Journal of Research Methodology, Volume 19, Number 1 (January-April 2006)

วารสารวิธีวิทยาการวิจัย ปีที่ 19 ฉบับที่ 1 (มกราคม-เมษายน 2549)

Effect Magnitude Estimators' Properties: A Comparison Between Classical Test Theory and Item Response Theory*

Chayut Piromsombat

ABSTRACT

There were 4 objectives for this research: (1) to compare the estimators' properties in biasedness, consistency, and relative efficiency aspects among the effect size derived from Classical Test Theory (d_{CTT}), effect size derived from Item Response Theory which estimation models fit for the data (d_{IRT1}), effect size derived from Item Response Theory which estimation models unfit for the data (d_{IRT2}), correlation coefficient derived from Classical Test Theory (r_{CTT}), correlation coefficient derived from Item Response Theory which estimation models fit for the data (r_{IRT1}), and correlation coefficient derived from Item Response Theory which estimation models unfit for the data (r_{IRT2}); (2) to compare the means among d_{CTT}, d_{IRT1}, and d_{IRT2} as well as among r_{CTT}, r_{IRT1}, and r_{IRT2}; (3) to study the relationship between d_{CTT} and d_{IRT1} and express the regression equation of d_{IRT1} on d_{CTT}; (4) to study the relationship between r_{CTT} and r_{IRT1} and express the regression equation of r_{IRT1} on r_{CTT}. The 540 examination situations were built up from the conditions of the true effect magnitudes (.2, .5, .8, 1.2, 2.6), sample sizes (20, 50, 500, 2,000), test lengths (10, 50, 90), based models (one-, two-, and three-parameter logistic model), and estimation models (classical test model, item response models which fit for the data, item response models which unfit for the data).

The summarized findings were: (1) in the overview, the lowest biased estimator was r_{IRT1}, the highest consistency estimator was r_{CTT}, and the highest relative efficiency estimator was r_{IRT1}, in addition, r_{IRT1} was the most appropriate estimator for all properties; (2) the means of d_{CTT}, d_{IRT1}, and d_{IRT} were different at the .05 significance

* Thesis of Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University under the advice of Assoc. Prof. Suchada Bowarnkitiwong, Ph.D.

level, in fact, the mean of d_{CTT} was the highest, in the same way, the means of r_{CTT}, r_{IRT1}, and r_{IRT} were different at the .05 significance level and the mean of r_{CTT} was the highest; (3) the correlation coefficient between d_{CTT} and d_{IRT1} was .626 and significant at the .05 level, the regression of d_{IRT1} on d_{CTT} could be expressed by $d_{IRT1} = .004 + .065d_{CTT}$ for the raw scores and $Z_{d_{IRT1}} = .626Z_{d_{IRT}}$ for the standardized scores; (4) the correlation coefficient between r_{CTT} and r_{IRT1} was .570 and significant at the .05 level, the regression of r_{IRT1} on r_{CTT} could be expressed by $r_{IRT1} = .003 + .079r_{CTT}$ for the raw scores and $Z_{r_{IRT1}} = .570Z_{CTT}$ for the standardized scores.

Journal of Research Methodology, Volume 19, Number 1 (January-April 2006)

วารสารวิธีวิทยาการวิจัย ปีที่ 19 ฉบับที่ 1 (มกราคม-เมษายน 2549)

คุณสมบัติของตัวประมาณค่าความเข้มของอิทธิพล: การเปรียบเทียบระหว่างทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบสนองข้อสอบ*

ชยุตม์ ภิรมย์สมบัติ

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ 4 ข้อ ได้แก่ (1) เพื่อเปรียบเทียบคุณสมบัติของตัวประมาณค่าในด้านความลำเอียง ความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์ระหว่างขนาดอิทธิพลที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิม (d_{CTT}) ขนาดอิทธิพลที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าสอดคล้องกับข้อมูล (d_{IRT_1}) ขนาดอิทธิพลที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าไม่สอดคล้องกับข้อมูล (d_{IRT_2}) สัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิม (r_{CTT}) สัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าสอดคล้องกับข้อมูล (r_{IRT_1}) และสัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าไม่สอดคล้องกับข้อมูล (r_{IRT_2}) (2) เพื่อเปรียบเทียบค่าเฉลี่ยระหว่าง d_{CTT}, d_{IRT_1} และ d_{IRT_2} และระหว่าง r_{CTT}, r_{IRT_1} และ r_{IRT_2} (3) เพื่อศึกษาความสัมพันธ์ระหว่าง d_{CTT} และ d_{IRT_1} และสร้างสมการถดถอยของ d_{IRT_1} บน d_{CTT} และ (4) เพื่อศึกษาความสัมพันธ์ระหว่าง r_{CTT} และ r_{IRT_1} และสร้างสมการถดถอยของ r_{IRT_1} บน r_{CTT} ภายใต้สถานการณ์การสอบ 540 สถานการณ์ ตามเงื่อนไขของค่าความเข้มของอิทธิพลที่แท้จริง (.2, .5, .8, 1.2, 2.6) ขนาดกลุ่มตัวอย่าง (20, 50, 500, 2,000) ความยาวแบบสอบ (10, 50, 90) โมเดลฐาน (โมเดลโลจิสติกแบบหนึ่ง, สอง และสามพารามิเตอร์) และโมเดลประมาณค่า (โมเดลการทดสอบแบบดั้งเดิม, โมเดลการตอบสนองข้อสอบที่สอดคล้องกับข้อมูล และโมเดลการตอบสนองข้อสอบที่ไม่สอดคล้องกับข้อมูล)

* อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร.สุชาดา บวรกิติวงศ์ วิทยานิพนธ์ครุศาสตรมหาบัณฑิต สาขาการวัดและประเมินผลการศึกษา ปีการศึกษา 2547

ผลการวิจัยโดยสรุปพบว่า (1) ในภาพรวมตัวประมาณค่าที่มีความลำเอียงต่ำที่สุดคือ rIRT1 ตัวประมาณค่าที่มีความคงเส้นคงวาสูงสุดคือ r_{CTT} และตัวประมาณค่าที่มีประสิทธิภาพสัมพัทธ์สูงสุดคือ r_{IRT1} นอกจากนี้ r_{IRT1} ยังเป็นตัวประมาณค่าที่มีคุณสมบัติทุกด้านเป็นที่น่าพอใจที่สุด (2) ค่าเฉลี่ยของ d_{CTT}, d_{IRT1} และ d_{IRT2} มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดย d_{CTT} มีค่าเฉลี่ยสูงสุด เช่นเดียวกับค่าเฉลี่ยของ r_{CTT}, r_{IRT1} และ r_{IRT2} ที่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดย r_{CTT} มีค่าเฉลี่ยสูงสุด (3) สัมประสิทธิ์สหสัมพันธ์ระหว่าง d_{CTT} และ d_{IRT1} มีค่า .626 และมีนัยสำคัญทางสถิติที่ระดับ .05 สมการถดถอยของ d_{IRT1} บน d_{CTT} ในรูปคะแนนดิบ คือ $d_{IRT1} = .004 + .065d_{CTT}$ สมการในรูปคะแนนมาตรฐาน คือ $Z_{d_{IRT1}} = .626z_{d_{CTT}}$ (4) สัมประสิทธิ์สหสัมพันธ์ระหว่าง r_{CTT} และ r_{IRT1} มีค่า .570 และมีนัยสำคัญทางสถิติที่ระดับ .05 สมการถดถอยของ r_{IRT1} บน r_{CTT} ในรูปคะแนนดิบ คือ $r_{IRT1} = .003 + .079r_{CTT}$ และสมการในรูปคะแนนมาตรฐาน คือ $z_{r_{IRT1}} = .570z_{r_{CTT}}$

ความเป็นมาและความสำคัญของปัญหา

ทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) และทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) มีจุดเน้นที่คล้ายกันประการหนึ่ง คือ การให้ความสำคัญกับการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและพารามิเตอร์ของข้อสอบ (ค่าอำนาจจำแนก ความยากและโอกาสในการเดาถูก) จุดเน้นที่คล้ายกันนี้เองที่ทำให้มีการศึกษาเปรียบเทียบค่าพารามิเตอร์ของผู้สอบและข้อสอบที่ได้จากทั้งสองทฤษฎี อาทิงานวิจัยของ Ndalichako และ Rogers (1997), Fan (1998), Stage (1998; 2003), MacDonald และ Paunonen (2002), นภดล ยิ่งยงสกุล (2539), เบญจพร ยนตจักรวิถี (2539), วีระพันธ์ พรหมบุตร (2536) และอรวรรณ สุขโต (2542) นอกจากนี้ยังมีการเปรียบเทียบค่าสถิติอื่นที่ใช้ค่าพารามิเตอร์ของผู้สอบและ/หรือค่าพารามิเตอร์ของข้อสอบเป็นพื้นฐาน เช่น การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (นิคม กีรติวรางกูร, 2542) การเปรียบเทียบคุณภาพของการวัดคะแนนพัฒนาการ (อวยพร เรืองตระกูล, 2544) และการเปรียบเทียบค่าขนาดอิทธิพล (Wang & Chen, 2005) เป็นต้น การเปรียบเทียบดังกล่าวล้วนมีประโยชน์ทั้งในเชิงวิชาการและการประยุกต์ใช้ทฤษฎีการทดสอบให้เหมาะสมกับสถานการณ์ที่แตกต่างกัน

สำหรับสารสนเทศจากการเปรียบเทียบค่าขนาดอิทธิพลที่ได้จากคะแนนความสามารถของผู้สอบตาม CTT และ IRT ของ Wang และ Chen (2004) เป็นประโยชน์อย่างยิ่งต่อการวิจัยเชิง

ปริมาณทางการศึกษา เนื่องจากการวิจัยเชิงปริมาณต้องการตอบคำถามที่สำคัญ 3 ข้อ คือ 1) อิทธิพล/ปรากฏการณ์ที่สนใจศึกษามีโอกาสเกิดขึ้นหรือไม่ 2) ถ้ามีโอกาสเกิดขึ้น อิทธิพลดังกล่าวมีปริมาณมากน้อยเพียงใด และ 3) อิทธิพลดังกล่าวมีปริมาณมากพอที่จะใช้ประโยชน์ในทางปฏิบัติหรือไม่ นักวิจัยสามารถใช้การทดสอบนัยสำคัญทางสถิติ เพื่อตอบคำถามแรกแต่ไม่สามารถใช้ในการตอบคำถามที่เหลือได้โดยตรง (Kirk, 2001; Paul & Plucker, 2004) ทำให้เกิดการพัฒนาวิธีการทางสถิติ เพื่อตอบคำถามเกี่ยวกับอิทธิพลของตัวแปรในงานวิจัยขึ้น

เนื่องจากอิทธิพลที่ต้องการศึกษาเป็นอิทธิพลที่เกิดขึ้นในกลุ่มประชากร ผู้วิจัยจึงใช้คำว่า "การประมาณค่าความเข้มของอิทธิพล" แทน "การวัดความเข้มของอิทธิพล" โดยพบว่าคำศัพท์ภาษาอังกฤษที่หมายถึงการประมาณค่าอิทธิพลที่ปรากฏในรายงานวิจัยและเอกสารทางวิชาการนั้นมีหลายคำ ดังที่ Snyder และ Lawson (1993) และ Ives (2003) ได้สำรวจไว้ เช่น estimates of magnitude of the effect, estimates of explained variance, effect size estimates, estimates of proportion of variance accounted for และ measure of association เป็นต้น จะเห็นได้ว่าในภาษาอังกฤษมีทั้งการใช้คำว่า effect magnitude (Hedges & Olkin, 1985; Snyder & Lawson, 1993; Kirk, 2001; Gliner, Leech & Morgan, 2002) และคำว่า effect size (Wilkinson & APA Task Force on Statistical Inference, 1999; Huberty, 2002; Trusty, Thompson & Petrocelli, 2004) นอกจากนี้ Shaver (1993) ยังได้เสนอคำว่า result size แต่ไม่ได้รับความนิยมโดย Shaver ได้กล่าวไว้ว่านักวิจัยทางสังคมศาสตร์นิยมใช้คำว่า effect size เกินกว่าจะเปลี่ยนแปลงได้ ปัจจุบันจึงพบว่านักวิจัยส่วนใหญ่ใช้คำว่า effect size ในความหมายเดียวกับ effect magnitude แต่ในงานวิจัยครั้งนี้ผู้วิจัยใช้คำว่า effect magnitude ในความหมายของความเข้มของอิทธิพลที่เกิดขึ้นในงานวิจัย และใช้คำว่า effect size ในความหมายของผลต่างมาตรฐานตามนิยามที่ Cohen ได้เริ่มใช้เป็นคนแรก (นงลักษณ์ วิรัชชัย, 2542) เนื่องจากยังมีสถิติอีกมากที่ไม่ได้อยู่ในรูปของผลต่างมาตรฐานแต่สามารถใช้ประมาณค่าความเข้มของอิทธิพลได้

นอกจากขนาดอิทธิพล (d) ยังมีสถิติที่สามารถใช้ประมาณค่าความเข้มของอิทธิพลได้อีกกว่า 40 ชนิด (Kirk, 1996) โดย Baugh (2002), Thompson (2002) และ Ives (2003) กล่าวไว้สอดคล้องกันว่าการประมาณค่าความเข้มของอิทธิพลควรใช้สถิติในกลุ่มความสัมพันธ์หรือความแปรปรวนที่ถูกอธิบาย (variance accounted for) เพราะสามารถนำมาใช้ได้กับการวิเคราะห์เชิงปริมาณที่ใช้โมเดลเชิงเส้นทั่วไป รวมทั้งงานวิจัยเชิงปริมาณทั้งที่เป็นเชิงทดลองและไม่ใช่เชิงทดลอง เนื่องจากเป้าหมายหนึ่งของการวิจัยเชิงปริมาณส่วนใหญ่คือต้องการศึกษาความสัมพันธ์หรือความผันแปรร่วมกันระหว่างตัวแปรนั่นเอง (ศิริชัย กาญจนวาสี, 2541; Thompson, 2000)

นอกจากนี้ Rosenthal และ DiMatteo (2001) ยังกล่าวไว้ว่าสถิติในกลุ่มความสัมพันธ์อย่างสัมประสิทธิ์สหสัมพันธ์มีข้อดีเหนือกว่าขนาดอิทธิพลบางประการ คือ การแปลงขนาดอิทธิพลให้อยู่ในรูปของสัมประสิทธิ์สหสัมพันธ์มีความสมเหตุสมผล เนื่องจากสัมประสิทธิ์สหสัมพันธ์แบบพอยต์ไบซีเรียล สามารถแสดงความสัมพันธ์ระหว่างตัวแปรต้นที่มีสองระดับกับตัวแปรตามที่มีค่าต่อเนื่องได้ ในขณะที่การแปลงสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันซึ่งตัวแปรมีค่าต่อเนื่องให้อยู่ในรูปของขนาดอิทธิพลซึ่งใช้ข้อมูลทวิภาคจะทำให้เสียสารสนเทศไป นอกจากนี้ยังไม่จำเป็นต้องปรับแก้วิธีการคำนวณสัมประสิทธิ์สหสัมพันธ์เมื่อใช้กลุ่มตัวอย่างมากกว่า 2 กลุ่ม และสัมประสิทธิ์สหสัมพันธ์แปลความหมายได้ง่ายกว่าอีกด้วย อย่างไรก็ตาม การเลือกใช้ตัวประมาณค่ายังต้องคำนึงถึงคุณสมบัติของตัวประมาณค่า อาทิ ความลำเอียง ความคงเส้นคงวา ความพอเพียง และประสิทธิภาพสัมพัทธ์ด้วย (Mandenhall & Beaver, 1994; Glass & Hopkins, 1995)

การเปรียบเทียบค่าสถิติที่ได้จาก CTT และ IRT นิยมใช้ข้อมูลจากการจำลองสถานการณ์การสอบต่าง ๆ (Fan, 1998; DeMars, 2001; MacDonald & Paunonen, 2002; Dawber, Rogers & Carbonaro, 2004; Wang & Chen, 2004) สำหรับการเปรียบเทียบตัวประมาณค่าในสถานการณ์ต่าง ๆ ควรใช้การจำลองข้อมูลเช่นกัน (Harwell et al., 1996) ผู้วิจัยจึงเลือกใช้การจำลองข้อมูลเป็นแนวทางในการวิจัยครั้งนี้

สำหรับเงื่อนไขในการเปรียบเทียบค่าสถิติที่ได้ CTT และ IRT ที่สำรวจจากงานวิจัยที่เกี่ยวข้องพบว่า มีการใช้เงื่อนไขที่แตกต่างกันตามเป้าหมายของการเปรียบเทียบ โดยเงื่อนไขที่พบมาก ได้แก่ ความยาวข้อสอบและขนาดกลุ่มผู้สอบ (Fan, 1998; DeMars, 2001; Roberts & Henson, 2002; Dawber, Rogers & Carbonaro, 2004; Stone & Yumoto, 2004; Wang & Chen, 2004) สอดคล้องกับผลสำรวจของ Harwell และคณะ (1996) ซึ่งยังพบเงื่อนไขอื่นอีก เช่น ชนิดของโมเดลการตอบสนองข้อสอบ การกระจายของค่าความยาก อำนาจจำแนกและโอกาสในการเดา เป็นต้น นอกจากนี้ Hambleton และ Swaminathan (1985), Embretson และ Reise (2000) และศิริชัย กาญจนวาสี (2545) กล่าวไว้สอดคล้องกันว่า โมเดลการวัดที่ไม่สอดคล้องกับข้อมูลย่อมทำให้ผลการวิเคราะห์ขาดความถูกต้อง ความสอดคล้องของโมเดลกับข้อมูลจึงเป็นอีกประเด็นที่น่าสนใจ ผู้วิจัยจึงประยุกต์แนวคิดของ DeMars (2001) ซึ่งกำหนดโมเดลที่ใช้ประมาณค่าความสามารถของผู้สอบมี 2 โมเดล โมเดลแรกเป็นโมเดลเดียวกับโมเดลฐานที่ใช้ในการจำลองคำตอบของผู้สอบ แทนสถานการณ์ที่โมเดลประมาณค่าสอดคล้องกับข้อมูล และโมเดลที่สองเป็นโมเดลที่ต่างจากโมเดลฐานแทนสถานการณ์ที่โมเดลประมาณค่าไม่สอดคล้องกับข้อมูล

การวิจัยครั้งนี้จึงเป็นการเปรียบเทียบคุณสมบัติของตัวประมาณค่าระหว่างขนาดอิทธิพลที่ได้จาก CTT (d_{CTT}) ขนาดอิทธิพลที่ได้จาก IRT (d_{IRT}) สัมประสิทธิ์สหสัมพันธ์ที่ได้จาก CTT (r_{CTT})

และสัมประสิทธิ์สหสัมพันธ์ที่ได้จาก IRT (r_{IRT}) โดยศึกษาจากข้อมูลที่จำลองขึ้นตาม 5 เงื่อนไขหลัก ได้แก่ ความเข้มของอิทธิพลที่แท้จริง ความยาวแบบสอบ ขนาดกลุ่มตัวอย่าง โมเดลฐาน และโมเดลประมาณค่า โดยใช้ค่าความยาก อำนาจจำแนก และโอกาสในการเดาที่สุ่มได้จากการกระจายที่เหมาะสมกับค่าทั้งสามตามที่ Pelton (2002) ได้สรุปไว้ นั่นคือ อำนาจจำแนกของข้อสอบ (a) ควรสุ่มค่าในช่วง 0 ถึง 1 จากการแจกแจงปกติมาตรฐาน โอกาสในการเดา (c) ควรสุ่มค่าในช่วง 0 ถึง 1 จากการแจกแจงยูนิฟอร์มและค่าความยากของข้อสอบ (b) สำหรับโมเดลโลจิสติกหนึ่งและสองพารามิเตอร์ควรสุ่มค่าในช่วง -3 ถึง 3 จากการแจกแจงปกติมาตรฐาน แต่โมเดลโลจิสติกแบบสามพารามิเตอร์นั้นค่าความยากขึ้นอยู่กับโอกาสในการเดา (b = (1 + c) ÷ 2)

วัตถุประสงค์การวิจัย

1. เพื่อเปรียบเทียบคุณสมบัติของตัวประมาณค่าในด้านความลำเอียง ความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์ ระหว่างขนาดอิทธิพลที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิม (d_{CTT}) สัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิม (r_{CTT}) ขนาดอิทธิพลที่ได้จากทฤษฎีการตอบสนองข้อสอบ (d_{IRT}) ซึ่งแบ่งย่อยเป็นขนาดอิทธิพลที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าสอดคล้องกับข้อมูล (d_{IRT_1}) และขนาดอิทธิพลที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าไม่สอดคล้องกับข้อมูล (d_{IRT_2}) และสัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการตอบสนองข้อสอบ (r_{IRT}) ซึ่งแบ่งย่อยเป็นสัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าสอดคล้องกับข้อมูล (r_{IRT_1}) และสัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าไม่สอดคล้องกับข้อมูล (r_{IRT_2})

2. เพื่อเปรียบเทียบค่าเฉลี่ยระหว่าง d_{CTT}, d_{IRT_1} และ d_{IRT_2} และระหว่าง r_{CTT}, r_{IRT_1} และ r_{IRT_2}

3. เพื่อศึกษาความสัมพันธ์ระหว่าง d_{CTT} และ d_{IRT_1} และสร้างสมการถดถอยของ d_{IRT_1} บน d_{CTT}

4. เพื่อศึกษาความสัมพันธ์ระหว่าง r_{CTT} และ r_{IRT_1} และสร้างสมการถดถอยของ r_{IRT_1} บน r_{CTT}

สมมติฐานการวิจัย

1. d_{CTT}, d_{IRT_1}, d_{IRT_2}, r_{CTT}, r_{IRT_1} และ r_{IRT_2} มีคุณสมบัติของตัวประมาณค่าในด้านความลำเอียง ความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์ แตกต่างกัน

2. ค่าเฉลี่ยของ d_{CTT}, d_{IRT_1}, d_{IRT_2}, r_{CTT}, r_{IRT_1} และ r_{IRT_2} ในการประมาณค่าความเข้มของอิทธิพลเดียวกันมีค่าแตกต่างกัน

3. d_{CTT} และ d_{IRT} มีความสัมพันธ์กันในทางบวกค่อนข้างสูง
4. r_{CTT} และ r_{IRT} มีความสัมพันธ์กันในทางบวกค่อนข้างสูง

ขอบเขตการวิจัย

1. การวิจัยในครั้งนี้มีขอบเขตการศึกษาเฉพาะการประมาณค่าความเข้มของอิทธิพลแบบจุด (point estimation of effect magnitude) ในงานวิจัยเชิงปริมาณทางการศึกษาที่ศึกษาอิทธิพลเจาะจง (fixed effect) ของตัวแปรต้นหรือตัวแปรจัดกระทำที่มีต่อตัวแปรความสามารถของบุคคล ซึ่งวัดด้วยข้อสอบที่มีการตรวจให้คะแนนแบบทวิภาค (dichotomous scoring)

2. โมเดลการวัดที่ใช้ในการวิจัยครั้งนี้เป็นโมเดลการตอบสนองข้อสอบในกลุ่มของโมเดลการตรวจให้คะแนนแบบทวิภาค 3 โมเดล คือ โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ โมเดลโลจิสติกแบบสองพารามิเตอร์ และโมเดลโลจิสติกแบบสามพารามิเตอร์

3. ทุกสถานการณ์ที่จำลองขึ้นจะทำซ้ำ (replicate) 100 รอบ

นิยามศัพท์

ค่าความเข้มของอิทธิพล หมายถึง ตัวเลขที่ให้สารสนเทศเกี่ยวกับขนาดของผลอันเกิดจากอิทธิพลของตัวแปรจัดกระทำที่ทำให้เกิดการเปลี่ยนแปลงในตัวแปรตาม หรือตัวเลขที่ให้สารสนเทศเกี่ยวกับความสัมพันธ์ระหว่างตัวแปร ในงานวิจัยนี้ค่าความเข้มของอิทธิพลจะประมาณค่าได้จากขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์แบบไบซีเรียล

คุณสมบัติของตัวประมาณค่า หมายถึง ลักษณะของตัวประมาณค่าหรือสถิติที่ใช้ประมาณค่าพารามิเตอร์ในประชากร ซึ่งตัวประมาณค่าที่ดีควรจะไม่มีความลำเอียง (unbiased) หรืออย่างน้อยที่สุดคือมีความลำเอียงต่ำเมื่อเทียบกับตัวประมาณค่าอื่น นอกจากนี้ยังควรมีความคงเส้นคงวาสูง (consistency) มีความพอเพียง (sufficiency) ในการใช้ข้อมูลจากทุกหน่วยตัวอย่างมาใช้ประมาณค่า และมีประสิทธิภาพสัมพัทธ์สูง (relative efficiency) ในงานวิจัยครั้งนี้ ผู้วิจัยไม่ได้พิจารณาความพอเพียงเนื่องจากตัวประมาณค่าทุกตัวใช้ข้อมูลจากผู้สอบทั้งหมดในการคำนวณอยู่แล้ว

ความลำเอียง หมายถึง คุณสมบัติของตัวประมาณค่าที่มีค่าเฉลี่ยของการแจกแจงการสุ่มของตัวประมาณค่าต่างจากค่าพารามิเตอร์ ในงานวิจัยนี้ความลำเอียงวัดได้จากค่าความแตกต่างระหว่างค่าคาดหวังของตัวประมาณค่ากับค่าที่แท้จริงหรือ ค่าความลำเอียง (BIAS) และส่วนเบี่ยงเบนของรากกำลังสองเฉลี่ยมาตรฐาน (Standardized Root Mean Square Deviation: SRMSD) ตามสูตรต่อไปนี้

BLAS = E(â) - a (Hay, 1963 อ้างถึงในสุกัญญรัตน์ คงงาม, 2539)

และ SRMSD = $\sqrt{\frac{\sum_{i=1}^{n}\left(Standardized\ BLAS\right)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n}\left\{\frac{\hat{a}_i - a}{SD_{\hat{a}_i - a}}\right\}^2}{N}}$

(ดัดแปลงจาก Pelton, 2002)

โดยที่ a หมายถึง พารามิเตอร์ที่ต้องการประมาณค่า

$\hat{a}$ หมายถึง ค่าประมาณของพารามิเตอร์

$E(\hat{a})$ หมายถึง ค่าคาดหวังของ $\hat{a}$ หรือค่าเฉลี่ยของการแจกแจงการสุ่มของ$\hat{a}$

$SD_{\hat{a}_i - a}$ หมายถึง ส่วนเบี่ยงเบนมาตรฐานของค่าความลำเอียง (BIAS)

i หมายถึง การประมาณค่าครั้งที่ i

N หมายถึง จำนวนครั้งในการประมาณค่า

เกณฑ์คือ ตัวประมาณค่าที่ดีควรมีความลำเอียงต่ำเมื่อเทียบกับตัวประมาณค่าชนิดอื่น นั่นคือ มีค่า SRMSD ต่ำที่สุด สำหรับค่า BIAS จะใช้ในการพิจารณาว่าตัวประมาณค่าให้ค่าสูงหรือต่ำกว่าค่าจริง (overestimate or underestimate) หากค่า BIAS เป็นลบแสดงว่าตัวประมาณค่าให้ค่าที่ต่ำกว่าค่าจริง แต่ถ้าค่า BIAS เป็นบวกแสดงว่าตัวประมาณค่าให้ค่าที่สูงกว่าค่าจริง

ความคงเส้นคงวา หมายถึง คุณลักษณะของตัวประมาณค่าที่มีแนวโน้มในการให้ค่าประมาณเข้าใกล้ค่าพารามิเตอร์เมื่อกลุ่มตัวอย่างมีขนาดเพิ่มขึ้น หรือเมื่อขนาดกลุ่มตัวอย่างลู่เข้าสู่อนันต์ ($N \rightarrow \infty$) ในการวิจัยครั้งนี้ ความคงเส้นคงวาของตัวประมาณค่าความเข้มของอิทธิพลจะคำนวณด้วยสูตรที่ปรับจาก สุกัญญรัตน์ คงงาม (2539) ดังต่อไปนี้

$$\triangle = SRMSD_{Nmin} - SRMSD_{Nmax}$$

โดยที่ $\triangle$ หมายถึง ผลต่างของส่วนเบี่ยงเบนของรากกำลังสองเฉลี่ยมาตรฐาน

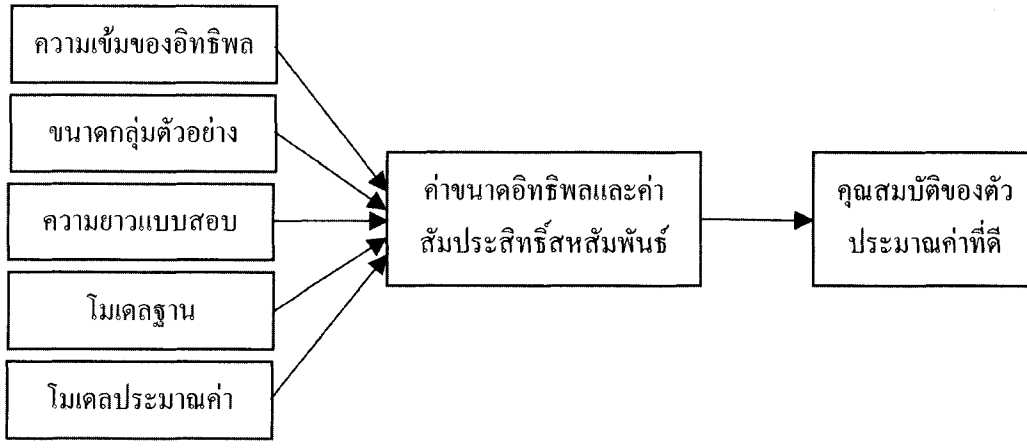
$SRMSD_{Nmin}$ หมายถึง ส่วนเบี่ยงเบนของรากกำลังสองเฉลี่ยมาตรฐานในกรณีที่กลุ่มตัวอย่างมีขนาดเล็กที่สุด (ในการวิจัยครั้งนี้ Nmin มีค่า 20)

$SRMSD_{Nmax}$ หมายถึง ส่วนเบี่ยงเบนของรากกำลังสองเฉลี่ยมาตรฐานในกรณีที่กลุ่มตัวอย่างมีขนาดใหญ่ที่สุด (ในการวิจัยครั้งนี้ Nmax มีค่า 2000)

เกณฑ์คือ ตัวประมาณค่าที่ดีควรมีความคงเส้นคงวาสูง นั่นคือ มีค่า △ สูงกว่าตัวประมาณค่าอื่น

ประสิทธิภาพสัมพัทธ์ หมายถึง คุณสมบัติของตัวประมาณค่าในด้านความถูกต้องสัมพัทธ์(relative precision) ของการประมาณค่า หรือระดับของความคลาดเคลื่อนของการสุ่มเมื่อเทียบกับตัวประมาณค่าชนิดอื่นเมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน (Glass & Hopkins, 1995) ในงานวิจัยครั้งนี้พิจารณาโดยการเปรียบเทียบความแปรปรวนของค่าประมาณความเข้มของอิทธิพลที่ได้จากตัวประมาณค่าแต่ละชนิดในสถานการณ์เดียวกัน ตัวประมาณค่าที่ดีควรมีประสิทธิภาพสัมพัทธ์สูง นั่นคือ มีค่าความแปรปรวนต่ำกว่าตัวประมาณค่าอื่น

กรอบแนวคิดในการวิจัย



ภาพที่ 1 กรอบแนวคิดในการวิจัย

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ใช้ระเบียบวิธีวิจัยเชิงทดลอง โดยศึกษาจากการจำลองสถานการณ์การสอบที่เป็นไปได้ต่าง ๆ กัน โดยมีรายละเอียดของเงื่อนไขที่ใช้ในการจำลองข้อมูล การจำลองข้อมูล และการวิเคราะห์ข้อมูล ดังนี้

1. **เงื่อนไขที่ใช้ในการจำลองข้อมูล** ประกอบด้วย 8 เงื่อนไข ดังนี้

1) ค่าขนาดอิทธิพลที่แท้จริง (TRUES) 5 ค่า คือ ขนาดอิทธิพลในประชากร (δ) มีค่า .2, .5, .8, 1.2 และ 2.6 ซึ่งมีค่าประมาณสัมประสิทธิ์สหสัมพันธ์ในประชากร (ρ) เท่ากับ .1, .2, .4, .5 และ .8 ตามลำดับ

2) ขนาดกลุ่มตัวอย่าง (NSAMP) 4 ค่า คือ 20, 50, 500 และ 2,000 คน

3) ความยาวแบบสอบ (NITEM) 3 ค่า คือ 10, 50 และ 90 ข้อ

4) โมเดลฐานสำหรับจำลองข้อมูล (MBASE) 3 โมเดล คือ โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (1PLM) โมเดลโลจิสติกแบบสองพารามิเตอร์ (2PLM) และโมเดลโลจิสติกแบบสามพารามิเตอร์ (3PLM)

5) โมเดลประมาณค่าความสามารถของผู้สอบ (MUSED) เพื่อศึกษาผลเมื่อความสอดคล้องระหว่างโมเดลกับข้อมูลมีลักษณะต่าง ๆ กัน ในการวิจัยครั้งนี้ผู้วิจัยประยุกต์วิธีการของ DeMars (2001) โดยใช้โมเดล 3 ลักษณะ ได้แก่ **โมเดลที่ 1** เป็นโมเดลการตอบสนองข้อสอบที่เป็นโมเดลฐาน แทนสถานการณ์ที่ขนาดอิทธิพลหรือสัมประสิทธิ์สหสัมพันธ์ได้มาจากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าสอดคล้องกับข้อมูล **โมเดลที่ 2** เป็นโมเดลการตอบสนองข้อสอบที่ไม่ใช่โมเดลฐาน (ได้โดยสุ่ม) แทนสถานการณ์ที่ขนาดอิทธิพลหรือสัมประสิทธิ์สหสัมพันธ์ได้มาจากทฤษฎีการตอบสนองข้อสอบที่โมเดลประมาณค่าไม่สอดคล้องกับข้อมูล และ **โมเดลที่ 3** เป็นโมเดลการทดสอบแบบดั้งเดิม เช่น กรณีที่ใช้ 1PLM เป็นฐานในการจำลองข้อมูล จะใช้ 1PLM 2PLM และโมเดลการทดสอบแบบดั้งเดิม (CTT) เป็นโมเดลประมาณค่า

6) อำนาจจำแนกของข้อสอบ (*a*) สุ่มค่าในช่วง 0 ถึง 1 จากการแจกแจงปกติมาตรฐาน

7) โอกาสในการเดา (c) สุ่มค่าในช่วง 0 ถึง 1 จากการแจกแจงยูนิฟอร์ม

8) ความยากของข้อสอบ (*b*) สำหรับ 1PLM และ 2PLM จะสุ่มค่าในช่วง -3 ถึง 3 จากการแจกแจงปกติมาตรฐาน แต่ 3PLM ค่าความยากขึ้นอยู่กับโอกาสในการเดา ($b = (1 + c) \div 2$)

เงื่อนไขที่ 1 - 5 เป็นเงื่อนไขตามกรอบแนวคิดการวิจัย เงื่อนไขที่ 6 - 8 เป็นเงื่อนไขเกี่ยวกับ ข้อสอบที่ใช้ประกอบการจำลองข้อมูลคำตอบของผู้สอบ ดังนั้น สถานการณ์ที่จำลองข้อมูลครั้งนี้จึงมีทั้งหมด 540 สถานการณ์ และเมื่อนำมาจัดกลุ่มเพื่อเปรียบเทียบคุณสมบัติของ d_{CTT}, d_{IRT_1}, d_{IRT_2}, r_{CTT}, r_{IRT_1} และ r_{IRT_2} จะได้ 180 สถานการณ์

2. การจำลองข้อมูล

ผู้วิจัยจำลองข้อมูลตามเงื่อนไขที่กำหนดไว้โดยทำซ้ำ 100 ครั้งด้วยขั้นตอนมาตรฐานที่ใช้ศึกษาทฤษฎีการตอบสนองข้อสอบด้วยการจำลองสถานการณ์ (Harwell et al., 1996; MacDonald & Paunonen, 2002; Pelton, 2002; Dawber, Rogers & Carbonaro, 2004) ดังนี้

1) สร้างข้อมูลค่าความสามารถที่แท้จริง (true ability) ของผู้สอบในกลุ่มควบคุม และกลุ่มทดลองในกรณีที่ขนาดอิทธิพลมีค่า .2, .5, .8, 1.2 และ 2.6 โดยผู้วิจัยอาศัยความสัมพันธ์ระหว่างขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์ตามสมการของ Hunter และ Schmidt (1990) จึงใช้ค่าขนาดอิทธิพล .5, 1.2 และ 2.6 ซึ่งมีค่าประมาณสัมประสิทธิ์สหสัมพันธ์ที่ .2, .5 และ .8 ตามลำดับ แทนสถานการณ์ที่สัมประสิทธิ์สหสัมพันธ์ในประชากรมีค่า .2, .5 และ .8 ตามลำดับ การกำหนดค่าลักษณะนี้ช่วยให้จำลองข้อมูลสะดวกยิ่งขึ้น โดยผู้วิจัยจำลองค่าความสามารถที่แท้จริงให้มีจำนวนเท่ากันทั้งหมดตามเงื่อนไขของขนาดกลุ่มตัวอย่าง (NSAMP) แต่ละค่า

สำหรับส่วนเบี่ยงเบนมาตรฐานของค่าความสามารถที่แท้จริงของประชากรทุกกลุ่มมีค่าเป็น 0 แต่ค่าเฉลี่ยกลุ่มควบคุมจะมีค่าเป็น 1 สำหรับค่าเฉลี่ยของประชากรของกลุ่มทดลองจะมีค่าเฉลี่ยเท่ากับค่าขนาดอิทธิพลของกลุ่มนั้น ๆ การกำหนดค่าดังกล่าวมีที่มาจากสมการต่อไปนี้

$$\delta = \frac{\mu_E - \mu_c}{\sigma_{pooled}}$$

โดยที่ δ แทนค่าขนาดอิทธิพล

μ_E, μ_c แทนค่าเฉลี่ยของประชากรกลุ่มทดลองและกลุ่มควบคุมตามลำดับ

และ σ_{pooled} แทนส่วนเบี่ยงเบนมาตรฐานรวมของกลุ่มทดลองและกลุ่มควบคุม

เนื่องจากผู้วิจัยกำหนดให้ส่วนเบี่ยงเบนมาตรฐานรวมของกลุ่มทดลอง และกลุ่มควบคุมมีค่าเท่ากับ 1 ทำให้ σ_{pooled} มีค่าเป็น 1 ด้วย สมการข้างต้นจึงลดรูปเป็น $\delta = \mu_E - \mu_c$ แต่เนื่องจาก μ_c มีค่าเท่ากับ 0 จะได้ว่า $\delta = \mu_E$ กล่าวคือ ค่าเฉลี่ยของประชากรกลุ่มทดลองมีค่าเท่ากับค่าขนาดอิทธิพลนั่นเอง

2) สร้างข้อมูลค่าพารามิเตอร์ของข้อสอบ อันได้แก่ ค่าความยากสำหรับ 1PLM ค่าความยากและค่าอำนาจจำแนกสำหรับ 2PLM และค่าความยาก ค่าอำนาจจำแนกและค่าโอกาสในการเดาสำหรับ 3PLM ให้มีจำนวน 10, 50 และ 90 ค่า ตามเงื่อนไขของความยาวแบบสอบ

3) นำข้อมูลที่สร้างไว้ในข้อ 1 และ 2 มาคำนวณค่าความน่าจะเป็นที่ผู้สอบคนที่ i จะตอบข้อสอบข้อที่ j ถูกหรือค่า $P_j(\theta_i)$ ตามโมเดลฐานที่ใช้ในการจำลองข้อมูล (1PLM, 2PLM, 3PLM)

4) นำข้อมูลจากขั้นตอนที่ 3 มาสร้างคำตอบของผู้สอบในแต่ละสถานการณ์ โดยสุ่มค่าความน่าจะเป็น ($PROB_{ij}$) สำหรับผู้สอบคนที่ i ในการทำข้อสอบข้อที่ j เพื่อใช้เป็นค่าเปรียบเทียบ กับค่าความน่าจะเป็นที่ผู้สอบคนที่ i จะตอบข้อสอบข้อที่ j ถูกหรือ $P_j(\theta_i)$ และกำหนดเป็นค่าคำตอบของผู้สอบ (U_{ij}) ซึ่งเป็นไปได้ 2 กรณีคือ กรณีที่ 1 $PROB_{ij}$ มีค่ามากกว่า $P_j(\theta_i)$ จะกำหนดให้ U_{ij} มีค่าเป็น 0 หรือ กรณีที่ 2 $PROB_{ij}$ มีค่าน้อยกว่าหรือเท่ากับ $P_j(\theta_i)$ จะกำหนดให้ U_{ij} มีค่าเป็น 1 โดยข้อมูลคำตอบในส่วนนี้จะจัดเก็บแยกเป็นแฟ้มข้อมูลแบบอักขระ (ASCII file) เพื่อนำมาประมาณค่าความสามารถตามทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบสนองข้อสอบ ด้วยโปรแกรม BILOG 3.0 และนำผลที่ได้ไปวิเคราะห์ข้อมูลเพื่อตอบถามวิจัยต่อไป

อนึ่ง ก่อนการวิเคราะห์ด้วยโปรแกรม BILOG ผู้วิจัยจะตรวจสอบความเป็นเอกมิติซึ่งเป็นข้อตกลงเบื้องต้นที่สำคัญของทฤษฎีการตอบสนองข้อสอบที่ตรวจให้คะแนนแบบทวิภาค (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000) โดยการคำนวณอัตราส่วนของค่าไอเกน (Eigen Ratio: ER) ขององค์ประกอบที่ 1 และ 2 ที่ได้จากการวิเคราะห์องค์ประกอบ หากอัตราส่วนของค่าไอเกนมีค่าต่ำกว่า 3 แสดงว่าแบบสอบไม่เป็นไปตามข้อตกลงว่าด้วยความเป็นเอกมิติ ผู้วิจัยจะทำการจำลองสถานการณ์นั้นใหม่

3. การวิเคราะห์ข้อมูล

1) นำคะแนนความสามารถของผู้สอบมาคำนวณ d_{CTT}, d_{IRT1}, d_{IRT2}, r_{CTT}, r_{IRT1} และ r_{IRT2}

2) คำนวณสถิติเชิงบรรยายของ d_{CTT}, d_{IRT1}, d_{IRT2}, r_{CTT}, r_{IRT1} และ r_{IRT2} พร้อมทั้งเปรียบเทียบค่าเฉลี่ยระหว่าง d_{CTT}, d_{IRT1} และ d_{IRT2} และระหว่าง r_{CTT}, r_{IRT1} และ r_{IRT2} โดยการวิเคราะห์ความแปรปรวนทางเดียว (one-way ANOVA)

3) คำนวณค่า RMSD, SRMSD, BIAS และ $\triangle$ เพื่อใช้ในการพิจารณาคุณสมบัติของตัวประมาณค่าที่ดีในแต่ละสถานการณ์โดยใช้เกณฑ์ตามที่กำหนดไว้ในนิยาม

4) วิเคราะห์ความสัมพันธ์ระหว่าง d_{CTT} และ d_{IRT1} และสร้างสมการถดถอยของ d_{IRT1} บน d_{CTT}

5) วิเคราะห์ความสัมพันธ์ระหว่าง r_{CTT} และ r_{IRT1} และสร้างสมการถดถอยของ r_{IRT1} บน r_{CTT}

ผลการวิเคราะห์ข้อมูล

สำหรับการเปรียบเทียบคุณสมบัติของตัวประมาณค่าระหว่าง d_{CTT}, d_{IRT1}, d_{IRT2}, r_{CTT}, r_{IRT1} และ r_{IRT2} พบว่า r_{IRT1} มีความลำเอียงต่ำสุด รองลงมาคือ r_{IRT2}, d_{IRT1}, d_{IRT2}, d_{CTT} และ r_{CTT} ตามลำดับ โดยตัวประมาณค่าทั้งหมดให้ค่าประมาณต่ำกว่าค่าความเข้มของอิทธิพลที่แท้จริง ในด้านความคงเส้นคงวา พบว่า r_{CTT} มีความคงเส้นคงวาสูงสุด รองลงมาคือ d_{CTT}, d_{IRT1}, r_{IRT1}, d_{IRT2} และ r_{IRT2} ตามลำดับ นอกจากนี้ยังพบว่า r_{IRT1} มีประสิทธิภาพสัมพัทธ์สูงสุด รองลงมาคือ r_{IRT2}, d_{IRT1}, d_{IRT2}, r_{CTT} และ d_{CTT} ตามลำดับ

ผลการพิจารณาความลำเอียง ความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์ของตัวประมาณค่าทั้ง 6 ชนิดไม่พบว่าตัวประมาณค่าชนิดใดมีคุณสมบัติทั้ง 3 ด้านดีที่สุด แต่พบว่า r_{IRT1} เป็นตัวประมาณค่าที่มีคุณสมบัติทั้ง 3 ด้านอยู่ในระดับที่น่าพอใจที่สุด เนื่องจาก r_{IRT1} เป็นตัวประมาณค่าที่มีความลำเอียงต่ำที่สุดและมีประสิทธิภาพสัมพัทธ์สูงสุด แม้ว่า r_{IRT1} จะมิได้เป็นตัวประมาณค่าที่มีความคงเส้นคงวาสูงสุดแต่เป็นตัวประมาณค่าที่มีความคงเส้นคงวาเป็นอันดับ 3 ในขณะที่ d_{CTT} และ r_{CTT} ซึ่งมีความคงเส้นคงวาสูงกว่า r_{IRT1} นั้นมีความลำเอียงสูงกว่า r_{IRT1} และประสิทธิภาพต่ำกว่า r_{IRT1} ร้อยละ 50 โดยประมาณ รายละเอียดดังตารางที่ 1

ตารางที่ 1 ผลการเปรียบเทียบคุณสมบัติของตัวประมาณค่า 6 ชนิดในภาพรวม

	สถิติ	ตัวประมาณค่า					
		d_{CTT}	d_{IRT1}	d_{IRT2}	r_{CTT}	r_{IRT1}	r_{IRT2}
ความลำเอียง	SRMSD	1.084	.937	.975	1.291	.777	.849
		5	3	4	6	1	1
	BIAS	-.743	-1.036	-1.053	-.248	-.389	-.399
ความคงเส้นคงวา	△	.129	.001	-.018	.187	.009	-.038
		2	4	5	1	3	6
ประสิทธิภาพสัมพัทธ์	VRA	.567	.006	.012	.124	.002	.004
		6	3	4	5	1	2

ผลการเปรียบเทียบคุณสมบัติของตัวประมาณค่าระหว่างขนาดอิทธิพล (d) และสัมประสิทธิ์สหสัมพันธ์ (r) ในภาพรวมตามตัวแปรเงื่อนไขในการจำลองข้อมูล พบว่า แม้ในบางเงื่อนไขค่าส่วนเบี่ยงเบนของรากกำลังสองเฉลี่ยมาตรฐาน (SRMSD) ของขนาดอิทธิพลและ

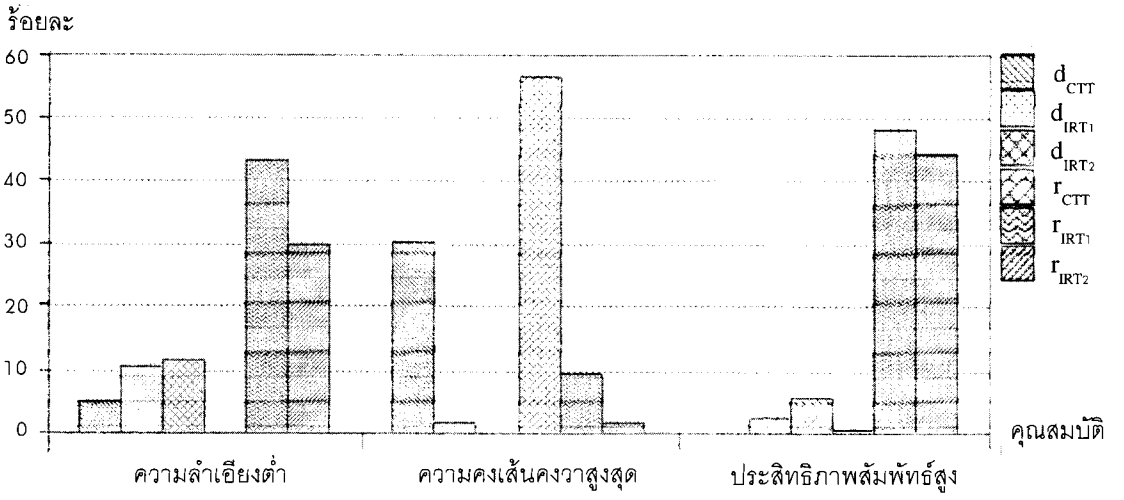
สัมประสิทธิ์สหสัมพันธ์จะมีค่าแตกต่างกัน แต่ยังคงมีค่าใกล้เคียงกัน ในบางเงื่อนไขจะพบว่าส่วนเบี่ยงเบนของรากกำลังสองเฉลี่ยมาตรฐานของขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากัน ขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์จึงมีความลำเอียงใกล้เคียงกันมาก อย่างไรก็ดีในเงื่อนไขส่วนใหญ่จะพบว่าสัมประสิทธิ์มีความคงเส้นคงวาและประสิทธิ์ภาพสัมพัทธ์สูงกว่าขนาดอิทธิพล ดังผลสรุปในตารางที่ 2

ตารางที่ 2 สรุปผลการเปรียบเทียบคุณสมบัติของขนาดอิทธิพล (d) และสัมประสิทธิ์สหสัมพันธ์ (r) ในภาพรวมของแต่ละเงื่อนไข

เงื่อนไข		ความลำเอียงต่ำ[1]	ความคงเส้นคงวาสูง	ประสิทธิภาพสัมพัทธ์สูง
ความเข้มของอิทธิพลที่แท้จริง	.2	d ≈ r	r	r
	.5	d ≈ r	r	r
	.8	d ≈ r	r	r
	1.2	d ≈ r	d	r
	2.6	d ≈ r	d	r
ขนาดกลุ่มตัวอย่าง	20	d ≈ r	ไม่สามารถคำนวณได้ เนื่องจากเป็นสถานการณ์ที่มีกลุ่มเพียงขนาดเดียว	r
	50	d ≈ r		r
	500	d ≈ r		r
	2000	d ≈ r		r
ความยาวแบบสอบ	10	d ≈ r	r	r
	50	d ≈ r	r	r
	90	d ≈ r	r	r
โมเดลฐาน	1PLM	d ≈ r	d	r
	2PIM	d ≈ r	r	r
	3PLM	d ≈ r	r	r
โมเดลประมาณค่า	CTT	d ≈ r	r	r
	IRT1	d ≈ r	r	r
	IRT2	d ≈ r	d	r

หมายเหตุ [1] d ≈ r หมายถึง ความลำเอียงของ d มีค่าใกล้เคียงกับความลำเอียงของ r
d = r หมายถึง ความลำเอียงของ d มีค่าเท่ากับความลำเอียงของ r

สำหรับการเปรียบเทียบคุณสมบัติของตัวประมาณค่าของ d_{CTT}, d_{IRT1}, d_{IRT2}, r_{CTT}, r_{IRT1} และ r_{IRT2} ในด้านความลำเอียง ความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์ในสถานการณ์ย่อย พบว่ามีความแตกต่างกันไปตามเงื่อนไขของแต่ละสถานการณ์ ในเบื้องต้นผู้วิจัยคำนวณร้อยละของสถานการณ์ย่อยที่พบคุณสมบัติของตัวประมาณค่าที่ดีในแต่ละด้านของตัวประมาณค่าแต่ละตัว จาก 180 สถานการณ์ย่อยพบว่า ร้อยละ 43 เป็นสถานการณ์ที่ r_{IRT1} มีความลำเอียงต่ำกว่าตัวประมาณค่าชนิดอื่น รองลงมาคือ r_{IRT2} และ d_{IRT2} คิดเป็นร้อยละ 30 และ 12 ตามลำดับ ในด้านความคงเส้นคงวา พบว่า ร้อยละ 57 เป็นสถานการณ์ที่ r_{CTT} มีความคงเส้นคงวาสูงสุด รองลงมาคือ d_{CTT} และ r_{IRT1} คิดเป็นร้อยละ 30 และ 9 ตามลำดับ สำหรับประสิทธิ์ภาพสัมพัทธ์ พบว่า ร้อยละ 48 เป็นสถานการณ์ที่ r_{IRT1} มีประสิทธิภาพสัมพัทธ์สูงสุด รองลงมาคือ r_{IRT2} และ d_{IRT2} คิดเป็นร้อยละ 44 และ 6 ตามลำดับ รายละเอียดดังภาพที่ 2



ภาพที่ 2 ร้อยละของสถานการณ์ที่พบคุณสมบัติของตัวประมาณค่าที่ดีในตัวประมาณค่า

สำหรับการเปรียบเทียบค่าเฉลี่ยระหว่าง d_{CTT}, d_{IRT1} และ d_{IRT2} พบว่า ค่าเฉลี่ยของ d_{CTT}, d_{IRT1} และ d_{IRT2} มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดย d_{CTT} มีค่าเฉลี่ยสูงสุด รองลงมาคือค่าเฉลี่ยของ d_{IRT1} และ d_{IRT2} ตามลำดับ ในทำนองเดียวกันพบว่า ค่าเฉลี่ยของ r_{CTT}, r_{IRT1} และ r_{IRT2} มีความแตกต่างกันที่ระดับนัยสำคัญทางสถิติ .05 โดย r_{CTT} มีค่าเฉลี่ยสูงสุด รองลงมาคือค่าเฉลี่ยของ r_{IRT1} และ r_{IRT2} ตามลำดับ รายละเอียดดังตารางที่ 3 และ 4

ตารางที่ 3 ผลการเปรียบเทียบค่าเฉลี่ยของตัวประมาณค่าระหว่างทฤษฎีการทดสอบ

ตัวแปรต้น	df_{BG}	df_{WG}	Levene.	sig.	SS_{BG}	SS_{WG}	MS_{BG}	MS_{WG}	F	sig.
MUSED	2	53997	14087.510	.000	1093.103	10532.708	546.552	.195	2801.953	.000
			18976.711	.000	257.396	2356.602	128.698	.044	2948.870	.000

หมายเหตุ ค่าสถิติเหนือเส้นประเป็นของขนาดอิทธิพล (d) ค่าสถิติใต้เส้นประเป็นของสัมประสิทธิ์สหสัมพันธ์ (r)

ตารางที่ 4 ผลการทดสอบภายหลัง (post hoc) การเปรียบเทียบค่าเฉลี่ยของตัวประมาณค่า

ตัวแปรต้น	คู่ที่แตกต่างกันอย่างมีนัยสำคัญที่ระดับ .05
MUSED	$d_{CTT} > d_{IRT1} > d_{IRT2}$ และ $r_{CTT} > r_{IRT1} > r_{IRT2}$

สำหรับการวิเคราะห์ความสัมพันธ์ระหว่าง d_{CTT} และ d_{IRT1} พบว่า เป็นความสัมพันธ์เชิงเส้นทางบวก โดยสัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากับ .626 ขณะที่ความสัมพันธ์ระหว่าง r_{CTT} และ r_{IRT1} เป็นความสัมพันธ์เชิงเส้นทางบวกเช่นกันแต่มีค่าสัมประสิทธิ์สหสัมพันธ์ .570 เมื่อทำการวิเคราะห์การถดถอยของ d_{IRT1} บน d_{CTT} และการถดถอยของ r_{IRT1} บน r_{CTT} พบว่า ความแปรปรวนของ d_{CTT} สามารถทำนายความแปรปรวนของ d_{IRT1} ได้ร้อยละ 39.2 ในขณะที่ความแปรปรวนของ r_{CTT} สามารถทำนายความแปรปรวนของ r_{IRT1} ได้ร้อยละ 32.5 โดยสมการถดถอยในรูปคะแนนดิบคือ $d_{IRT1} = .004 + .065d_{CTT}$ และ $r_{IRT1} = .003 + .079r_{CTT}$ สำหรับสมการในรูปคะแนนมาตรฐานคือ $Z_{d_{IRT1}} = .626Z_{d_{CTT}}$ และ $Z_{r_{IRT1}} = .570Z_{r_{CTT}}$ ตามลำดับ รายละเอียดดังตารางที่ 5

ตารางที่ 5 สรุปผลการวิเคราะห์ความสัมพันธ์และการถดถอย

ตัวแปรเกณฑ์		b	SE	β	T	sig.	r	R^2
d_{IRT1}	ค่าคงที่	.004	.000	-	7.618	.000	.626	.392
	d_{CTT}	.065	.001	.626	107.640	.000		
r_{IRT1}	ค่าคงที่	.003	.000	-	8.994	.003	.570	.325
	r_{CTT}	.079	.001	.570	93.019	.079		

อภิปรายผล

1. ผลการเปรียบเทียบคุณสมบัติของ d_{CTT}, d_{IRT1}, d_{IRT2}, r_{CTT}, r_{IRT1} และ r_{IRT2} ซึ่งพบว่ามีคุณสมบัติในด้านความลำเอียง ความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์แตกต่างกัน โดยตัวประมาณค่าทุกตัวมีความลำเอียงและให้ค่าประมาณความเข้มของอิทธิพลที่ต่ำกว่าค่าความเข้มของอิทธิพลที่แท้จริง สอดคล้องกับผลการวิจัยของ **Roberts** และ **Henson** (2002) ที่พบว่าขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์ล้วนเป็นตัวประมาณค่าที่มีความลำเอียง และช่วยขยายข้อค้นพบของ **DeMars** (2001) ที่พบว่าขนาดอิทธิพลที่ได้จากโมเดลการตอบสนองข้อสอบที่โมเดลประมาณค่าสอดคล้องและไม่สอดคล้องกับข้อมูลมีความลำเอียงที่แตกต่างกันด้วย โดยพบว่าโมเดลการทดสอบแบบดั้งเดิม โมเดลการตอบสนองข้อสอบที่โมเดลประมาณค่าสอดคล้องกับข้อมูลและโมเดลการตอบสนองข้อสอบที่โมเดลประมาณค่าไม่สอดคล้องกับข้อมูลจะให้ค่าขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์ที่มีความลำเอียงต่างกัน สำหรับคุณสมบัติในด้านความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์ พบว่า r_{CTT} มีความคงเส้นคงวาสูงสุด และ r_{IRT1} มีประสิทธิภาพสัมพัทธ์สูงสุด นอกจากนี้ยังพบว่าสัมประสิทธิ์สหสัมพันธ์มีความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์สูงกว่าขนาดอิทธิพล เมื่อใช้โมเดลการทดสอบแบบดั้งเดิมหรือโมเดลการตอบสนองข้อสอบที่สอดคล้องกับข้อมูลในการประมาณค่าความสามารถของกลุ่มตัวอย่าง ผลการวิจัยในส่วนนี้ช่วยสนับสนุนแนวคิดของ **Rosenthal** และ **DiMatteo** (2001), **Baugh** (2002), **Thompson** (2002) และ **Ives** (2003) ที่เสนอว่าการประมาณค่าความเข้มของอิทธิพลในงานวิจัยเชิงปริมาณทั่วไปควรใช้สถิติในกลุ่มความสัมพันธ์

2. ผลการวิจัยในสถานการณ์ที่ใช้โมเดลการตอบสนองข้อสอบที่ไม่สอดคล้องกับข้อมูล ค่าผลต่างของส่วนเบี่ยงเบนของรากกำลังสองเฉลี่ยมาตรฐาน ($\triangle$) มีค่าเป็นลบ ทั้งนี้อาจเป็นผลมาจากการเลือกใช้โมเดลการประมาณค่าที่ไม่เหมาะสม ทำให้การประมาณค่าความสามารถของผู้สอบมีความคลาดเคลื่อนเกิดขึ้น (**ศิริชัย กาญจนวาสี**, 2545; **Embretson & Reise**, 2000) และเมื่อนำค่าประมาณความสามารถดังกล่าวมาประมาณค่าความเข้มของอิทธิพลจึงเกิดความคลาดเคลื่อนขึ้น และความคลาดเคลื่อนนี้มิได้ลดลงเมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น ทำให้ความคงเส้นคงวาของตัวประมาณค่าต่ำจนติดลบ

3. แม้ว่าผลการพิจารณาความลำเอียง ความคงเส้นคงวา และประสิทธิภาพสัมพัทธ์ของตัวประมาณค่าทั้ง 6 ชนิดจะไม่พบว่าตัวประมาณค่าชนิดใดมีคุณสมบัติทั้ง 3 ด้านดีที่สุด แต่พบว่า r_{IRT1} เป็นตัวประมาณค่าที่มีคุณสมบัติทั้ง 3 ด้านอยู่ในระดับที่น่าพอใจที่สุด เนื่องจาก r_{IRT1} เป็นตัวประมาณค่าที่มีความลำเอียงต่ำที่สุดและมีประสิทธิภาพสัมพัทธ์สูงสุด แม้ว่า r_{IRT1} จะมิได้เป็นตัวประมาณค่าที่มีความคงเส้นคงวาสูงสุดแต่เป็นตัวประมาณค่าที่มีความคงเส้นคงวาเป็นอันดับ 3 ใน

ขณะที่ d_{CTT} และ r_{CTT} ซึ่งมีความคงเส้นคงวาสูงกว่า r_{IRT} นั้นมีความลำเอียงสูงกว่า r_{IRT} และประสิทธิภาพต่ำกว่า r_{IRT} ร้อยละ 50 โดยประมาณ

4. ในการจำลองสถานการณ์ตามเงื่อนไข 5 เงื่อนไข ผู้วิจัยพบว่า แนวโน้มของค่าสถิติที่บ่งชี้ความลำเอียง (SRMSD) ของขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์ในภาพรวมของทุกเงื่อนไขเป็นไปในทางเดียวกัน คือขนาดอิทธิพลมีความลำเอียงมีค่าใกล้เคียงกับความลำเอียงของสัมประสิทธิ์สหสัมพันธ์ ทั้งนี้สอดคล้องกับแนวคิดของ Hedges และ Olkin (1985), Cohen (1988 อ้างถึงใน Thompson, 2002), Friedman (1968 อ้างถึงใน Thompson, 2002) และ Hunter และ Schmidt (1990) ที่แสดงให้เห็นว่าขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์มีความเกี่ยวข้องสัมพันธ์กันและสามารถแปลงรูประหว่างกันและกันได้ โดยผลการวิจัยในครั้งนี้นอกจากจะสนับสนุนแนวคิดดังกล่าวยังได้ขยายแนวคิดออกไปอีกด้วย เพราะผลการวิจัยชี้ให้เห็นว่าแม้จะใช้โมเดลการทดสอบที่แตกต่างกัน ขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์ที่ได้ยังคงมีความสัมพันธ์กันในทางบวกค่อนข้างสูง

5. ผลการวิจัยที่พบว่า d_{CTT} และ d_{IRT} มีค่าเฉลี่ยแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 สอดคล้องกับสมมติฐานที่ผู้วิจัยตั้งไว้ แต่ไม่สอดคล้องกับผลการวิจัยของ Wang และ Chen (2004) ทั้งนี้อาจเป็นเพราะ Wang และ Chen ศึกษาจากขนาดกลุ่มตัวอย่างเพียง 50 คนซึ่งอาจจะส่งผลต่อการประมาณค่าตาม IRT เนื่องจากการประมาณค่าด้วยโมเดลการตอบสนองข้อสอบจำเป็นต้องใช้ข้อมูลขนาดใหญ่ ดังที่ Hambleton และ Swaminathan (1985) เสนอว่าในกรณีที่มีผู้สอบน้อยกว่า 200 คนเหมาะที่จะใช้โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์เท่านั้น แต่การศึกษาของ Wang และ Chen ใช้โมเดลโลจิสติกแบบสามพารามิเตอร์ทั้งที่มีผู้สอบเพียง 50 คน

6. สำหรับความสัมพันธ์ระหว่างขนาดอิทธิพลที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิม และขนาดอิทธิพลที่ได้จากทฤษฎีการตอบสนองข้อสอบ และความสัมพันธ์ระหว่างสัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิม และสัมประสิทธิ์สหสัมพันธ์ที่ได้จากทฤษฎีการตอบสนองข้อสอบ พบว่าเป็นความสัมพันธ์เชิงเส้นในทางบวก และมีนัยสำคัญทางสถิติที่ระดับ .05 ผลที่ได้นี้สอดคล้องกับผลการวิจัยของ Fan (1998), MacDonald และ Paunonen (2002) และ Ndalichako และ Rogers (1997) ที่พบว่าค่าสถิติที่ได้จากการวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบสนองข้อสอบ อาทิ ค่าความสามารถของผู้สอบ และค่าความยากมีความสัมพันธ์กันระหว่างทฤษฎีการทดสอบในทางบวกค่อนข้างสูง

ข้อเสนอแนะสำหรับการนำไปใช้

1. ผลการวิจัยในครั้งนี้พบว่า ในสถานการณ์ทั่วไปสัมประสิทธิ์สหสัมพันธ์เป็นตัวประมาณค่าความเข้มของอิทธิพลที่มีคุณสมบัติการเป็นตัวประมาณค่าที่ดีเหนือกว่าขนาดอิทธิพล ดังนั้นนักวิจัยที่ต้องการศึกษาอิทธิพลของตัวแปรต้นที่มีต่อตัวแปรตามประเภทคะแนนความสามารถของผู้สอบที่มีการให้คะแนนแบบสองค่า แต่ไม่สามารถออกแบบการวิจัยด้วยระเบียบวิธีวิจัยเชิงทดลองได้ อาจใช้ระเบียบวิธีวิจัยเชิงสหสัมพันธ์และเลือกใช้สัมประสิทธิ์สหสัมพันธ์เป็นตัวประมาณค่าความเข้มของอิทธิพลแทน

2. ความสัมพันธ์ของขนาดอิทธิพลที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิมและขนาดอิทธิพลที่ได้จากทฤษฎีการตอบสนองข้อสอบ และความสัมพันธ์ของสหสัมพันธ์ที่ได้จากทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบสนองข้อสอบเป็นความสัมพันธ์ทางบวก ช่วยสนับสนุนให้การวิจัยเชิงปริมาณทางการศึกษาที่มีข้อจำกัดเกี่ยวกับขนาดของกลุ่มผู้สอบ สามารถศึกษาอิทธิพลได้จากการวิเคราะห์ข้อมูลที่วัดตามทฤษฎีการทดสอบแบบดั้งเดิม ซึ่งไม่ต้องใช้จำนวนกลุ่มผู้สอบขนาดใหญ่ได้อย่างสะดวก โดยผลการวิจัยยังคงสอดคล้องกับการวิเคราะห์ด้วยความสามารถของผู้สอบตามทฤษฎีการตอบสนองข้อสอบ

ข้อเสนอแนะสำหรับการวิจัยในอนาคต

1. การจำลองค่าความสามารถของกลุ่มทดลองในกรณีของสัมประสิทธิ์สหสัมพันธ์ ผู้วิจัยใช้ความสัมพันธ์ระหว่างขนาดอิทธิพลและสัมประสิทธิ์สหสัมพันธ์ตามที่ Hunter และ Schmidt (1990) เสนอ ผู้ที่สนใจอาจจำลองข้อมูลโดยไม่ใช้ความสัมพันธ์ดังกล่าว อาจจำลองข้อมูลให้คะแนนความสามารถของผู้สอบสองกลุ่มมีความสัมพันธ์กันเท่ากับค่าสัมประสิทธิ์สหสัมพันธ์ที่ต้องการศึกษา เช่น .2, .5 และ .8 ซึ่งเป็นการกำหนดค่าความเข้มของอิทธิพลที่แท้จริงในรูปของสัมประสิทธิ์สหสัมพันธ์ก่อน จากนั้นจึงแปลงให้เป็นขนาดอิทธิพล ผลการวิจัยตามแนวทางดังกล่าวสามารถนำมาเปรียบเทียบกับผลการวิจัยครั้งนี้

2. การเปรียบเทียบคุณสมบัติของตัวประมาณค่าความเข้มของอิทธิพลสำหรับการวิจัยที่มีกลุ่มตัวอย่างมากกว่า 2 กลุ่ม เช่น d (West & Wiratchai, 1984 อ้างถึงใน นงลักษณ์ วิรัชชัย, 2542) η^2 และ ω^2 หรือการเปรียบเทียบระหว่างตัวประมาณค่าในกลุ่มความแตกต่างระหว่างกลุ่ม (group difference) อาทิ Cohen's d กับ ตัวประมาณค่าในกลุ่ม (group overlap) อาทิ พื้นที่ร่วมโค้งการกระจาย และดัชนี I ของ Huberty (2002) ซึ่งตัวประมาณค่าทั้ง 2 กลุ่มนี้มีแนวคิดที่ตรงกันข้ามกันชัดเจน การเปรียบเทียบเหล่านี้เป็นประเด็นที่ควรทำการศึกษาต่อไป

3. ในการทบทวนเอกสารและงานวิจัยที่เกี่ยวข้อง ผู้วิจัยพบว่ามีงานวิจัยที่เปรียบเทียบค่าสถิติที่ได้จากการวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม และทฤษฎีการตอบสนองข้อสอบจำนวนมาก การสังเคราะห์งานวิจัยดังกล่าวด้วยการวิเคราะห์อภิมานน่าจะก่อให้เกิดประโยชน์อย่างมาก โดยเฉพาะการหาข้อสรุปรวมเกี่ยวกับผลการเปรียบเทียบหรืออาจจะทำให้ได้ตัวแปรปรับที่ทำให้ผลการเปรียบเทียบแตกต่างกัน

เอกสารอ้างอิง

นงลักษณ์ วิรัชชัย. (2542). *การวิเคราะห์อภิมาน: META-ANALYSIS.* ปทุมวัน, กทม.: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

นภดล ยิ่งยงสกุล. (2539). *การศึกษาค่าความเที่ยงความตรงและความสัมพันธ์ของคะแนนสอบระหว่างการให้คะแนนตามทฤษฎีการทดสอบดั้งเดิมกับการให้คะแนนตามทฤษฎีการตอบสนองข้อสอบ.* วิทยานิพนธ์ปริญญามหาบัณฑิต บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลา-นครินทร์.

นิคม กีรติวรางกูร. (2542). *การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด แมนเทล-แฮนส์เซลและการตอบสนองข้อสอบ.* วิทยานิพนธ์ปริญญามหาบัณฑิต บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

เบญจพร ยนตจักรวิถี. (2539). *การเปรียบเทียบผลการวิเคราะห์ข้อสอบวัดผลสัมฤทธิ์วิชาคณิตศาสตร์ชั้นมัธยมศึกษาปีที่ 1 ระหว่างทฤษฎีการทดสอบดั้งเดิมและทฤษฎีการตอบสนองข้อสอบ.* วิทยานิพนธ์ปริญญามหาบัณฑิต บัณฑิตวิทยาลัย มหาวิทยาลัย สงขลานครินทร์.

วีระพันธ์ พรหมบุตร. (2536). *การศึกษาความสัมพันธ์ระหว่างค่าพารามิเตอร์ที่ได้จากการวิเคราะห์ข้อสอบโดยทฤษฎีดั้งเดิมกับทฤษฎีการตอบสนองข้อสอบ.* ปริญญานิพนธ์มหาบัณฑิต บัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร.

ศิริชัย กาญจนวาสี. (2541). การวิเคราะห์ส่วนประกอบความแปรปรวนทางการศึกษา. วิธีวิทยาการวิจัย 1. (มกราคม-มิถุนายน 2541): 20-26.

ศิริชัย กาญจนวาสี. (2545). *ทฤษฎีการทดสอบแนวใหม่ (MODERN TEST THEORIES).* ปทุมวัน, กทม.: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สุกัญญรัตน์ คงงาม. (2539). *การเปรียบเทียบคุณสมบัติของตัวประมาณค่าพารามิเตอร์ที่ได้จากกลุ่มตัวอย่างสุ่มแบบหลายขั้นตอนระหว่างวิธีสุ่มแบบง่ายกับแบบมีระบบ.* วิทยานิพนธ์ปริญญามหาบัณฑิต บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

อรวรรณ สุขโต. (2542). *การเปรียบเทียบผลการวิเคราะห์ข้อสอบของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาคณิตศาสตร์ที่มีรูปแบบการตอบและวิธีการวิเคราะห์ข้อสอบต่างกัน.* วิทยานิพนธ์ปริญญามหาบัณฑิต บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.

อวยพร เรืองตระกูล. (2544). *การพัฒนาและวิเคราะห์คุณภาพของวิธีการวัดคะแนนพัฒนาการตามทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบสนองข้อสอบ.* วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

Baugh, F. (2002). Correcting effect sizes for score reliability: a reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement,* 62: 254-263.

Dawber, T., Rogers, W. T., & Carbonaro, M. (2004). *Robustness of lord's formulas for item difficulty and discrimination conversions between classical and item response theory models [Online].* Available from http://www.education.ualberta.ca/educ/psych/crame/ files/AERA2004TD.pdf [2004, August, 1]

DeMars, C. (2001). Group differences based on IRT scores: Does the model matter? *Educational and Psychological Measurement,* 62: 783-801.

Embretson, E. S., & Reise, P. S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement,* 58: 357-381.

Glass, G. V., & Hopkins, K. D. (1995). *Statistical methods in education and psychology* (3nd ed.). Needham Heights, MA: Allyn & Bacon.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing: what do the textbooks say? *The Journal of Experimental Education,* 71: 83-92.

Hambleton, K. R., & Swaminathan, H. (1985). *Item response theory.* Norwell, MA: Kluwer Academic.

Harwell, M., et al. (1996). Monte Carlo studies in item response theory. *Applied psychological Measurement,* 20: 101-125.

Hedges, V. L., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement,* 62: 227-240.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: correcting error and bias in research findings.* Newbury Park, CA: Sage.

Ives, B. (2003). Effect size use in studies of learning disabilities. *Journal of Learning Disabilities,* 36: 490-501.

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement,* 56: 746-759.

Kirk, R. E. (2001). Promoting good statistical practices: some suggestions [Online]. Available from http://www.d.umn.edu/~rvaidyan/mba8211/Promoting_Good_Statistical.pdf [2004, July 14]

MacDonald, P., & Paunonen, S., V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement,* 62: 921-943.

Mandenhall, W., & Beaver, R. J. (1994). *Introduction to probability and statistics* (9th ed.). Belmont, CA: Wadsworth Publishing.

Ndalichako, J. L., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. Educational and Psychological Measurement, 57: 580-589.

Paul, M. K., & Plucker, A. J. (2004). Two steps forward, one step back: Effect size reporting in gifted education research from 1995-2000. *Roeper Review,* 26: 68-72.

Pelton, W. T. (2002). *The accuracy of unidimensional measurement models in the presence of deviations from the underlying assumptions.* A Dissertation Presented to the Department of Instructional Psychology and Technology In Partial Fulfillment of the Requirements For the Degree of Doctor of Philosophy, Brigham Young University [Online]. Available from http://web.uvic.ca/~tpelton/timdissertation.pdf [2004, July 26]

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement,* 62: 241-253.

Rosenthal, R., & Dimatteo, M. R. (2001). Meta-analysis: recent developments in quantitative methods for literature reviews. *Annual Review of Psychology,* 52: 59-82.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61:* 293-316.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61:* 334-349.

Stage, C. (1998). *A comparison between item analysis based on item response theory and classical test theory: a study of the SweSAT subtest WORD* [Online]. Available from http://www.umu.se/edmeas/publikationer/pdf/enr2998sec.pdf [2004, June, 12]

Stage, C. (2003). *Classical test theory or item response theory: the SWEDICH experience* [Online]. Available from http://www.umu.se/edmeas/publikationer/pdf/em%20no%20 42.pdf [2004, June, 12]

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal for applied measurement,* 5: 48-61.

Thompson, B. (2000). A suggested revision to the forthcoming 5th edition of the APA Publication Manual [Online]. Available from http://www.coe.tamu.edu/~bthomson/apaeffec.htm [2004, June 12]

Thompson, B. (2002). "Statistical," "practical," and "clinical": how many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80:* 64-71.

Trusty, J., Thompson, B., & Petrocelli, V. J. (2004). Practical guide for reporting effect size in quantitative research in the journal of counseling & development. *Journal of Counseling & Development, 82:* 107-110.

Wang, W.-C., & Chen, H.-C. (2004). The standardized mean difference within the framework of item response theory. *Educational and Psychological Measurement, 64:* 201-223.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). *Statistical methods in psychology journals: Guidelines and explanations* [Online]. Available from http://www.apa.org/journals/amp/amp548594.html [2004, June 9]