัย แต่ในหลักการทั่วไปคล้าย ๆ กัน ด้านงานวิจัยที่เกี่ยวกับการตรวจให้คะแนนแบบสอบชนิดเลือกตอบมีการทำกันมากทั้งในประเทศและต่างประเทศ โดยเฉพาะต่างประเทศมีการศึกษาค้นคว้ากันมาก ซึ่งจะขอเสนองานวิจัยที่ทำกันภายในประเทศก่อนดังนี้

อรวรรณ ตันฑ์เจริญรัตน์ (2517 : 23 - 28) ได้ทำการศึกษาเปรียบเทียบวิธีการตอบและการตรวจให้คะแนนวิธี Ø-1 วิธีของคูมบ์ส และวิธีใหม่ตามแนวความคิดของอนันต์ ศรีโสภา โดยใช้แบบสอบวิชาภาษาไทย กลุ่มตัวอย่างที่ใช้เป็นนักศึกษาวิทยาลัยหมู่บ้านจอมบึง จำนวน 214 คน โดยแบ่งออกเป็น 3 กลุ่ม ดังนี้

กลุ่ม 1 จำนวน 83 คน ใช้วิธีการตอบและตรวจให้คะแนนด้วยวิธี Ø-1

กลุ่ม 2 จำนวน 74 คน ใช้วิธีการตอบและตรวจให้คะแนนด้วยวิธีของคูมบ์ส

กลุ่ม 3 จำนวน 57 คน ใช้วิธีการตอบและตรวจให้คะแนนด้วยวิธีของอนันต์

ทั้งนี้โดยศึกษาเฉพาะความเชื่อมั่นและเปอร์เซนต์การเดา ตามเงื่อนไขวิธีการตอบและตรวจให้คะแนนทั้ง 3 วิธี ดังกล่าวผลการศึกษาพบว่า

1. เปอร์เซนต์การเดาจากการตอบตามเงื่อนไขวิธีของคูมบ์สมีค่าสูงสุด

เปอร์เซนต์การเดาจากการตอบตามเงื่อนไข วิธี Ø-1 มีค่าต่ำสุด ส่วนเปอร์เซนต์การเดาจากการตอบตามเงื่อนไขวิธีใหม่มีค่าสูงกว่าเปอร์เซนต์การเดาจากการตอบตามเงื่อนไข วิธี Ø-1 แต่ต่ำกว่าเปอร์เซนต์การเดาที่ได้จากการตอบตามเงื่อนไขวิธีของคูมบ์ส

2. ค่าความเชื่อมั่นของแบบสอบที่ได้จากการตอบและตรวจให้คะแนนตามเงื่อนไขวิธีใหม่ค่าสูงที่สุด ค่าความเชื่อมั่นของแบบสอบที่ได้จากการตอบและการตรวจให้คะแนนตามเงื่อนไขวิธี Ø-1 มีค่าต่ำสุด ส่วนค่าความเชื่อมั่นของแบบสอบที่ได้จากการตอบและการตรวจให้คะแนนตามเงื่อนไขวิธีของคูมบ์สมีค่าสูงกว่าค่าความเชื่อมั่นที่ได้จากการตอบและการตรวจให้คะแนนตามเงื่อนไขวิธี Ø-1 แต่ต่ำกว่าค่าความเชื่อมั่นที่ได้จากการตอบและตรวจให้คะแนนตามเงื่อนไขวิธีใหม่

งานวิจัยที่มีลักษณะเดียวกันกับของอรวรรณที่กล่าวมาซึ่ง กาญจนา คิริวัฒนพงษ์ (2520 : 66-72) ทำวิจัยเรื่อง การศึกษาเปรียบเทียบวิธีการตอบและการตรวจให้คะแนนแบบสอบปรนัยชนิดเลือกตอบที่มีลักษณะแตกต่างกัน โดยมีจุดมุ่งหมายเพื่อเปรียบเทียบค่าความเชื่อมั่น ค่าความตรง ค่าอำนาจจำแนก และเปอร์เซนต์การเดา ของแบบสอบปรนัยชนิดเลือกตอบ 3 ลักษณะ คือตัวเลือกถูกตัวเดียว ตัวเลือกถูกที่สุดเพียงตัวเดียว และตัวเลือกถูกหลายตัว เมื่อตอบและตรวจให้คะแนนด้วยวิธีการ 3 วิธีนี้คือวิธี Ø-1 วิธีของคูมบ์ส และวิธีของอนันต์ แบบสอบที่ใช้เป็นแบบสอบวิชาวิทยาศาสตร์ชั้นมัธยมศึกษาปีที่ 3 และนำไปสอบกับกลุ่มตัวอย่างซึ่งเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2519 ในจังหวัดปทุมธานี 4 โรงเรียน จำนวน 810 คน โดยแบ่งเป็น 9 กลุ่ม ๆ ละ 90 คน เพื่อให้ทำแบบสอบทั้ง 3 ลักษณะ ด้วยวิธีการตอบและการตรวจให้คะแนนทั้ง 3 วิธีผลการศึกษาพบว่า

1. เมื่อตอบและตรวจให้คะแนนด้วยวิธี Ø-1 และวิธีของคูมบ์ส ไม่พบว่าความเชื่อมั่น ค่าความเที่ยงตรง และค่าอำนาจจำแนกของแบบสอบทั้ง 3 ลักษณะแตกต่างกัน แต่เมื่อตอบและตรวจให้คะแนนด้วยวิธีของอนันต์ ค่าความเชื่อมั่นของแบบสอบปรนัยชนิดเลือกตอบที่มีตัวเลือกถูกตัวเดียว สูงกว่าที่มีตัวเลือกถูกที่สุดเพียงตัวเดียว และตัวเลือกถูกหลายตัว ส่วนค่าความเที่ยงตรง และค่าอำนาจจำแนก ไม่พบว่าแตกต่างกัน

2. แบบสอบปรนัยชนิดเลือกตอบที่มีตัวเลือกถูกตัวเดียว และตัวเลือกถูกที่สุดเพียงตัวเดียว เมื่อตอบและตรวจให้คะแนนด้วยวิธีของอนันต์ ให้ค่าความเชื่อมั่นและค่าความเที่ยงตรงสูงกว่าวิธี 0-1 และวิธีของคูมบ์ส ส่วนค่าอำนาจจำแนกไม่พบว่าแตกต่าง นอกจากนี้วิธี 0-1 ให้ค่าความเชื่อมั่นสูงกว่าวิธีของคูมบ์ส แต่ค่าความเที่ยงตรงและค่าอำนาจจำแนกไม่พบว่าแตกต่างกัน ส่วนแบบสอบปรนัยชนิดเลือกตอบที่มีตัวเลือกถูกหลายตัว เมื่อตอบและตรวจให้คะแนนด้วยวิธีทั้ง 3 ไม่พบว่าคุณภาพทั้ง 3 ด้านแตกต่างกัน

3. เมื่อตอบและตรวจให้คะแนนทั้ง 3 วิธีแบบสอบปรนัยชนิดเลือกตอบที่มีตัวเลือกถูกตัวเดียว และตัวเลือกถูกหลายตัว มีการเดาน้อยกว่าตัวเลือกถูกที่สุดเพียงตัวเดียว ส่วนตัวเลือกตัวเดียวมีการเดาน้อยกว่าตัวเลือกถูกหลายตัว เมื่อตอบและตรวจให้คะแนนด้วยวิธีของอนันต์ แต่ไม่พบว่าแตกต่างกันเมื่อตอบและตรวจให้คะแนนด้วยวิธี 0-1 และวิธีของคูมบ์ส

4. แบบสอบปรนัยชนิดเลือกตอบทั้ง 3 ลักษณะ ทุกลักษณะเมื่อตอบและตรวจให้คะแนนด้วยวิธีของอนันต์ มีการเดาน้อยกว่าเมื่อตอบและตรวจให้คะแนนด้วยวิธี 0-1 และวิธีของคูมบ์สสำหรับแบบสอบปรนัยชนิดเลือกตอบที่มีตัวเลือกถูกตัวเดียวและตัวเลือกถูกหลายตัวนั้น วิธี 0-1 มีการเดาน้อยกว่าวิธีของคูมบ์ส แต่ไม่พบว่าแตกต่างกันในแบบสอบปรนัยชนิดเลือกตอบที่มีตัวเลือกถูกที่สุดเพียงตัวเดียว

งานวิจัยอีกเรื่องหนึ่งที่มีวิธีการตรวจให้คะแนนคล้ายกันในการตอบแบบสอบเพื่อเพิ่มคุณภาพของการวัด ซึ่ง ธีรศักดิ์ อินทรมาตย์ (2520 : 63-69) ได้ทำการวิจัยเรื่องอิทธิพลของวิธีการตอบแบบสอบชนิดเลือกตอบวิธีต่าง ๆ ที่มีค่าความเชื่อมั่น ค่าความเที่ยงตรงและปริมาณการเดา โดยมีจุดมุ่งหมายเพื่อจะศึกษาผลของวิธีการตอบแบบสอบปรนัยชนิดเลือกตอบที่มีวิธีตอบธรรมดา วิธีตอบโดยบอกความเชื่อมั่นในการตอบ และวิธีตอบทุกตัวเลือกที่มีต่อค่าความเชื่อมั่น ค่าความเที่ยงตรง และปริมาณการเดาของแบบสอบ โดยใช้แบบสอบวิชาวิทยาศาสตร์ไปทดสอบกับกลุ่มตัวอย่างที่เป็นนักเรียนชั้นประถมปีที่ 7 ของโรงเรียนเมืองสมุทรสงคราม จำนวน 216 คน ในการดำเนินการสอบ ให้กลุ่มตัวอย่างที่หนึ่ง

ได้รับแบบสอบที่มีวิธีตอบธรรมดา กลุ่มตัวอย่างที่สองได้รับแบบสอบที่มีวิธีตอบโดยบอกความมั่นใจในการตอบ กลุ่มตัวอย่างที่สามได้รับแบบสอบที่มีวิธีตอบทุกตัวเลือก

ผลการวิจัยพบว่า วิธีตอบโดยบอกความมั่นใจในการตอบมีค่าความเชื่อมั่นสูงที่สุดกว่าวิธีตอบธรรมดา และวิธีตอบทุกตัวเลือกอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนวิธีตอบทุกตัวเลือกมีค่าความเชื่อมั่นสูงกว่าวิธีตอบธรรมดาอย่างไม่มีนัยสำคัญทางสถิติ

ในด้านความเที่ยงตรงของแบบสอบพบว่า แบบสอบที่มีวิธีตอบทุกตัวเลือกมีค่าความเที่ยงตรงสูงสุด แบบสอบที่มีวิธีตอบโดยบอกความมั่นใจในการตอบมีค่าความเที่ยงต่ำสุด แต่แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติทั้ง 3 วิธีตอบ

สำหรับปริมาณการเดานั้นแบบสอบที่มีวิธีตอบแตกต่างกันทั้ง 3 วิธี ปริมาณการเดาแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยแบบสอบที่มีวิธีตอบธรรมดามีปริมาณการเดาสูงสุด แบบสอบที่มีวิธีตอบทุกตัวเลือกมีปริมาณการเดาต่ำที่สุด และนักเรียนที่มีความสามารถในระดับต่ำจะตอบแบบสอบโดยการเดามากกว่านักเรียนที่มีความสามารถในระดับสูงทุกวิธีตอบ

งานวิจัยในประเทศที่เกี่ยวกับการตอบและการตรวจให้คะแนนด้วยวิธีทดสอบความมั่นใจอีกเรื่องหนึ่งคืองานวิจัยของเกริกชัย อวบเจริญ (2525 : 36-37) เขาได้ศึกษาเปรียบเทียบค่าความเที่ยงและค่าคะแนนการเดาของการตอบและตรวจให้คะแนนแบบบสอบชนิดเลือกตอบแบบธรรมดา แบบสอบความมั่นใจ และวิธีตอบและตรวจให้คะแนนแบบใหม่ ซึ่งวิธีตอบและตรวจให้คะแนนแบบใหม่นี้ก็คือวิธีตอบที่มีรูปแบบของกระดาษคำตอบชนิดพิเศษ เมื่อตอบแล้วนักเรียนสามารถทราบทันทีว่าตนเองทำถูกหรือผิดในข้อนั้น ๆ และถ้านักเรียนทำผิดนักเรียนสามารถตัดสินใจอย่างเสรีว่า จะทำข้อเดิมต่อไปหรือไม่ทำ หรือทำจนกระทั่งถูก เมื่อทำผิดอีกก็ได้ ผลการศึกษาปรากฏว่าแบบสอบเลือกตอบ เมื่อใช้วิธีตอบและตรวจให้คะแนนด้วยวิธีใหม่มีค่าความเที่ยงสูงสุดเท่ากับ .894 เมื่อตอบและตรวจให้คะแนนด้วยวิธีทดสอบความมั่นใจมีค่าความเที่ยงเท่ากับ .885 และเมื่อตอบและตรวจให้คะแนนด้วยวิธี

ธรรมดามีค่าความเที่ยงต่ำสุดเท่ากับ .716 เมื่อทดสอบความแตกต่างของค่าความเที่ยงพบว่าแบบสอบที่มีวิธีตอบและตรวจให้คะแนนแบบใหม่ และแบบสอบที่มีวิธีตรวจให้คะแนนแบบวิธีทดสอบความมั่นใจ มีค่าความเที่ยงแตกต่างจากแบบสอบที่มีวิธีตอบและตรวจให้คะแนนแบบธรรมดาอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนแบบสอบที่มีวิธีตอบและตรวจให้คะแนนแบบใหม่ กับแบบสอบที่มีวิธีตอบและตรวจให้คะแนนแบบวิธีทดสอบความมั่นใจ มีค่าความเที่ยงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

สำหรับค่าคะแนนการเดาในการตอบแบบสอบนั้น พบว่าเมื่อใช้วิธีตอบแบบใหม่และวิธีตอบแบบวิธีทดสอบความมั่นใจ ให้ค่าคะแนนการเดาของนักเรียนน้อยกว่าวิธีตอบแบบธรรมดาอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แต่วิธีตอบแบบใหม่แลบะวิธีตอบแบบวิธีทดสอบความมั่นใจให้ค่าคะแนนการเดาของนักเรียนแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

งานวิจัยต่างประเทศที่เกี่ยวกับการเพิ่มคุณภาพของเครื่องมือวัดหรือเพิ่มค่าความเที่ยงและความตรงของแบบสอบในการทำแบบสอบดังนี้

ผู้ที่ให้ความสนใจทำวิจัยเพื่อแก้ไขข้อบกพร่องในการสอบในระยะแรกคือเอฟเนอร์ (Hevner 1932 : 359 - 362) ได้ใช้แบบสอบชนิดถูกผิดและกำหนดระดับความมั่นใจไว้ 3 ระดับ ระบบการให้คะแนนของเขามี 4 วิธี คือ (1) นับจำนวนข้อถูก (2) จำนวนข้อถูกหักออกด้วยจำนวนข้อผิด (3) คิดคะแนนโดยวิธีถ่วงน้ำหนักถ้าตอบถูกและมั่นใจสูงให้คะแนน 3 มั่นใจปานกลางให้คะแนน 2 ไม่มั่นใจให้ 1 และ (4) คะแนนที่ถ่วงน้ำหนักของข้อถูกหักออกด้วยคะแนนที่ถ่วงน้ำหนักของข้อผิด การคิดคะแนนที่ถ่วงน้ำหนักข้อผิดมีวิธีคิดคือ ถ้าตอบผิดและมั่นใจสูงให้คะแนน -3 มั่นใจปานกลางให้คะแนน -2 ไม่มั่นใจให้ -1 ผลการวิจัยพบว่าวิธีที่ 3 คือ วิธีคิดคะแนนโดยวิธีถ่วงน้ำหนักของข้อถูกให้หาค่าความเที่ยงสูงสุด

ต่อมาโซเดอร์ควิสต์ (Soderquist 1936 : 290-292) ได้ทำการศึกษาเกี่ยวกับเรื่องนี้ เครื่องมือที่ใช้เป็นแบบสอบชนิดถูก-ผิดเหมือนกับของเอฟเนอร์ แต่ต่างกันตรงวิธีกำหนดน้ำหนักคะแนน ถ้าตอบถูกผู้สอบจะได้คะแนนสำหรับข้อนั้นเป็น 4,3,2 หรือ 1

ตามระดับความมั่นใจ ถ้าตอบผิดผู้สอบจะได้คะแนนติดลบสำหรับข้อนั้นเป็นสองเท่าของข้อถูก คือ -8, -6, -4 หรือ -2 ตามระดับความมั่นใจ แล้วศึกษาเปรียบเทียบคะแนนรวม ซึ่งวิธีคิดคะแนนรวมมี 2 วิธี คือ (1) คิดจากคะแนนที่ถ่วงน้ำหนักของข้อถูกหักออกด้วยคะแนนที่ถ่วงน้ำหนักของข้อผิด และ (2) คิดคะแนนจากจำนวนข้อที่ถูกหักออกด้วยจำนวนข้อผิดโดยไม่มีการถ่วงน้ำหนัก ผลการวิจัยพบว่าคะแนนถ่วงน้ำหนักของข้อถูกหักออกด้วยคะแนนที่ถ่วงน้ำหนักของข้อผิดให้ความเที่ยงสูงกว่า ซึ่งเป็นข้อค้นพบที่ต่างไปจากเอฟเนอร์ ทั้งนี้อาจเป็นเพราะกลุ่มตัวอย่างของโซเดอร์ควิสต์ ได้รับคำสั่งชี้แจงอธิบายวิธีให้คะแนน ขณะที่วิธีให้คะแนนเป็นความลับสำหรับกลุ่มตัวอย่างของเอฟเนอร์

หลังจากที่โซเดอร์ควิสต์ได้ทำการศึกษาเกี่ยวกับเรื่องนี้มาแล้ว ต่อมาเดรสเซลและชมิด (Dressel and Schmid 1953:574-595 quoted in Echternacht 1972:217-234) ได้ทำการวิจัยโดยใช้ข้อคำถามและระบุการให้คะแนนแบบต่าง ๆ เพื่อที่จะปรับปรุงอำนาจจำแนกของแบบสอบแบบเลือกตอบ แบบสอบแบบเลือกตอบที่เขาใช้มี 4 แบบด้วยกันคือ (1) แบบเลือกตอบเสรี (Free choice test) เป็นแบบสอบที่ผู้ตอบจะเลือกคำตอบกี่ข้อก็ได้ตามที่เขาคิดว่าถูกต้อง (2) แบบสอบแบบบอกความแน่ใจ (A degree - of certainly test) เป็นแบบสอบที่ผู้สอบเลือกคำตอบได้หนึ่งข้อและบอกด้วยว่าตนเองแน่ใจในคำตอบนั้นอยู่ในระดับใด ซึ่งความแน่ใจแบ่งออกเป็น 4 ระดับ (3) แบบมีคำตอบถูกมากกว่าหนึ่ง (A multiple - answer test) เป็นแบบสอบที่มีจำนวนคำตอบถูกในแต่ละข้อไม่จำกัด ผู้เข้าสอบต้องกาเครื่องหมายทุกข้อที่ถูก และ (4) แบบมีคำตอบถูก 2 คำตอบ (A two - answer test) จากตัวเลือก 5 ตัวเลือกในแต่ละข้อจะมีคำตอบอยู่ 2 ข้อ ในแต่ละกรณีผู้เข้าสอบจะต้องเข้าใจถึงชนิดของแบบสอบที่ตนทำ และเพราะเหตุผลว่าแบบสอบสองแบบแรกเท่านั้นที่เกี่ยวข้องกับการทดสอบความมั่นใจ จึงขอแสดงวิธีการให้คะแนนเฉพาะ 2 แบบแรก ดังนี้

วิธีการให้คะแนนแบบที่ 1 แบบเลือกตอบเสรี

จำนวนคำตอบที่ทำเครื่องหมาย	ตอบถูกได้คะแนน	ตอบผิดได้คะแนน
1	4	-1
2	3	-2
3	2	-3
4	1	-4
5	0	

วิธีการให้คะแนนแบบที่ 2 แบบบอกระดับความมั่นใจ

ระดับความมั่นใจ	ตอบถูกได้คะแนน	ตอบผิดได้คะแนน
แน่ใจ	4	-4
ค่อนข้างในใจ	3	-3
เดาอย่างมีเหตุผล	2	-2
ไม่แน่ใจ	1	-1

เดรสเซลและชมิดพบว่าการตอบข้อสอบของนักเรียนเก่ง ซึ่งพิจารณาจากคะแนนการสอบแตกต่างกับนักเรียนประเภทปานกลาง และอ่อนอย่างมีนัยสำคัญ เมื่อใช้แบบสอบชนิดเลือกตอบเสรี คือกลุ่มเก่งจะทำเครื่องหมายคำตอบบนตัวเลือกต่าง ๆ น้อยกว่ากลุ่มอ่อน ทั้งข้อสอบที่มีระดับความยาก (Difficulty level) สูง ปานกลาง และต่ำ ในทางตรงกันข้ามแบบสอบแบบบอกระดับความมั่นใจไม่ได้จำแนกนักเรียนเก่ง ปานกลาง และอ่อนได้ดีเท่าวิธีแรก เพราะเมื่อพิจารณาการตอบข้อสอบแล้ว พบว่าผู้เข้าสอบตามข้อคำถามที่มีความยากปานกลาง และยากมากด้วยระดับความมั่นใจอย่างเดียวกัน เดรสเซลและชมิด ได้สรุปว่าองค์ประกอบของความมั่นใจที่วัดโดยการใช้แบบสอบชนิดเลือกตอบเสรีนั้นแตกต่างไปจากที่วัดโดยแบบให้ผู้เข้าสอบบอกระดับความมั่นใจ

คูมบ์ส, มิลฮอลแลนด์และโวเนอร์ (Coombs, Milholland and Vomer 1956 : 13-37) ได้เปรียบเทียบคุณภาพของแบบสอบที่มีวิธีการตรวจให้คะแนนต่างกัน โดยใช้แบบสอบ 3 ฉบับ ๆ ละ 40 ข้อ คือ

1. แบบสอบเกี่ยวกับศัพท์
2. แบบสอบเกี่ยวกับความรู้ของนักขับรถ
3. แบบสอบเกี่ยวกับการใช้สายตาในด้านมิติสัมพันธ์

วิธีการตรวจให้คะแนนมี 3 วิธี คือ

1. วิธี Ø-1 (วิธีธรรมดา)
2. วิธีทดลอง (วิธีของคูมบ์ส)
3. วิธี 1 และ 2 รวมกัน

กลุ่มตัวอย่างที่ใช้คือนักเรียนชั้นปีที่ 3 และชั้นปีที่ 4 ของแจคสัน ไฮสคูล (Jackson High School) รัฐมิชิแกน จำนวน 855 คน แบ่งออกเป็น 3 กลุ่ม แต่ละกลุ่มได้ทำแบบสอบทั้ง 3 ฉบับ โดยแต่ละกลุ่มมีวิธีการตอบและตรวจที่แตกต่างกันกลุ่มละ 1 วิธี

ผลการวิจัยพบว่า

1. ค่าความเที่ยงของคะแนนที่ให้ตามวิธีการทดลองของคูมบ์สสูงกว่าวิธีธรรมดา
2. ค่าความตรงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ
3. การวิเคราะห์ข้อสอบทั้ง 2 วิธีให้คุณลักษณะของข้อสอบที่ดีพอ ๆ กัน
4. มาตรฐานความแน่ใจในการทำแบบสอบวิธีการทดลองมีมากกว่าวิธีธรรมดา คือป้องกันการเดาได้มากกว่า
5. วิธีทดลองให้ค่าอำนาจจำแนกสูงกว่าวิธีธรรมดา
6. ทัศนคติของกลุ่มตัวอย่างชอบวิธีทดลองมากกว่าวิธีธรรมดา โดยให้เหตุผลว่าวิธีทดลองยุติธรรมและง่ายกว่าวิธีธรรมดา

เดวิส และฟิฟเฟอร์ (Davis and Fifer 1959 : 159-170) ได้ทำการทดลองเกี่ยวกับค่าความเที่ยงและค่าความตรงของการตรวจให้คะแนน แบบสอบความถนัดและแบบสอบผลสัมฤทธิ์ โดยให้คะแนนทุกตัวเลือก กลุ่มตัวอย่างที่ใช้เป็นนักเรียนการบินที่ Lackland Air Force Rase จำนวน 65 คน โดยใช้แบบสอบเกี่ยวกับเหตุทาง

ต้นฉบับ หน้าขาดหาย

ต่อมาในปี ค.ศ. 1965 อีเบล (Ebel 1965 : 134-135) ได้รายงานผลการทดลองเปรียบเทียบวิธีการให้คะแนนแบบสอบแบบถูกผิดธรรมดา และแบบถูกผิดที่มีการถ่วงน้ำหนักความมั่นใจ กลุ่มตัวอย่างที่ใช้เป็นนักศึกษาที่เรียนรายวิชาการวัดผลการศึกษา มหาวิทยาลัยเซ้าเทิร์น แคลิฟอร์เนีย (The University of Southern California) ปี ค.ศ. 1960 แบ่งนักศึกษาออกเป็น 3 กลุ่ม เครื่องมือที่ใช้เป็นแบบสอบ 3 ฉบับ ให้นักศึกษาสอบกลุ่มละฉบับ แต่ละฉบับจะได้รับการตรวจให้คะแนนสองวิธีคือ แบบธรรมดา และแบบถ่วงน้ำหนักความมั่นใจ ปรากฏว่าวิธีการให้คะแนนแบบถ่วงน้ำหนักความมั่นใจ มีความเที่ยงสูงกว่าทั้ง 3 ฉบับ

อัลเบิร์ท (Albert 1970 : 1619 - A) ได้ทำการศึกษาวิธีการให้คะแนนความรู้บางส่วนในวิชาคณิตศาสตร์ โดยใช้โจทย์ปัญหาทางคณิตศาสตร์จำนวน 25 ปัญหา ศึกษากับเด็กเกรด 11 โดยศึกษาถึงการตรวจให้คะแนน 2 ลักษณะ คือ

1. แบบเลือกตอบธรรมดา มีวิธีตรวจ 3 วิธี คือ
   1.1 แบบธรรมดาคือ 0-1
   1.2 กำหนดน้ำหนักตัวเลือก โดยดูโอกาสที่ผิดพลาดมากหรือน้อย
   1.3 กำหนดน้ำหนักตัวเลือก โดยครูเป็นผู้กำหนด

2. แบบขยายคำตอบ เป็นแบบที่มีการแก้ปัญหาหลายขั้นตอนมีวิธีการตรวจ 2 วิธี คือ
   2.1 ให้คะแนนขั้นละ 1 คะแนน ในการตอบที่ถูกต้อง
   2.2 ให้คะแนนขั้นละ 1 คะแนน ถ้านักเรียนเริ่มจากข้อมูลที่ถูกต้องถึงคำตอบที่ถูกต้อง

กลุ่มตัวอย่างที่ใช้ในแต่ละลักษณะมีจำนวนประมาณ 146 คน ผลการทดลองปรากฏว่า

1. แบบขยายคำตอบสามารถพิจารณาความรู้บางส่วน

2. แบบขยายคำตอบมีค่าอำนาจจำแนกสูงกว่าแบบธรรมดา

3. แบบขยายคำตอบให้ค่าความเที่ยงสูงกว่าแบบธรรมดา

4. แบบขยายคำตอบมีระดับความยากมากกว่าแบบธรรมดา และสามารถกำจัดความคลาดเคลื่อนอันเกิดจากการเดาได้

จากการทดลองจะเห็นได้ว่า แบบขยายคำตอบสามารถพิจารณาความรู้บางส่วนของนักเรียนในการตอบ จะทำให้ค่าอำนาจจำแนกและความเที่ยงสูงขึ้นเปอร์เซนต์การเดาลดลง และยังให้ความตรงตามสภาพอีกด้วย ซึ่งผลการทดลองครั้งนี้สอดคล้องกับของคูมบ์ส มิลฮอลแลนด์ และโวเมอร์ ดังกล่าวแล้ว

นอกจากนี้ แพทเนอิค และทรอบ (Patnaic and Traub 1973 : 281-285) ได้ทำการทดลองศึกษา วิธีตอบและตรวจให้คะแนนที่คล้าย ๆ กับวิธีของเดวิส และฟิฟเฟอร์ที่กล่าวมาข้างต้น โดยกำหนดคะแนนของตัวเลือกถูกในแต่ละข้อให้มีคะแนนแตกต่างกันเปรียบเทียบกับวิธีตอบแบบธรรมดา โดยใช้แบบสอบ Dominion Group Test of Learning Capacity (DLC) จำนวน 60 ข้อ ให้เวลาทำแบบสอบ 30 นาที ทดลองกับกลุ่มตัวอย่างที่เป็นเด็กเกรด 10-13 จำนวน 1,054 คน ในการกำหนดน้ำหนักคะแนนของตัวเลือกที่ถูกในแต่ละข้อนั้น ใช้การตัดสินใจของครูและผู้เชี่ยวชาญรวม 61 คน จัดอันดับน้ำหนักคะแนนตัวเลือกที่ถูกในแต่ละข้อ แล้วหาค่าเฉลี่ยออกมาเป็นคะแนนในแต่ละข้อนั้น ผลจากการศึกษาครั้งนี้พบว่าสอดคล้องกับการศึกษาของเดวิส (Davis) กล่าวคือค่าความเที่ยงของแบบสอบที่มีวิธีตรวจให้คะแนนแบบธรรมดา และแบบให้คะแนนข้อถูกแต่ละข้อไม่เท่ากัน เมื่อคำนวณหาโดยวิธีแบ่งครึ่ง (Split - half) มีสัมประสิทธิ์ความเที่ยงเป็น 0.881 และ 0.915 ตามลำดับ ส่วนค่าความตรงของแบบสอบซึ่งหาแบบความตรงเชิงพยากรณ์ (Predictive validity) โดยคำนวณหาจากค่าสหสัมพันธ์ระหว่างคะแนนของแบบสอบกับคะแนนผลสัมฤทธิ์ทางการเรียนเฉลี่ย พบว่าวิธีตอบแบบให้ค่าน้ำหนักคะแนนแต่ละข้อไม่เท่ากัน มีความตรงน้อยกว่าวิธีตอบและตรวจให้คะแนนแบบธรรมดา

พัฟและบรันซา (Pugh and Brunza 1975 : 75-78) ได้เสนอความคิดให้มีวิธีตอบโดยบอกความมั่นใจในการตอบและนำน้ำหนักความมั่นใจในการตอบมาเป็นส่วนในการให้คะแนนด้วย พัฟและบรันซา จึงได้ทำการศึกษาหาความเที่ยงของแบบสอบที่มีวิธีตอบและตรวจให้คะแนน 2 วิธี คือวิธีตอบแบบธรรมดาและวิธีตอบโดยให้ผู้ตอบบอกความมั่นใจในการเลือกตอบข้อคำถามเหล่านั้นด้วย โดยใช้ข้อสอบชนิดเลือกตอบเป็นข้อสอบวัดสติปัญญา ชื่อ I.E.R. Intelligence Scales จำนวน 48 ข้อ แบ่งออกเป็น 2 ส่วน ๆ ละ 24 ข้อ ทดลองกับกลุ่มตัวอย่างที่เป็นนักศึกษามหาวิทยาลัยอินเดียน่า จำนวน 84 คน แบ่งออกเป็น 3 กลุ่ม ๆ ละ 28 คน โดยให้แต่ละกลุ่มได้รับแบบสอบที่แตกต่างกันดังนี้

กลุ่มแรก ได้รับแบบสอบชุดเอ ซึ่งมีวิธีตอบแบบธรรมดาทั้ง 2 ส่วน

กลุ่มที่สอง ได้รับแบบสอบชุดบี ซึ่งมีวิธีตอบแบบธรรมดาในส่วนที่หนึ่งและมีวิธีตอบโดยบอกความมั่นใจในการตอบในส่วนที่ 2

กลุ่มที่สาม ได้รับแบบสอบชุดซี ซึ่งมีวิธีตอบโดยบอกความมั่นใจในการตอบทั้ง 2 ส่วน

แบบสอบทั้ง 3 ชุดนี้มีเนื้อหาเหมือนกัน ผิดกันแต่คำชี้แจงเกี่ยวกับวิธีการตอบเท่านั้น ผลการศึกษาพบว่า

ค่าความเที่ยงของแบบสอบซึ่งหาโดยวิธีวิเคราะห์ความแปรปรวนของแบบสอบที่มีวิธีตอบ โดยบอกความมั่นใจสูงกว่าแบบสอบที่มีวิธีตอบแบบธรรมดา ดังนี้ Ø.85 และ Ø.57 ตามลำดับ

ต่อมาอาบูซาอีฟและไดมอน (Abu-sayfand Diamond 1976 : 62-65) ได้ทำการศึกษาเกี่ยวกับค่าความเที่ยงและค่าความตรงของแบบสอบเพื่อให้นักเรียนตอบแบบสอบชนิดเลือกตอบ โดยบอกความมั่นใจในการตอบจากข้อมูลที่ อาบูซาอีฟ ได้ศึกษาไว้แล้วจากกลุ่มตัวอย่าง 152 คน ใช้แบบสอบภาษาอังกฤษชนิดเลือกตอบ 4 ตัวเลือก จำนวน 60 ข้อ โดยการตรวจให้คะแนน 3 วิธีคือ ให้คะแนนเฉพาะข้อถูกและตอบอย่างมั่นใจ ให้คะแนน

ข้อที่ถูกและไม่ได้เดา และให้คะแนนทุกข้อที่ตอบถูกโดยไม่คำนึงว่ามีความมั่นใจในการตอบหรือไม่เพียงใด แล้วหาค่าความเที่ยงของแบบสอบ ผลปรากฏว่าการให้คะแนนเฉพาะข้อที่ตอบถูกและมีความมั่นใจ ได้ค่าความเที่ยงสูงสุด คือ .898 ส่วนการให้คะแนนข้อที่ตอบถูกโดยไม่ได้เดาได้ค่าความเที่ยงสูงกว่าการให้คะแนนทุกข้อที่ตอบถูก คือได้ค่าความเที่ยง .871 และ .774 ตามลำดับ ส่วนค่าความตรงนั้นวิธีการให้คะแนนข้อที่ตอบถูกและไม่ได้เดามีค่าความตรงสูงสุดคือ .260 ส่วนการให้คะแนนเฉพาะข้อที่ตอบถูกและมีความมั่นใจให้ค่าความตรงสูงกว่าการให้คะแนนทุกข้อที่ตอบถูกคือ .236 และ .219 ตามลำดับ

ในปี 1980 ลอร์ด (Lord 1980 : 73) ได้ศึกษาเปรียบเทียบผลของการให้คะแนนข้อสอบแต่ละข้อต่างกัน ใช้แบบสอบทางภาษาชื่อ SCAT II Form 2 A มี 4 ตัวเลือก 40 ข้อ ผลการศึกษาพบว่าการให้คะแนนข้อสอบแต่ละข้อโดยวิธีให้ค่าน้ำหนักคะแนนที่เหมาะสม (w_i) ให้สาระของการสอบมากที่สุด แต่สำหรับกลุ่มผู้สอบที่มีความสามารถในระดับต่ำการให้คะแนนวิธีประเพณีนิยมให้สาระน้อยที่สุด

จากการวิจัยทั้งหมดที่กล่าวมานี้เป็นการพยายามของนักวัดผลและนักวิจัยที่จะหาวิธีการลดค่าความคลาดเคลื่อนของการวัดอันเนื่องมาจาก การทำแบบสอบโดยไม่ตั้งใจ หรือตอบโดยไม่ได้ใช้ความรู้ความสามารถอย่างเพียงพอ อันเป็นพฤติกรรมที่ผู้วิจัยคิดว่าน่าจะเกิดจากการรับรู้ผลการสอบว่าจะนำผลการสอบไปทำอะไรที่จะก่อให้เกิดความคลาดเคลื่อนของการวัดตามมา ซึ่งเป็นสิ่งที่ผู้วิจัยสนใจที่จะใช้วิธีการทางสถิติเข้าไปแก้ไข จึงได้ทำการวิจัยเรื่องนี้ขึ้น