บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

เอกสารและงานวิจัยที่เกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์ ผลสัมฤทธิ์ทางการเรียน และค่าความเที่ยงของแบบทดสอบ ผู้วิจัยได้เสนอเป็น 5 ตอน ดังนี้

ตอนที่ 1 การทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์

ตอนที่ 3 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์

ตอนที่ 4 ผลสัมฤทธิ์ทางการเรียน

ตอนที่ 5 ค่าความเที่ยงของแบบทดสอบ

ตอนที่ 1 การทำหน้าที่ต่างกันของข้อสอบ

ความเป็นมาของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาความลำเอียงหรือการทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบนั้น เริ่มมาจากการสังเกตปรากฏการณ์ของผลการสอบคัดเลือก เช่น การสอบคัดเลือกเพื่อเข้าศึกษาต่อ เพื่อบรรจุเข้าทำงาน หรือการสอบเพื่อเลื่อนตำแหน่ง หรือในการสอบอื่น ๆ ซึ่งพบว่าไม่เป็นไปตามสัดส่วนระดับสติปัญญาของผู้เข้าสอบหรือโครงสร้างของประชากร จึงเป็นเหตุให้เกิดความลำเอียงหรือความไม่ยุติธรรมในการสอบและส่งผลให้เกิดการศึกษาความลำเอียงของแบบทดสอบ (Test Bias) และข้อสอบ (Item Bias) ในระยะต่อ ๆ มาแนวคิดในการวิเคราะห์ความลำเอียงและการตรวจสอบความยุติธรรมของการใช้ข้อสอบ ได้เริ่มปรากฏและได้รับความสนใจตั้งแต่ปี ค.ศ. 1910 โดยอัลเฟรด บิเนต์และไซมอน ได้ทดสอบเด็กที่มีฐานะทางเศรษฐกิจและสังคมที่แตกต่างกัน พบว่า มีข้อสอบบางข้อที่น่าจะวัดในสิ่งที่เป็นการกระทบกับการฝึกฝนทางด้านวัฒนธรรม (เช่นที่บ้านหรือที่โรงเรียน) มากกว่าที่จะวัดความสามารถทางสมอง หลังจากศึกษาครั้งนี้ บิเนต์และไซมอน ได้มีการพิจารณาอย่างละเอียดและได้เอาข้อสอบเหล่านี้ออกจากแบบทดสอบ เช่น ถ้าเป็นเด็กอายุ 7 ปี ข้อคำถามที่เกี่ยวกับนิ้วมือ การคัดลอกประโยคและการเรียกชื่อเหรียญ จะถูกคัดออกจากแบบทดสอบทั่วไป โดยในการทดสอบจะพิจารณาเฉพาะข้อคำถามที่ไม่ขึ้นกับการฝึกฝนของทางบ้าน (Home Training) ความตั้งใจ (Attention) ภาษา (Language) นิสัยในการพูด (Habit of

มหาวิทยาลัยบูรพา Burapha University

Looking at Picture) และแบบฝึกหัดเชิงวิชาการ (Scholastic Exercise) ในปี ค.ศ. 1912 วิลเลียม สเตอร์น (William Stern) ผู้ที่ได้ริเริ่มในนิยาม I.Q. ในรูปของอัตราส่วน ระหว่างอายุสมองกับอายุปฏิทิน ได้ศึกษาความแตกต่างระหว่างชนชั้นต่าง ๆ (Class Difference) ในประเทศเยอรมัน ผลที่ได้มีความคล้ายคลึงกับที่บีเนต์พบ คือ ความแตกต่างระหว่างชนชั้นมีนัยสำคัญ เสตอร์นพยายามที่จะค้นหาและทำความเข้าใจเกี่ยวกับแหล่งที่ทำให้เกิดความแตกต่างอย่างจริงจัง เพื่อที่จะเสนอแนะว่า แบบทดสอบต่าง ๆ ที่ได้พิสูจน์มาแล้วนั้นเข้าข้างชนชั้นหนึ่งมากกว่าชนชั้นอื่น ๆ อย่างชัดเจน

ต่อมาในปี ค.ศ. 1951 ได้มีการเริ่มต้นศึกษาเกี่ยวกับความลำเอียงของข้อสอบแนวใหม่ โดยเอลส์ (Eells) และคณะสนใจความแตกต่างของการวัดบางอย่าง ซึ่งไม่มีความแม่นยำพอที่จะสะท้อนให้เห็นความสามารถได้ เช่น ความแตกต่างทางด้านสติปัญญาของนักเรียน อาจขึ้นอยู่กับเนื้อหาเฉพาะของข้อสอบแต่ละข้อในแบบทดสอบ และไม่ได้สะท้อนความสามารถที่แท้จริงของนักเรียน จุดมุ่งหมายหลักเพื่อที่จะอธิบายความแตกต่างระหว่างกลุ่มนักเรียนโดยใช้สติปัญญา โดยมุ่งไปที่ข้อคำถามแต่ละข้อเช่นเดียวกับบิเนต์และสเตอร์น คือดูว่าข้อคำถามใดมีความลำเอียงหรือไม่ แล้ววิเคราะห์ด้วยวิธีหาค่าความยากของข้อสอบระหว่างกลุ่มที่แตกต่างกัน ผลการศึกษาพบว่ายังมีความแตกต่างระหว่างกลุ่มนักเรียนที่มีระดับชั้นทางสังคม และภูมิหลังของเชื้อชาติและวัฒนธรรมต่างกันในการตอบแบบทดสอบ เอลส์และคณะได้พยายามพัฒนาวิธีการปฏิบัติและแก้ไขอิทธิพลของระดับชั้นทางสังคมและภูมิหลังของเชื้อชาติ ที่มีผลต่อการตอบข้อสอบ เพื่อใช้ขจัดข้อสอบที่มีความลำเอียงในแบบทดสอบต่าง ๆ โดยเปรียบเทียบกลุ่มของเชื้อชาติที่อยู่ในชั้นทางสังคม

ผลจากการศึกษาของเอลส์ได้นำไปสู่การศึกษาอย่างกว้างขวางในช่วง 30 กว่าปีที่ผ่านมา เมื่อรัฐมีความมุ่งหวังที่จะทำให้เกิดความยุติธรรมในการสอบเข้าศึกษาและโอกาสในการมีงานทำ ดังนั้นในช่วงปลายปี ค.ศ. 1960 และ ค.ศ. 1970 ได้มีความตื่นตัวเกี่ยวกับความลำเอียงของข้อสอบ โดยมุ่งไปที่แบบทดสอบวัดสติปัญญา เพราะเป็นแบบทดสอบที่ถูกนำไปใช้ในทางด้านการศึกษา การคัดเลือก (Selection) และจัดบุคคลเข้าสู่ตำแหน่ง (Placement) อย่างแพร่หลายจนกระทั่งปี ค.ศ.1971 บริษัทกริกกัส วี ดุ๊ค เพาเวอร์ จำกัด (Griggs V. Duke Power) ได้วินิจฉัยแบบสอบต่าง ๆ แล้วชี้ให้เห็นว่าผลการทดสอบเหล่านั้นใช้ในการคัดเลือกอาจไม่เหมาะสม อาจพิจารณาตามเชื้อชาติของผู้สมัครงาน ยกเว้นกรณีที่ผลของการปฏิบัติงานมีความสอดคล้องกับการทดสอบ

ในช่วงเวลาเดียวกันมีการโต้แย้งในเรื่องความลำเอียงของข้อสอบมากขึ้น เมื่อเจนเซน (Jensen, 1969) ได้เสนอบทความลงในสารทางการศึกษาของฮาร์วาร์ด ฉบับที่ 39 ปี ค.ศ. 1969 โดยกล่าวว่าไอคิวเป็นผลมาจากการถ่ายทอดทางพันธุกรรมมากกว่าสิ่งแวดล้อม ความแตกต่างของไอคิว

มหาวิทยาลัยบูรพา Burapha University

ระหว่างคนผิวขาวกับคนผิวดำนั้นไม่ได้มาจากสิ่งแวดล้อมเพียงอย่างเดียว และพันธุกรรมอธิบายความแปรปรวนผลการปฏิบัติงานของแต่ละคนได้ประมาณร้อยละ 80 จากบทความของเจนเซนนี้เองทำให้สมาคมนักจิตวิทยาคนผิวดำเสนอว่า การทดสอบคนผิวดำทุกคนควรทำให้เท่าเทียมกันมากที่สุดเท่าที่จะเป็นไปได้ ยังมีบทความและวิธีการทางสถิติในการวิเคราะห์ความลำเอียงของข้อสอบเกิดขึ้นมากมาย ซึ่งแสดงให้เห็นว่าปัญหาเรื่องความลำเอียงของข้อสอบมีความสำคัญ เพื่อจะสร้างแบบทดสอบที่มีความยุติธรรมทางวัฒนธรรม และให้ผลการวิเคราะห์มีความถูกต้องมากที่สุด (Camilli & Shepard, 1994, pp. 4-7) นอกจากนี้ยังมีเหตุการณ์ที่ทำให้มีการตื่นตัวเป็นอันมากในเรื่องความลำเอียงของข้อสอบ คือ กรณีที่มาร์โค เดอฟูนิส (Marco Defunis) และคณะซึ่งถูกปฏิเสธจากโรงเรียนกฎหมายของมหาวิทยาลัยแห่งวอชิงตัน ได้ฟ้องร้องว่าเขาได้คะแนนการสอบสูงกว่าผู้ได้รับการคัดเลือกบางคนที่มหาวิทยารับเข้าศึกษา และได้ยื่นฎีกาฟ้อง ชาร์ล โอเดการ์ด (Charles Odcgard) อธิการบดีมหาวิทยาลัยดังกล่าวและคณะ เพื่อให้พิจารณาทบทวนการคัดเลือกนักศึกษาใหม่ หลังจากนั้นการพิจารณาตรวจสอบความลำเอียงของข้อสอบระหว่างผู้สอบกลุ่มย่อยที่มีลักษณะแตกต่างกันในเรื่องเพศ เชื้อชาติ ศาสนาหรือวัฒนธรรม และได้ปฏิบัติกันมาจนปัจจุบัน ซึ่งเป็นเสมือนส่วนหนึ่งของขั้นตอนการพัฒนาแบบทดสอบโดยทั่ว ๆ ไป โดยเฉพาะการพัฒนาแบบทดสอบมาตรฐาน เช่น ศูนย์บริการการทดสอบทางการศึกษา (Education Testing Service) ของประเทศสหรัฐอเมริกา ได้มีการตรวจสอบความลำเอียงทั้งก่อนนำไปทดลองใช้ จะใช้การตรวจสอบโดยผู้เชี่ยวชาญ ซึ่งประกอบด้วยบุคคลหลายฝ่ายตรวจสอบความไว (Sensitivity) ของแบบทดสอบโดยพิจารณาถึงรูปแบบของข้อสอบ เนื้อหา ภาพประกอบ คำที่ใช้ และอื่น ๆ เพื่อไม่ให้มีความลำเอียงหรือเกิดการได้เปรียบเสียเปรียบระหว่างผู้สอบกลุ่มย่อยที่มีลักษณะแตกต่างกัน ส่วนการตรวจสอบความลำเอียงของข้อสอบหลังจากการทดลองใช้นั้นจะใช้วิธีการทางสถิติ

"ความตรง" เป็นหัวใจสำคัญของคุณภาพแบบสอบ ในการสร้างและการตรวจสอบคุณภาพของแบบสอบจะต้องคำนึงถึงคุณภาพด้านความตรงเป็นสำคัญ ทั้งนี้เพราะว่าความตรงเป็นคุณสมบัติของแบบสอบที่แสดงถึงความสามารถในการวัดได้ถูกต้องแม่นยำ ถ้าผลการวัดได้ค่าที่ใกล้เคียงกันของข้อสอบและแบบสอบ ก็เป็นอีกลักษณะหนึ่งของการตรวจสอบคุณภาพด้านความตรง โดยเป็นการตรวจสอบในประเด็นของความยุติธรรมของข้อสอบและแบบสอบ (Item and Test Unfairness)

ความอยุติธรรมของข้อสอบ หรือแบบสอบเกิดขึ้นในกรณีที่ผู้สอบกลุ่มย่อยต่างกลุ่มกันและมีลัษณะเฉพาะบางอย่างแตกต่างกัน มีความได้เปรียบหรือเสียเปรียบกันทั้งที่มีความสามารถจริงเท่าเทียมกัน การศึกษาคุณภาพด้านความยุติธรรมของข้อสอบหรือแบบสอบระหว่างผู้สอบกลุ่ม

มหาวิทยาลัยบูรพา Burapha University

ต่าง ๆ เริ่มศึกษากันอย่างจริงจังในช่วงปลายทศวรรษของปี ค.ศ. 1960 มีการเสนอวิธีการต่าง ๆ เพื่อตรวจสอบความลำเอียงของข้อสอบ (Item Bias) ความลำเอียงของแบบสอบ (Test Bias) และความลำเอียงในการคัดเลือก (Selection Bias) โดยนิยามความลำเอียงว่าเป็น ความคลาดเคลื่อนอย่างเป็นระบบ (Systematic Error) ที่เกิดขึ้นจากการวัด ความพยายามในการตรวจสอบความลำเอียงดังกล่าวดำเนินไปเพื่อจำแนกข้อสอบที่ทำหน้าที่ไม่เหมาะสมหรือไม่ยุติธรรมสำหรับปรับปรุงหรือตัดข้อสอบข้อนั้นออกจากแบบสอบถาม เป็นการขจัดข้อสอบที่ทำให้เกิดปัญหาความยุติธรรมระหว่างกลุ่มข้อสอบกลุ่มต่าง ๆ ที่มีลักษณะบางอย่างแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ ประสบการณ์ เป็นต้น เพื่อพัฒนาแบบสอบให้มีคุณภาพเหมาะสมสำหรับนำไปใช้ทดสอบต่อไป

ในเวลาต่อมานักวัดผลการศึกษาได้ทำการศึกษาความลำเอียงของข้อสอบกันอย่างกว้างขวาง ทำให้เกิดความสับสนของการใช้คำและความหมาย มีประเด็นโต้แย้งกันว่าความลำเอียงของข้อสอบ เป็นผลการตัดสินว่าข้อสอบมีความยุติธรรมหรือไม่ อันส่งผลต่อการบรรลุจุดมุ่งหมายของการใช้แบบสอบ หรือความลำเอียงของข้อสอบเป็นสารสนเทศทางสถิติที่ได้จากข้อสอบ เกี่ยวกับความสัมพันธ์ระหว่างคุณลักษณะที่ข้อสอบมุ่งวัดกับประสบการณ์ของผู้สอบกลุ่มต่าง ๆ ที่ทำการสอบ เมื่อกลุ่มผู้สอบต่างกลุ่มกันตอบข้อสอบข้อเดียวกัน ความแตกต่างที่เกิดขึ้นอาจมาจากความไม่เหมาะสมของข้อคำถาม ซึ่งสามารถเกิดขึ้นได้หลายลักษณะ หรือประสบการณ์ของผู้สอบซึ่งอาจมีลักษณะพื้นฐานเดิมแตกต่างกัน ในหลายสถานการณ์จึงไม่เหมาะสมที่จะใช้คำว่า ข้อสอบลำเอียง (Biased Item) เนื่องจากเป็นภาษาที่มีความหมายในเชิงลบ ประกอบกับเกณฑ์ที่ใช้สำหรับตัดสินความลำเอียงยังมีความคลุมเครือและค่อนข้างสับสน ดังนั้นจึงควรเปลี่ยนมาใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบ ซึ่งเป็นคำที่มีความเป็นกลางและเหมาะสมกว่า

การทำหน้าที่ต่างกันของข้อสอบกับความลำเอียงของข้อสอบมีแนวคิดที่แตกต่างกัน สำหรับการทำหน้าที่ต่างกันของข้อสอบ เป็นกระบวนการที่เน้นการใช้วิธีการทางสถิติสำหรับตรวจสอบ เพื่อให้ได้สารสนเทศเกี่ยวกับการทำหน้าที่ของข้อสอบสำหรับกลุ่มผู้สอบกลุ่มย่อยที่มีลักษณะเฉพาะบางอย่างแตกต่างกัน ส่วนความลำเอียงของข้อสอบเป็นกระบวนการตัดสินความยุติธรรมของข้อสอบ โดยนำสารสนเทศการทำหน้าที่ต่างกันของข้อสอบมาวิเคราะห์เชิงตรรกะ (Logical Analysis) โดยผู้เชี่ยวชาญพิจารณาถึงการเขียนข้อสอบ เนื้อหาสาระของข้อสอบและจุดมุ่งหมายของการวัด เพื่อระบุว่าข้อสอบข้อนั้นลำเอียงเข้าข้างกลุ่มใดหรือไม่ เพราะเหตุใดจึงเป็นการตัดสินความลำเอียงของข้อสอบ (Camilli & Shapard, 1994)

มหาวิทยาลัยบูรพา Burapha University

ความหมายของความลำเอียงและการทำหน้าที่ต่างกันของข้อสอบ

นักวิจัยทางการวัดผลหลายท่านได้ให้ความหมายของความลำเอียงของข้อสอบและการทำหน้าที่ต่างกันของข้อสอบไว้ดังนี้

ความลำเอียงของข้อสอบ หมายถึง สัดส่วนของผู้สอบที่ตอบข้อสอบได้ถูกต้องไม่เท่ากันในแต่ละกลุ่มประชากรที่ใช้ในการศึกษา เมื่อกลุ่มผู้สอบมีคะแนนเท่ากันและข้อสอบมีความเป็นเอกพันธ์ (Scheuneman, 1979)

ความลำเอียงของข้อสอบ หมายถึง ข้อสอบที่มีค่าความยากสัมพัทธ์สำหรับสมาชิกของผู้สอบกลุ่มหนึ่งมากกว่าสมาชิกของผู้สอบอีกกลุ่มหนึ่ง (Rudner, Getson, & Knight, 1980)

ข้อสอบที่มีความลำเอียงด้านวัฒนธรรมว่า หมายถึง ข้อสอบที่ทำให้ผู้สอบซึ่งคุ้นเคยกับวิชาเฉพาะหรือกระบวนการเฉพาะมีโอกาสตอบถูกมากกว่าผู้สอบคนอื่น ๆ ที่มีลักษณะของกลุ่มที่แตกต่างกัน (Eells, 1951 citing Jensen, 1980)

ความลำเอียงของข้อสอบ หมายถึง ความโน้มเอียงของข้อสอบที่เมื่อใช้คะแนนจากข้อสอบนั้นแล้วทำให้การตัดสินใจผลเป็นไปอย่างไม่ยุติธรรม (Popham, 1981)

ความลำเอียงของข้อสอบ หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันสำหรับการวัดความสามารถ หรือโอกาสในการตอบข้อสอบในทางบวกแตกต่างกันสำหรับการวัดเจตคติ เมื่อผู้สอบที่มีคุณลักษณะของการวัดในปริมาณเท่ากัน แต่มาจากกลุ่มประชากรย่อยที่แตกต่างกัน (Hulin, Drasgow, & Parson, 1983)

ความลำเอียงของข้อสอบ หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบกลุ่มหนึ่ง มีค่าต่ำกว่าหรือสูงกว่าผู้สอบอีกกลุ่มหนึ่งที่มีระดับความสามารถเตียวกัน (Dorans & Kulick, 1986)

ความลำเอียงของข้อสอบ หมายถึง คะแนนข้อสอบของกลุ่มผู้สอบที่มีความสามารถเท่ากันแต่มาจากต่างกลุ่มกัน มีความแตกต่างกันอย่างเป็นระบบ (Kederman, 1990)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง สารสนเทศทางสถิติของข้อสอบที่ได้จากผลการดอบของผู้สอบต่างกลุ่มกันและมีความสามารถเท่ากัน แต่มีโอกาสในการดอบข้อสอบได้ถูกต้องแตกต่างกัน (Holland & Wainer, 1993)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความเป็นพหุมิติในการวัดของข้อสอบ ซึ่งแสดงได้จากการแจกแจงความสามารถหลัก (Primary Ability) ของผู้สอบตั้งแต่ 2 กลุ่มขึ้นไปมีความเท่ากัน แต่มีการแจกแจงความสามารถรอง (Secondary Abiliry) แตกต่างกัน (Camilli & Shepard, 1994)

มหาวิทยาลัยบูรพา Burapha University

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ฟังก์ชันการตอบสนองข้อสอบซึ่งคำนวณจากกลุ่มผู้สอบย่อยที่ต่างกัน มีค่าไม่เท่ากัน (Naryanan & Swaminathan, 1996)

ข้อสอบทำหน้าที่ต่างกัน เป็นข้อสอบที่วัดความสามารถหรือคุณลักษณะทางจิตวิทยาของแต่ละกลุ่มไม่ตรงกัน (Allen & Yen, 1979 อ้างถึงใน อารี วัชรโสดถิกุล, 2543, หน้า 8)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความแตกต่างในการทำหน้าที่ของข้อสอบหลังจากกลุ่มของผู้สอบได้ถูกจับคู่ตามความสามารถหรือคุณลักษณะที่ข้อสอบนั้นวัด (Millsap & Everson, 1993 อ้างถึงใน อารี วัชรโสดถิกุล, 2543, หน้า 8)

ดังนั้นจึงสรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การที่ข้อสอบทำให้ผู้สอบจากต่างกลุ่มกันที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน หรือมีฟังก์ชันการตอบสนองข้อสอบแตกต่างกัน การทำหน้าที่ต่างกันของข้อสอบเกิดขึ้นเมื่อนำข้อสอบไปทดสอบกับผู้สอบกลุ่มย่อยต่างกัน ที่มีความสามารถหลัก (Primary Abiliry) ระดับเดียวกันหรือมีคุณลักษณะแฝง (Latent Trait) ที่ต้องการวัดเท่ากัน แต่มีความสามารถรอง (Secondary Ability) แตกต่างกัน ทำให้ผู้สอบต่างกลุ่มที่นำมาจับคู่เปรียบเทียบมีโอกาสตอบข้อสอบถูกแตกต่างกัน

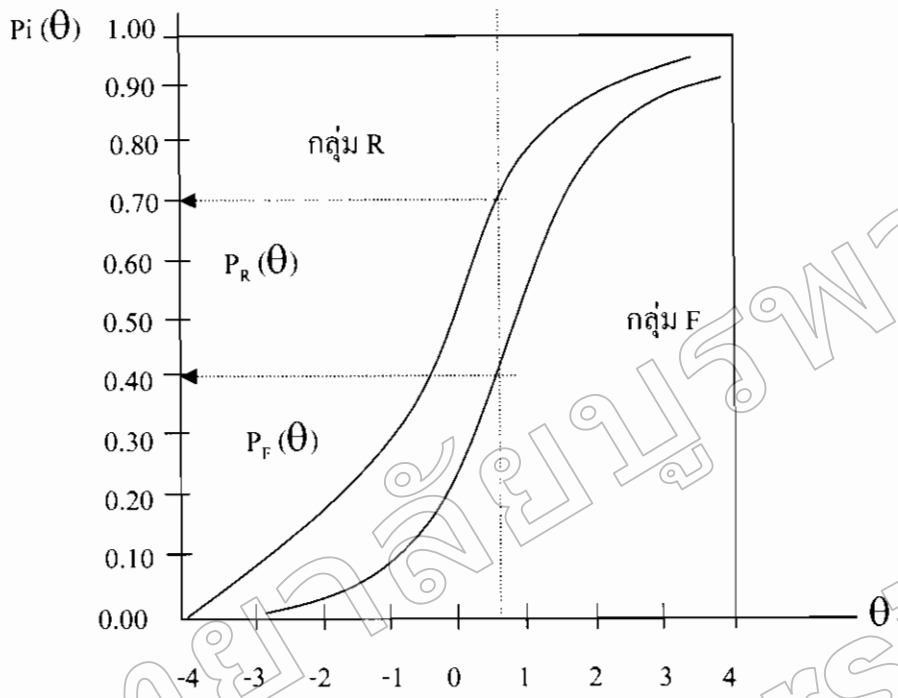
ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป ปกตินิยมทำการเปรียบเทียบ 2 กลุ่ม ประกอบด้วยกลุ่มแรก เรียกว่า กลุ่มเปรียบเทียบ (Focal Group หรือกลุ่ม F) เป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ และกลุ่มที่สอง เรียกว่า กลุ่มอ้างอิง (Reference Group หรือกลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้อง

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ พบว่า ข้อสอบสามารถทำหน้าที่แตกต่างกันได้ 2 ประเภท (Mellenbergh, 1982) ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และแบบอเนกรูป (Nonuniform)

1. ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หมายถึง ข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่ง มีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งสม่ำเสมอกัน ในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม จะพบว่าไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกกลุ่ม (Group Membership) ดังภาพที่ 1

มหาวิทยาลัยบูรพา Burapha University



ภาพที่ 1 ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF)

2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มแตกต่างกันอย่างไม่สม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกกลุ่ม เช่น ที่ระดับความสามารถอีกระดับหนึ่ง กลุ่มผู้สอบกลุ่ม R มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม F แต่ที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้สอบกลุ่ม F มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม R

ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) สามารถพิจารณา "ปฏิสัมพันธ์" ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูปแล้ว โค้งลักษณะข้อสอบ (Item Characteristic Curves: ICCs) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบ (Item Response Theory: IRT) เหมือนกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแล้ว โค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะไม่ขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบต่างกัน ดังนั้นความแตกต่างระหว่างโค้งลักษณะข้อสอบทั้งสอง

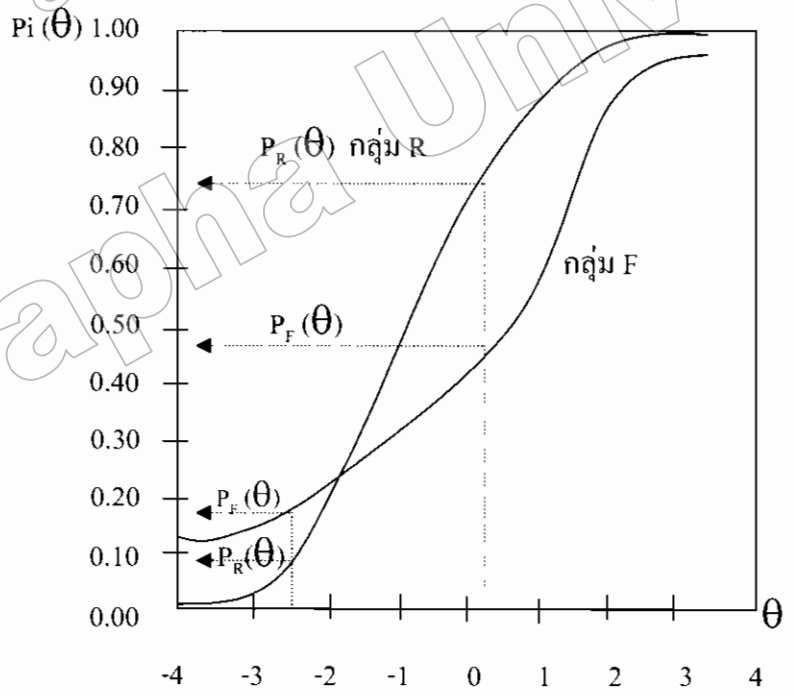
มหาวิทยาลัยบูรพา Burapha University

แบบจะบ่งบอกถึงขนาดและทิศทางของข้อสอบทำหน้าที่ต่างกัน ซึ่งสามารถคำนวณได้โดยใช้สูตรการคำนวณพื้นที่ของ Raju (1990)

ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปสามารถจำแนกได้เป็น 2 ลักษณะ (Swaminathan & Rogers, 1990) ดังนี้

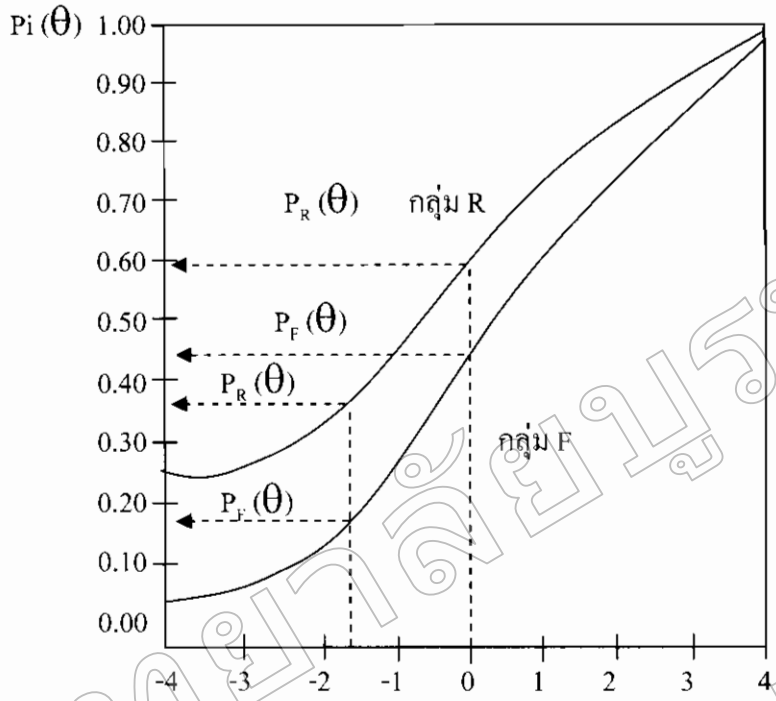
1. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal Interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบตัดกันระหว่างช่วงความสามารถของผู้สอบหรือเรียกว่าข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non - Directional DIF) ดังภาพที่ 2

2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์เป็นลำดับ (Ordinal Interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบต่างกันอย่างไม่สม่ำเสมอแต่ไม่ตัดกัน หรืออาจตัดกันนอกช่วงความสามารถของผู้สอบตรงปลายสุดของช่วงความสามารถต่ำหรือสูง อาจเรียกข้อสอบลักษณะนี้ว่า ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirectional DIF) ดังภาพที่ 3



ภาพที่ 2 ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non - Directional DIF)

มหาวิทยาลัยบูรพา Burapha University



ภาพที่ 3 ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียวกัน (Unidirectional DIF)

หลักการและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

1. หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่ม ที่มีความสามารถหลัก (Primary Ability) หรือความสามารถเป้าหมายที่ต้องการวัดเท่ากัน แต่คาดว่าจะมีความได้เปรียบหรือเสียเปรียบกัน โดยกลุ่มหนึ่งถือเป็นกลุ่มอ้างอิงซึ่งตอบข้อสอบได้ถูกต้องมากกว่า ส่วนอีกกลุ่มคือกลุ่มเปรียบเทียบ ซึ่งเป็นกลุ่มที่สนใจศึกษาและคาดว่าน่าจะเป็นกลุ่มที่เสียเปรียบ

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำเป็นต้องจับคู่ (Matching) ผู้สอบตามความสามารถ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมี 2 วิธี ดังนี้

1.1 เกณฑ์ภายนอก (External Criterion)

การวิเคราะห์การทำหน้าที่ต่างกันโดยใช้เกณฑ์ภายนอกนี้ สามารถนำไปใช้ได้ทั้งข้อสอบรายข้อและแบบสอบทั้งฉบับ โดยการใช้คะแนนจากแบบสอบสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิคการวิเคราะห์การถดถอย (Regression Analysis) เพื่อทำการเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรทำนาย ระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

มหาวิทยาลัยบูรพา Burapha University

หลักการนี้มีจุดมุ่งหมายเพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบสอบอื่นจากตัวแปรทำนายซึ่งเป็นคะแนนรายข้อ หรือคะแนนสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะใช้คะแนนรายข้อเป็นตัวแปรทำนาย แต่ถ้าการวิเคราะห์การทำหน้าที่ต่างกันของแบบสอบ จะใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอก อาจใช้คะแนนรวมทั้งฉบับหรือเกรดเฉลี่ย หรือผลสัมฤทธิ์ในงานที่เกี่ยวข้องของผู้สอบ (Cronbach & Furby, 1970) สมการทำนายสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแสดงได้ดังนี้

กลุ่มอ้างอิง (R) $R_i = A_R + B_R X_i$

กลุ่มเปรียบเทียบ (F) $Y_i = A_F + B_F X_i$

เมื่อ Y_i = คะแนนของตัวแปรเกณฑ์ภายนอก

X_i = คะแนนของตัวแปรทำนาย

A = ค่าคงที่หรือค่าตัดแกน (Intercept)

B = ค่าความชัน (Slope)

ภาพที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและแบบสอบโดยใช้วิธีการวิเคราะห์สมการถดถอย

จากฟังก์ชันการทำนายทั้ง 2 สมการ สามารถเปรียบเทียบค่าตัดแกน (A) และค่าความชัน (B) ของเส้นกราฟระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบได้ ถ้าเส้นกราฟดังกล่าวมีค่าความชันหรือค่าตัดแกนแตกต่างกันสำหรับข้อสอบใดหรือแบบสอบใด แสดงว่าข้อสอบหรือแบบสอบนั้นมีการทำหน้าที่ต่างกัน โดยเข้าข้างกลุ่มผู้สอบที่มีค่าตัดแกนหรือค่าความชันที่สูงกว่า

มหาวิทยาลัยบูรพา Burapha University

การใช้เกณฑ์ภายนอกมีข้อดี คือ เกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบและแบบสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อนตรงที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ ในทางปฏิบัติเป็นการยากที่จะหาตัวแปรเกณฑ์ภายนอกจากแบบสอบฉบับอื่นที่มีความตรงเชิงทำนาย ซึ่งจะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบ ขาดความแม่นยำและความสมบูรณ์

1.2 เกณฑ์ภายใน (Internal Criterion)

การวิเคราะห์การทำหน้าที่ต่างกันโดยใช้เกณฑ์ภายใน เป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบ โดยเน้นการพิจารณาจากโครงสร้างภายในของแบบสอบเป็นหลัก ด้วยการวิเคราะห์ผลจากการตอบข้อสอบ และความสามารถหรือคะแนนจริงของผู้สอบที่ได้จากแบบสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ การวิเคราะห์ในลักษณะนี้นิยมใช้ค่าสถิติต่าง ๆ เป็นตัวบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ ค่าสถิติทดสอบที่นิยมนำมาใช้พอสรุปได้ดังนี้

1.2.1 การทดสอบปฏิสัมพันธ์ (Interaction)

ในระยะเริ่มแรกของการศึกษาความลำเอียงของข้อสอบ มีการใช้สถิติทดสอบเอฟ (F- Test) ในการวิเคราะห์ความแปรปรวน เพื่อทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับข้อสอบ ถ้าการทดสอบมีนัยสำคัญ เป็นสัญญาณของการทำหน้าที่ต่างกันของข้อสอบ (Cleary & Hilton, 1968, Jensen, 1974) จากนั้นจึงทำการวิเคราะห์ต่อด้วยวิธีการ Post Hoc เพื่อระบุข้อสอบที่มีผลต่อการเกิดปฏิสัมพันธ์ซึ่งเป็นข้อที่ทำหน้าที่ต่างกัน

วิธีการนี้มีข้อดีที่สามารถศึกษาผู้สอบหลาย ๆ กลุ่มได้สะดวก แต่มีจุดอ่อนในเรื่องการควบคุมกลุ่มต่าง ๆ ให้มีความสามารถที่ทัดเทียมกัน ขนาดกลุ่มตัวอย่างของกลุ่มต่าง ๆ และอัตราความคลาดเคลื่อนประเภทที่ 1 จะสูงขึ้น ถ้าจำนวนข้อสอบเพิ่มมากขึ้น

1.2.2 การวัดความเบี่ยงเบนสัมพันธ์ (Relative Deviation)

การคำนวณค่าความยากของข้อสอบ เช่น p, b เป็นต้น เมื่อคำนวณแยกระหว่างกลุ่ม และแปลงให้เป็นค่าความยากมาตรฐาน (Δ) สามารถนำมาพล็อตเปรียบเทียบเป็นรายข้อ ถ้าข้อใดเบี่ยงเบนไปจากแกนหลักที่คาดหมาย หรือเบี่ยงเบนเกินความคลาดเคลื่อนมาตรฐานของค่าความยากที่กำหนด ย่อมแสดงถึงการกระทำหน้าที่ต่างกันของข้อสอบ (Cleary & Hilton, 1968; Angoff & Ford, 1973) รวมทั้งสามารถคำนวณค่าสหสัมพันธ์ระหว่างค่าความยากรายข้อระหว่างกลุ่ม เพื่อแสดงถึงการทำหน้าที่ต่างกันของข้อสอบ ถ้าสหสัมพันธ์เข้าใกล้ 1.00 แสดงว่าค่าความ

มหาวิทยาลัยบูรพา Burapha University

ยากสัมพัทธ์ของข้อสอบมีค่าใกล้เคียงกันระหว่างกลุ่ม ดังนั้นแบบสอบวัดคุณลักษณะคล้ายกันระหว่างกลุ่ม

วิธีการนี้มีข้อดีและข้อเสียคล้ายการทดสอบปฏิสัมพันธ์ นอกจากนี้ค่าความยากของข้อสอบ มิใช่ตัวแทนของค่าความยากจริงของข้อสอบ และได้รับอิทธิพลจากค่าแทรกซ้อนอื่น ได้แก่ ค่าอำนาจจำแนกและความสามารถของผู้สอบ

1.2.3 การเปรียบเทียบน้ำหนักตัวประกอบ (Factor Loading)

การวิเคราะห์ตัวประกอบ (Factor Analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ในการตรวจสอบความตรงเชิงทฤษฎีหรือโครงสร้าง (Construct Validity) เมื่อนำการวิเคราะห์ตัวประกอบมาใช้ในการวิเคราะห์โครงสร้างของแบบสอบแยกตามกลุ่มผู้สอบ ความไม่สอดคล้องกันระหว่างน้ำหนักตัวประกอบบนคุณลักษณะสำคัญที่มุ่งวัด หรือความแตกต่างของค่าเฉลี่ยคะแนนตัวประกอบ (Factor Scores) ระหว่างกลุ่มผู้สอบ ย่อมสะท้อนการทำหน้าที่ต่างกันของข้อสอบและแบบสอบ

การใช้เทคนิคการวิเคราะห์ตัวประกอบเชิงสำรวจ (Exporatory Factor Analysis: EFA) สำหรับศึกษาการทำหน้าที่ต่างกันของข้อสอบ จะมีจุดอ่อนในเรื่องความไม่สอดคล้องระหว่างน้ำหนักตัวประกอบ อาจเกิดจากความต้องการของความสามารถระหว่างกลุ่มก็ได้ แนวทางที่เหมาะสมจึงควรใช้เทคนิคการวิเคราะห์ตัวประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) นอกจากนี้ยังสามารถใช้ CFA สำหรับตรวจสอบความแตกต่างระหว่างกลุ่ม ในด้านคุณลักษณะหรือความสามารถหลักและความสามารถรองได้อีกด้วย (Camilli & Shepard, 1994)

1.2.4 การเปรียบเทียบโอกาสตอบข้อสอบถูก

การวิเคราะห์โอกาสตอบข้อสอบถูกของผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน เป็นแนวทางสำคัญที่นิยมใช้กันและเป็นที่ยอมรับในปัจจุบัน สำหรับบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ มีการคำนวณค่าสถิติ 2 แนวทางดังนี้

เปรียบเทียบค่าสัดส่วนหรือความน่าจะเป็นในการตอบข้อสอบถูกของผู้สอบต่างกลุ่มที่มีความสามารถเท่ากัน เช่น วิธีแมนเทล - แฮนส์เซล เป็นต้น

เปรียบเทียบค่าฟังก์ชันการตองสนองข้อสอบ หรือโค้งลักษณะข้อสอบระหว่างกลุ่มที่มีระดับความสามารถเท่ากัน เป็นวิธีที่อยู่บนพื้นฐานของทฤษฎี IRT เช่น วิธีวัดความแตกต่างของพื้นที่ วิธีวัดความแตกต่างของค่าพารามิเตอร์ความยาก วิธีการทดสอบไค-สแควร์ของลอร์ด เป็นต้น

วิธีการนี้มีข้อดีที่สำคัญ ได้แก่ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือ มีกลไกลควบคุมความสามารถของผู้สอบโดยการจับคู่กลุ่มความสามารถ เพื่อทำการเปรียบเทียบ

มหาวิทยาลัยบูรพา Burapha University

ณ ตำแหน่งต่าง ๆ ที่มีความสามารถเท่ากัน จึงเป็นวิธีการที่ยอมรับกันทั่วไปแต่มีข้อจำกัดในด้านความสลับซับซ้อนของแนวคิดพื้นฐาน และการวิเคราะห์มีความจำเป็นต้องใช้โปรแกรมคอมพิวเตอร์โดยเฉพาะ

2. วิธีการตรวจสอบกันทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำแนกตามลักษณะการตรวจให้คะแนนได้เป็น 2 ประเภท คือ ข้อสอบที่มีการตรวจให้คะแนนแบบทวิภาคหรือ 2 ค่า และข้อสอบที่มีการให้คะแนนแบบพหุวิภาคหรือหลายค่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแต่ละประเภทยังสามารถจำแนกได้อีก 2 มิติ ได้แก่ มิติลักษณะของตัวแปรเกณฑ์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้คะแนนสังเกตได้ และกลุ่มวิธีที่ใช้คะแนนสังเกตไม่ได้หรือคะแนนของตัวแปรแฝง และมิติลักษณะของสถิติวิเคราะห์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้สถิติพาราเมตริก และกลุ่มวิธีที่ใช้สถิตินันพาราเมตริก

2.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิภาค

2.1.1 กลุ่มวิธีที่ใช้คะแนนสังเกตได้

วิธีในกลุ่มนี้มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม หรือกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่ของกลุ่มข้อสอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่

2.1.1.1 วิธีวิเคราะห์ความแปรปรวน (Cleary & Hilton, 1968)

2.1.1.2 วิธีวิเคราะห์การถดถอยโลจิสติก (Swaminathan & Rogers, 1990)

2.1.1.3 วิธีการแปลงค่าความยากของข้อสอบ (Clear & Hilton, 1968; Angoff & Ford, 1973)

2.1.1.4 วิธีแมนเทล – แฮนส์เซล (Holland & Thayer, 1989)

2.1.1.5 วิธีดัชนีมาตรฐาน การปรับให้เป็นมาตรฐานด้วยน้ำหนักตัวประกอบ (Dorans & Kulick, 1986)

2.1.2 กลุ่มที่ใช้คุณลักษณะแฝง

วิธีในกลุ่มนี้ใช้คุณลักษณะหรือตัวแปรแฝง ซึ่งวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ สำหรับใช้เป็นเกณฑ์การจัดกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่

2.1.2.1 วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (Linn et al., 1981; Shepard et al.,1984; Raju, 1990; Kim & Cohen, 1991)

2.1.2.2 วิธีไค – สแควร์ของลอร์ด (Lord, 1980)

2.1.2.3 วิธีอัตราส่วนไลค์ลิฮูดทั่วไป (Thissen, Steinberg, & Wainer, 1993)

2.1.2.4 วิธีอัตราส่วนไลค์ลิฮูด ลอกลิเนียร์ (Loglinear IRT Likelihood Ratio) (Thissen, Steinberg, & wainer, 1993)

2.1.2.6 วิธีซิปเทสท์ (Shealy & Stout, 1993)

2.2 วิธีการตรวจสอบการทำหน้าที่แตกต่างกันของข้อสอบที่ให้คะแนนแบบพหุวิภาค

2.2.1 กลุ่มวิธีที่ใช้คะแนนสังเกตได้

2.2.1.1 วิธีการวิเคราะห์ความแปรปรวน (Clear & Hilton, 1968)

2.2.1.2 วิธีการวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Swaminathan & Rogers, 1990)

2.2.1.3 วิธีดัชนีมาตรฐานหุวิภาค (Dorans & Kulick, 1968)

2.2.1.4 วิธีแมนเทล – แฮนส์เซลทั่วไป (Holland & Thayer, 1988, 1989)

2.2.2 กลุ่มที่ใช้คะแนนลักษณะแฝง

2.2.2.1 วิธีอัตราส่วนไลค์ลิฮูดในรูปทั่วไป (Thissen, Steinberg, & Wainer, 1993)

2.2.2.2 วิธีการให้คะแนนบางส่วน (Master, 1982)

2.2.2.3 วิธีซิปเทสท์พหุวิภาค (Shealy & Stout, 1993)

2.2.2.4 วิธีการใช้คะแนนบางส่วนทั่วไป (Muraki, 1992, 1993)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิธีนี้ จะพิจารณาเปรียบเทียบฟังก์ชันการตอบข้อสอบ (Item Responsc Function) หรือเส้นคุณลักษณะข้อสอบ (Item Characteristic Curves) ระหว่างกลุ่มตั้งแต่ 2 กลุ่มขึ้นไปว่ามีความแตกต่างกันหรือไม่

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดทฤษฎีการตอบข้อสอบมีจุดเด่นกว่าวิธีตามแนวคิดทฤษฎีการทดสอบแบบดั้งเดิม ดังนี้

1. ทฤษฎีการตอบข้อสอบสามารถประมาณค่าพารามิเตอร์ของข้อสอบ เช่น ค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าการเดา (c) ทำให้ลดความปนกันของพารามิเตอร์ข้อสอบลงได้

2. คุณสมบัติทางสถิติของข้อสอบ (Statistical Properties of Item) สามารถอธิบายได้ละเอียดและชัดเจนกว่า เมื่อข้อสอบทำหน้าที่ต่างกันระหว่าง 2 กลุ่ม

3. คุณสมบัติทางสถิติของข้อสอบ สามารถแสดงด้วยแผนภาพที่เป็นกราฟ ทำให้เข้าใจง่ายและนำไปใช้ได้อย่างกว้างขวาง

มหาวิทยาลัยบูรพา Burapha University

การวิเคราะห์การทำหน้าที่ต่างกันของแบบสอบและข้อสอบ

1. แนวคิดพื้นฐานของการวิเคราะห์การทำหน้าที่ต่างกันของแบบสอบ

ถ้าให้ $P_i(\theta_s)$ เป็นความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้องของผู้สอบคนที่ S ซึ่งมีความสามารถ θ ในกรณีที่เรามีผู้สอบร่วมกัน 2 กลุ่ม ได้แก่ กลุ่มอ้างอิง (R) และกลุ่มเปรียบเทียบ (F) แต่ละกลุ่มทำแบบสอบฉบับเดียวกันซึ่งมีจำนวน k ข้อ เมื่อทำการแยกวิเคราะห์ตามกลุ่ม แต่ละกลุ่มต่างก็มีชุดของค่าพารามิเตอร์ของข้อสอบ (a , b และ c) ซึ่งสามารถคำนวณหาค่าความน่าจะเป็นในการตอบถูกได้ ดังนี้

$P_{iR}(\theta_s)$ = ความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้อง ของผู้สอบคนที่ S ซึ่งเป็นสมาชิกของกลุ่มอ้างอิงและมีความสามารถ θ

$P_{iF}(\theta_s)$ = ความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้อง ของผู้สอบคนที่ S ซึ่งเป็นสมาชิกของกลุ่มเปรียบเทียบและมีความสามารถ θ

ตามทฤษฎี IRT เราสามารถคำนวณค่าคะแนนจริง (True Scores) ของผู้สอบคนที่ S ซึ่งมีความสามารถ θ ได้ดังนี้

$$T_S = \sum_{i=1}^{K} P_i(\theta_S)$$

ผู้สอบแต่ละคนไม่ว่าจะเป็นสมาชิกของกลุ่มใด มี T_S ได้ 2 ค่า ในฐานะสมาชิกของกลุ่ม R และ F จึงคำนวณค่าได้ทั้ง T_{SR} และ T_{SF}

ถ้า $T_{SR} = T_{SF}$ แสดงว่า ผู้สอบ S เป็นอิสระจากกลุ่มสมาชิก ดังนั้นแบบสอบจึงทำหน้าที่ไม่แตกต่างกัน

$T_{SR} \neq T_{SF}$ แสดงว่า การเป็นสมาชิกกลุ่มของผู้สอบคนที่ S มีผลทำให้ T_S แตกต่างกันจึงเป็นสัญญาณของการทำหน้าที่ต่างกันของแบบสอบ

2. การวิเคราะห์การทำหน้าที่ต่างกันของแบบสอบและการทำหน้าที่ต่างกันของข้อสอบ

ราจูและคณะ (Raju et al., 1995) ได้เสนอกระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและแบบสอบ (DIF and DTF Proccdures) บนพื้นฐานของทฤษฎี IRT มีการคำนวณที่ซับซ้อนแต่ให้ผลที่น่าเชื่อถือ โดยนำวิธีทางสถิติมาใช้ในการตรวจสอบโครงสร้างแบบสอบถามด้วยเกณฑ์ภายใน กระบวนการตรวจสอบประกอบด้วยการคำนวณดัชนีการทำหน้าที่ต่างกัน 2ระดับ

มหาวิทยาลัยบูรพา Burapha University

ได้แก่ ดัชนีการทำหน้าที่ต่างกันระดับแบบสอบ (DTF Index) และดัชนีการทำหน้าที่ต่างกันระดับข้อสอบ (DIF Index)

2.1 ดัชนีการทำหน้าที่ต่างกันระดับข้อสอบ (DIF Index)

ถ้าให้ $d_{is} = P_{iR}(\theta_s) - P_{iF}(\theta_s)$

และ $D_S = T_{SR} - T_{SF}$

ดังนั้นดัชนีการทำหน้าที่ต่างกันของข้อสอบแบบชดเชย (Compensatory DIF Index: CDIF) เป็นดังนี้

$$\begin{aligned} \text{CDIF}_i &= E(d_i, D) \\ &= Cov(d_i, D) + \mu_{di}\mu_D \end{aligned}$$

เมื่อ E = ค่าความคาดหมาย (Expectation)

$Cov(d_i, D)$ = ความแปรปรวนระหว่างความแตกต่างของความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูก กับความแตกต่างของคะแนนจริงระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีค่า θ เท่ากัน

μ_{di} = ค่าเฉลี่ยของความแตกต่างของความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้อง ระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีค่า θ เท่ากัน

μ_D = ค่าเฉลี่ยความแตกต่างของคะแนนจริงระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มี θ เท่ากัน

2.2 ดัชนีการทำหน้าที่ต่างกันระดับแบบสอบ (DTF Index)

$$\begin{aligned} \text{DTF} &= E(T_{SR} - T_{SF})^2 \\ &= E(D_S)^2 \\ &= E(D_S)^2 - \mu_D^2 + \mu_D^2 \\ &= \sigma_D^2 + (\mu_{TR} - \mu_{TF})^2 \\ &= \sigma_D^2 + \mu_D^2 \end{aligned}$$

เมื่อ $\text{D} = T_{SF} - T_{SR} = \sum_{i=1}^{K} P_{iR}(\theta_S) - \sum_{i=1}^{K} P_{iF}(\theta_S)$

$$= \sum_{i=1}^{K} P_{iR}(\theta_S) - \sum_{i=1}^{K} P_{iF}(\theta_S)$$

มหาวิทยาลัยบูรพา Burapha University

$$= \sum_{i=1}^{K} d_i$$

$$\therefore \quad \text{DTF} = E_{I\cdot} D^2$$

$$= E_{I\cdot} \left(\sum_{i=1}^{K} d_i \right)^2$$

$$= \sum_{i=1}^{K} \left[\text{cov}(d_i, D) + \mu_{di} \mu_{D} \right]$$

สาเหตุของการทำหน้าที่ต่างกันของข้อสอบ

สาเหตุที่น่าจะก่อให้เกิดการทำหน้าที่ต่างกันของข้อมีหลายสาเหตุด้วยกัน ซึ่งสามารถสรุปสาเหตุที่ก่อให้เกิดการทำหน้าที่ต่างกันของข้อสอบมากที่สุด ได้ดังนี้

1. การเดา (Guessing) อาจจะเกิดจากข้อสอบนั้นยากเกินไปหรือเวลาไม่เพียงพอ จะก่อให้เกิดความไม่เท่าเทียมกันในโอกาสการตอบข้อสอบถูกของผู้สอบแต่ละคน

2. ความเร็ว (Speed) หรือเวลาในการตอบ จะทำให้เกิดการเดาหรือในกรณีเวลาน้อยอาจจะทำข้อสอบไม่ทัน ซึ่งจะมีผลกับข้อสอบข้อหลัง ๆ โดยเฉพาะในการศึกษาความลำเอียงของข้อสอบวัดความถนัด

3. ความกำกวมหรือความไม่ชัดเจน (Unclear) ของข้อคำถาม นั่นคือ ข้อคำถามขาดความเป็นปรนัย การใช้ภาษาท้องถิ่นหรือใช้คำที่ไม่เป็นภาษากลางในการสื่อสารความหมาย ซึ่งจะก่อให้เกิดความลำเอียงกับกลุ่มภาษาใดภาษาหนึ่งขึ้นได้

4. ลำดับขั้นของคำถาม (Scries) อาจจะเป็นสิ่งที่ก่อให้เกิดความสับสน หรือชี้แนะคำตอบของข้อสอบบางข้อได้

5. สถานภาพทางสังคมหรือความเกี่ยวข้องทางสังคม (Social Implication) ก็เป็นสิ่งที่ก่อให้เกิดความแตกต่างระหว่างกลุ่มผู้สอบแต่ละกลุ่มได้

6. ประสบการณ์หรือการฝึกฝนของแต่ละกลุ่มย่อย เป็นสิ่งที่ก่อให้เกิดการได้เปรียบเสียเปรียบของแต่ละกลุ่มค่อนข้างชัดเจน

7. องค์ประกอบทางวัฒนธรรม ความเป็นอยู่ ขนบธรรมเนียมประเพณี เชื้อชาติ ศาสนา ก็จะเอื้อให้กับบางกลุ่มย่อย จึงทำให้เกิดการได้เปรียบเสียเปรียบในบางเนื้อหาวิชาได้

นอกจากนี้การทำหน้าที่ต่างกันของข้อสอบอาจจะเกิดจากสาเหตุหรือแหล่งที่สำคัญ 2 แหล่งคือ

1. การเลือกเนื้อหา (Bias in Selection) คือผู้สร้างข้อสอบเลือกเนื้อหาเฉพาะส่วนใดส่วนหนึ่งมาสร้างข้อสอบ ทำให้ได้ข้อสอบที่มีเนื้อหาไม่ครอบคลุมและไม่ได้สัดส่วนที่สมดุลกัน

มหาวิทยาลัยบูรพา Burapha University

2. การสร้างข้อสอบ (Bias in Construction) คือการใช้ภาษาหรือข้อความบางอย่างที่เอื้อให้เกิดประโยชน์กับผู้สอบกลุ่มใดกลุ่มหนึ่ง

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์

วิธีซิปเทสท์ พัฒนาโดยเชียลีและสเตาท์ (Shealy & Stout, 1993) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ การทำหน้าที่ต่างกันของแบบสอบ และการทำหน้าที่ต่างกันของกลุ่มข้อสอบ (Differential Bundle Functioning: DBF) วิธีนี้สามารถวิเคราะห์ได้ทั้งในแบบสอบเอกมิติและแบบสอบพหุมิติ (Stout, Li, & Nandakumar, 1997) วิธีซิปเทสท์ใช้สถิติ ทดสอบแบบนันพารา เมตริก ซึ่งพัฒนาบนพื้นฐานของทฤษฎี IRT ชนิดพหุมิติ แต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบหรือการประมาณค่าความสามารถแฝง วิธีซิปเทสท์ถูกออกแบบมาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว ดังนั้นจึงไม่มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่มีทิศทาง (Li & Stout, 1996) จุดเด่นของวิธีซิปเทสท์ คือ สามารถคำนวณได้ง่ายไม่ซับซ้อน ประหยัดค่าใช้จ่าย และไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ ทั้งยังใช้สถิติทดสอบนัยสำคัญ (Narayanan & Swaminathan, 1996) นอกจากนี้ยังสามารถนำไปประยุกต์ใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่มีการให้คะแนนแบบพหุวิภาค (Polytomous DIF) (Chang, Mazzeo, & Roussos, 1995)

ในการศึกษาการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ของแบบสอบเอกมิติถือว่าข้อสอบในแบบสอบจะต้องมุ่งวัดคุณลักษณะ หรือความสามารถแฝงเพียงลักษณะเดียว ความสามารถแฝงสามารถจำแนกเป็นความสามารถเป้าหมายที่ต้องการวัด (θ) กับความสามารถแทรกซ้อนที่ไม่ใช่เป้าหมายของการวัด (η) ตัวอย่าง เช่น แบบสอบคำศัพท์ในวิชาภาษาอังกฤษ ข้อสอบบางข้ออาจถามความรู้สำหรับผู้ชายเป็นพิเศษ เช่น ความรู้เกี่ยวกับกีฬา ในขณะที่ข้อสอบบางข้ออาจถามความรู้เกี่ยวกับผู้หญิงโดยเฉพาะ เช่น ความรู้เกี่ยวกับงานในบ้าน จากสถานการณ์ดังกล่าวทักษะความรู้เกี่ยวกับคำศัพท์ในวิชาภาษาอังกฤษ เป็นความสามารถเป้าหมายที่ต้องการวัด (θ) ส่วนความรู้ทางด้านกีฬาและงานในบ้านเป็นความสามารถแทรกซ้อน ที่ไม่ใช่เป้าหมายของการวัด (η_1 และ η_2) ข้อสอบทุกข้อในแบบสอบจะวัดความสามารถเป้าหมาย ส่วนข้อสอบบางข้อทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมาย และความสามารถแทรกซ้อน (Nandakumar, 1993)

ถ้าให้ฟังก์ชันการตอบสนองข้อสอบ (IRF) ข้อที่ i ซึ่งขึ้นอยู่กับความสามารถ θ เพียงอย่างเดียวแทนด้วย $P_i(\theta)$ ส่วน IRF ข้อที่ i ที่ขึ้นอยู่กับความสามารถทั้ง θ และ η แทนด้วย

$P_i(\theta,\eta)$ ฟังก์ชันการตอบสนองข้อสอบของข้อสอบดังกล่าวแบบ 3 พารามิเตอร์จะเป็นดังนี้ (Shealy & Stout, 1993)

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp[-1.7a_{i\theta}(\theta - b_{i\theta})]} \quad , i = 1,\ldots,N$$

$$P_i(\theta,\eta) = c_i + \frac{(1-c_i)}{1+\exp\{-1.7[a_{i\theta}(\theta - b_{i\theta}) + a_{i\eta}(\eta - b_{i\eta})]\}} \quad , i = 1,\ldots,N$$

ดังนั้นฟังก์ชันความน่าจะเป็นอย่างมีเงื่อนไขของแบบแผนการตอบข้อสอบทั้งฉบับเป็นดังนี้

$$P[U/(\Theta=\theta,\eta=\eta)] = \prod_{i=1}^{N} P_i(\theta,\eta)^{u_i}[1-P_i(\theta,\eta)]^{1-u_i}$$

เชียลีและสเตาท์ (Shealy & Stout, 1993) ได้ใช้ Marginal IRFs อธิบายการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

$$M_{ig}(\theta) = \int_n P_i(\theta,\eta) f_g(\eta/\theta) d\eta$$

เมื่อ $M_{gi}(\theta)$ = Marginal IRF สำหรับความสามารถเป้าหมาย θ ที่ต้องการวัดของผู้สอบกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบ

$P_i(\theta,\eta)$ = IRF ของข้อสอบข้อที่ i

$f_g(\eta \mid \theta)$ = การแจกแจงแบบมีเงื่อนไขของกลุ่มผู้สอบ

การเปรียบเทียบ Marginal IRF ระหว่างกลุ่มอ้างอิง (R) กับกลุ่มเปรียบเทียบ (F) จะทำให้ทราบถึงทิศทางของการได้เปรียบหรือเสียเปรียบ กล่าวคือ ถ้า $M_{iF}(\theta) < M_{iR}(\theta)$ ทุกค่าของ θ แสดงว่า ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มอ้างอิงและถ้า $M_{iF}(\theta) > M_{iR}(\theta)$ ทุกค่าของ θ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบ การทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียวอาจเรียกอีกอย่างหนึ่งว่า "การทำหน้าที่ต่างกันแบบไม่ตัดกัน" (Noncrossing DIF)

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามวิธีซิปเทสท์ จะแบ่งข้อสอบออกเป็น 2 ชุดย่อย (Subtests) คือ (1) ชุดแบบสอบที่มีความตรง (Valid Subtests) หรือชุดแบบสอบที่ใช้ในการจับคู่เปรียบเทียบ (Matching Subtests) แบบสอบชุดนี้ประกอบด้วยข้อสอบที่ไม่ทำหน้าที่ต่างกัน และ (2) ชุดแบบสอบที่ต้องการศึกษา (Studied Subtests)

มหาวิทยาลัยบูรพา Burapha University

ประกอบด้วยข้อสอบที่สงสัยว่าทำหน้าที่ต่างกัน ถ้าแบบสอบชุดแรกมีจำนวน n ข้อ (ข้อที่ 1 ถึง n) แล้วแบบสอบชุดที่ 2 จะมีจำนวน N-n ข้อ (ข้อที่ n+1 ถึง N) เมื่อ N เป็นจำนวนข้อสอบทั้งหมด

ฟังก์ชันการตอบสนองข้อสอบของแบบสอบที่ต้องการศึกษา จากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ กำหนดในรูปฟังก์ชัน Marginal ดังนี้

$$M_{SR}(\theta) = \sum_{i=n+1}^{N} M_{iR}(\theta)$$

$$M_{SF}(\theta) = \sum_{i=n+1}^{N} M_{iF}(\theta)$$

เมื่อ $M_{SR}(\theta)$ = ผลรวมของ Marginal IRFs ของข้อสอบที่ต้องการศึกษาจากผู้สอบกลุ่มอ้างอิง ณ ระดับความสามารถ θ

$M_{SF}(\theta)$ = ผลรวมของ Marginal IRFs ของข้อสอบที่ต้องการศึกษาจากผู้สอบกลุ่มเปรียบเทียบ ณ ระดับความสามารถ θ

ขนาดของความแตกต่างระหว่าง $M_{SR}(\theta)$ และ $M_{SF}(\theta)$ แสดงถึงปริมาณความเข้มของการทดสอบการทำหน้าที่ต่างกันของข้อทดสอบแบบมีทิศทางเดียว หรือการทำหน้าที่แบบไม่ตัดกันจากชุดแบบสอบถามที่ต้องการศึกษา ณ ระดับความสามารถ θ ซึ่งสามารถคำนวณคลอดช่วงความสามารถ θ ของผู้สอบ ด้วยการอินทิเกรท ดังนี้

$$\beta_{uni} = \int_{\theta} \left[M_{SR}(\theta) - M_{SF}(\theta) \right] f_p(\theta) d\theta$$

เมื่อ β_{uni} = ดัชนีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว

$f_p(\theta)$ = ฟังก์ชันความหนาแน่นของโอกาสการแจกแจงความสามารถ θ ของผู้สอบทั้ง 2 กลุ่ม

ดัชนี β_{uni} ที่คำนวณได้จากสูตรดังกล่าว นำมาทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว ดังนี้

$$H_0 : \beta_{uni} = 0$$

$$H_1 : \beta_{uni} > 0$$

มหาวิทยาลัยบูรพา Burapha University

สมมติฐานอื่น (H_1) มีลักษณะทิศทางเดียว ซึ่งใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบที่เข้าข้างผู้สอบกลุ่มอ้างอิง สำหรับค่าประมาณของ β_{uni} คำนวณได้จากคะแนนรวมของชุดแบบสอบที่มีความตรงและชุดแบบสอบที่ต้องการศึกษา ซึ่งกำหนดด้วยสัญลักษณ์ ดังนี้

$$X = \sum_{i=1}^{n} U_i$$

$$Y = \sum_{i=n+1}^{N} U_i$$

เมื่อ X = คะแนนรวมของชุดแบบสอบที่มีความตรง

Y = คะแนนรวมของชุดแบบสอบที่ต้องการศึกษา

U_i = ผลการตอบข้อสอบข้อที่ i (ตอบถูกได้ 1 คะแนน และตอบผิดได้ 0 คะแนน)

นำคะแนนรวมของชุดแบบสอบที่มีความตรง (X) เป็นเกณฑ์ในการจับคู่ผู้ที่มีความสามารถระดับเดียวกัน แล้วคำนวณคะแนนเฉลี่ยรายข้อจากผลการตอบข้อสอบชุดแบบสอบที่ต้องการศึกษา ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกันมาจับคู่เปรียบเทียบกัน ณ $X = k$ โดยเขียนในรูปสัญลักษณ์ได้ ดังนี้

$$\bar{Y}_{Rk} - \bar{Y}_{I\cdot k} : \quad k = 0, 1, 2, ..., n$$

เมื่อ $\bar{Y}_{Rk}$ = ค่าเฉลี่ยของคะแนนรายข้อจากชุดแบบสอบที่ต้องการศึกษาของผู้สอบกลุ่มอ้างอิง ซึ่งได้คะแนน $X = k$

$\bar{Y}_{I\cdot k}$ = ค่าเฉลี่ยของคะแนนรายข้อจากชุดแบบสอบที่ต้องการศึกษาของผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้คะแนน $X = k$

k = คะแนนรวมจากชุดแบบสอบถามที่มีความตรง

ค่า $\bar{Y}_{Rk} - \bar{Y}_{I\cdot k}$ ดังกล่าวเป็นความแตกต่างของผลการตอบข้อสอบในชุดแบบสอบที่ต้องการศึกษาระหว่างผู้สอบกลุ่มอ้างอิง และกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน ถ้า $\bar{Y}_{Rk} - \bar{Y}_{I\cdot k} = 0$ ทุกคะแนน k แสดงว่า ข้อสอบที่ต้องการศึกษาทำหน้าที่ไม่ต่างกัน แต่ถ้า $\bar{Y}_{Rk} - \bar{Y}_{I\cdot k} > 0$ ทุกคะแนน k แสดงว่า ข้อสอบทำหน้าต่างกันแบบมีทิศทางเดียว โดยจะเข้าข้างผู้สอบกลุ่มอ้างอิง ค่าความแตกต่างของผลการตอบข้อสอบสามารถประมาณค่าในรูป β_{uni} ได้ดังนี้

$$\hat{\beta}_{uni} = \sum_{k=0}^{n} \hat{P}_k \left(\bar{Y}_{Rk} - \bar{Y}_{I\cdot k} \right)$$

มหาวิทยาลัยบูรพา Burapha University

โดยที่ $\hat{P}_k = \dfrac{(J_{Rk} + J_{Fk})}{\sum_{k=0}^{n}(J_{Rk} + J_{Fk})}$

เมื่อ P_k = สัดส่วนของจำนวนผู้สอบกลุ่มเปรียบเทียบและกลุ่มอ้างอิงที่ได้คะแนนรวม $X = k$ จากจำนวนผู้สอบทั้งหมด

J_{Fk} = จำนวนผู้สอบกลุ่มเปรียบเทียบที่ได้คะแนนรวม $X = k$

J_{Rk} = แทนจำนวนผู้สอบกลุ่มอ้างอิงที่ได้คะแนนรวม $X = k$

สำหรับการทดสอบสมมติฐานศูนย์ของ no DIF ใช้สถิติ β_{uni} ดังนี้

$$\beta_{uni} = \frac{\hat{\beta}_{uni}}{\sigma(\beta_{uni})}$$

โดยที่ $\hat{\sigma}(\hat{\beta}_{uni}) = \sqrt{\sum_{k=0}^{n} \hat{P}_k^2 \left[\frac{1}{J_{Rk}} \hat{\sigma}^2(Y|k,R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y|k,F) \right]}$

เมื่อ $\sigma(\beta_{uni})$ = ค่าประมาณความคลาดเคลื่อนมาตรฐานของ β_{uni}

$\sigma^2(Y \mid k, g)$ = ค่าประมาณความแปรปรวนของคะแนนจากชุดแบบสอบที่ต้องการศึกษา สำหรับผู้สอบกลุ่ม g (R หรือ F) ซึ่งมีคะแนนรวมเท่ากับ k

J_{gk} = จำนวนผู้สอบกลุ่ม g (R หรือ F) ซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

β_{uni} มีการแจกแจงในลักษณะปกติมาตรฐาน $[N(0,1)]$ จึงสามารถทดสอบนัยสำคัญด้วยสถิติทดสอบ Z ถ้าผลการทดสอบปรากฏว่า $\beta_{uni} > Z_{\alpha}$ แสดงว่าการทดสอบมีนัยสำคัญจึงปฏิเสธ H_0 นั่นคือ ข้อสอบที่นำมาตรวจสอบทำหน้าที่ต่างกัน โดยจะเข้าข้างผู้สอบกลุ่มอ้างอิง เมื่อ β_{uni} มีค่าเป็นบวก และจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบเมื่อ β_{uni} มีค่าเป็นลบ

สถิติที่ใช้ในการทดสอบสำหรับสรุปอ้างอิงการทำหน้าที่ต่างกันของข้อสอบดังกล่าวมักจะมีปัญหาในกรณีที่มีความแตกต่างของการแจกแจงความสามารถเป้าหมาย ระหว่างกลุ่มผู้สอบกล่าวคือ ถ้าผู้สอบกลุ่มอ้างอิงมีความสามารถเป้าหมายสูงกว่าผู้สอบกลุ่มเปรียบเทียบ จะเกิดผลกระทบทำให้สถิติ β_{uni} มีค่าเฟ้อ (Inflate) หรือมีค่าสูงผิดปกติ ถึงแม้ว่าในความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน จึงควรแก้ไขความแตกต่างของการแจกแจงความสามารถเป้าหมาย ด้วยการ

มหาวิทยาลัยบูรพา Burapha University

ปรับแก้ค่าความถดถอย เพื่อขจัดผลกระทบดังกล่าวโดยปรับแก้ค่า $\overline{Y}_{Rk}$ และ $\overline{Y}_{I\cdot k}$ เป็นรายคู่ ก่อนการคำนวณ β_{uni}

ตอนที่ 3 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยวิธีซิปเทสท์

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีวิวัฒนาการมาเป็นเวลานาน โดยพยายามแก้ไขข้อพกพร่องของวิธีการที่มีมาก่อน ประกอบกับความก้าวหน้าของเทคโนโลยีคอมพิวเตอร์ ทำให้มีการพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหลายวิธี และได้มีผลการศึกษาค้นคว้าวิจัยโดยใช้ข้อมูลจำลองจำนวนมาก เพื่อหลีกเลี่ยงข้อจำกัดบางประการของข้อมูลเชิงประจักษ์ งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบมีดังนี้

กาญจนา วัธนสุนทร (2538) ได้พัฒนาเกณฑ์ในการตัดสินข้อสอบลำเอียงทางเพศ โดยใช้ข้อมูลเชิงประจักษ์ จากการสอบคัดเลือกเข้าระดับอุดมศึกษาปีการศึกษา 2535 ใช้วิธีการตรวจสอบ 3 วิธี คือ วิธีตามแนวคิดทฤษฎีการตอบข้อสอบ วิธีแมนแทลและแฮนส์เซล และวิธีซิปเทสท์ ดัชนีที่พัฒนาเพื่อเป็นเกณฑ์ในการตัดสินข้อสอบลำเอียงคือ SA, UA, α_{MH}, β_{SIB} ตามลำดับซึ่งในการวิจัยครั้งนี้ได้จากการวิเคราะห์ค่าเฉลี่ยของดัชนีแต่ละตัว ตัวแปรที่ใช้ในการศึกษา ได้แก่ความยาวของแบบทดสอบขนาด 20, 30, 40 ข้อ สำหรับวิชาคณิตศาสตร์ และ 50, 60, 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ ขนาดกลุ่มตัวอย่างผู้ตอบข้อสอบ 100, 200, 400, 600, 800 และ 1,000 คน ผลการวิจัย พบว่า ขนาดผู้ตอบข้อสอบมีอิทธิพลต่อค่าเฉลี่ยของดัชนีทุกตัว ความยาวของแบบทดสอบมีผลต่อค่าเฉลี่ยของดัชนี SA และ UA แต่ไม่อิทธิพลต่อค่าเฉลี่ย ของดัชนี α_{MH} และ β_{SIB} เกณฑ์ที่พัฒนาขึ้นเพื่อใช้ตัดสินความลำเอียง ระหว่างผู้ตอบข้อสอบเพศหญิงและเพศชาย มีดังนี้

1. $|SA| > .80$ และ $UA > .50$ เมื่อความยาวแบบทดสอบน้อยกว่า 50 ข้อ
2. $|SA| > .40$ และ $UA > 1.20$ เมื่อความยาวแบบทดสอบ 50 ข้อ ขึ้นไป
3. $\alpha_{MH} > .60$ และ $\alpha_{MH} < 1.40$ สำหรับทุกขนาดของผู้ตอบข้อสอบและความยาวแบบทดสอบ
4. $|\beta_{SIB}| > .06$ สำหรับทุกขนาดผู้ตอบข้อสอบ และความยาวแบบทดสอบ

ทั้งนี้ในการใช้ดัชนี SA หรือ UA ควรใช้ผู้ตอบข้อสอบขนาด 800 คนขึ้นไป ส่วนดัชนี α_{MH} และ β_{SIB} ควรใช้ผู้ตอบข้อสอบอย่างน้อย 600 คน

มหาวิทยาลัยบูรพา Burapha University

จิติมา วรรณศรี (2539) ได้เปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีแมลเทล - แฮนส์เซลและวิธีซิปเทสท์ โดยศึกษาจากข้อมูลจำลอง ตัวแปรที่ศึกษาได้แก่ ความยาวของแบบทดสอบ 3 ขนาด คือ 30, 60 และ 90 ข้อ ขนาดกลุ่มตัวอย่าง 3 ขนาด คือ 200, 600 และ 1,000 คน โดยแต่ละขนาดมีอัตราส่วนระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบต่างกัน คือ 1:1, 1: 0.9, 1:0.75 และ 1:0.5 ผลการศึกษาปรากฏว่า วิธีแมนเทล - แฮนส์เซลและวิธีซิปเทสท์มีประสิทธิภาพเท่าเทียมกัน ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในทุกขนาดกลุ่มตัวอย่างและทุกอัตราส่วนภายใต้ความยาวของแบบทดสอบเดียวกัน และเมื่อใช้แบบทดสอบที่มีความยาวปานกลาง (60 ข้อ) ทั้งสองสามารถตรวจสอบได้อย่างมีประสิทธิภาพที่สุด นอกจากนี้เมื่อใช้ขนาดกลุ่มตัวอย่างมากขึ้นจะสามารถตรวจสอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องมากขึ้น โดยส่วนมากวิธีซิปเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากกว่าวิธีแมนเทล-แฮนส์เซลเล็กน้อย

พรรณี จินตมาส (2540) ได้ศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับเพศ จากแบบทดสอบคณิตศาสตร์โจทย์ปัญหาที่ผู้วิจัยสร้างขึ้นเอง โดยใช้วิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธีแปลงค่าความยาก วิธีแมลเทล - แฮนส์เซลและวิธีซิปเทสท์ ในขนาดกลุ่มผู้เข้าสอบ 500 คน และขนาดกลุ่มผู้เข้าสอบ 1,000 คน โดยเปรียบเทียบจำนวนข้อสอบทำหน้าที่ต่างกัน และเปรียบเทียบค่าความเที่ยงแบ่งครึ่งฉบับของแบบทดสอบหลังคัดเลือกข้อสอบทำหน้าที่ต่างกันออกแล้ว กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ภาคเรียนที่ 2 ปีการศึกษา 2539 ของโรงเรียนสังกัดกรมสามัญศึกษาส่วนกลางจำนวน 2,200 คน ซึ่งได้มาโดยการสุ่มแบบชั้นมีขนาดโรงเรียนเป็นชั้นและโรงเรียนเป็นหน่วยการสุ่ม ผลการศึกษาปรากฏว่าเมื่อวิเคราะห์จากกลุ่มผู้เข้าสอบขนาด 500 คน วิธีซิปเทสท์พบข้อสอบทำหน้าที่ต่างกันมากที่สุด และวิธีแปลงค่าความยากพบข้อสอบทำหน้าที่ต่างกันน้อยที่สุด โดยจำนวนข้อสอบทำหน้าที่ด่างกันแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติทุกวิธีวิเคราะห์ และเมื่อวิเคราะห์จากกลุ่มผู้เข้าสอบ 1,000 คน วิธีแมนเทล - แฮนส์เซล พบข้อสอบทำหน้าที่ต่างกันมากที่สุด วิธีแปลงค่าความยากไม่พบข้อสอบทำหน้าที่ต่างกัน โดยจำนวนข้อสอบทำหน้าที่ต่างกัน จากการวิเคราะห์ด้วยวิธีแปลงค่าความยากกับวิธีแมลเทล - แฮนส์เซล และวิธีแปลงค่าความยากกับวิธีซิปเทสท์แตกต่างกันอย่างนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับจำนวนข้อสอบทำหน้าที่ต่างกัน จากการวิเคราะห์ด้วยวิธีแปลงค่าความยาก ระหว่างกลุ่มผู้เข้าสอบ 500 คนและขนาดกลุ่มผู้เข้าสอบ 1,000 คน จะมีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

วลีมาศ แซ่อึ้ง (2543) เปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อน

มหาวิทยาลัยบูรพา Burapha University

ประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอเนกรูป ระหว่างวิธีซิบเทสท์ปรับใหม่ และวิธีการถดถอยโลจิสติก โดยการศึกษาจำลองด้วยโมเดลโลจิสติกชนิด 3 พารามิเตอร์ ในเงื่อนไขต่าง ๆ 324 เงื่อนไข ผลการศึกษาพบว่าอำนาจการทดสอบของวิธีซิบเทสท์ปรับใหม่และวิธีการถดถอยโลจิสติกมีค่าเท่าเทียมกันทุกเงื่อนไข และทั้งสองวิธีดังกล่าวมีอำนาจการทดสอบสูงกว่าวิธีซิปเทสท์และแมนแทล - แฮนส์เซลเกือบทุกเงื่อนไข อัตราความคลาดเคลื่อนประเภทที่ 1 ที่ตรวจสอบด้วยวิธีการทั้ง 4 ข้างต้น มีค่าอยู่ภายในเกณฑ์ของอัตราความคลาดเคลื่อน ที่ระดับ 10 % เกือบทุกเงื่อนไข

นารายานาน และสวามิธาน (Narayanan & Swaminathan, 1994, pp. 315- 328 อ้างถึงในอารี วัชรโสตถิกุล, 2543, หน้า 28) ได้เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ด้วยวิธีแมนเทล - แฮนส์เซลและวิธีซิปเทสท์ โดยใช้ข้อมูลที่จำลองภายใต้เงื่อนไข 1,296 เงื่อนไข (9x3x2x4x6) แล้วคำนวณซ้ำ 100 ครั้ง แต่ละเงื่อนไขประกอบด้วย

1. ขนาดกลุ่มตัวอย่างที่แตกต่างกัน 3 ขนาด คือกลุ่มเปรียบเทียบใช้ขนาด 100, 200 และ 300 คน กลุ่มอ้างอิงใช้ขนาด 300, 500 และ 1,000 คน วางไขว้กันได้กลุ่มตัวอย่างทั้งหมด 9 กลุ่ม

2. ความแตกต่างในการแจกแจงความสามารถ 3 ระดับ ระดับที่ 1 ให้ทั้ง 2 กลุ่ม มีค่าเฉลี่ยการแจกแจงความสามารถที่ 0.0 ระดับที่ 2 ให้ค่าเฉลี่ยของความสามารถในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็น 0.0 และ 0.5 ตามลำดับ และระดับที่ 3 ค่าเฉลี่ยของความสามารถในกลุ่มอ้างอิงและเปรียบเทียบเป็น 0.0 และ -1.0 ตามลำดับ ส่วนเบี่ยงเบนมาตรฐานของทั้ง 3 ระดับเป็น 1.0

3. ร้อยละของข้อสอบทำหน้าที่ต่างกันมี 2 ขนาด คือ ร้อยละ 10 และ ร้อยละ 20 จากแบบทดสอบจำนวน 40 ข้อ

4. ขนาดอิทธิพลของข้อสอบทำหน้าที่ต่างกัน 4 ระดับ คือ 0.4, 0.6, 0.8 และ 1.0

5. ลักษณะของข้อสอบ 6 ระดับ (ลักษณะ b ต่ำกับ a ปานกลาง, b ต่ำกับ a สูง, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง, b สูงกับ a ต่ำ, b สูงกับ a ปานกลาง)

ผลการศึกษาปรากฏว่า วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันได้ใกล้เคียงกัน โดยตรวจพบได้ดีในขนาดกลุ่มตัวอย่าง 600 คน (กลุ่มอ้างอิงและกลุ่มเปรียบเทียบกลุ่มละ 300 คน) ขนาดของข้อสอบทำหน้าที่ต่างกันทั้ง 4 ขนาดมีผลต่อการตรวจพบการทำหน้าที่ต่างกันของข้อสอบทั้งสองวิธี อย่างมีนัยสำคัญทางสถิติ ความแตกต่างในการแจกแจงความสามารถไม่มีผลต่อวิธีการวิเคราะห์การทำหน้าที่ต่างกันโดยวิธีซิปเทสท์ แต่มีผลต่อวิธีแมนเทล-แฮนส์เซล ที่เป็นเช่นนี้อาจเป็นเพราะวิธีซิปเทสท์ได้มีการปรับแก้การถดถอย ขนาดของแบบทดสอบและร้อยละของข้อสอบทำหน้าที่ต่างกันไม่มีผลต่อวิธีทั้งสอง สำหรับการ

มหาวิทยาลัยบูรพา Burapha University

เกิดความคลาดเคลื่อนแบบที่ 1 วิธีแมนเทล-แฮนส์เซล จะมีอัตราการเกิดอยู่ในระดับปกติ แต่วิธีซิปเทสท์จะเพิ่มเล็กน้อยในกลุ่มที่มีความสามารถเท่ากัน แต่ในกลุ่มที่มีความสามารถไม่เท่ากันการเกิดความคลาดเคลื่อนแบบที่ 1 จะมีอัตราเกิดสูงขึ้นทั้งสองวิธี ซึ่งการเพิ่มความยาวของแบบทดสอบจะทำให้เกิดความเที่ยงสูงขึ้น มีผลให้ความคลาดเคลื่อนประเภทที่ 1 น่าจะมีอัตราการเกิดลดลง

นารายานันและสวามินาทาน (Narayanan & Swaminathan, 1996, หน้า 257-274 อ้างถึงใน นุชริน ใบโพธิ์, 2544, หน้า 46) ได้เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธีแมนเทล - แฮนส์เซล วิธีฉดถอยโลจิสติกและวิธีซิปเทสท์ โดยศึกษาอำนาจการตรวจสอบและการจำแนกผิดพลาดโดยการจำลองข้อมูล ตัวแปรที่ศึกษา ได้แก่ ขนาดกลุ่มตัวอย่าง ความแตกต่างของการกระจายความสามารถระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ สัดส่วนของข้อสอบทำหน้าที่ต่างกันที่มีภายในแบบทดสอบ ขนาดพื้นที่ระหว่างโค้งลักษณะข้อสอบของผู้เข้าสอบสองกลุ่ม ค่าความยากและค่าอำนาจจำแนกของแบบทดสอบ ผลการศึกษาปรากฏว่าวิธีซิปเทสท์และวิธีฉดถอยโลจิสติก มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอนเกรูปได้เท่าเทียมกันในทุกเงื่อนไขที่ศึกษา ส่วนวิธีแมนเทล - แฮนส์เซลไม่สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป และการจำแนกผิดพลาดวิธีแมนเทล - แฮนส์เซลจะสูงกว่าวิธีฉดถอยโลจิสติกและวิธีซิปเทสท์

นุชริน ใบโพธิ์ (2544) เปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบทำหน้าที่ต่างกันระหว่างวิธี ไค - สแควร์ของลอร์ด วิธีแมนเทล - แฮนส์เซล และวิธีซิปเทสท์ โดยใช้กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ปีการศึกษา 2543 ในเขตกรุงเทพมหานคร จำนวน 790 คน เครื่องมือที่ใช้วัดเป็นแบบทดสอบวัดความถนัดทางคณิตศาสตร์ จำนวน 60 ข้อ ผลการวิจัยพบว่า

1. ข้อสอบทำหน้าที่ต่างกันจากการตรวจสอบด้วย วิธี ไค - สแควร์ของลอร์ด, วิธีแมนเทล - แฮนส์เซล และวิธีซิปเทสท์ มีจำนวนข้อสอบที่ตรวจพบแตกต่างกัน เมื่อขนาดกลุ่มตัวอย่างและจำนวนข้อสอบต่างกัน ได้แก่ กรณีแบบทดสอบ 30 ข้อ ที่ใช้กลุ่มตัวอย่างที่มีขนาด 300 คน และ 500 คน แบบทดสอบ 40 ข้อ ใช้กลุ่มตัวอย่าง 500 คน และ 700 คน

2. ความสอดคล้องของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อพิจารณาจากข้อสอบจำนวน 20 ข้อ และ 30 ข้อ ที่ตรวจสอบโดยวิธีไค - สแควร์ของลอร์ด วิธีแมนเทล - แฮนส์เซลและวิธีซิปเทสท์ พบว่ากลุ่มตัวอย่างขนาด 500 คน และ 700 คน มีจำนวนข้อสอบที่ตรวจพบสอดคล้องกันสูงสุด ส่วนข้อสอบจำนวน 40 ข้อ ซึ่งตรวจด้วยวิธีไค - สแควร์ของลอร์ด พบว่ากลุ่มตัวอย่างขนาด 300 คน และ 500 คน มีจำนวนข้อสอบที่ตรวจพบสอดคล้องกันสูงสุด

มหาวิทยาลัยบูรพา Burapha University

แต่เมื่อตรวจสอบด้วยวิธีแมนเทล - แฮนส์เซลและวิธีซิปเทสท์ พบว่ากลุ่มตัวอย่างขนาด 500 และ 700 คน มีจำนวนข้อสอบที่ตรวจพบสอดคล้องกันสูงสุด

สิริรัตน์ วิภาสศิลป์ (2545) ได้เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมวดข้อสอบและแบบทดสอบ จากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติระหว่างวิธีซิปเทสท์กับวิธีดีเอฟไอที ภายใต้เงื่อนไขความยาวของแบบทดสอบ 30, 40 และ 50 ข้อ กลุ่มตัวอย่างขนาด 50, 100, 200, 500, และ 1,000 คน กลุ่มตัวอย่างการศึกษาได้มาจากการสุ่มแบบใส่คืนจากประชาชนเทียม ซึ่งกำหนดจากนักเรียนชายและนักเรียนหญิงชั้นมัธยมศึกษาปีที่ 1 จังหวัดนนทบุรี แต่ละขนาดสุ่มกลุ่มตัวอย่าง 50 ครั้ง เครื่องมือที่ใช้ในการวิจัยเป็นแบบทดสอบวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างขึ้น ซึ่งเป็นข้อสอบแบบหลายตัวเลือก 5 ตัวเลือก จำนวน 50 ข้อ มีข้อสอบที่ผู้เชี่ยวชาญพิจารณาว่าเป็นข้อสอบทำหน้าที่ต่างกันต่อเพศชายจำนวน 16 ข้อ หลังจากเก็บรวบรวมข้อมูลแล้วคัดเลือกข้อสอบตามสัดส่วนในตารางสัดส่วนกำหนด จัดเป็นแบบทดสอบที่มีความยาว 30 และ 40 ข้อ แล้วตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมวดข้อสอบ และแบบทดสอบในเงื่อนไขต่าง ๆ โดยโปรแกรมซิปเทสท์และดีเอฟทีไอ จากนั้นนำผลที่ได้ไปเปรียบเทียบความถูกต้อง และการระบุผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีเดียวกันและต่างกัน โดยการวิเคราะห์ความแปรปรวนแบบตัวแปรพหุ คำนวณความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ และแบบทดสอบด้วยวิธีซิปเทสท์และวิธีดีเอฟไอที โดยใช้สถิติ Z-test ผลการศึกษาปรากฏว่า

1. เมื่อแบบทดสอบประกอบด้วยข้อสอบ 30, 40 และ 50 ข้อ กลุ่มตัวอย่างขนาด 50, 100 และ 200 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ไม่แตกต่างกัน กลุ่มตัวอย่างขนาด 500 และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์สูงกว่ากลุ่มตัวอย่างขนาด 50, 100 และ 200 คน แต่การระบุผิดพลาดในการตรวจสอบสูงกว่าด้วย เมื่อตรวจสอบด้วยวิธีดีเอฟไอที พบว่ากลุ่มตัวอย่างขนาด 50, 100, 200, 500 และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่แตกต่างกัน

2. ทุกเงื่อนไขความยาวแบบทดสอบและกลุ่มตัวอย่างขนาดแตกต่างกัน วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบน้อยกว่าวิธีดีเอฟไอที และพบว่าความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทั้งสองมีค่าต่ำกว่าร้อยละ 1

3. วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบมากกว่าดีเอฟไอที เมื่อแบบทดสอบมีข้อสอบ 30 ข้อ กลุ่มตัวอย่าง 1,000 คน และเมื่อแบบทดสอบแบบมี 40 ข้อ กลุ่มตัวอย่างขนาต 500 คน

มหาวิทยาลัยบูรพา Burapha University

4. วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบมากกว่าวิธีดีเอฟไอที เมื่อแบบทดสอบมีข้อสอบ 50 ข้อ กลุ่มตัวอย่างขนาด 100, 200 และ 1,000 คน

ราชู แวนเดอร์ ลินเดน และฟลีร์ (Raju, Vander, Linden, & Fleer, 1995, pp. 353 – 368 อ้างถึงใน สิริรัตน์ วิภาสศิลป์, 2545, หน้า40) เสนอวิธีการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบและแบบทดสอบ (Internal Measures of Difficential Functioning of Items and Tests: DFIT) โดยอาศัยกรอบความคิดของทฤษฎีการตอบข้อสอบ ศึกษาโดยใช้ข้อมูลจำลองที่สร้างขึ้นมาใช้แบบจำลองโลจิสติก ชนิด 2 พารามิเตอร์ ด้วยโปรแกรม RANDGEN (Fleer, Kiley, & Raju, 1991) สร้างแบบทดสอบที่มีความยาว 40 ข้อ ประกอบด้วยข้อสอบที่แสดง DIF 0 % , 5 %, 10 % และ 20 % ขนาดของกลุ่มตัวอย่างกลุ่มละ 500 คน สำหรับเงื่อนไขกลุ่มตัวอย่างขนาดเล็ก และกลุ่มละ 1000 คน สำหรับกลุ่มตัวอย่างขนาดใหญ่ ตรวจสอบการทำหน้าที่เบี่ยงเบนในเงื่อนไขต่าง ๆ ด้วยวิธีการของไค – สแควร์ของลอร์ด การวัดพื้นที่แบบคิดเครื่องหมาย การวัดพื้นที่โดยไม่คิดเครื่องหมายและวิธีดีเอฟไอที ผลที่ได้พบว่าวิธีการดีเอฟไอทีสามารถตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบทั้งแบบที่มีการชดเชยและแบบที่ไม่มีการชดเชย (CDIF และ NCDIF) ได้อย่างถูกต้อง

ดักล๊าส รุสโซ และ สเตาต์ (Douglas, Roussos, & Stout, 1996, pp. 465 – 484 อ้างถึงใน สิริรัตน์ วิภาสศิลป์, 2545, หน้า 40) ได้ตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบ โดยพิจารณาข้อสอบพร้อม ๆ กันครั้งละหลายข้อ โดยเสนอวิธีการตรวจสอบหมวดข้อสอบที่คาดว่าจะแสดงการทำหน้าที่เบี่ยงเบนของข้อสอบเหล่านั้นเข้ามาพิจารณาพร้อมกัน 2 วิธี คือ ใช้เฉพาะความคิดเห็นของผู้เชี่ยวชาญ และใช้วิธีทางสถิติและตรวจสอบซ้ำด้วยความเห็นของผู้เชี่ยวชาญ พร้อมทั้งแสดงตัวอย่างตรวจสอบการทำหน้าที่เบี่ยงเบน 3 ตัวอย่าง ดังนี้ ตัวอย่างที่ 1 เลือกหมวดข้อสอบโดยใช้ความเห็นของผู้เชี่ยวชาญอย่างเดียว ให้ผู้เชี่ยวชาญชาย 3 คน และหญิง 1 คน พิจารณาแบบทดสอบย่อยใช้เหตุผลเชิงตรรกศาสตร์ (Logical Reasoning Subtest) ที่ได้จากการดำเนินการสอบในเดือนธันวาคม ค.ศ. 1991 (December 1991 Administation of the Low School Admission Test: LSAT) จำนวน 49 ข้อ ในการตรวจให้เพศชายเป็นกลุ่มอ้างอิงและเพศหญิงเป็นกลุ่มสนใจ ผู้เชี่ยวชาญจัดหมวดข้อสอบออกเป็น 8 หมวด และพิจารณาว่าหมวดใดน่าจะให้ประโยชน์แก่เพศชายหรือเพศหญิง หลังจากนั้นวิเคราะห์ด้วยโปรแกรมคอมพิวเตอร์สำเร็จรูปซิปเทสท์ ใช้กลุ่มตัวอย่างเพศชาย 3,000 คน พบว่าความคิดเห็นของผู้เชี่ยวชาญและผลวิเคราะห์ด้วยโปรแกรมซิปเทสท์สอดคล้องกัน จึงทำการวิเคราะห์ปริมาณ DIF ของแต่ละข้อ พบว่าแต่ละข้อแสดง DIF ด้วยปริมาณที่น้อยมาก จึงไม่มีการคัดเลือกข้อสอบออกจากแบบทดสอบ ตัวอย่างที่ 2 ใช้วิธีการทางสถิติในการตรวจสอบหมวดข้อสอบ คือ วิธี HCA (Agglomerative Hierachical Cluster Analysis;

มหาวิทยาลัยบูรพา Burapha University

Jian & Dubes, 1988) และ DIMTEST (Nandacumar & Stout, 1993; Stout, 1987) แล้วจึงตรวจสอบด้วยโปรแกรมซิปเทสท์ ในการตรวจสอบใช้ข้อสอบ (National Assessment of Educational Progress (NEAP) จำนวน 36 ข้อ ใช้กลุ่มตัวอย่างจำนวน 500 คน ผลการตรวจสอบพบว่าการตรวจสอบการทำหน้าที่เบี่ยงเบนของหมวดข้อสอบ ประสบความสำเร็จเป็นอย่างดี ตัวอย่างที่ 3 ใช้แบบทดสอบความเข้าใจในการอ่าน ที่สอบในเดือนธันวาคม ค.ศ. 1991 ซึ่งเป็นส่วนหนึ่งของการทดสอบ LSAT แบ่งเป็นบทความที่กำหนดให้อ่านเป็น 4 บทความ แต่ละบทความมีข้อสอบให้อ่าน 5-8 ข้อ แล้วทำการตรวจสอบหมวดข้อสอบซ้ำโดยใช้วิธี เอชซีเอ และ ดิมเทสท์ หลังจากนั้นจึงวิเคราะห์ด้วยโปรแกรมซิปเทสท์ ใช้กลุ่มตัวอย่างกลุ่มละ 1,000 คน พบว่าทั้ง 4 บทความ แสดงการทำหน้าที่เบี่ยงเบนของหมวดข้อสอบ ข้อสอบ 2 หมวดให้ประโยชน์แก่เพศชาย และ 2 หมวด ให้ประโยชน์แก่เพศหญิง

สุมาลี แกล้วทนงค์ (2547) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบด้านเพศ ภาษาพูด และเชื้อชาติ ของแบบสอบสาระการเรียนรู้ภาษาไทย และสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรม และตรวจสอบสาเหตุของการทำหน้าที่ต่างกันของข้อสอบด้านเพศ ภาษาพูด และเชื้อชาติ ของแบบสอบกลุ่มสาระการเรียนรู้ภาษาไทย และกลุ่มสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรม กลุ่มตัวอย่างที่ใช้ในการวิจัย คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 1 ปีการศึกษา 2546 สำนักงานเขตพื้นที่การศึกษาสงขลา พัทลุง ตรัง และสตูล จำนวน 1,320 คน เครื่องมือที่ใช้ในการวิจัยประกอบด้วยแบบสอบถามเพื่อสำรวจความสนใจในการอ่านของผู้เรียน แบบสอบในกลุ่มสาระการเรียนรู้ภาษาไทยและกลุ่มสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรม และแบบวินิจฉัยสาเหตุการทำหน้าที่ต่างกันของข้อสอบ ในกลุ่มสาระการเรียนรู้ภาษาไทยและสังคมศึกษา ศาสนา และวัฒนธรรม วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทล - แฮนส์เซล และโปรแกรมซิปเทสท์ และตรวจสอบสาเหตุการทำหน้าที่ต่างกันโดยผู้เชี่ยวชาญ ผลการวิจัยพบว่า

1. แบบสอบกลุ่มสาระการเรียนรู้ภาษาไทยมี 3 ฉบับ ประกอบด้วยข้อสอบจำนวน 120 ข้อ มีข้อสอบที่ทำหน้าที่ต่างกันด้านเพศ 9 ข้อ ภาษาพูด 15 ข้อ และเชื้อชาติ 28 ข้อ และกลุ่มสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรมมีข้อสอบที่ทำหน้าที่ต่างกันด้านเพศ 22 ข้อ ด้านภาษาพูด 52 ข้อ และเชื้อชาติ 20 ข้อ

2. ข้อสอบที่ทำหน้าที่ต่างกันด้านเพศของแบบสอบสาระการเรียนรู้ภาษาไทย ส่วนใหญ่มีสาเหตุมาจากการเนื้อเรื่องที่น่าสนใจและภาษาที่ใช้ในแบบสอบ โดยมีลักษณะของข้อสอบที่ออกตามความสนใจและมีการใช้ภาษาในแบบสอบที่เป็นสำนวนและคำศัพท์ และข้อสอบที่ทำหน้าที่ต่างกันด้านเพศ ของแบบสอบสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรม มีสาเหตุมาจาก

มหาวิทยาลัยบูรพา Burapha University

เนื้อเรื่องที่สนใจและเนื้อเรื่องเกี่ยวกับวัฒนธรรม โดยมีลักษณะของข้อสอบที่มีเนื้อเรื่องที่สนใจและมีเนื้อเรื่องเกี่ยวกับวัฒนธรรมของผู้สอบ

3. ข้อสอบที่ทำหน้าที่ต่างกันด้านภาษาพูดของแบบสอบสาระการเรียนรู้ภาษาไทย ส่วนใหญ่มีสาเหตุมาจาก เนื้อเรื่องที่สนใจ และภาษาที่ใช้ในแบบสอบ โดยมีลักษณะข้อสอบที่ออกตามความสนใจและมีการใช้ภาษาที่เป็นคำศัพท์เฉพาะ เช่น ราชาศัพท์ และข้อสอบที่ทำหน้าที่ต่างกันด้านภาษาพูดของแบบสอบสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรม มีสาเหตุมาจากเนื้อเรื่องที่สนใจ และเนื้อเรื่องเกี่ยวกับวัฒนธรรม โดยมีลักษณะของข้อสอบที่ออกตามความสนใจ เช่น ภูมิศาสตร์ และในเรื่องเกี่ยวกับวัฒนธรรมไทย

4. ข้อสอบที่ทำหน้าที่ต่างกันในด้านเชื้อชาติของแบบสอบสาระการเรียนรู้ภาษาไทย ส่วนใหญ่มีสาเหตุมาจากเนื้อเรื่องที่สนใจ และภาษาที่ใช้ในแบบสอบถาม โดยมีลักษณะข้อสอบที่ออกตามความสนใจ และมีการใช้คำราชาศัพท์หรือคำศัพท์ยากมาออกข้อสอบ และข้อสอบที่ทำหน้าที่ต่างกันในด้านเชื้อชาติของแบบสอบสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรม มีสาเหตุจากเนื้อเรื่องที่สนใจและเนื้อเรื่องเกี่ยวกับวัฒนธรรมและประเพณี

ศุภวัฒน์ มะลิเผือก (2548) ได้ศึกษาอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบที่ส่งผลต่อคุณภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติวิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 ด้วยวิธีซิปเทสท์กับวิธีถดถอยโลจิสติก โดยการเปรียบเทียบค่าความเที่ยง ความตรงเชิงโครงสร้าง ความคงที่ของโครงสร้างองค์ประกอบ และค่าสัมประสิทธิ์สหสัมพันธ์อันดับของผู้สอบ ระหว่างแบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออก ผลปรากฏว่า พบข้อสอบที่ต่างกันเมื่อจำแนกกลุ่มผู้สอบตามตัวแปรเพศจำนวน 12 ข้อ คิดเป็นร้อยละ 30 แบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันมีค่าความเที่ยงแตกต่างกัน โครงสร้างองค์ประกอบคงที่และค่าสัมประสิทธิ์สหสัมพันธ์อันดับของผู้สอบ มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.5

รุสโซ และสเตาท์ (Roussos & Stout, 1996, pp. 215-230 อ้างถึงใน ปิยะทิพย์ ตินวร, 2549, หน้า 34) ได้ศึกษาสถานการณ์จำลองของผลกระทบของกลุ่มตัวอย่างขนาดเล็ก และค่าพารามิเตอร์ของข้อสอบที่มีต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ด้วยวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล โดยจำลองข้อมูล 2 ครั้ง ครั้งแรกศึกษากลุ่มตัวอย่างขนาดเล็ก ใช้กลุ่มตัวอย่าง 4 ขนาด (100, 200, 500, และ 1,000) และความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ 3 ขนาด (0, 0.5 และ 1) ใช้แบบทดสอบ 25 ข้อ ครั้งที่ 2 ศึกษาค่าพารามิเตอร์ของข้อสอบ ใช้กลุ่มตัวอย่าง 3 ขนาด (500, 1,000 และ 3,000) และความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ 2 ขนาด

มหาวิทยาลัยบูรพา Burapha University

(0 และ 1) เลือกค่าพารามิเตอร์อำนาจจำแนก 3 ค่า (0.4, 1.0 และ 2.5) พารามิเตอร์ความยาก 5 ค่า (-1.5, -0.5, 0, 0.5 และ 1.5) และค่าพารามิเตอร์การเดา 1 ค่า (.20) ผลการศึกษาพบว่า ครั้งที่ 1 เมื่อศึกษากลุ่มตัวอย่างขนาดเล็ก พบว่าอัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล มีค่าไม่แตกต่างกัน ครั้งที่ 2 เมื่อศึกษาค่าพารามิเตอร์ของข้อสอบ พบว่าเมื่อความแตกต่างของการแจกแจงค่าความสามารถ ระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบไม่แตกต่างกัน และอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์มีค่าต่ำกว่าวิธีแมนเทล - แฮนส์เซลในทุกเงื่อนไข

ตอนที่ 4 ผลสัมฤทธิ์ทางการเรียน

ผลสัมฤทธิ์ทางการเรียน หมายถึงคุณลักษณะรวมถึงความรู้ความสามารถของบุคคลอันเป็นผลมาจากการเรียนการสอน หรือมวลประสบการณ์ทั้งปวงที่บุคคลได้รับจากการเรียนการสอน ทำให้บุคคลเกิดการเปลี่ยนแปลงพฤติกรรมในด้านต่าง ๆ ของสมรรถภาพสมอง แสดงว่าเป็นการตรวจสอบพฤติกรรมของผู้เรียนในด้านพุทธิพิสัยนั่นเอง โดยการวัด 2 องค์ประกอบ คือ

1. การวัดด้านปฏิบัติ โดยให้ผู้เรียนได้ลงมือปฏิบัติจริงให้เป็นผลงานปรากฏออกมา ทำให้สังเกตและวัดได้ เช่น ศิลปศึกษา พลศึกษา เป็นต้น การวัดแบบนี้จึงต้องใช้ข้อสอบภาคปฏิบัติ ซึ่งเป็นการประเมินผลและพิจารณาที่วิธีปฏิบัติและผลงานที่ปฏิบัติ

2. การวัดด้านเนื้อหาเป็นการตรวจสอบความรู้ความสามารถเกี่ยวกับเนื้อหาวิชา รวมถึงพฤติกรรมความสามารถในด้านต่าง ๆ อันเป็นผลมาจากการทำกิจกรรมการเรียนการสอน มีวิธีสอบวัดได้ 2 ลักษณะ คือ การสอบปากเปล่า และการสอบแบบให้เขียนตอบ ในการสอบแบบเขียนตอบนี้จะใช้กันแพร่หลาย เครื่องมือที่ใช้ในการสอบวัด เรียกว่า “ข้อสอบวัดผลสัมฤทธิ์” หรือ “แบบทดสอบวัดผลสัมฤทธิ์” เพื่อให้การวัดผลสัมฤทธิ์ทางการเรียนเกิดประสิทธิภาพสามารถแยกพฤติกรรมด้านต่าง ๆได้ชัดเจน จึงต้องมีการสร้างแบบทดสอบที่เหมาะสมกับจุดมุ่งหมายของเนื้อหาขึ้น

จะเห็นว่าผลสัมฤทธิ์ทางการเรียน เป็นผลที่เกิดขึ้นจากการเรียนการสอนหรือมวลประสบการณ์ที่ได้รับจากสิ่งแวดล้อม ซึ่งทำให้บุคคลเกิดการเปลี่ยนแปลงพฤติกรรมในด้านต่าง ๆ ของสมรรถภาพทางสมอง ซึ่งในการศึกษาครั้งนี้ผู้รายงานใช้แบบทดสอบในลักษณะเขียนตอบ โดยเรียกว่าแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน เพราะสามารถทราบสมรรถภาพทางสมองและสามารถวัดพฤติกรรมต่าง ๆ ได้ชัดเจนครบถ้วนตามจุดมุ่งหมายที่ต้องการวัด

จุดมุ่งหมายของการวัดผลสัมฤทธิ์ทางการเรียน

การวัดผลสัมฤทธิ์ทางการเรียนมีจุดมุ่งหมายเพื่อ

มหาวิทยาลัยบูรพา Burapha University

1. ตรวจสอบความสามารถด้านพุทธิพิสัย เพื่อเปรียบเทียบการเปลี่ยนแปลงในการเรียนรู้ โดยวัดทางด้านเนื้อหา

2. ปรับปรุงแก้ไขกระบวนการเรียนการสอน โดยใช้การวัดผลเป็นองค์ประกอบหนึ่งของการเรียนการสอน

3. ประเมินผลผู้เรียนว่าบรรลุเป้าหมายในสิ่งที่สอนไปหรือไม่ โดยการประเมินผลผู้เรียน มีประโยชน์ต่อการจัดการศึกษาคือ

3.1 ช่วยสำรวจความต้องการด้านการศึกษาของแต่ละบุคคล

3.2 ช่วยตรวจสอบประสิทธิผลของวิธีการจัดการเรียนการสอน

3.3 ช่วยให้สารสนเทศแก่ผู้เรียน อาจารย์ที่ปรึกษาและนักแนะแนว

3.4 ช่วยให้สารสนเทศแก่ผู้บริหารการศึกษา เพื่อนำไปใช้ในการจัดทรัพยากรของโรงเรียน

3.5 ช่วยให้สารสนเทศแก่ผู้บริหารการศึกษา เพื่อนำไปใช้จัดการบริหารงานบุคคล

3.6 เพื่อเลือกวิธีหรือกระบวนการสอบที่เหมาะสม ที่จะเป็นแรงกระตุ้นให้ผู้เรียนมีความสนใจในการเรียนยิ่งขึ้น

วิธีการทดสอบวัดผลสัมฤทธิ์ทางการเรียน

วิธีการทดสอบวัดผลสัมฤทธิ์ทางการเรียนมี 2 วิธี

1. การทดสอบแบบอิงกลุ่ม เป็นการทดสอบที่แปลความหมายของคะแนนโดยการนำเอาผลการปฏิบัติงาน ไปเปรียบเทียบกับผลการปฏิบัติงานของคนอื่น ๆ ภายในกลุ่ม การรายงานผลการทดสอบใช้คะแนนมาตรฐานในการบ่งบอก เช่น เปอร์เซ็นต์ไทล์ เกรดเทียบเท่า (Grade Equivalent)

2. การทดสอบแบบอิงเกณฑ์ เป็นการทดสอบที่แปลความหมายของคะแนนโดยการนำเอาผลการปฏิบัติงานไปเปรียบเทียบกับมาตรฐานที่แท้จริง (Absolute Standard) ซึ่งเป็นเกณฑ์ภายนอกกลุ่มที่กำหนดไว้อย่างรอบคอบโดยไม่เปรียบเทียบกับผลงานคนอื่นภายในกลุ่ม ดังนั้นผลงานของนักเรียนจะอยู่ในระดับมาตรฐานหรือไม่ ต้องพิจารณาและเปรียบเทียบกับมาตรฐานที่แท้จริงเท่านั้น การรายงานผลการทดสอบเสนอในพจน์ของจำนวน หรือเปอร์เซ็นต์การตอบถูกของแต่ละบุคคล แบบทดสอบแบบอิงเกณฑ์จำแนกเป็น 2 ประเภท ดังนี้

2.1 แบบทดสอบอิงจุดประสงค์ เป็นแบบทดสอบอิงเกณฑ์ที่สร้างขึ้นโดยใช้จุดประสงค์รายวิชา ส่วนมากจะมีการกำหนดเกณฑ์เพื่อบ่งชี้ระดับความรอบรู้ของผู้สอบ ซึ่งมักจะใช้คะแนนจุดตัดของแบบทดสอบ

2.2 แบบทดสอบอิงความรู้ เป็นแบบทดสอบอิงเกณฑ์ที่สร้างขึ้นโดยยึดหลัก

มหาวิทยาลัยบูรพา Burapha University

เฉพาะของมวลความรู้ ใช้การประมวลความสามารถของผู้สอบที่สามารถตอบข้อสอบถูกในประชากรข้อสอบ

แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน (Achievement Test)

แนวคิดของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน (เยาวดี วิบูลย์ศรี, 2545, หน้า 14-27)

1. แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน (Achievement Test) เป็นเครื่องมือสำหรับช่วยให้ครูสามารถตัดสินผลสัมฤทธิ์ทางการเรียนได้อย่างมีประสิทธิภาพ เพราะเป็นวิธีการประเมินพฤติกรรมของนักเรียนที่มีความเป็นอิสระได้มากกว่าวิธีอื่น ๆ แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนที่ใช้ในโรงเรียนมุ่งวัดความรู้ในแต่ละวิชา และทักษะต่าง ๆ โดยมีวัตถุประสงค์พื้นฐานที่สำคัญ 2 ประการ คือ ประการแรก เพื่อเป็นเครื่องมือในการวัดผลสัมฤทธิ์ทางการเรียน อันเป็นข้อมูลที่ได้รับการประเมินผลการเรียนการสอนเป็นรายบุคคล ประการที่สอง เพื่อเป็นการตรวจสอบความสามารถของนักเรียนแต่ละคนซึ่งแตกต่างกันโดยธรรมชาติ เมื่อความสามารถของแต่ละคนมีความแตกต่างกันทั้งเด็กและผู้ใหญ่ ดังนั้นการที่จะให้บุคคลต่าง ๆ ได้รับการพัฒนาความสามารถเฉพาะตนที่มีอยู่อย่างเหมาะสม จึงก่อให้เกิดการพัฒนาแบบทดสอบวัดผลสัมฤทธิ์ที่มีประสิทธิภาพขึ้น เพื่อใช้เป็นเครื่องมือในการวัดระดับความสามารถของบุคคล หรือเพื่อจำแนกความสามารถของบุคคลที่แตกต่างกัน ทำให้เราสามารถศึกษาให้สอดคล้องกับระดับความสามารถของบุคคลต่าง ๆ ได้

2. ประเภทของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน เราสามารถจำแนกตามมิติต่าง ๆ ได้หลายมิติ ดังนี้

2.1 จำแนกตามขอบข่ายเนื้อหาวิชาของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน อาจกำหนดให้กว้างได้หรือกำจัดให้แคบลง ตามปกติแล้วยังไม่มีมาตรฐานอ้างอิงสากลที่จะนำไปใช้ในการกำหนดเนื้อหาวิชาสำหรับแบบทดสอบวัดผลสัมฤทธิ์ ผู้ใช้แบบทดสอบเท่านั้นจะต้องกำหนดเนื้อหาวิชาขึ้นเอง โดยให้สอดคล้องกับวัตถุประสงค์ของการสอบ ผู้สร้างแบบทดสอบสามารถที่จะพัฒนาแบบทดสอบให้มีเนื้อหาได้ตามขอบข่ายที่ต้องการ สำหรับแบบทดสอบมาตรฐานวัดผลสัมฤทธิ์ ซึ่งจัดพิมพ์ในต่างประเทศจำนวนมากมักจะรวมแบบทดสอบต่าง ๆ ไว้เป็นชุด (Batteries) แต่ละชุดจะครอบคลุมเนื้อหาวิชาสำหรับระดับชั้นเรียนที่ต่าง ๆ กัน ทั้งนี้ก็เพื่อให้โรงเรียนทั้งหลายสามารถวัดผลสัมฤทธิ์ของนักเรียนได้อย่างต่อเนื่อง ตามความเจริญงอกงามทางวิชาการในแต่ละยุคแต่ละสมัย ตัวอย่างเช่น แบบทดสอบวัดผลสัมฤทธิ์ทั่วไป (General Achievement Test Batteries) หรือเรียกย่อ ๆ ว่า GATB ซึ่งเป็นแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานที่นิยมใช้กันในระดับประถมศึกษาและมัธยมศึกษาเป็นส่วนใหญ่ ทั้งนี้เพราะเนื้อหาที่

เหมาะสมกับช่วงชั้นดังกล่าวเป็นอย่างดี เช่น ในระดับประถมศึกษานั้นลักษณะของแบบทดสอบจะประกอบด้วยเนื้อหาซึ่งสัมพันธ์กับทักษะสามประการ (3R) คือ การอ่าน (Reading) การเขียน (Writing) และการคำนวณ (Arithmetic) ส่วนในระดับมัธยมศึกษาแบบทดสอบ GATB มีความมุ่งหมายเพื่อวัดผลสัมฤทธิ์ทั่วไป รวมทั้งเพื่อวัดพื้นฐานความรู้ของแต่ละวิชา เช่น พื้นฐานความรู้ทางสังคมศาสตร์ พื้นฐานความรู้ของวิชาวิทยาศาสตร์ธรรมชาติ เป็นต้น นอกจากนั้นก็จะเน้นการวัดความสามารถด้านการอ่าน ความเข้าใจในสาขาวิชา รวมทั้งความคิดเกี่ยวกับจำนวนและคำศัพท์ทั่วไป ซึ่งเป็นการวัดในด้านการพัฒนาความสามารถเชิงวิชาการ ดังนั้นแบบทดสอบ GATB ในระดับนี้ จึงมักจะใช้ในการสอบคัดเลือกบุคคลเข้าศึกษาต่อในสถาบันอุดมศึกษา หรือในการสอบคัดเลือกเข้าแข่งขันชิงทุนประเภทต่าง ๆ เป็นต้น

2.2 จำแนกตามลักษณะหน้าที่ทั่วไปของแบบทดสอบ ซึ่งแบ่งออกเป็น 3 ลักษณะ ดังนี้

2.2.1 แบบทดสอบเพื่อการสำรวจผลสัมฤทธิ์ (Survey Tests) เป็นแบบทดสอบวัดผลสัมฤทธิ์ที่ทำหน้าที่ในการสำรวจความสามารถทั่ว ๆ ไปของนักเรียน โดยประเมินความรู้ในเนื้อหาวิชาหรือทักษะต่าง ๆ เพื่อแสดงระดับความสามารถของนักเรียน ดังนั้นแบบทดสอบวัดผลสัมฤทธิ์จึงมักจะครอบคลุมเนื้อหาทั้งในระดับกว้างและระดับทั่วไป โดยถือคะแนนรวมที่ได้จากแบบทดสอบเป็นตัวบ่งชี้ถึงระดับความสามารถที่วัดได้

2.2.2 แบบทดสอบเพื่อวินิจฉัยผลสัมฤทธิ์ (Diagnostic Test) เป็นแบบทดสอบวัดผลสัมฤทธิ์ที่ทำหน้าที่วินิจฉัยเกี่ยวกับจุดเด่นและจุดด้อยขององค์ประกอบสำคัญทางด้านทักษะต่าง ๆ ของนักเรียน ตัวอย่างเช่น แบบทดสอบเพื่อวินิจฉัยผลสัมฤทธิ์ทางด้านภาษา อาจจะรวมแบบทดสอบย่อยหลายชุด ซึ่งครอบคลุมเนื้อหาเกี่ยวกับการรู้จักใช้คำ ความเข้าใจเกี่ยวกับถ้อยคำต่าง ๆ รวมทั้งคำศัพท์ ตลอดจนอัตราการอ่าน ความเข้าใจในเรื่องที่อ่าน การจำแนกเสียงและการจำแนกพยางค์ ฯลฯ คะแนนที่ได้จากแต่ละองค์ประกอบของแบบทดสอบวินิจฉัยดังกล่าว จะช่วยให้นักจิตวิทยาหรือครูสามารถตัดสินใจได้ว่า อะไรคือจุดพกพร่องของผู้สอบ ซึ่งจะช่วยให้สามารถสอนเสริมในส่วนของเนื้อหาวิชาหรือทักษะที่ยังขาดอยู่ได้อย่างมีประสิทธิภาพ

2.2.3 แบบทดสอบเพื่อวัดความพร้อม (Readiness Test) เป็นแบบทดสอบวัดผลสัมฤทธิ์ ซึ่งทำหน้าที่ในการวัดทักษะที่จำเป็นสำหรับการเรียนในชั้นที่สูงขึ้น แบบทดสอบที่วัดความพร้อม ใช้สำหรับทำนายการกระทำในอนาคต จึงทำหน้าที่เป็นเครื่องมือในการวัดความถนัด ไปในตัวด้วย

2.3 จำแนกตามคำตอบที่ใช้ โดยทั่วไปแล้วแบบทดสอบวัดผลสัมฤทธิ์ส่วนใหญ่ที่ใช้กันมักจะเป็นประเภทข้อเขียน และที่ใช้กันค่อนข้างมาก คือแบบทดสอบภาคปฏิบัติ (Performance

Test) เป็นแบบทดสอบที่ต้องการให้นักเรียนหรือผู้เข้าสอบได้สาธิตทักษะของเขาเอง เช่น ให้แสดงทักษะในการแก้ไขเครื่องยนต์กลไกลที่ไม่ทำงาน หรือแสดงทักษะในการเล่นดนตรี เป็นต้น

สำหรับแบบทดสอบประเภทข้อเขียนนั้น ยังแยกออกได้อย่างกว้าง ๆ อีก 2 ระดับ คือ (1) ระดับการเลือกคำตอบจากที่กำหนดไว้แล้ว (Recognition) และ (2) ระดับของการเขียนคำตอบจากความรู้หรือความทรงจำที่มีอยู่เดิม (Recall) ในแบบทดสอบระดับที่ 1 แต่ละข้อจะมีคำตอบที่ตายตัว และจะประกอบด้วยตัวเลือกหลาย ๆ ตัวที่เป็นไปได้รวมอยู่ในคำตอบที่เกี่ยวข้อง ผู้เข้าสอบจะต้องตัดสินใจเลือกคำตอบอย่างรอบคอบและถูกต้องให้สอดคล้องกับชนิดของคำถามที่ระบุไว้ ตัวอย่างของข้อสอบระดับนี้ เช่น แบบทดสอบหลายตัวเลือก (Muliple Choice) แบบทดสอบประเภทถูก – ผิด (True - False) และแบบทดสอบประเภทจับคู่ (Matching)

ส่วนแบบทดสอบระดับที่ 2 ซึ่งต้องใช้ความรู้และความทรงจำที่มีอยู่เดิมมาเขียนตอบนั้นลักษณะของคำตอบอาจจะไม่ตายตัว ขึ้นอยู่กับเหตุผลและความถูกต้องในเชิงวิชาการผสมผสานกับความคิดริเริ่มสร้างสรรค์ของผู้เข้าสอบเป็นสำคัญ แบบทดสอบระดับนี้ ได้แก่ แบบทดสอบประเภทเติมคำหรือข้อความในช่องว่าง (Completion) แบบทดสอบประเภทตอบสั้น (Short Answer) และแบบทดสอบประเภทความเรียง

3. แบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน

เป็นแบบทดสอบที่สร้างขึ้น โดยกลุ่มผู้เชี่ยวชาญมากกว่าที่จะสร้างขึ้นโดยบุคคลใดบุคคลหนึ่งเพียงบุคคลเดียวเท่านั้น ตามปกติแล้วผู้สร้างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานมักจะประกอบด้วยผู้เชี่ยวชาญทางด้านการวัดและประเมินผล รวมทั้งผู้เชี่ยวชาญในสาขาวิชานั้น ๆ ตลอดจนครูในโรงเรียนต่าง ๆ ซึ่งมีบทบาทในการกำหนดขอบข่ายเนื้อหาวิชาที่ต้องการทดสอบให้เหมาะสม แบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน ไม่จำเป็นต้องครอบคลุมเนื้อหาและทักษะที่มีในหลักสูตร เนื้อหาและทักษะของแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานส่วนมากมักจะได้จากตำราเรียนและความคิดเห็นของผู้เชี่ยวชาญทางด้านหลักสูตร เนื้อหาโดยทั่วไปจะเป็นความรู้และทักษะในระดับที่กว้าง ๆ เพื่อให้สามารถนำไปใช้กับนักเรียนโรงเรียนต่าง ๆ ได้ สำหรับขั้นตอนในการสร้างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน จะต้องมีการวางแผนสร้างอย่างเป็นระบบ คือ มีการระบุหลักเกณฑ์และเหตุผลของการสร้างแบบทดสอบ มีการกำหนดวัตถุประสงค์ของการสร้างที่ชัดเจน มีการทดลองใช้แบบทดสอบที่สร้างขึ้น เพื่อตรวจสอบความเป็นมาตรฐาน โดยการวิเคราะห์ระดับความยากง่าย และอำนาจจำแนกของข้อสอบ มีการหาค่าความตรง และความเที่ยงของแบบทดสอบ พร้อมทั้งพัฒนาตารางปกติวิสัย (Norm Table) เพื่อใช้ในการเปรียบเทียบ มีการกำหนดเวลาของการทดสอบและวิธีดำเนินการสอบ ตลอดจนมีคู่มือประกอบการใช้แบบทดสอบ

มหาวิทยาลัยบูรพา Burapha University

ซึ่งจะระบุจุดมุ่งหมายของแบบทดสอบ ประสิทธิภาพของแบบทดสอบรวมทั้งวิธีการใช้แบบทดสอบ และวิธีการตรวจหรือวิธีการให้คะแนน พร้อมทั้งตารางปกติวิสัยของกลุ่ม

4. ความแตกต่างระหว่างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐานกับแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้น สามารถจำแนกความแตกต่างที่ชัดเจนได้ 5 ประการ ดังนี้

4.1 การจำกัดเนื้อหาวิชาที่สอบ แบบทดสอบวัดผลสัมฤทธิ์มาตรฐานจะสุ่มเนื้อหาสำหรับนำมาสอบในระดับที่กว้างและทั่วไป เพื่อใช้กับโรงเรียนต่าง ๆ ตลอดจนมีการกลั่นกรองเนื้อหาในการสร้างข้อกระทงโดยผู้เชี่ยวชาญทางเนื้อหาและหลักสูตร สำหรับแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้น จะเน้นเนื้อหาเฉพาะที่เกี่ยวกับการเรียนการสอนในชั้นเรียน ครูจะทำหน้าที่เป็นผู้เชี่ยวชาญ ซึ่งอาจประกอบด้วยครูคนเดียวหรือคณะครูเป็นผู้กำหนดเนื้อหาที่เหมาะสมในการสอบ

4.2 การทดลองใช้แบบทดสอบ แบบทดสอบวัดผลสัมฤทธิ์มาตรฐานเมื่อสร้างขึ้นแล้วจะต้องมีการทดลองใช้ เพื่อทำการวิเคราะห์ประสิทธิภาพของแบบทดสอบด้วยค่าสถิติต่าง ๆ ต่อจากนั้นก็จะรายงานในคู่มือการใช้แบบทดสอบ เช่น ค่าความตรง ความเที่ยง ระดับความยากง่ายและค่าอำนาจจำแนก ในทำนองตรงกันข้าม สำหรับแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้นโดยทั่วไปจะไม่มีการทดลองใช้เพื่อวิเคราะห์ค่าสถิติในการตรวจสอบประสิทธิภาพของแบบทดสอบมาก่อน

4.3 วิธีดำเนินการสอบ แบบทดสอบวัดผลสัมฤทธิ์มาตรฐานโดยปกติจะต้องมีคู่มืออธิบายวิธีดำเนินการสอบอย่างเป็นมาตรฐาน เช่น วิธีการตอบ เวลาในการสอบ ฯลฯ ผู้ใช้แบบทดสอบต้องปฏิบัติตามอย่างเคร่งครัด สำหรับแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้นโดยทั่วไปไม่มีคู่มือประกอบการใช้ เพราะตัวครูเองเป็นผู้กำหนดมาตรฐานในการปฏิบัติเกี่ยวกับวิธีการสอบ

4.4 วิธีการให้คะแนน แบบทดสอบวัดผลสัมฤทธิ์มาตรฐานจะต้องมีคำเฉลยสำหรับการตรวจสอบให้คะแนนตามที่ระบุอยู่ในคู่มือการใช้แบบทดสอบ ส่วนแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้น ครูจะเป็นผู้ให้คะแนนตามมาตรฐานที่กำหนดขึ้นเอง

4.5 ตารางปกติวิสัยเพื่อการเปรียบเทียบ โดยปกติแล้วการสร้างแบบทดสอบวัดผลสัมฤทธิ์มาตรฐาน จะมีการนำไปใช้กับกลุ่มอ้างอิงหรือที่เรียกว่า (Norm Group) เพื่อทำตารางปกติวิสัย (Norm Table) ไว้ในคู่มือของการใช้แบบทดสอบมาตรฐาน โดยมีจุดมุ่งหมายให้ผู้ใช้แบบทดสอบสามารถนำไปใช้ดีความสำหรับคะแนนสอบที่ได้รับ รวมทั้งใช้เป็นตารางเพื่อเปรียบเทียบของคะแนนดังกล่าวด้วย ส่วนแบบทดสอบวัดผลสัมฤทธิ์ที่ครูสร้างขึ้นจะมีเพียงคะแนนของกลุ่มผู้เข้าสอบด้วยกัน ซึ่งอาจใช้เปรียบเทียบได้เฉพาะภายในกลุ่มเท่านั้น

มหาวิทยาลัยบูรพา Burapha University

แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 (สำนักทดสอบทางการศึกษา, 2546 อ้างถึงใน ศุภวัฒน์ มะลิเผือก, 2548, หน้า 53-55)

กระทรวงศึกษาธิการเห็นความสำคัญและความจำเป็นในการรักษาและยกระดับคุณภาพการศึกษาของสถานศึกษาต่าง ๆ ให้มีมาตรฐานทัดเทียมกัน จึงกำหนดให้มีการประเมินคุณภาพการศึกษาระดับชาติในทุก ๆ ปี โดยจะประเมินในปลายปีการศึกษา โดยมอบหมายให้สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการดำเนินการจัดสอบวัดผลสัมฤทธิ์ทางการเรียนของนักเรียนชั้นประถมศึกษาปีที่ 6 ให้ได้ผลการประเมินที่น่าเชื่อถือและสามารถบ่งชี้คุณภาพการศึกษาระดับชาติ ระดับเขตพื้นที่การศึกษา ระดับสถานศึกษา และระดับผู้เรียนเป็นรายบุคคลได้อย่างสมเหตุสมผล

วัตถุประสงค์องการวัดผลสัมฤทธิ์ทางการเรียน

1. เพื่อนำผลไปใช้กำกับ ดูแล ติดตามและประเมินการจัดการศึกษาในระดับประเทศสำหรับพัฒนาคุณภาพทางการศึกษา

2. เพื่อนักเรียนได้นำผลจากการสอบวัดไปพัฒนาความรู้ ความสามารถและการศึกษาต่อของนักเรียนในระดับที่สูงขึ้น

ขอบข่ายของการประเมิน

ดำเนินการประเมินนักเรียนทุกคนในสังกัดคณะกรรมการการศึกษาขั้นพื้นฐานและสังกัดอื่นที่จัดการศึกษาตามระบบ ดามหลักสูตรประถมศึกษา พุทธศักราช 2521 (ฉบับปรับปรุง 2533) ทุกจังหวัดทั่วประเทศ ได้แก่ สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน สำนักงานคณะกรรมการการศึกษาเอกชน สำนักงานคณะอนุกรรมการการอุดมศึกษา สำนักการศึกษากรุงเทพมหานคร สำนักบริหารการศึกษาท้องถิ่น กองบังคับการตำรวจตระเวณชายแดน สำนักงานตำรวจแห่งชาติ และสำนักงานพัฒนาการศึกษาและนันทนาการ

ลักษณะทั่วไปของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน

เป็นแบบทดสอบที่สร้างและพัฒนาขึ้นจนเป็นแบบทดสอบมาตรฐาน เพื่อใช้ประเมินผลการเรียนรู้ของนักเรียนที่ได้สะสมมาตลอดการเรียน ตามหลักสูตรประถมศึกษา พุทธศักราช 2521 (ฉบับปรับปรุง 2533) ของกระทรวงศึกษาธิการ

โครงสร้างของแบบทดสอบ

1. เนื้อหาสาระของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนชั้นประถมศึกษาปีที่ 6 ประกอบด้วย เนื้อหาตามหลักสูตรประถมศึกษา พุทธศักราช 2521 (ฉบับปรับปรุง 2533) วิชาที่สอบในครั้งนี้ ได้แก่ วิชาภาษาไทย คณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ

มหาวิทยาลัยบูรพา Burapha University

2. แบบทดสอบแต่ละวิชาเน้นวัดความรู้ความเข้าใจ ความคิดวิเคราะห์ และทักษะเชิงกระบวนการเฉพาะวิชา

การเตรียมการสอบ

1. แต่งตั้งคณะกรรมการคุมสอบ

ให้สำนักงานเขตพื้นที่การศึกษาแต่งตั้งคณะกรรมการคุมสอบสลับโรงเรียนภายในเขตพื้นที่การศึกษานั้น ๆ โดยในการคุมสอบแต่ละห้องต้องมีคณะกรรมการคุมสอบ 2 คน

2. การเตรียมตัวนักเรียน

ก่อนดำเนินการสอบให้ผู้บริหารโรงเรียน และครู – อาจารย์ ชี้แจงทำความเข้าใจกับนักเรียนให้ตระหนักถึงประโยชน์และความสำคัญของการสอบ ตั้งใจตอบแบบทดสอบให้เต็มความสามารถที่แท้จริง ทั้งนี้ข้อมูลที่ได้จากการประเมินจะเกิดประโยชน์ต่อนักเรียนและส่วนรวม ดังนี้

2.1 ประโยชน์ต่อนักเรียน นักเรียนจะได้ประโยชน์จากการสอบวัดผลสัมฤทธิ์ทางการเรียนของนักเรียนในระดับชาติ ทำให้เกิดความชำนาญในการคิด ได้เห็นรูปแบบของแบบทดสอบมาตรฐาน เป็นการเพิ่มพูนประสบการณ์ในการสอบ และคะแนนที่ได้จะเป็นข้อมูลประกอบในการพิจารณาเข้าศึกษาต่อของนักเรียน รวมทั้งได้รับทราบความสามารถของตนว่ามีจุดเด่นด้านใด ควรปรับปรุงและพัฒนาด้านใด เพราะในการสอบครั้งนี้เป็นการวัดความสามารถหลายด้าน คือ ภาษาไทย คณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ผลจากการสอบต้องมาบันทึกลงในเอกสารหลักฐานการศึกษา

2.2 ประโยชน์ต่อสถานศึกษา คณะครู – อาจารย์ นำผลการประเมินไปใช้เป็นข้อมูลในการแนะแนวทางการศึกษาให้กับผู้เรียน และพัฒนาการจัดการเรียนการสอนให้ได้คุณภาพมาตรฐานยิ่งขึ้น

2.3 ประโยชน์ต่อประเทศ รัฐบาลและหน่วยงานที่เกี่ยวข้องกับการจัดการศึกษานำผลการประเมินไปใช้เป็นข้อมูลในการกำหนดนโยบายทางการศึกษา และส่งเสริมสนับสนุนให้สถานศึกษาและหน่วยงานที่เกี่ยวข้อง พัฒนาคุณภาพการศึกษาให้ได้มาตรฐานทัดเทียมนานาอารยประเทศ

แนวปฏิบัติกรณีจำเป็นในการดำเนินการสอบ

กระทรวงศึกษาธิการต้องการให้นักเรียนชั้นที่ประเมิน ได้รับการประเมินทุกคนพร้อมกันทั่วประเทศ เพื่อควบคุมความเป็นมาตรฐานในการดำเนินงาน และผลการประเมินมีความตรงเป็นที่น่าเชื่อถือ แต่ในกรณีที่โรงเรียนไม่สามารถดำเนินการสอบในวันที่กำหนด อาจเนื่องมาจาก

มหาวิทยาลัยบูรพา Burapha University

เหตุจำเป็นต้องปฏิบัติ ถ้าไม่ดำเนินการจะทำให้ราชการเสียหาย หรือเหตุสุดวิสัยอื่น โรงเรียนอาจเลื่อนการสอบไปได้ โดยเสนอประธานคณะกรรมการดำเนินงานระดับเขตพื้นที่การศึกษาพิจารณาแล้วแจ้งสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานทราบ

การดำเนินการสอบให้เป็นไปตามระเบียบการสอบของกระทรวงศึกษาธิการอย่างเคร่งครัด กรณีที่นักเรียนขาดสอบให้โรงเรียนพิจารณาถึงสาเหตุและความจำเป็น แล้วจึงพิจารณาให้นักเรียนเข้าสอบชดเชย โดยเก็บรักษาข้อสอบเป็นความลับ และดำเนินการสอบตามระเบียบของกระทรวงศึกษาธิการ

การดำเนินการหลังการสอบ

การดำเนินการหลังการสอบ แบ่งเป็น 3 ส่วน ดังนี้

1. เมื่อกรรมการดำเนินการสอบเก็บและตรวจความเรียบร้อยของกระดาษคำตอบแล้ว ให้นำส่งคณะกรรมการดำเนินการสอบระดับโรงเรียน จากนั้นโรงเรียนรวบรวมกระดาษคำตอบทั้งหมดส่งคณะกรรมการที่สำนักงานเขตพื้นที่การศึกษากำหนดไว้ สำหรับส่วนกลางให้รวบรวมกระดาษคำตอบส่งที่สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กรุงเทพมหานคร

2. โรงเรียนรวบรวมแบบทดสอบทั้งหมดส่งคืนคณะกรรมการที่สำนักงานเขตพื้นที่การศึกษาแต่งตั้ง เพื่อดำเนินการย่อย/เผาทำลายต่อไป สำหรับส่วนกลาง นำส่งที่โรงพิมพ์คุรุสภาลาดพร้าว

3. คณะกรรมการดำเนินงานระดับเขตพื้นที่การศึกษา รวบรวมกระดาษคำตอบทั้งหมดส่งสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการทันที เนื่องจากสำนักทดสอบทางการศึกษาจะต้องตรวจกระดาษคำตอบให้ทั้งหมด ซึ่งมีปริมาณมากให้เสร็จสิ้นโดยเร็ว เพื่อให้โรงเรียนได้ทำบันทึกลงในเอกสารหลักฐานการศึกษาได้ทันก่อนปิดภาคเรียน

ตอนที่ 5 ค่าความเที่ยงของแบบทดสอบ

ความเที่ยงของแบบทดสอบ หมายถึง ความคงที่ของคะแนนที่ได้จากการสอบนักเรียนคนเดียวกันหลายครั้งในแบบทดสอบชุดเดิม ถ้าคะแนนเปลี่ยนไปจากเดิม แสดงว่าแบบทดสอบขาดความเที่ยง ทำให้ผลการสอบมีความคลาดเคลื่อนไปจากความรู้จริงของนักเรียน ความคลาดเคลื่อนชนิดนี้ เรียกว่า ความคลาดเคลื่อนในการวัด (Error of Measurement) ในการวัดผลนั้นจะต้องสร้างเครื่องมือที่ต้องการนำไปวัดผลให้มีคุณภาพที่เชื่อมั่นได้ เพื่อผลการวัดที่ออกมาจะได้เป็นคะแนนความรู้จริงของนักเรียนที่ปราศจากความคลาดเคลื่อนในการวัด และค่าเที่ยงจะมีค่าอยู่

มหาวิทยาลัยบูรพา Burapha University

ระหว่าง –1 ถึง +1 และจะพิจารณาเฉพาะค่าที่เป็นบวกเท่านั้น ซึ่งควรมีค่ามากกว่า 0.70 จึงจะเป็นแบบทดสอบที่มีความเชื่อถือได้

การคำนวณหาค่าความเที่ยงนั้นจะคำนวณในรูปของการประมาณค่า ในรูปของค่าสัมประสิทธ์ของความเที่ยง มักใช้สัญลักษณ์ว่า r_{tt}, r_{xx} หรือ r_{cc} โดยทั่วไปแล้วความเที่ยงแบ่งออกเป็น 2 ประเภท คือ ความเที่ยงของแบบทดสอบอิงกลุ่ม (Reliability of Norm-Referenced Test) กับความเที่ยงของแบบทดสอบอิงเกณฑ์ (Reliability of Criterion-Referenced Test) และในแต่ละประเภทของการคำนวณหาค่าความเที่ยงนั้น จะมีสูตรที่ใช้ประมาณค่าความเที่ยงอยู่หลายสูตร ซึ่งวิธีคำนวณจำแนกตามลักษณะของแบบทดสอบอิงกลุ่มและอิงเกณฑ์ สำหรับค่าความเที่ยงที่เกี่ยวข้องกับงานวิจัยนี้ เป็นค่าความเที่ยงที่คำนวณโดยใช้ความสอดคล้องภายใน (Internal Consistency Reliability) เป็นการหาความเที่ยงที่ใช้แบบทดสอบฉบับเดียวทำการทดสอบเพียงครั้งเดียว วิธีหาค่าความเที่ยงที่เกี่ยวข้องกับงานวิจัยนี้ได้แก่ วิธีของครอนบัค (Cronbach Alpha Procedure)

ครอนบัคได้พัฒนาสูตรหาความเที่ยงในรูปสัมประสิทธ์แอลฟา (α-Coefficient) ในปี ค.ศ. 1951 โดยพัฒนามาจากสูตร KR-20 ทั้งนี้เป็นเพราะว่าจะได้ใช้หาความเที่ยงกับเครื่องมือที่ไม่ได้ตรวจให้คะแนนเป็น 1 กับ 0 จะตรวจให้คะแนนลักษณะใดก็ได้ เช่น ถ้าทำถูกได้คะแนนเป็น 10 , 8 หรือในลักษณะแบบสอบถามที่ให้คะแนนแต่ละข้อเป็น 3 , 2 , 1 หรือ 5, 4 , 3, 2 , 1 ก็ได้ สูตรที่ใช้คือ

$$\alpha = \frac{k}{k-1}\left\{1-\frac{\sum \sigma_i^2}{\sigma^2}\right\}$$

เมื่อ α คือ สัมประสิทธ์ความเที่ยง

k คือ จำนวนข้อสอบ

σ_i^2 คือ คะแนนความแปรปรวนเป็นรายข้อ

σ^2 คือ คะแนนความแปรปรวนของทั้งฉบับ

โดยที่

$$\sigma_i^2 = \frac{N\sum X_i^2 - \left(\sum X_i\right)^2}{N^2}$$

เมื่อ $\sum X_i^2$ คือ ผลรวมกำลังสองของคะแนนในข้อที่ i

$\left(\sum X_i\right)^2$ คือ ผลรวมทั้งหมดของคะแนนในข้อที่ i ยกกำลังสอง

มหาวิทยาลัยบูรพา Burapha University

N คือ จำนวนคนเข้าสอบ

และ

$$\sigma^2 = \frac{N\sum X^2 - \left(\sum X\right)^2}{N^2}$$

เมื่อ $\sum X^2$ คือ ผลรวมกำลังสองของคะแนนของข้อสอบทั้งฉบับ

$\left(\sum X\right)^2$ คือ ผลรวมทั้งหมดของคะแนนของข้อสอบทั้งฉบับยกกำลังสอง

มหาวิทยาลัยบูรพา Burapha University