บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดผลทางการศึกษาและจิตวิทยาส่วนมากเป็นการวัดความสามารถหรือคุณลักษณะภายใน เช่น ความสามารถในการคำนวณเลข เจตคติต่อวิชาคณิตศาสตร์ ซึ่งสิ่งเหล่านี้ไม่สามารถสังเกตหรือวัดได้โดยตรงเหมือนกับการวัดทางกายภาพ ต้องอาศัยจากการวัดทางอ้อม โดยการใช้ข้อคำถามหรือข้อความในแบบทดสอบเป็นสิ่งเร้ากระตุ้นให้บุคคลแสดงพฤติกรรมออกมาภายนอก สนองต่อข้อคำถาม หรือข้อความเหล่านั้น แล้วจึงอนุมานผลการตอบจากคะแนนที่ได้สรุปอ้างอิงกลับไปอธิบายคุณลักษณะที่ต้องการวัด (สงบ ลักษณะ, 2525) เครื่องมือวัดผลทางการศึกษาที่สำคัญประเภทหนึ่งคือมาตรประมาณค่า

มาตรประมาณค่า ประกอบด้วย ข้อคำถามหรือสิ่งเร้ากับตัวเลือกที่ให้ผู้ตอบตอบสนองเพื่อที่จะใช้ประมาณค่าคุณลักษณะสิ่งใดสิ่งหนึ่งของบุคคลหรือสิ่งของ (Wiersma and Jurs, 1990) โดยมีลักษณะเป็นเส้นตรงที่มีตัวเลขหรือตัวอักษรกำกับ อาจแบ่งเป็นช่วงได้ตั้งแต่ 2 ช่วงขึ้นไป เป็นเลขคู่หรือเลขคี่ก็ได้ มีความหมายทั้งทางบวกและทางลบอยู่ในมาตรเดียวกันก็ได้ หรือจะมีเฉพาะทางบวกแยกจากลบก็ได้ มาตรประมาณค่ามีลักษณะเด่นคือ เป็นการกำหนดลักษณะเฉพาะ เดี่ยว ทีละลักษณะเพื่อทำการประมาณค่า (อุทุมพร จามรมาน, 2537) มาตรประมาณค่าที่นิยมใช้กันในการวิจัยการศึกษาสามารถจำแนกตามวิธีสร้างได้ 4 แบบ คือ

1. มาตรประมาณค่าตามวิธีการของเทอร์สโตน
2. มาตรประมาณค่าตามวิธีการของออสกูด
3. มาตรประมาณค่าตามวิธีการของกัทแมน
4. มาตรประมาณค่าตามวิธีการของลิเคิร์ท (Guilford, 1954; Anderson, 1983)

วิธีที่นิยมใช้ในทางปฏิบัติคือ มาตรประมาณค่าแบบลิเคิร์ท เพราะเป็นมาตรวัดที่สร้างขึ้นด้วยวิธีการที่ไม่ยุ่งยากซับซ้อนและยังสามารถนำไปปรับใช้กับการวัดคุณลักษณะจิตพิสัยด้านอื่นๆ ได้เป็นอย่างดีอีกด้วย (สุชาติ ประสิทธิ์รัฐสินธุ์, 2524; Koch, 1983)

โดยทั่วไปสามารถแบ่งมาตรประมาณค่าได้เป็น 5 ประเภท (Guilford, 1954) คือ มาตรประมาณค่าที่กำหนดตัวเลข (numerical rating scale) เป็นมาตรที่ระบุตัวเลขให้กับการตอบ มาตรประมาณค่าแบบกราฟ (graphic rating scale) หรือแบบกำหนดเส้น มีผู้ใช้มาตรประเภทนี้กันมาก มักใช้สำหรับการประเมินตนเอง มาตรประมาณค่าแบบสเกลมาตรฐาน (standard scale) มาตรประมาณค่าประเภทนี้กำหนดเกณฑ์มาตรฐานไว้เปรียบเทียบ มาตรประมาณค่าแบบแต้มสะสม (cumulated points scale) เป็นมาตรประมาณค่าที่ผู้ตอบเป็นผู้เลือกคุณลักษณะจากตัวเลือกหลาย ๆ ตัว พิจารณาว่าตัวเลือกใดเหมาะสมกับคุณลักษณะที่ต้องการจะวัด และมาตรประมาณค่าแบบตัวเลือกบังคับตอบ (forced-choice rating) เป็นมาตรประมาณค่าที่กำหนดตัวเลือกและค่าคะแนนให้กับตัวเลือก โดยผู้ตอบจะต้องเลือกตัวใดตัวหนึ่งโดยตัวเลือกนั้นอาจใช้ข้อความทางบวกหรือทางลบก็ได้ เครื่องมือประเภทนี้มีการดัดแปลงและนำมาใช้ในการวัดคุณลักษณะอย่างแพร่หลายในปัจจุบัน

มาตรประมาณค่าโดยทั่วไปมีวิธีการตรวจให้คะแนน 2 วิธี คือ การตรวจให้คะแนนแบบทวิวิภาค (dichotomous) และการตรวจให้คะแนนแบบพหุวิภาค (polytomous) ตัวอย่างเช่น

(ข้อ 0) ถ้านักเรียนลืมเอาสีมาในชั่วโมงศิลปะ แล้วนักเรียนจะเลือกทำอย่างไร

ก. ขอยืมเพื่อนแล้วหยิบสีมาใช้เลย

ข. หยิบสีของเพื่อนสนิทมาใช้แล้วค่อยบอก

ค. บอกคุณครูว่าลืมเอาสีมาขอทำเป็นการบ้าน

ง. บอกขอยืมแล้วรอให้เพื่อนอนุญาตจึงค่อยนำสีมาใช้

(แบบวัดความมีระเบียบวินัยสำหรับนักเรียนชั้นประถมศึกษาปีที่ 6 ของ สุพัตรา เทียนอุดม, 2536)

การตรวจให้คะแนนวิธีแรกเป็นวิธีการตรวจให้คะแนนแบบทวิวิภาค (dichotomous) โดยตรวจให้คะแนนตัวเลือกที่ถูกต้องเหมาะสมที่สุดเป็น 1 คะแนน และให้คะแนนตัวเลือกอื่น ๆ ที่มีความถูกต้องเหมาะสมน้อยกว่าเป็น 0 คะแนน การตรวจให้คะแนนวิธีนี้พบมากในงานวิจัยที่ผ่านมา (บุษรินทร์ บุญรอด, 2536; วารี นิยมธรรม, 2536) สำหรับการตรวจให้คะแนนวิธีที่สอง เป็นการตรวจให้คะแนนแบบพหุวิภาค (polytomous) โดยตรวจให้คะแนนกับตัวเลือกทุกตัวตามระดับของคุณลักษณะและความเหมาะสม โดยในตัวอย่างนี้กำหนดให้เป็น 2, 1, 3, และ 4 ตามลำดับ

การตรวจให้คะแนนวิธีนี้พบว่านิยมใช้ในทางวัดผลการศึกษาเช่นกัน (นิภาพรรณ แก่นคง, 2531; วลัยรัตน์ องค์ศิริมงคล, 2533; วิมลรัตน์ สิริอาภรณ์, 2536; แหวนไพลิน เย็นสุข, 2538)

การศึกษาเชิงเปรียบเทียบวิธีการให้คะแนนทั้ง 2 วิธีดังกล่าวที่ผ่านมาส่วนใหญ่เป็นการศึกษากับแบบสอบผลสัมฤทธิ์ (Donoghue, 1994; Zin and Williams, 1991; Smith, 1987) ผลการวิจัยพบว่าในการวัดความสามารถของนักเรียน การตรวจให้คะแนนแบบพหุวิภาคมีความแม่นตรงในการวัดความสามารถของนักเรียนมากกว่าการตรวจให้คะแนนแบบทวิวิภาค มูราคิ (Muraki, 1993) ได้ให้ข้อสังเกตว่าการเพิ่มจำนวนลำดับขั้นของคะแนน (category) ในแต่ละข้อกระทง หรือในการตรวจให้คะแนนแบบพหุวิภาคไม่จำเป็นที่จะต้องทำให้ข้อกระทงมีค่าฟังก์ชันสารสนเทศสูงเสมอไป แต่การตรวจให้คะแนนแบบพหุวิภาคจะให้พิสัยของการวัด (range of θ scale) กว้างกว่าการตรวจให้คะแนนแบบทวิวิภาค การวัดที่ตรงและเที่ยงที่สุดนั้นนอกจากจะต้องพิจารณาถึงวิธีการตรวจให้คะแนนที่เหมาะสมแล้ว การวิเคราะห์เพื่อประมาณค่าความสามารถของผู้ตอบและการวิเคราะห์คุณภาพของข้อกระทงก็เป็นกระบวนการหนึ่งที่จะทำให้สารสนเทศที่เกี่ยวกับบุคคลและข้อกระทงมีความถูกต้องและแม่นตรง

การวิเคราะห์คุณภาพของแบบวัดชนิดมาตรประมาณค่าที่ผ่านมาส่วนมากยังเป็นทฤษฎีแบบดั้งเดิม (classical test thory : CTT) (อุทุมพร (ทองอุไทย) จามรมาน, 2532) แต่การวิเคราะห์คุณภาพแบบวัดตามทฤษฎีนี้มีข้อจำกัดด้านค่าสถิติจากการวิเคราะห์ที่แปรเปลี่ยนไปตามกลุ่มตัวอย่างที่ใช้ เช่น กลุ่มตัวอย่างที่มีลักษณะแตกต่างกันมาก (heterogeneous) ค่าอำนาจจำแนกมีแนวโน้มที่จะสูงกว่ากลุ่มตัวอย่างที่มีลักษณะแตกต่างกันน้อย (homogeneous) และกลุ่มตัวอย่างที่มีระดับความสามารถต่ำก็จะทำให้ข้อกระทงมีแนวโน้มที่จะให้ค่าความยากต่ำ (ข้อกระทงไม่ง่าย) กลุ่มตัวอย่างที่มีระดับความสามารถสูงก็จะทำให้ข้อกระทงนั้นมีแนวโน้มที่จะให้ค่าความยากสูง (ข้อกระทงมีความง่ายมาก)

นอกจากนี้ ไรท์ และมาสเตอร์ (Wright and Masters, 1982) ได้เสนอว่าการแปลความหมายจากคะแนนซึ่งอยู่ในรูปของคะแนนดิบไม่เหมาะสมเนื่องจากไม่สามารถสรุปได้ว่าระดับคะแนนที่ได้อยู่ในระดับอันตรภาคชั้น (interval) บนมาตรเชิงเส้น (linear scale) การแปลความหมายจากคะแนนรวมดังกล่าวแปรเปลี่ยนไปตามจำนวนข้อคำถามและจำนวนผู้ตอบ โดยพบว่าเมื่อมีการลดจำนวนข้อความจะทำให้คะแนนรวมของผู้ตอบต่ำลง และถ้าลดจำนวนผู้ตอบในกลุ่ม

ตัวอย่างที่ทำการวิเคราะห์ลงจะทำให้คะแนนของข้อความต่ำลง จึงยากที่จะอธิบายคุณลักษณะที่ต้องการได้ชัดเจน

ต่อมาได้มีการพัฒนาทฤษฎีการตอบสนองข้อสอบ (item response theory : IRT) เพื่อแก้ปัญหาการวิเคราะห์คุณภาพแบบวัดตามแนวทฤษฎีการทดสอบแบบดั้งเดิม การพัฒนาในระยะแรกมุ่งนำไปใช้วิเคราะห์กับแบบสอบที่ตรวจให้คะแนนแบบทวิวิภาค ซึ่งผลการวิเคราะห์ให้ค่าพารามิเตอร์ของผู้สอบและข้อสอบที่มีค่าคงที่มากกว่าผลการวิเคราะห์ที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิม (Hambleton, Swaminathan and Roger, 1991) ในระยะต่อมาได้มีการประยุกต์ใช้กับแบบวัดเจตคติและแบบสอบผลสัมฤทธิ์ที่มีวิธีการตรวจให้คะแนนแบบพหุวิภาค (polytomous) ผลการพัฒนาโมเดลจากทฤษฎีการตอบสนองข้อสอบให้ใช้กับแบบวัดและแบบสอบที่มีวิธีการตรวจให้คะแนนแบบพหุวิภาคทั้งหมดเรียกว่า Polytomous Item Response Models ซึ่งมีอยู่หลายโมเดล (Muraki, 1993; Donoghue, 1994) มีโมเดลที่สำคัญ (ธนวัฒน์ แสนสุข, 2539) ได้แก่ Graded Response Model (GRM) พัฒนาโดย ซาเมจิมา (Samejima) ในปี ค.ศ. 1960, Nominal Response Model (NRM) พัฒนาโดย บอค (Bock) ในปี ค.ศ. 1972, Rating Scale Model (RSM) พัฒนาโดย แอนดริช (Andrich) ในปี ค.ศ. 1978, Partial Cradit Model (PCM) พัฒนาโดย มาสเตอร์ (Masters) ในปี ค.ศ. 1982, Successive Interval Model (SIM) พัฒนาโดย รอส์ท (Rost) ในปี ค.ศ. 1988, และ Generalized Partial Cradit Model (GPCM) พัฒนามาจาก PCM โดย มูราคิ (Muraki) ในปี ค.ศ. 1992

โมเดลส่วนใหญ่พัฒนามาจากโมเดลราส์ช (rasch model) และโมเดลโลจิสติก (logistic model) การพัฒนาในระยะแรกมีจุดประสงค์เพื่อนำไปใช้กับแบบสอบและแบบวัดบางชนิดโดยเฉพาะ เช่น RSM ใช้วิเคราะห์กับมาตรวัดแบบลิเคิร์ท (Likert scale) แนวทฤษฎีของของโมเดลนี้จึงเข้มงวดเกี่ยวกับข้อตกลงเบื้องต้นเรื่องความเป็นเอกมิติของแบบวัดและคะแนนต้องมีระดับการวัดเป็นอันตรภาค (interval) แต่ถ้าคะแนนเป็นแบบจัดลำดับ (category) ใช้ NRM จะเหมาะสมกว่า (Muraki, 1990) โมเดลที่ได้รับการพัฒนาและใช้กันอย่างแพร่หลายในปัจจุบันคือ GRM และ GPCM (Donoghue, 1994; De Ayala, 1994; Murski, 1992, 1993; Dodd and De Ayala, 1989; Koch, 1983) เนื่องจากโมเดลดังกล่าวไม่เข้มงวดกับข้อตกลงเบื้องต้นของความเป็นเอกมิติของแบบวัดและการประมาณค่าพารามิเตอร์มีค่าอำนาจจำแนกรายข้อในฟังก์ชันด้วย ซึ่งโมเดลอื่นจะไม่มีหรือกำหนดให้คงที่ในฟังก์ชัน

ทั้ง GRM และ GPCM จะอยู่ในรูปฟังก์ชันทางคณิตศาสตร์ โดยปรับความสามารถของบุคคลให้อยู่ในรูปคะแนนมาตรฐานมีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 โมเดลดังกล่าวเป็นฟังก์ชันความสัมพันธ์ระหว่างผลการตอบกับความสามารถหรือคุณลักษณะของผู้ตอบในรูปโค้งฟังก์ชันสารสนเทศ (information function) และรวมค่าพารามิเตอร์คุณลักษณะข้อสอบหลายลักษณะ เช่น ค่าความยาก, ค่าอำนาจจำแนก เป็นต้น เมื่อพิจารณาถึงแนวคิดและลักษณะทั่วไปของฟังก์ชันแล้วจะเห็นว่า โมเดลทั้งสองสามารถวิเคราะห์ได้ครอบคลุมถึง Logistic Dichotomous Response Model (Muraki, 1993) การวิเคราะห์โมเดลทั้ง 2 จะให้สารสนเทศที่สำคัญ คือ ฟังก์ชันสารสนเทศของแบบวัด (test information function : TIF) ฟังก์ชันสารสนเทศของข้อกระทง (item information function : IIF) และฟังก์ชันสารสนเทศของลำดับขั้นคะแนน (item-category information function : ICF) ลักษณะทั่วไปของ GRM, GPCM และ โมเดลโลจิสติก สรุปได้ดังตารางที่ 1 (Donoghue, 1994; De Ayala, 1994; Muraki, 1992, 1993; Dodd and De Ayala, 1989; Koch, 1983; ศิริชัย กาญจนวาสี, 2538; ธนวัฒน์ แสนสุข, 2539)

ตารางที่ 1 ลักษณะทั่วไปของ GRM, GPCM และ โมเดลโลจิสติก

ลักษณะทั่วไป	GRM	GPCM	Logistic Model
1. ลักษณะทั่วไปของฟังก์ชัน	$P_{jk}(\theta)=\frac{\exp^{\left[Da_j(\theta-b_{jk})\right]}}{1+\exp^{\left[Da_j(\theta-b_{jk})\right]}}$	$P_{jk/k-1,k}(\theta)=\frac{\exp^{\left[Da_j(\theta-b_{jk})\right]}}{1+\exp^{\left[Da_j(\theta-b_{jk})\right]}}$	$P_i(\theta)=C_i+\frac{(1-C_i)}{1+e^{-Da_i(\theta-b_i)}}$
2. ฟังก์ชันสารสนเทศของแบบวัด (TIF)	$I_j(\theta)=\sum_{k=0}^{k}\frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)}$	$I(\theta)=\sum_{j=1}^{m}I_j(\theta)$	$I(\theta)=\sum_{j=1}^{m}I_j(\theta)$
3. ฟังก์ชันสารสนเทศของข้อกระทง (IIF)	$I_j(\theta)=\sum_{k=0}^{c}\frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)}$	$I_j=D^2a_j^2\sum_{k=1}^{m}\left[T_k-\overline{T}_j(\theta)\right]^2P_{jk}(\theta)$ เมื่อ $\overline{T}_j(\theta)=\Sigma T_kP_{jk}(\theta)$	$I_i(\theta)=\frac{(P_i)^2}{P_iQ_i}$
4. ฟังก์ชันสารสนเทศของลำดับขั้นคะแนน (ICFs)	$I_{jk}(\theta)=\frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)}$	$I_{jk}(\theta)=P_{jk}(\theta)I_j(\theta)$	.....
5. การตรวจให้คะแนน	แบบพหุวิภาค	แบบพหุวิภาค	แบบทวิวิภาค
6. ความเป็นเอกมิติของแบบวัด	ไม่เข้มงวด	ไม่เข้มงวด	เข้มงวด
7. ค่าพารามิเตอร์เทรชโฮลด์ในแต่ละลำดับขั้นเดียวกันของแต่ละข้อ	แตกต่างกันได้	แตกต่างกันได้	.....
8. ค่าพารามิเตอร์ของอำนาจจำแนก	แตกต่างกันได้ (การนำไปใช้มักกำหนดให้มีค่าคงที่)	แตกต่างกันได้	.....
9. ลักษณะการเรียงลำดับขั้นคะแนนแต่ละข้อ	เรียงตามลำดับค่าความยาก	เรียงตามลำดับขั้นของความสำเร็จ	.....
10. ความเป็นอันตรภาคของมาตรประมาณค่า	ไม่จำเป็น	ไม่จำเป็น	.....
11. ความต่อเนื่องของลำดับขั้นคะแนน	ต่อเนื่องกัน	ต่อเนื่องกัน	.....
12. พัฒนาการของโมเดล	พัฒนาจาก Logistic Model (แบบ 2 พารามิเตอร์)	พัฒนาจาก Rasch Model (ของ PCM) (1 พารามิเตอร์)	เป็น Logistic Model (1, 2, 3 พารามิเตอร์)
13. โปรแกรมที่ใช้วิเคราะห์	MULTILOG (Thissen)	PARSCALE (Muraki and Bock)	MULTILOG (Thissen)

(ปรับปรุงจาก ธนวัฒน์ แสนสุข, 2539)

การศึกษาเกี่ยวกับการประยุกต์ใช้โมเดลการตรวจให้คะแนนแบบพหุวิภาค GRM และ GPCM จำแนกได้เป็น 4 ประเด็น (ธนวัฒน์ แสนสุข, 2539) คือ

ประเด็นแรกเป็นการนำไปใช้วิเคราะห์แบบสอบหรือแบบวัดเพื่อศึกษาค่าฟังก์ชันสารสนเทศและปรับระดับคะแนนให้เหมาะสม (Muraki, 1992; Reise and Yu, 1994)

ประเด็นที่สองเป็นการประยุกต์โมเดลใช้กับแบบสอบปรับเหมาะในคอมพิวเตอร์ (computerized adaptive testing : CAT) (Dodd, Koch and De Ayala, 1989; Koch, 1983)

ประเด็นที่สามเป็นการศึกษาเชิงเปรียบเทียบค่าฟังก์ชันสารสนเทศของแบบสอบที่ตรวจให้คะแนนแบบทวิวิภาคกับแบบพหุวิภาค ผลการศึกษาส่วนใหญ่ให้ข้อสรุปว่าการตรวจให้คะแนนแบบพหุวิภาคให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค (Donoghue, 1994) ส่วนผลที่ขัดแย้งคือ การศึกษาของ ยามาโมโต กับ คูลิก (Yamamoto and Kulick, 1992 cited by Donoghue, 1994) พบว่า การตรวจให้คะแนนแบบทวิวิภาคให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบพหุวิภาค แต่โดนัฟ (Donoghue, 1994) ให้ข้อสังเกตว่าการศึกษาของยามาโมโต และคูลิก มีวิธีการตรวจให้คะแนนที่ไม่เหมาะสม เนื่องจากใช้แบบสอบที่มีวิธีการตรวจให้คะแนนแบบทวิวิภาคมาปรับเป็นการตรวจให้คะแนนแบบพหุวิภาค

ประเด็นสุดท้ายเป็นการศึกษาเชิงเปรียบเทียบประสิทธิภาพของโมเดลในการนำไปประยุกต์ใช้ มีการวิจัยน้อยมาก เจนเซน และรอสคัม (Jansen and Roskam, 1968 cited by Muraki, 1993) ได้ศึกษาเปรียบเทียบ PCM กับ GRM และสรุปได้ว่า GRM มีความเหมาะสมกับข้อมูลที่มีลักษณะเป็นมาตรประมาณค่ามากกว่า เนื่องจากการวิเคราะห์ตาม PCM ค่าพารามิเตอร์ของผู้สอบจะไม่คงที่หลังจากที่มีการปรับเปลี่ยนลำดับขั้นของคะแนน มูราคิ (Muraki, 1993) ให้ข้อเสนอแนะว่าสามารถแก้ไขความคลาดเคลื่อนดังกล่าวได้ โดยให้มีการปรับลำดับค่าคะแนนเพียงข้อเดียวในการวัดแต่ละครั้ง และใช้วิธีการประมาณค่าแบบมาร์จินอลแมกซิมัมไลลิฮูด (MML) ซึ่งวิธีดังกล่าวพัฒนามาเป็น GPCM นั่นเอง

การศึกษาเปรียบเทียบความเหมาะสมของวิธีการตรวจให้คะแนนแบบวัดชนิดมาตรประมาณค่าในงานวิจัยของ ธนวัฒน์ แสนสุข (2539) ซึ่งได้ทำการศึกษาเปรียบเทียบ GRM , GPCM และโมเดลโลจิสติก ในการเปรียบเทียบฟังก์ชันสารสนเทศของแบบวัดที่มีวิธีการให้คะแนนต่างกัน พบว่าในการนำไปใช้กับแบบวัดคุณลักษณะซึ่งเป็นมาตรประมาณค่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM มีประสิทธิภาพสัมพัทธ์สูงกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตาม โมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ แต่การตรวจให้

คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GPCM มีประสิทธิภาพสัมพัทธ์ต่ำกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ ซึ่งผลที่ได้ไม่เป็นไปตามสมมุติฐานที่ ธนวัฒน์ แสนสุข (2539) ตั้งไว้

เนื่องจากการวิจัยของ ธนวัฒน์ แสนสุข (2539) ใช้ข้อมูลทุติยภูมิภายใต้แบบวัดคุณลักษณะที่มีลักษณะเป็นมาตรประมาณค่า ที่มีวิธีการตรวจให้คะแนนแบบพหุวิภาค(1, 2, 3, 4) เมื่อ นำวิธีดังกล่าวมาเปรียบเทียบกับการให้คะแนนแบบทวิวิภาค โดยปรับให้คะแนนในลำดับขั้นที่ 4 เป็น 1 คะแนน และลำดับขั้นอื่นๆเป็น 0 คะแนน การวิจัยดังกล่าวมีประเด็นที่น่าสงสัยว่าการปรับยุบคะแนนเป็น 1, 0 นั้นเหมาะสมหรือไม่ คุณภาพของตัวเลือกที่ถูกปรับคะแนนเป็น 0 มีความ เหมาะสมเพียงใด ผลการวิจัยของ ธนวัฒน์ แสนสุขนี้ยังไม่สามารถให้ข้อชี้แจงในประเด็นนี้ได้ จึง เป็นเรื่องที่น่าสนใจที่จะทำการศึกษาในประเด็นนี้ซ้ำ แต่เก็บข้อมูลจริง (ข้อมูลปฐมภูมิ)

จากความเป็นมาดังกล่าวข้างต้น ผู้วิจัยจึงทำการศึกษาซ้ำในประเด็นการศึกษาวิจัยเดียวกันกับของ ธนวัฒน์ แสนสุข (2539) โดยใช้มาตรประมาณค่าที่มีการสร้างให้มีการตรวจให้คะแนนแบบทวิวิภาค (0, 1) จริง ๆ ไม่ใช่เกิดจากการปรับยุบคะแนนมาจากมาตรประมาณค่าที่มีวิธีการตรวจให้คะแนนแบบพหุวิภาค ใช้มาตรประมาณค่าที่มีลักษณะคำตอบที่ชัดเจนเหมาะสมกับลำดับขั้นของคะแนนในการตรวจให้คะแนนแบบพหุวิภาค ใช้การเก็บข้อมูลจริง จากนั้นทำการเปรียบเทียบผลการวิจัยที่ได้

วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบฟังก์ชันสารสนเทศของมาตรประมาณค่าระหว่างการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคเมื่อวิเคราะห์ตาม โมเดลจิสติก, GRM และ GPCM

2. เพื่อเปรียบเทียบอัตราส่วนสารสนเทศเฉลี่ย (ratio of average information) ของมาตรประมาณค่าระหว่างการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคเมื่อวิเคราะห์ตามโมเดลโลจิสติก, GRM และ GPCM

3. เพื่อตรวจสอบความสอดคล้องของฟังก์ชันสารสนเทศระหว่างมาตรประมาณค่าที่ให้คะแนนแบบทวิวิภาคและแบบพหุวิภาค เมื่อวิเคราะห์ตาม โมเดลโลจิสติก, GRM และ GPCM

สมมุติฐานของการวิจัย

จากการศึกษาเอกสารและงานวิจัยที่ผ่านมาพบว่า ในแบบสอบที่ตรวจให้คะแนนแบบพหุวิภาคส่วนใหญ่จะให้ความแม่นยำ (precision) ในการประมาณค่าพารามิเตอร์บุคคล และข้อกระทงมากกว่าการตรวจให้คะแนนแบบทวิวิภาค (Samejima, 1976;Thissen, 1976; Muraki, 1993; Donoghue, 1994, ธนวัฒน์ แสนสุข, 2539) มีเพียงผลการศึกษาของ ยามาโมโต และคูลิก (Yamamoto and Kulick) ที่ให้ผลขัดแย้งกับผู้อื่น โดนัฟ (Donoghue, 1994) ได้อธิบายว่าที่เป็นเช่นนี้เพราะเกิดจากความไม่เหมาะสมของการใช้แบบสอบที่ตรวจให้คะแนนแบบทวิวิภาคมาปรับให้เป็นแบบสอบที่ตรวจให้คะแนนแบบพหุวิภาค ส่วนในมาตรประมาณค่ายังไม่มีการศึกษามากนัก

กล่าวโดยสรุปว่าทั้งในแบบสอบและแบบวัดที่เป็นมาตรประมาณค่า การตรวจให้คะแนนแบบพหุวิภาคน่าจะให้คุณภาพด้านความตรง ความเที่ยงสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค แต่ในงานของ ธนวัฒน์ แสนสุข (2539) ในการศึกษาเปรียบเทียบการตรวจให้คะแนนแบบพหุวิภาคโดย GRM และ GPCM กับการตรวจให้คะแนนแบบทวิวิภาคโดย โมเดลโลจิสติก ในแบบวัดคุณลักษณะ ผลปรากฏว่า GRM ให้ค่าฟังก์ชันสารสนเทศสูงที่สุด รองลงมาคือ โมเดลโลจิสติก และ GPCM ให้ค่าฟังก์ชันสารสนเทศต่ำที่สุด ซึ่งผลการวิจัยที่เกิดขึ้นนี้อธิบายได้ว่าอาจเกิดจากข้อจำกัดของการวิจัยคือ ใช้แบบวัดคุณลักษณะที่มีลักษณะเป็นมาตรประมาณค่าที่มีการตรวจให้คะแนนเป็นแบบพหุวิภาค (1, 2, 3, 4) อยู่แล้วมาปรับให้เป็นการตรวจให้คะแนนแบบทวิวิภาค (0, 1) นอกจากนี้ยังมีข้อโต้แย้งว่าการให้คะแนนในแต่ละลำดับขั้นยังไม่เหมาะสม

แต่ในงานวิจัยนี้ใช้มาตรประมาณค่าที่มีลักษณะคำตอบที่ชัดเจนเหมาะสมกับลำดับขั้นของคะแนนคือ มาตรวัดเจตคติที่สร้างตามวิธีการของลิเคอร์ท และมาตรประมาณค่าแบบตัวเลือกบังคับตอบที่มีวิธีการกำหนดให้คะแนนตัวเลือกที่เหมาะสม ใช้มาตรประมาณค่าที่มีวิธีการตรวจให้คะแนนแบบทวิวิภาค (0, 1) จริง ๆ ไม่ได้เกิดจากการปรับยุบคะแนนจากมาตรประมาณค่าที่มีการตรวจให้คะแนนแบบพหุวิภาค และใช้การเก็บข้อมูลจริง

ดังนั้นจึงตั้งสมมุติฐานในการวิจัยนี้ว่า

1. ผลการวิเคราะห์การตรวจให้คะแนนในมาตรประมาณค่าเมื่อวิเคราะห์ตาม GRM GPCM และโมเดลโลจิสติก จะให้ผลที่สอดคล้องกันคือ การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM และ GPCM จะให้ค่าฟังก์ชันสารสนเทศ (TIF) สูงกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก

2. ค่าสารสนเทศเฉลี่ย (RAI) ของมาตรประมาณค่าทั้ง 2 แบบที่ใช้วิธีการตรวจให้คะแนนแบบพหุวิภาคเมื่อวิเคราะห์ตาม GRM จะให้ค่าสูงกว่าเมื่อวิเคราะห์ตาม GPCM และสูงกว่าการตรวจให้คะแนนแบบทวิวิภาคเมื่อวิเคราะห์ตามโมเดลโลจิสติก ตามลำดับ

3. การวิเคราะห์ค่าฟังก์ชันสารสนเทศระหว่างมาตรประมาณค่าที่ให้คะแนนแบบทวิวิภาคและแบบพหุวิภาค เมื่อวิเคราะห์ตามโมเดลโลจิสติก, GRM และ GPCM ให้ผลที่สอดคล้องกัน

ขอบเขตของการวิจัย

การวัดทางการศึกษาจำแนกตามจุดมุ่งหมายเชิงพฤติกรรมได้ 3 ปริเขต (domain) คือ การวัดความรู้ด้านพุทธิปริเขต (cognitive domain) การวัดด้านจิตปริเขต (affective) และการวัดด้านทักษะปริเขต (psychomotor domain) แต่ในการวิจัยนี้ผู้วิจัยได้กำหนดขอบเขตของการศึกษาโดย ศึกษาเครื่องมือที่วัดเฉพาะด้านจิตปริเขตเท่านั้น เนื่องจากในงานวิจัยที่ศึกษาเกี่ยวกับการเปรียบเทียบวิธีการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคที่ผ่านมาส่วนใหญ่เป็นการศึกษาในด้านพุทธิปริเขตและผลการวิจัยที่ได้สอดคล้องกัน แต่ในด้านจิตปริเขตยังมีผู้สนใจศึกษากันน้อย

การวิจัยนี้ศึกษาเฉพาะค่าฟังก์ชันสารสนเทศของแบบวัด (TIF) เท่านั้น เพราะมีวัตถุประสงค์ต้องการเปรียบเทียบความเหมาะสมของวิธีการตรวจให้คะแนน และโมเดลการวิเคราะห์ซึ่งต้องพิจารณาโดยภาพรวมมากกว่าจะมุ่งพิจารณาค่าฟังก์ชันสารสนเทศรายข้อกระทง (IIF)

เครื่องมือที่ใช้ในการวิจัยนี้ เป็นแบบวัดเจตคติทางวิทยาศาสตร์ของนักเรียนชั้นประถมศึกษาปีที่ 6 ชนิดมาตรประมาณค่าที่สร้างขึ้นตามวิธีการของลิเคิร์ท ของ วิมลรัตน์ สิริอาภรณ์ (2538) ซึ่งมีวิธีการให้คะแนนแบบพหุวิภาคและแบบทวิวิภาคแท้ และแบบวัดความมีระเบียบวินัยสำหรับนักเรียนชั้นประถมศึกษาปีที่ 6 ชนิดมาตรประมาณค่าแบบตัวเลือกบังคับตอบ ของ สุพัตรา เทียนอุดม (2536) ซึ่งมีวิธีการให้คะแนนแบบพหุวิภาคและแบบทวิวิภาคไม่แท้

ตัวแปรที่ศึกษา ได้แก่

ตัวแปรต้น

1. วิธีการตรวจให้คะแนน ในการศึกษานี้ใช้ 2 วิธี คือ การตรวจให้คะแนนแบบทวิวิภาค (0-1) และการตรวจให้คะแนนแบบพหุวิภาค

2. โมเดลที่ใช้ในการวิเคราะห์เปรียบเทียบ มี 3 โมเดล คือ GRM, GPCM และโมเดลโลจิสติก

ตัวแปรตาม คือ ค่าฟังก์ชันสารสนเทศ

ข้อจำกัดของการวิจัย

1. ในการวิจัยนี้ไม่ได้ควบคุมให้ลำดับขั้นคะแนนของมาตรประมาณค่ามีจำนวนคงที่โดยในมาตรประมาณค่าแบบลิเคิร์ทและมาตรประมาณค่าแบบตัวเลือกบังคับตอบมีจำนวนลำดับขั้นคะแนนไม่เท่ากัน

2. การเปรียบเทียบค่าฟังก์ชันสารสนเทศของมาตรประมาณค่าแบบลิเคิร์ทและมาตรประมาณค่าแบบตัวเลือกบังคับตอบ จะเปรียบเทียบในช่วง θ ของผู้สอบตั้งแต่ -2 ถึง 2 เท่านั้น เนื่องจากเป็นข้อจำกัดของโปรแกรมคอมพิวเตอร์ MULTILOG ที่ใช้วิเคราะห์ตาม GRM

คำจำกัดความที่ใช้ในการวิจัย

1. การตรวจให้คะแนน หมายถึง การกำหนดคะแนนให้กับคำตอบของนักเรียน ในที่นี้แบ่งออกเป็น 2 ลักษณะ คือ การตรวจให้คะแนนแบบพหุวิภาค และการตรวจให้คะแนนแบบทวิวิภาค

2. การตรวจให้คะแนนแบบทวิวิภาค หมายถึง การกำหนดคะแนนให้กับคำตอบที่ถูกต้องเหมาะสมหรือคำตอบที่ให้คะแนนสูงสุดเป็น 1 คะแนน และให้คำตอบอื่น ๆ เป็น 0 คะแนน ซึ่งในการวิจัยนี้จะแบ่งออกเป็น 2 วิธี คือ

2.1 การตรวจให้คะแนนแบบทวิวิภาคแท้ หมายถึง การกำหนดน้ำหนักคะแนนให้กับตัวเลือกที่สามารถบอกได้ชัดเจนว่าตัวเลือกใดที่ควรจะเป็น 1 คะแนน และตัวเลือกใดควรจะเป็น 0 คะแนน ผู้ตอบสามารถที่จะทราบได้ว่าตนเองตอบแล้วได้ระดับคะแนนเท่าใด ซึ่งในกรณีนี้ใช้ในมาตรประมาณค่าแบบลิเคิร์ท (แบบวัดเจตคติทางวิทยาศาสตร์)

2.2 การตรวจให้คะแนนแบบทวิวิภาคไม่แท้ หมายถึง การกำหนดน้ำหนักคะแนนให้กับตัวเลือกโดยปรับมาจากการให้คะแนนแบบพหุวิภาค ที่ไม่สามารถบอกได้ชัดเจนว่าตัวเลือกใดที่ควรจะเป็น 1 คะแนน และตัวเลือกใดควรจะเป็น 0 คะแนน เพราะว่าแต่ละตัวเลือกก็มีระดับคะแนนซึ่งแต่ละบุคคลจะกำหนดน้ำหนักคะแนนตัวเลือกไม่เท่ากัน จึงใช้เกณฑ์ปกติของคนทั่วไปมาเป็นเกณฑ์การกำหนดน้ำหนักคะแนน โดยให้ตัวเลือกที่มีระดับคุณลักษณะสูงสุดเป็น 1 คะแนน และตัวเลือกที่มีระดับคุณลักษณะรองลงไปเป็น 0 คะแนน การตรวจให้คะแนนวิธีนี้ ผู้ตอบไม่สามารถที่จะทราบได้ว่าตนเองตอบแล้วได้ระดับคะแนนเท่าใด ซึ่งในกรณีนี้ใช้ในมาตรประมาณค่าแบบตัวเลือกบังคับตอบ (แบบวัดความมีระเบียบวินัย)

3. การตรวจให้คะแนนแบบพหุวิภาค หมายถึง การกำหนดคะแนนให้กับคำตอบของนักเรียนตามระดับของความถูกต้องเหมาะสม โดยกำหนดเป็น 5, 4, 3, 2 และ 1 ตามลำดับ ซึ่งในการวิจัยนี้จะแบ่งออกเป็น 2 วิธี คือ

3.1 การตรวจให้คะแนนแบบพหุวิภาคแท้ หมายถึง การกำหนดน้ำหนักคะแนนให้กับตัวเลือกที่สามารถบอกได้ชัดเจนว่าตัวเลือกใดที่ควรจะมีระดับคะแนนเป็น 1, 2, 3, 4 และ 5 คะแนน ผู้ตอบสามารถที่จะทราบได้ว่าตนเองตอบแล้วได้ระดับคะแนนเท่าใด ซึ่งในกรณีนี้ใช้ในมาตรประมาณค่าแบบลิเคิร์ท (แบบวัดเจตคติทางวิทยาศาสตร์)

3.2 การตรวจให้คะแนนแบบพหุวิภาคไม่แท้ หมายถึง การกำหนดน้ำหนักคะแนนให้กับตัวเลือกที่ไม่สามารถบอกได้ชัดเจนว่าตัวเลือกใดที่ควรจะมีระดับคะแนนเป็น 1, 2, 3, และ 4 คะแนน เพราะว่าแต่ละตัวเลือกก็มีระดับคะแนนซึ่งแต่ละบุคคลจะกำหนดน้ำหนักคะแนนตัวเลือกไม่เท่ากัน จึงใช้เกณฑ์ปกติของคนทั่วไปมาเป็นเกณฑ์การกำหนดน้ำหนักคะแนน โดยให้ตัวเลือกที่มีระดับคุณลักษณะสูงที่สุดเป็น 4 คะแนน และตัวเลือกที่มีระดับคุณลักษณะรองลงไปก็กำหนดน้ำหนักคะแนนเป็น 3, 2 และ 1 ตามลำดับ การตรวจให้คะแนนวิธีนี้ ผู้ตอบไม่สามารถที่จะทราบได้ว่าตนเองตอบแล้วได้ระดับคะแนนเท่าใด ซึ่งในกรณีนี้ใช้ในมาตรประมาณค่าแบบตัวเลือกบังคับตอบ (แบบวัดความมีระเบียบวินัย)

4. มาตรประมาณค่า หมายถึง มาตรประมาณค่าที่ใช้ในการวิจัยนี้ มีรายละเอียดดังนี้

4.1 มาตรประมาณค่าที่สร้างขึ้นตามวิธีการของลิเคิร์ท หมายถึง มาตรประมาณค่าที่สร้างขึ้นตามวิธีการของลิเคิร์ท ซึ่งมีวิธีการให้คะแนนแบบทวิวิภาคแท้และแบบพหุวิภาค ในการวิจัยนี้ใช้แบบวัดเจตคติทางวิทยาศาสตร์ชั้นประถมศึกษาปีที่ 6 ของ วิมลรัตน์ สิริอาภรณ์ (2538)

4.2 มาตรประมาณค่าแบบตัวเลือกบังคับตอบ หมายถึง มาตรประมาณค่าแบบตัวเลือกบังคับตอบ ซึ่งนำมาปรับให้มีวิธีการให้คะแนนแบบพหุวิภาคและแบบทวิวิภาคไม่แท้ ในการวิจัยนี้ใช้แบบวัดความมีระเบียบวินัย สำหรับนักเรียนชั้นประถมศึกษาปีที่ 6 ของ สุพัตรา เทียนอุดม (2536)

5. การวิเคราะห์มาตรประมาณค่า หมายถึง การวิเคราะห์เพื่อประมาณค่าพารามิเตอร์ของมาตรประมาณค่าของผู้สอบ และค่าฟังก์ชันสารสนเทศของมาตรประมาณค่า โดยการจำแนกเป็น 5 วิธี คือ

5.1 การวิเคราะห์ตามโมเดลโลจิสติก 1 พารามิเตอร์ หมายถึง การวิเคราะห์มาตรประมาณค่าจากการตรวจให้คะแนนแบบทวิวิภาคตามโมเดลโลจิสติก โดยประมาณค่าพารามิเตอร์ความยาก (b_j) ของข้อกระทง และค่าฟังก์ชันสารสนเทศของแบบวัด (TIF) โดยใช้โปรแกรม MULTILOG

5.2 การวิเคราะห์ตามโมเดลโลจิสติก 2 พารามิเตอร์ หมายถึง การวิเคราะห์มาตรประมาณค่าจากการตรวจให้คะแนนแบบทวิวิภาคตามโมเดลโลจิสติก โดยการประมาณค่าพารามิเตอร์ความยาก (b_j) ค่าอำนาจจำแนก (a_j) ของข้อกระทง และค่าฟังก์ชันสารสนเทศของมาตรประมาณค่า (TIF) โดยใช้โปรแกรม MULTILOG

5.3 การวิเคราะห์ตามโมเดลโลจิสติก 3 พารามิเตอร์ หมายถึง การวิเคราะห์มาตรประมาณค่าจากการตรวจให้คะแนนแบบทวิวิภาคตามโมเดลโลจิสติก โดยการประมาณค่าพารามิเตอร์ความยาก (b_j) ค่าอำนาจจำแนก (a_j) ค่าโอกาสการเดา (c_j) ของข้อกระทง และค่าฟังก์ชันสารสนเทศของมาตรประมาณค่า (TIF) โดยใช้โปรแกรม MULTILOG

5.4 การวิเคราะห์ตาม GRM หมายถึง การวิเคราะห์มาตรประมาณค่าจากการตรวจให้คะแนนแบบพหุวิภาค โดยการประมาณค่าพารามิเตอร์ความยาก (b_j) อำนาจจำแนก (a_j) และค่าฟังก์ชันสารสนเทศของมาตรประมาณค่า (TIF) ตาม GRM โดยใช้โปรแกรม MULTILOG

5.5 การวิเคราะห์ตาม GPCM หมายถึง การวิเคราะห์มาตรประมาณค่าจากการตรวจให้คะแนนแบบพหุวิภาค โดยการประมาณค่าพารามิเตอร์ความยาก (b_j) อำนาจจำแนก (a_j) และค่าฟังก์ชันสารสนเทศของมาตรประมาณค่า (TIF) ตาม GPCM โดยใช้โปรแกรม PARSCALE

6. ค่าพารามิเตอร์ของข้อกระทง หมายถึง ค่าพารามิเตอร์อำนาจจำแนก (a_j) ค่าความยากรายข้อ (b_j) และค่าโอกาสการเดาข้อกระทง (c_j)

7. ค่าฟังก์ชันสารสนเทศของแบบวัด (TIF) หมายถึง ผลรวมของค่าสารสนเทศของข้อกระทงทุกข้อในมาตรประมาณค่าที่มีวิธีการตรวจให้คะแนนอย่างเดียวกันทั้งฉบับ โดยแสดงค่าฟังก์ชันสารสนเทศตามช่วงพิสัยของระดับคุณลักษณะผู้ตอบที่แตกต่างกัน ถ้าสูงที่ระดับใดก็แสดงว่ามีความแม่นยำสูงในการจำแนกคุณลักษณะผู้ตอบ ณ ระดับ θ นั้น ๆ

8. อัตราส่วนสารสนเทศเฉลี่ย (ratio of average information : RAI) หมายถึง ดัชนีบ่งชี้ประสิทธิภาพ มาตรประมาณค่าที่ใช้วิธีการตรวจให้คะแนนที่ต่างกัน 2 แบบ และวิเคราะห์ด้วยโมเดลที่ต่างกัน 3 โมเดล สามารถใช้พิจารณาโดยภาพรวมว่าเครื่องมือที่วิเคราะห์ด้วยวิธีใดจะมีประสิทธิภาพมากกว่ากัน

ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้ได้ข้อความรู้เกี่ยวกับค่าฟังก์ชันสารสนเทศของแบบวัดที่มีลักษณะเป็นมาตรประมาณค่าเมื่อใช้วิธีการตรวจให้คะแนนแบบทวิวิภาคกับแบบพหุวิภาค ในการวิเคราะห์ด้วย GRM GPCM และ โมเดลโลจิสติก

2. ผลการวิจัยสามารถใช้เป็นประโยชน์ในการเลือกใช้วิธีการตรวจให้คะแนนแบบวัดที่มีลักษณะเป็นมาตรประมาณค่าและเป็นแนวทางในการเลือกใช้โมเดลที่เหมาะสม

3. ผลการวิจัยนี้สามารถใช้เป็นแนวทางในการศึกษาเกี่ยวกับวิธีการตรวจให้คะแนนแบบวัดชนิดอื่น ๆ และสามารถใช้เป็นแนวทางในการศึกษา Polytomous Item Response Model อื่น ๆ อีก