

การพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการตัดสินของผู้เชี่ยวชาญ

A Development of the Differential Item Functioning Detection Methods by Expert Judgment

ชัยวัฒน์ หฤทัยพันธ์ และคณะ

บทคัดย่อ

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง มีวัตถุประสงค์ 2 ประการ คือ ประการแรก เพื่อพัฒนาวิธีการสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ และประการที่สองเพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในด้านอัตราความถูกต้อง และอัตราความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับข้อสอบที่คัดสรรมาสำหรับการทดลองเมื่อใช้วิธีการวิเคราะห์ด้วยแบบวินิจัยโดยผู้เชี่ยวชาญ วิธีการประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ และวิธีการประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ม ตัวอย่างที่ใช้ในการวิจัย คือ ผู้เชี่ยวชาญจำนวน 21 คน และนักเรียนระดับมัธยมศึกษาปีที่ 6 ปีการศึกษา 2556 จำนวน 139 คน ซึ่งได้จากการเลือกตัวอย่างแบบเจาะจง เครื่องมือที่ใช้ในการวิจัยประกอบด้วยแบบวินิจัยการทำหน้าที่ต่างกันของข้อสอบจากผู้เชี่ยวชาญ แบบยืนยันการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยประยุกต์ใช้เทคนิคการประชุมแบบเดลฟาย แบบสอบถามสำหรับการตรวจสอบความลำเอียงของข้อสอบสำหรับนักเรียน ชุดข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับผู้เชี่ยวชาญ ข้อสอบที่คัดสรรมาได้นำมาผ่านการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทิล-แฮนส์เชล ด้วยโปรแกรม DDFS 1.0 และโปรแกรม DIFAS 5.0

ผลการวิจัยพบว่า 1) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการตัดสินของผู้เชี่ยวชาญที่สำคัญมี 3 วิธี ได้แก่ วิธีที่ 1 การวินิจัยการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ และวิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ม 2) ข้อสอบที่ทำหน้าที่ต่างกันด้านเพศของแบบสอบถามการเรียนรู้สุขศึกษาและพลศึกษาจากผลการวิเคราะห์เปรียบเทียบอัตราความถูกต้องระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติกับการตรวจสอบโดยผู้เชี่ยวชาญพบว่า วิธีที่ 1 การตรวจสอบด้วยแบบวินิจัยโดยผู้เชี่ยวชาญ มีอัตราความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 50 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ

50 วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ มีอัตราความถูกต้องตามฉันทามติจากกลุ่มผู้เชี่ยวชาญคิดเป็นร้อยละ 0 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบคิดเป็นร้อยละ 100 วิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด มีอัตราความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 25 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ 75

คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ / แบบวินิจฉัยโดยผู้เชี่ยวชาญ / การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ / การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด

Abstract

This study is an experimental research conducted with two objectives. The first objective was to develop a differential item functioning (DIF) detected method by experts and the second objective was to compared the effectiveness of DIF detected in terms of accuracy rates and error rates in the outcomes from DIF detected and test questions selected for the experiment when using analysis by expert diagnosis and implementation of the Delphi Conference technique by a group of experts with the Think Aloud Protocols (TAPs). The sample group for this study was composed of 21 experts and 139 Grade 12 students during the 2013 academic year selected through purposive sampling. The research instrumentation was conducted by diagnosing DIF by implementing the Delphi Conference technique, a questionnaire for testing question bias for students, the substance health education and the physical education learning test for DIF detection for experts. The selected questions were obtained through DIF analysis by the Mantel – Haenszel method using the DDFS 1.0 and DIFAS 5.0 programs.

According to the research findings: 1) A DIF detection method by experts judgment 3 methods such as the first method, detected with the diagnosis by experts the second method, implementation of the Delphi Conference technique by experts the third method, implementation of TAPs 2) Concerning the test questions on gender with DIF in the substance of health education and physical education learning test, according to the analysis to compare accuracy rates between DIF detection methods by statistical means and detection by experts. The first method, detection with the diagnosis form by experts, was found to have a mean accuracy rate of 50% with a mean error rate of DIF detection at 50%. Furthermore, in the second method, the implementation of the Delphi Conference technique by experts, showed an accuracy rate rendered by the consensus of a group of

experts at 0% and an error rate of DIF detection at 100%, while the third method, implementation of TAPs, had a mean accuracy rate of 25% and a mean error rate of DIF detection of 75%.

KEYWORDS: DIFFERENTIAL ITEM FUNCTIONING (DIF) / DETECTION BY EXPERTS / APPLICATION OF THE DELPHI CONFERENCE BY EXPERTS / APPLICATION OF THINK ALOUD PROTOCOLS

บทนำ

ในการสอบต่างๆ เช่น การสอบคัดเลือกเข้าเรียนในระดับอุดมศึกษา การสอบคัดเลือกเข้าเรียนในสถาบันการศึกษาต่างๆ และการสอบคัดเลือกเพื่อบรรจุเข้าทำงานในหน่วยงานทั้งภาครัฐ และภาคเอกชนนั้น แบบสอบปรนัยเป็นเครื่องมือสำคัญอย่างหนึ่งที่นิยมใช้กันอย่างแพร่หลาย เนื่องจากมีลักษณะเด่นที่สามารถใช้กับผู้เข้าสอบได้ครั้งละเป็นจำนวนมาก ประหยัดเวลา และสามารถตรวจสอบความสามารถหรือคุณลักษณะแฝง (latent trait) ของผู้สอบได้ค่อนข้างตรงตามความต้องการ แต่การใช้แบบสอบก็ยังมีข้อจำกัดและเกิดปัญหาตามมา เนื่องจากแบบสอบบางฉบับประกอบด้วยข้อสอบที่ขาดความตรงหรือมีลักษณะที่แสดงถึงการทำหน้าที่ต่างกันของข้อสอบ คือมีผลต่อการตอบของผู้เข้าสอบแตกต่างกัน ซึ่งเมื่อนำแบบสอบไปทดสอบกับกลุ่มย่อยตั้งแต่ 2 กลุ่มขึ้นไปแล้วพบว่าผลการวิเคราะห์จากการตอบของผู้สอบที่มีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน แต่มีโอกาสในการตอบข้อสอบได้ถูกต้องไม่เท่ากัน นั่นคือ แบบสอบหรือข้อสอบเกิดปัจจัยเอื้อต่อผู้สอบบางกลุ่มเท่านั้น โดยแบบสอบมักจะประกอบด้วยคำถามบางข้อที่มีความแตกต่างออกไปจากสิ่งที่คนในกลุ่มบางกลุ่มคุ้นเคยหรือเป็นสิ่งที่ไม่

เคยรู้จักทำให้เกิดการได้เปรียบหรือเสียเปรียบกันขึ้น เช่น การให้นักเรียนหญิงคาดคะเนน้ำหนักของลูกฟุตบอล ซึ่งพบว่าข้อสอบลักษณะนี้เป็นข้อสอบที่ยาก และมีความลำเอียงเข้าข้างนักเรียนชายมากกว่านักเรียนหญิง เนื่องจากนักเรียนหญิงโดยทั่วไปมักจะไม่มีควมคุ้นเคยกับลูกฟุตบอลมากเท่ากับนักเรียนชาย (Hambleton & Rogers, 1989) หรือกรณีที่เมื่อนำข้อสอบหรือแบบสอบไปสอบกับกลุ่มผู้สอบ 2 กลุ่มขึ้นไปซึ่งมีความสามารถในระดับเดียวกันหรือมีคุณลักษณะแฝงที่ต้องการวัดเท่ากัน แต่อยู่ต่างกลุ่มกันแล้วผู้สอบกลุ่มใดกลุ่มหนึ่งได้รับโอกาสหรือการตอบสนองจากการตอบข้อสอบในข้อเดียวกันไม่เท่ากันทำให้เกิดการได้เปรียบหรือเสียเปรียบระหว่างกลุ่มย่อย จึงเป็นสาเหตุทำให้ผู้สอบกลุ่มนั้นได้รับประโยชน์มากกว่าทั้งที่มีความสามารถเท่ากันหรืออยู่ในระดับเดียวกัน ซึ่งสิ่งเหล่านี้ส่งผลโดยตรงต่อความตรงของข้อสอบซึ่งอาจสร้างปัญหาให้ทั้งผู้ออกข้อสอบและผู้เข้าสอบได้

ช่วงที่ผ่านมาเรามีหลายประเทศ เมื่อมีการสอบเกิดขึ้นผู้ที่เข้าสอบมักประกอบไปด้วยคนที่มีความหลากหลายกันทั้งทางด้านเชื้อชาติ ศาสนา ภาษา ประเพณี วัฒนธรรม ภูมิสำเนา สังคม เพศ อายุ ประสบการณ์ และเมื่อมีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะพบว่ามีข้อสอบที่ทำ

หน้าที่ต่างกันเสมอ ดังนั้น ปัญหาที่เกิดขึ้นตามมา คือ มักจะมีความขัดแย้งระหว่างหน่วยงานที่จัดสอบกับผู้เข้าสอบ และประเด็นที่สำคัญ คือ เป็นการลดความน่าเชื่อถือทั้งหน่วยงานที่จัดสอบ ข้อสอบ รวมถึงผู้ออกข้อสอบ จากสาเหตุดังกล่าว ทำให้นักการศึกษาหลายท่านพยายามแก้ไขปัญหาด้วยการศึกษาวิธีการทำหน้าที่ต่างกันของข้อสอบ และทดลองใช้วิธีการต่างๆ กันอย่างมากมาย แต่ถึงแม้ว่านักการศึกษาจะพยายามช่วยกันคิดหาวิธีการแก้ไขปัญหาด้วยพยายามตัดข้อสอบที่ทำหน้าที่ต่างกันออกไปเพื่อให้มีความยุติธรรมแก่ผู้เข้าสอบมากที่สุดก็ตามก็ยังเกิดปัญหาของการทำหน้าที่ต่างกันของข้อสอบเช่นเดิมอีก ในต่างประเทศมีการศึกษาเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) โดยทำการศึกษาข้อมูลการใช้ชุดโปรแกรมสำหรับการประเมินนักเรียนนานาชาติ (PISA) 2003 จำนวนข้อสอบที่ได้รับการตรวจพบว่ามีสถานะเกิดการทำหน้าที่ต่างกันของข้อสอบ DIF โดยผ่านวิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis: RFA), วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel: MH) และทฤษฎีการตอบสนองข้อสอบ ในการวิเคราะห์อัตราส่วน likelihood (IRT-LR) ได้ใช้วิธีการดังกล่าวตรวจสอบไปพร้อมกัน สำหรับการประเมินเนื้อหาโดยกลุ่มของผู้เชี่ยวชาญได้ข้อค้นพบ และรายงานว่ (1) ทักษะการคิดถูกวัดจากข้อสอบ (2) เกิดข้อผิดพลาดในการแปล และ (3) การใช้คำเชิงปริมาณซึ่งมีสามแหล่งสำคัญของการเกิดข้อผิดพลาดที่อาจจะเป็นสาเหตุให้เกิด DIF ในการศึกษา PISA 2003 ผลการวิจัยพบว่ากรณีวิเคราะห์ที่เป็นระบบในลำดับแรก ๆ ของเนื้อหาของข้อสอบอาจช่วยให้คาดการณ์ได้ว่าจะมีข้อสอบที่มี DIF ในการเปรียบเทียบ

ข้ามวัฒนธรรม ข้อตกลงระหว่างผู้เชี่ยวชาญเกี่ยวกับการตัดสินใจเกิด DIF และผลการวิเคราะห์การเกิด DIF ทางสถิติ เป็นการใช้องการตรวจสอบ DIF จากผู้เชี่ยวชาญและวิธีการตรวจสอบด้วยวิธีทางสถิติคู่ขนานกันไป (Hüseyin Hüsnü Yildirim & Giray Berberoğlu, 2009)

ในประเทศไทยแม้ว่าการเกิดปัญหาที่เกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบยังไม่เกิดความเสี่ยงมากนัก อีกทั้งผู้เข้าสอบเองก็ไม่ได้มีความสนใจเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบมากเท่าที่ควร แต่ก็ถือว่าเป็นสิ่งที่จะละเลยไม่ได้เนื่องจากเป็นประเทศหนึ่งที่มีความหลากหลายทางด้านเชื้อชาติ ศาสนา ภาษา วัฒนธรรม ภูมิสำนาลังคม เพศ อายุ ประสบการณ์ เช่นเดียวกัน (ศิริชัย กาญจนวาสี, 2550) ทั้งยังมีการจัดสอบหลายครั้งซึ่งผู้เข้าสอบมีความหลากหลายในด้านต่างๆ ข้างต้น คล้ายกับประเทศต่างๆ ในภูมิภาคเดียวกัน เช่น การสอบคัดเลือกเข้าศึกษาต่อในระดับต่างๆ เช่น ประถมศึกษา มัธยมศึกษา และในระดับอุดมศึกษา การสอบคัดเลือกเพื่อเข้ารับราชการ การสอบคัดเลือกเข้าเรียนต่อในโรงเรียนเตรียมทหารหรือการสอบคัดเลือกด้านต่างๆ จากการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอาจพบว่าเกิดปัญหาดังกล่าวคล้ายกับประเทศต่างๆ ประเด็นปัญหาที่กำลังร้อนระอุขึ้นในช่วงเวลาที่ผ่านมานี้ ก็คือกรณีที่นักเรียนชั้นมัธยมศึกษาปีที่ 6 นำตัวอย่างข้อสอบการทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน หรือ O-Net ปีการศึกษา 2555 ขึ้นโพสต์ผ่านเว็บไซต์ศูนย์ข่าวการศึกษาไทยโดยมีผู้เข้าสอบส่วนใหญ่สับสนกับข้อสอบ ซึ่งมีทั้งข้อสอบวิชาวิทยาศาสตร์ และวิชาสุศึกษา จากแหล่งข่าวอื่นๆ อีก เช่น Mthai news โดยมีการพาดหัวข่าวผ่านสื่อออนไลน์ ดังนี้ “ม. 6 อึ้ง!

ข้อสอบโอเน็ต ถ้ามเกิดอารมณ์ทางเพศต้องทำอย่างไร” บรรดาเด็กนักเรียนผู้เข้ารับการทดสอบทางการศึกษาระดับชาตินี้สามัญ (Ordinary National Educational Test: O-NET) กำลังวิพากษ์วิจารณ์ถึงลักษณะการออกข้อสอบที่มีการกำหนดข้อเลือกตอบที่ยากในการเลือกกว่าข้อไหนกันแน่ที่เป็นคำตอบที่ถูกต้อง อย่างเช่น ผู้ใช้ในเว็บไซต์เด็กดี ต่างปวดหัวกุมขมับกับแต่ละวิชา ในส่วนวิชาที่มีการวิพากษ์มากที่สุด คือวิชาสุขศึกษาและพลศึกษาที่ถามว่าหากเกิดอารมณ์ทางเพศขึ้นมาต้องทำอะไรโดยมีให้เลือก 4 ตัวเลือก คือ 1) ขวนเพื่อนไปเตะบอล 2) ปรีกษาครอบครัว 3) ไปเที่ยวกับเพื่อนต่างเพศ และ 4) ขวนเพื่อนสนิทไปดูหนัง ซึ่งเด็กส่วนใหญ่ไม่คาดคิดว่าจะมีการออกข้อสอบในลักษณะนี้ พร้อมกับตั้งคำถามว่าการตั้งคำถามลักษณะนี้เป็นการชี้นำอะไร ซึ่งคำตอบที่ถูกต้องที่นักเรียนต้องตอบได้แก่อข้อ 1) ขวนเพื่อนไปเตะบอล จากคำตอบดังกล่าวก็จะเป็นข้อสงสัยของนักเรียนหญิงทันทีว่านักเรียนหญิงโดยส่วนใหญ่แล้วก็จะไม่ชอบการเล่นกีฬาฟุตบอล จึงถือได้ว่าข้อสอบลักษณะดังกล่าวนี้เกิดการทำหน้าที่ต่างกันของข้อสอบทางด้านเพศ คือ เกิดความลำเอียงเข้าข้างเพศชายมากกว่าเพศหญิง ซึ่งมีการวิพากษ์วิจารณ์กันว่า ในช่วง 2-3 ปีมานี้ข้อสอบโอเน็ตมักจะแปลกประหลาดมากขึ้นทุกปี จึงต้องการคำอธิบายจากผู้ออกข้อสอบ ขณะที่บางคนก็มองเป็นเรื่องขำขัน คิดว่าผู้ออกข้อสอบคงไม่ยากให้เด็กเครียดกับการสอบมากเกินไป แต่กระนั้นกลับทำให้เครียดยิ่งกว่าเดิม จากประเด็นปัญหาดังกล่าว พบว่าการออกข้อสอบดังกล่าวเกิดปัญหาการทำหน้าที่ต่างกันของข้อสอบเกิดขึ้นโดยไม่มีกลุ่มผู้เชี่ยวชาญช่วยในการตรวจสอบข้อสอบที่เกิดปัญหาดังกล่าวก่อนนำมาใช้งานจริง ดังนั้น

ผู้วิจัยเห็นว่าควรมีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยอาศัยกลุ่มผู้เชี่ยวชาญได้ตรวจสอบข้อสอบไม่ให้เกิดปัญหาดังกล่าวก่อนนำมาจัดสอบให้กับนักเรียนจะได้ไม่ต้องเกิดปัญหากับเยาวชนของชาติอีกต่อไป

เมื่อได้ทำการศึกษาในเรื่องการทำหน้าที่ต่างกันของข้อสอบของประเทศไทย พบว่านักวิจัยในประเทศไทยได้ให้ความสำคัญในเรื่องการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาจากผลงานวิจัยฉบับต่างๆ ที่เกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบโดยศึกษาในเรื่อง ดังนี้ การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทิล-แฮนส์เซลกับวิธีชิปเทสต์เมื่อความยาวของแบบสอบขนาดกลุ่มตัวอย่าง และอัตราส่วนของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบต่างกัน (จิตติมา วรณศรี, 2539) การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด แมนเทิล-แฮนส์เซลและการตอบสนองข้อสอบ (นิคม กิรติวางกูร, 2542) การเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอนุกรูระหว่างวิธีชิปเทสต์ปรับใหม่ วิธีชิปเทสต์ วิธีแมนเทิล-แฮนส์เซล และวิธีการถดถอยโลจิสติก (วลีมาศ แซ่อึ้ง, 2543) สาเหตุของการทำหน้าที่ต่างกันของข้อสอบสาระการเรียนรู้ภาษาไทยและสาระการเรียนรู้สังคมศึกษา ศาสนา และวัฒนธรรม (สุมาลี แก้วทนต์, 2547) การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ตรวจให้คะแนนแบบพหุภาคระหว่างวิธีโครงสร้างความแปรปรวนร่วมและค่าเฉลี่ยกับวิธีการวิเคราะห์ฟังก์ชันเชิงจำแนกแบบ

โลจิสติก (วริษา ชะม้อย, 2550)

จากงานวิจัยในข้างต้นพบว่า นักวิจัยโดยส่วนใหญ่จะให้ความสำคัญและมุ่งประเด็นไปที่สิ่งที่ผู้สอบได้รับจากการตอบสนองร่วมกันจากการทำหน้าที่ต่างกันของข้อสอบ เช่น ลำเอียงเข้าข้างเพศใด เชื้อชาติใด ภาษาใด ภูมิศาสตร์ใด และค้นหาวิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เกิดปัญหาขึ้นหลังจากได้จัดสอบลุล่วงไปแล้ว แต่ยังไม่ได้ศึกษาเกี่ยวกับหลักการพิจารณาว่าข้อสอบข้อใดข้อหนึ่งลำเอียงหรือทำหน้าที่ต่างกันต่อผู้สอบกลุ่มใด จึงควรต้องพิจารณาข้อสอบข้อนั้นตามจุดมุ่งหมายในการวัดของแบบสอบเนื่องจากความไม่ยุติธรรมหรือความลำเอียงของข้อสอบขึ้นอยู่กับจุดมุ่งหมายในการวัดของแบบสอบเป็นสำคัญ ดังนั้น ในการพิจารณาว่าข้อสอบข้อใดที่ทำหน้าที่ต่างกัน ผู้พิจารณาต้องมีความรู้ความสามารถและความเข้าใจเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบอย่างแท้จริง เนื่องจากการพิจารณาอย่างผิวเผินอาจทำให้เกิดความผิดพลาดได้ ข้อสอบบางข้อถ้าพิจารณาโดยคำนึงถึงจุดมุ่งหมายในการวัดอาจไม่ได้มีความลำเอียง เช่น ข้อสอบที่ถามเกี่ยวกับเรื่องการทำอาหารเมื่อพิจารณาอย่างผิวเผินอาจเห็นว่าเป็นข้อสอบที่มีความลำเอียงเข้ากับเพศหญิงทำให้ผู้สอบเพศชายเสียเปรียบ แต่ถ้าแบบสอบฉบับนี้มีจุดมุ่งหมายในการคัดเลือกพ่อครัวหรือแม่ครัวในการทำอาหารหรือพ่อครัวแม่ครัวในการทำขนมข้อสอบข้อนี้อาจเป็นข้อสอบที่เหมาะสม และเป็นข้อสอบที่ไม่มีความลำเอียง หรือในกรณีข้อสอบวัดความเข้าใจในการอ่านซึ่งมีเนื้อเรื่องเกี่ยวกับกีฬาชนิดต่างๆ อาจถูกพิจารณาว่าเป็นข้อสอบที่ลำเอียงเข้าข้างเพศชายทำให้เพศหญิงเสียเปรียบ เนื่องจากผู้สอบเพศหญิงขาดความคุ้นเคยในการเล่นกีฬา แต่ถ้า

แบบสอบฉบับนี้มีจุดมุ่งหมายจะวัดความรู้เรื่องกีฬาของผู้สอบหรือใช้สำหรับคัดเลือกผู้ตัดสินกีฬา ก็อาจเป็นข้อสอบที่เหมาะสมได้เช่นกัน ปัญหาที่กล่าวมาจึงเป็นประเด็นให้ผู้วิจัยสนใจศึกษา และนำมาเป็นปัญหาในการวิจัยครั้งนี้เพื่อต้องการทราบว่าผู้เชี่ยวชาญที่ทำการตัดสินข้อสอบว่าข้อสอบข้อใดที่มีการทำหน้าที่ต่างกันของข้อสอบหรือไม่นั้น ควรได้รับการพิจารณาจากผู้เชี่ยวชาญที่มีความสามารถเฉพาะด้านและมีความรู้ทางด้านการวัดผลทางการศึกษา อาทิ ผู้เชี่ยวชาญด้านกีฬา ผู้เชี่ยวชาญด้านภาษา และผู้เชี่ยวชาญด้านหลักสูตรในกลุ่มสาระการเรียนรู้ต่างๆ ที่เกี่ยวข้อง เป็นต้น

เนื่องจากในอดีตข้อสอบจะได้รับการตรวจสอบ DIF ด้วยวิธีทางสถิติหลังจากผ่านการตอบจากผู้สอบก่อนซึ่งถือได้ว่าเป็นการตรวจพบปัญหาจากปลายทาง ไม่ได้เป็นการแก้ไขปัญหามาจากต้นทาง ดังนั้น ผู้วิจัยจึงสนใจศึกษาเพื่อพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญเป็นการตรวจสอบ DIF ตั้งแต่ต้นทางเพื่อปรับปรุงและแก้ไขข้อสอบที่เกิดปัญหาจากสาเหตุต่างๆ เป็นการลดข้อบกพร่องในการออกข้อสอบก่อนส่งไปให้ผู้เข้าสอบได้ทำจริง

วัตถุประสงค์

1. เพื่อพัฒนาวิธีการสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ
2. เพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในด้านอัตราความถูกต้อง และอัตราความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับข้อสอบที่คัดสรรมาสำหรับการทดลองเมื่อใช้วิธีการวิเคราะห์ด้วยแบบวินิจัยโดยผู้เชี่ยวชาญ วิธี

การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟาย จากกลุ่มผู้เชี่ยวชาญ และวิธีการประยุกต์ใช้เทคนิค โปรโตคอลอะลาร์วาร์ด

ขอบเขตของการวิจัย

1. ตัวแปรที่ศึกษาประกอบด้วย ตัวแปรอิสระ ได้แก่ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธีการตรวจสอบ DIF โดยแบบวินิจัยจากกลุ่มผู้เชี่ยวชาญ วิธีตรวจสอบ DIF ด้วยเทคนิคประยุกต์การประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ และวิธีประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์วาร์ด (TAPs) ตัวแปรตาม คือ ประสิทธิภาพในการตรวจสอบ DIF โดยพิจารณาจากความถูกต้องของการตรวจสอบ และอัตราความคลาดเคลื่อนในการตรวจสอบกับข้อสอบที่ได้รับการคัดสรรมาโดยผู้วิจัย

2. เกณฑ์ที่ใช้ในการตัดสินผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบของผู้เชี่ยวชาญ ดังนี้

2.1 การใช้แบบวินิจัยข้อสอบจากกลุ่มผู้เชี่ยวชาญโดยใช้เกณฑ์การพิจารณาการเกิด DIF ที่ผู้เชี่ยวชาญตอบได้สอดคล้องกันโดยมีค่าเฉลี่ยประมาณ 60 % ตรงกัน

2.2 วิธีประยุกต์ใช้เทคนิคเดลฟาย ใช้เกณฑ์ในการพิจารณาฉันทามติ (สวลี ทวีบุตร, 2540) ได้แก่

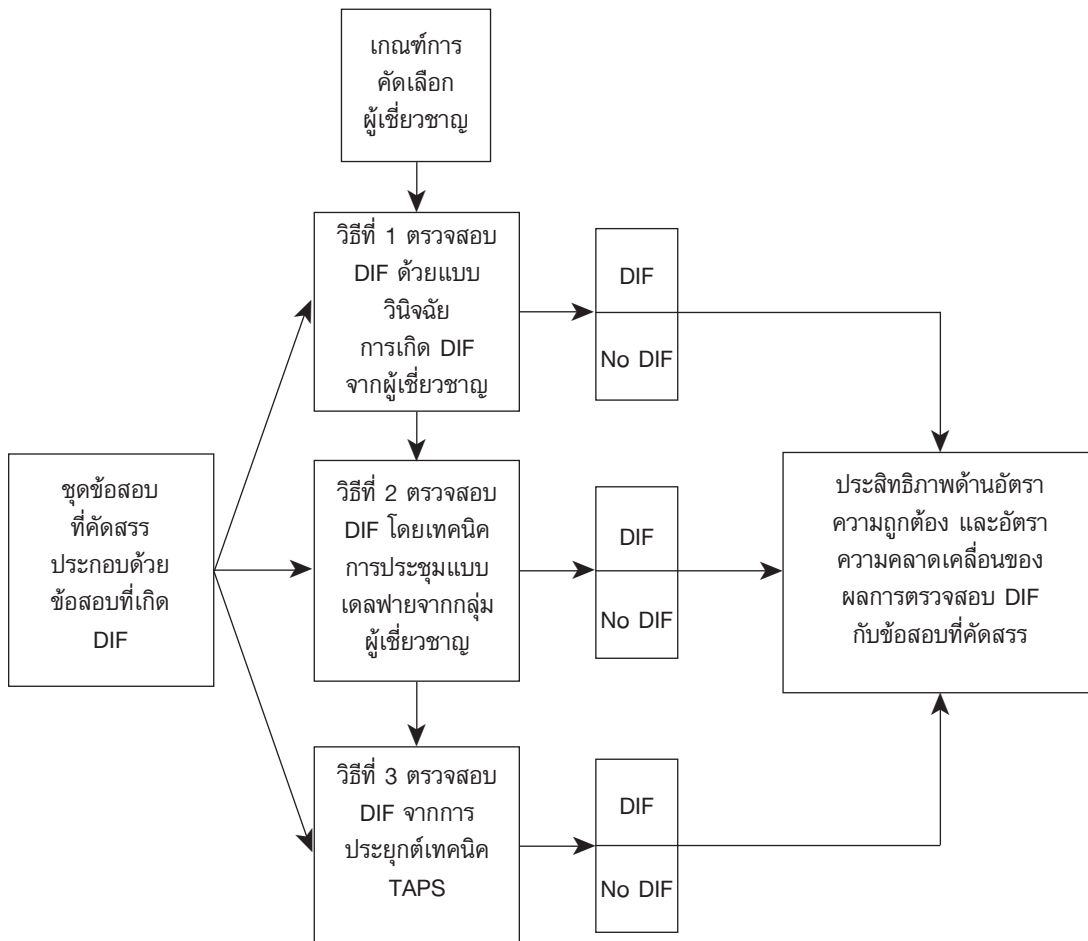
ระดับคะแนนฉันทามติ (degree of consensus) และความคงที่ (stability) ของระดับคะแนนฉันทามติ ซึ่งมีเกณฑ์ในการพิจารณา ดังนี้ เนื่องจากยังไม่ปรากฏข้อตกลงที่ชัดเจนในการกำหนดระดับ ฉันทามติ และความคงที่ของระดับฉันทามติขั้นต่ำที่เหมาะสม ดังนั้น จึงปรากฏ

เกณฑ์ที่นำไปใช้ในการพิจารณาแตกต่างกันไป เช่น ในด้านระดับคะแนนฉันทามติ กำหนดไว้ว่า ข้อความใดจะถือว่าได้รับฉันทามติก็ต่อเมื่อข้อความนั้นมีผู้ตอบสอดคล้องกัน 75% (Murry & Hammons, 1995) ส่วน Flanders (1989) กำหนดไว้ว่าข้อความใดจะถือว่าได้รับฉันทามติก็ต่อเมื่อข้อความนั้นมีผู้ตอบสอดคล้องกัน 60% ในด้านความคงที่ของระดับฉันทามติ มีผู้กำหนดเกณฑ์ที่ใช้ในการพิจารณาไว้หลายท่าน เช่น Murry & Hammons, 1995 ได้ให้ข้อเสนอว่า ควรยุติการวิจัยเมื่อระดับฉันทามติที่ได้เพิ่มขึ้นหรือลดลงจากรอบที่ผ่านมาน้อยกว่า 20% ส่วน Linstone and Turoff (2011) ให้ข้อเสนอแนะว่า ควรยุติการวิจัยในรอบต่อไปเมื่อระดับฉันทามติที่ได้เพิ่มขึ้นหรือลดลงจากรอบที่ผ่านมาน้อยกว่า 15%

2.3 วิธีการประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์วาร์ด (TAPs) สำหรับการตรวจสอบและยืนยันแหล่งที่มาของการทำหน้าที่ต่างกันของข้อสอบที่ถูกระบุโดยความคิดเห็นของผู้เชี่ยวชาญ โดยแนวคิดนี้ได้แสดงให้เห็นแนวทางในการตรวจสอบ และยืนยันความแตกต่างของแหล่งที่มาของการทำหน้าที่ต่างกันของข้อสอบ DIF เทคนิค TAPs ต้องใช้การแสดงความคิดเห็นของผู้เชี่ยวชาญประกอบกับกระบวนการคิดของผู้เข้าสอบ นักวิจัยมีการตรวจสอบแหล่งที่มาของการเกิด DIF โดยการแสดงความคิดเห็นของผู้เชี่ยวชาญด้านหลักสูตร การตรวจสอบเนื้อหาของข้อสอบ ภาษา และรูปแบบสำหรับตัวชี้วัดของแหล่งที่มาของ DIF โดยพิจารณาความจำเป็นที่จะต้องเข้าใจปฏิสัมพันธ์ระหว่างข้อสอบ แบบสอบ และกระบวนการคิดของผู้สอบ เครื่องมือที่ใช้ในการรวบรวมข้อมูลประกอบด้วยแบบสอบถามข้อมูลเกี่ยวกับกลุ่มผู้เข้าร่วม เกณฑ์การประเมินการเกิด

DIF ที่ผู้เชี่ยวชาญตอบได้สอดคล้องกันโดยมีค่าเฉลี่ยประมาณ 60 % ตรงกัน

กรอบแนวคิดในการวิจัย



ภาพที่ 1 แนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ

วิธีดำเนินการวิจัย

ขั้นตอนการวิจัย

1. ศึกษาค้นคว้าเกี่ยวกับทฤษฎี และหลักการพร้อมทั้งวิธีการที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ และงานวิจัยที่เกี่ยวข้องจากเอกสาร หนังสือ วารสาร และรายงานการวิจัยด้านการวัดผลการศึกษาทั้งในประเทศ และต่างประเทศ

2. กำหนดเกณฑ์ในการคัดเลือกผู้เชี่ยวชาญตามคุณสมบัติที่สำคัญในด้านต่างๆ

3. ออกแบบ และคัดเลือกเครื่องมือที่ใช้ในการวินิจฉัยซึ่งเป็นวิธีการในการตรวจสอบโดยผู้เชี่ยวชาญ และการจัดทำโดยประยุกต์ใช้เทคนิคการประชุมแบบเดลฟาย (Delphi Conference) จากกลุ่มผู้เชี่ยวชาญเพื่อหาแนวทางการตรวจสอบ DIF และการประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด (TAPs) มาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

4. คัดเลือกเครื่องมือ และคัดสรรข้อสอบที่ใช้ในการเก็บรวบรวมข้อมูล

5. ศึกษารายละเอียดการใช้โปรแกรมสำเร็จรูปทางสถิติเพื่อใช้ในการวิเคราะห์ข้อมูลตามเงื่อนไข ต่างๆ ที่ต้องการศึกษาการทำหน้าที่ต่างกันของข้อสอบโดยการนำผลการพิจารณาจากผู้เชี่ยวชาญมาวิเคราะห์ โดยใช้โปรแกรมสำเร็จรูปทางสถิติที่เป็นที่ยอมรับโดยทั่วไป การสร้างชุดเครื่องมือในการตรวจสอบด้วยวิธีทางสถิติ โดยใช้โปรแกรม DDFS 1.0 และโปรแกรม DIFAS 5.0 (Penfield, 2010)

6. นำข้อสอบที่ผ่านกระบวนการตรวจสอบ DIF ทางสถิติทั้ง 40 ข้อ มาพัฒนาสร้างเครื่องมือ

จากนั้นนำเครื่องมือที่ได้สร้างขึ้นในระยะแรกทั้งหมดส่งให้กลุ่มผู้เชี่ยวชาญตรวจสอบคุณภาพเครื่องมือและนำข้อเสนอแนะในการปรับปรุงเครื่องมือมาพัฒนาเป็นเครื่องมือสำหรับใช้กับการตรวจสอบ DIF เพื่อใช้กับวิธีการทั้ง 3 วิธี โดยมีเครื่องมือที่ได้รับการพัฒนาปรับปรุงสร้างขึ้นเพื่อใช้ในการวิจัย ดังนี้ แบบวินิจฉัยการทำหน้าที่ต่างกันของข้อสอบจากผู้เชี่ยวชาญ แบบยืนยันการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยประยุกต์เทคนิคการประชุมแบบเดลฟายแบบสอบถามสำหรับการตรวจสอบความลำเอียงของข้อสอบสำหรับนักเรียน ชุดข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับผู้เชี่ยวชาญ จากนั้นนำเครื่องมือที่ได้ทั้งหมดไปใช้ในการทดลองโดยวิธีการทั้ง 3 รูปแบบ ประกอบด้วยผู้เชี่ยวชาญทั้งหมด 21 คน แบ่งเป็น 3 กลุ่ม กลุ่มละ 7 คน ดังนี้วิธีที่ 1 นำเครื่องมือชุดแบบวินิจฉัยการทำหน้าที่ต่างกันของข้อสอบพร้อมด้วยคู่มือสำหรับการใช้เครื่องมือจัดส่งให้ผู้เชี่ยวชาญกลุ่มที่ 1 โดยผู้เชี่ยวชาญทั้ง 7 คน จะตอบเครื่องมือโดยเป็นอิสระจากกัน วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟาย ผู้เชี่ยวชาญกลุ่มที่ 2 ระยะแรกจะตอบเครื่องมือแบบวินิจฉัยการตรวจสอบ DIF โดยเป็นอิสระจากกันก่อน จากนั้นผู้วิจัยได้ทำการจัดประชุมผู้เชี่ยวชาญเพื่อให้เหตุผลในการตรวจสอบ DIF และหาฉันทมติจากกลุ่มผู้เชี่ยวชาญโดยใช้เครื่องมือเพิ่มเติมได้แก่ แบบยืนยันการตรวจสอบ DIF จากกลุ่มผู้เชี่ยวชาญ วิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด (TAPs) วิธีนี้ผู้วิจัยจะเก็บข้อมูลจากนักเรียนมัธยมศึกษาปีที่ 6 จากชุดเครื่องมือแบบสอบถามสำหรับการตรวจสอบความลำเอียงของข้อสอบสำหรับนักเรียน

จากนั้นนำข้อมูลจากนักเรียนสร้างเป็นเครื่องมือได้แก่ ชุดข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับผู้เชี่ยวชาญ โดยผู้เชี่ยวชาญในกลุ่มที่ 3 นี้จะตอบเครื่องมือแบบวิธีวินิจฉัยการตรวจสอบ DIF ก่อน จากนั้นจึงตอบเครื่องมือชุดข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับผู้เชี่ยวชาญซึ่งเครื่องมือชุดนี้จะมีข้อมูลสรุปความเห็นของนักเรียนให้ผู้เชี่ยวชาญพิจารณาประกอบการตัดสินใจในการตรวจสอบ DIF ในแต่ละข้อ

7. วิเคราะห์ข้อมูลของวิธีที่ 1 การวินิจฉัยการเกิด DIF จากผู้เชี่ยวชาญ วิธีที่ 2 ตรวจสอบ DIF โดยประยุกต์ใช้เทคนิคการประชุมแบบเดลฟาย (Delphi Conference) จากกลุ่มผู้เชี่ยวชาญ และวิธีที่ 3 ตรวจสอบ DIF จากการประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด (TAPs)

8. สรุปผลการวิเคราะห์ข้อมูลเพื่อเปรียบเทียบคุณภาพด้านความถูกต้องและความคลาดเคลื่อนของวิธีที่ 1 การวินิจฉัยการเกิด DIF จากผู้เชี่ยวชาญ วิธีที่ 2 ตรวจสอบ DIF โดยประยุกต์ใช้เทคนิคการประชุมแบบเดลฟาย (Delphi Conference) จากกลุ่มผู้เชี่ยวชาญ และวิธีที่ 3 ตรวจสอบ DIF จากการประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด (TAPs) ที่ทำการศึกษา

ผลการวิจัย

ผลการวิจัยในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาจากตัวแปรเพศ ใช้วิธีการวิเคราะห์สถิติแมนเทิล-แฮนส์เซลจากโปรแกรม DDFS 1.0 และโปรแกรม DIFAS 5.0 (Penfield, 2010) จากผลการตรวจสอบ DIF ทางสถิตินี้ ได้นำข้อสอบทั้ง 40 ข้อพัฒนาสร้างเป็นเครื่องมือแบบวิธีวินิจฉัยโดยผู้เชี่ยวชาญขึ้น และได้กระบวนการในการตรวจสอบทั้ง 3 วิธี ดังตารางที่ 1

ตารางที่ 1 ผลการวิเคราะห์ DIF สำหรับการเรียนรู้สู่ศึกษาและพลศึกษาจำแนกตามเพศ

ข้อ	MH CHI	MH LOR	LOR SE	LOR Z	BD	ETS	ผลสรุป DIF
1	87.5138	-.378	.0404	-9.3564	.976	A	No DIF
2	7.3071	.1123	.0413	2.7191	8.127	A	No DIF
3	35.3965	-.2148	.0361	-5.9501	96.389	A	No DIF
4	16.4382	-.1433	.0353	-4.0595	40.329	A	No DIF
5	2.7331	-.0802	.0478	-1.6778	1.701	A	No DIF
6	.2819	-.022	.04	-.55	5.573	A	No DIF
7	35.8273	-.2907	.0485	-5.9938	14.739	A	No DIF
8	.0007	.0061	.0872	.07	29.27	A	No DIF
9	2.4422	.1145	.0717	1.5969	4.42	A	No DIF
10	311.023	.7253	.0414	17.5193	.187	C	DIF
11	43.8509	.2133	.0322	6.6242	9.139	A	No DIF
12	443.1673	.9559	.0464	20.6013	202.034	C	DIF
13	6.1621	.2048	.0813	2.5191	7.316	A	No DIF
14	50.0064	.5222	.0743	7.0283	5.814	B	DIF
15	.3443	-.0261	.043	-.607	24.189	A	No DIF
16	68.1915	-.2839	.0344	-8.2529	38.064	A	No DIF
17	88.1051	-.3106	.0331	-9.3837	3.323	A	No DIF
18	6.5951	-.0871	.0337	-2.5846	1.95	A	No DIF
19	4.6414	.1583	.0724	2.1865	21.891	A	No DIF
20	13.598	-.1685	.0455	-3.7033	3.391	A	No DIF
21	5.8801	-.1058	.0433	-2.4434	28.463	A	No DIF
22	4.1686	.14	.0676	2.071	8.62	A	No DIF
23	1.1424	-.0603	.0551	-1.0944	21.99	A	No DIF
24	7.6259	.1073	.0386	2.7798	1.39	A	No DIF
25	8.4332	-.1559	.0533	-2.925	1.236	A	No DIF
26	4.0432	-.1244	.0611	-2.036	19.588	A	No DIF
27	19.4353	-.2458	.0555	-4.4288	3.867	A	No DIF
28	3.4036	-.0684	.0367	-1.8638	31.878	A	No DIF
29	2.4358	-.0703	.0445	-1.5798	39.854	A	No DIF
30	46.2086	-.2288	.0338	-6.7692	152.29	A	No DIF
31	10.0615	-.1374	.0431	-3.1879	19.792	A	No DIF
32	11.0201	-.1324	.0397	-3.335	60.092	A	No DIF
33	48.2685	.2571	.0373	6.8928	223.119	A	No DIF
34	3.1708	-.0757	.042	-1.8024	5.444	A	No DIF
35	41.3848	.3633	.0564	6.4415	3.819	A	No DIF
36	415.2477	.8424	.0422	19.9621	216.808	C	DIF
37	35.8053	.2007	.0335	5.991	4.557	A	No DIF
38	13.2244	.1255	.0344	3.6483	5.519	A	No DIF
39	47.2181	-.2761	.0401	-6.8853	2.792	A	No DIF
40	84.1455	-.3292	.0359	-9.1699	.093	A	No DIF

หมายเหตุ กลุ่มอ้างอิงคือ เพศชาย กลุ่มเปรียบเทียบคือ เพศหญิง

จากตารางที่ 1 ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาจำแนกตามเพศ ตรวจสอบด้วยวิธีแมนเทิล-แฮนส์เซล (MH) ซึ่งพิจารณาจากค่า The ETS Categorization Scheme (ETS) โดยจำแนกข้อสอบออกเป็นประเภท (A) หมายถึง ข้อสอบข้อนั้นมีขนาดของการเกิด DIF เล็กมากหรือถือได้ว่าข้อสอบข้อนั้นไม่เกิด DIF ข้อสอบประเภท

(B) หมายถึง ข้อสอบข้อนั้นเกิด DIF ในระดับปานกลาง ถือว่าข้อสอบข้อนั้นเกิด DIF แล้ว และข้อสอบประเภท (C) หมายถึง ข้อสอบข้อนั้นเกิด DIF มาก (Zieky, 1993) พบว่าแบบสอบสาระการเรียนรู้สุขศึกษาและพลศึกษานี้มีข้อสอบที่เกิดความลำเอียงจำนวน 4 ข้อ (ร้อยละ 10) โดยแบบสอบเกิดความลำเอียงในข้อต่อไปนี้ ข้อที่ 10, 12, 14 และ 36

ตารางที่ 2 อัตราความถูกต้องและอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีที่ 1 การวินิจฉัยโดยผู้เชี่ยวชาญเปรียบเทียบกับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติ

ผู้เชี่ยวชาญ	ผลการตรวจสอบข้อสอบ DIF ที่ถูกต้อง (4 ข้อ)		ผลการตรวจสอบข้อสอบ No DIF ที่ถูกต้อง (36 ข้อ)		สรุปผลการตรวจสอบข้อสอบทั้งหมด ที่ถูกต้อง (40 ข้อ)		สรุปผลการตรวจสอบข้อสอบทั้งหมด ที่ไม่ถูกต้อง (40 ข้อ)	
	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ
ท่านที่ 1	1	25	33	91.67	34	85	6	15
ท่านที่ 2	2	50	30	83.33	32	80	8	20
ท่านที่ 3	1	25	28	77.78	29	72.5	11	27.5
ท่านที่ 4	0	0	33	91.67	33	82.5	7	17.5
ท่านที่ 5	2	50	28	77.78	30	75	10	25
ท่านที่ 6	2	50	28	77.78	30	75	10	25
ท่านที่ 7	4	100	30	83.33	34	85	6	15
ค่าเฉลี่ย	2	50	30	83.33	32	80	8	20

จากตารางที่ 2 พบว่าอัตราความถูกต้องและอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีที่ 1 การวินิจฉัยโดยผู้เชี่ยวชาญเปรียบเทียบกับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติ มีอัตราความถูกต้องจากการตรวจสอบจากผู้เชี่ยวชาญเฉลี่ย จำนวน 2 ข้อ (ร้อยละ 50) และ

มีอัตราความคลาดเคลื่อนของการตรวจสอบโดยเฉลี่ยจำนวน 2 ข้อ (ร้อยละ 50) มีอัตราความถูกต้องจากการตรวจสอบจากผู้เชี่ยวชาญท่านที่ 1 และผู้เชี่ยวชาญท่านที่ 3 จำนวน 1 ข้อ (ร้อยละ 25) และมีอัตราความคลาดเคลื่อนจำนวน 3 ข้อ (ร้อยละ 75) เท่ากัน ผู้เชี่ยวชาญท่านที่ 2 ผู้เชี่ยวชาญท่านที่ 5 และผู้เชี่ยวชาญท่านที่ 6

มีอัตราความถูกต้องจากการตรวจสอบจำนวน 2 ข้อ (ร้อยละ 50) และมีอัตราความคลาดเคลื่อนจำนวน 2 ข้อ (ร้อยละ 50) เท่ากัน ผู้เชี่ยวชาญท่านที่ 4 มีอัตราความถูกต้องจากการตรวจสอบจำนวน 0 ข้อ (ร้อยละ 0) และมีอัตราความคลาดเคลื่อน

จำนวน 4 ข้อ (ร้อยละ 100) ผู้เชี่ยวชาญท่านที่ 7 มีอัตราความถูกต้องจากการตรวจสอบจำนวน 4 ข้อ (ร้อยละ 100) และมีอัตราความคลาดเคลื่อนจำนวน 0 ข้อ (ร้อยละ 0)

ตารางที่ 3 อัตราความถูกต้องและอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีที่ 2 การประยุกต์ใช้การประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญเปรียบเทียบกับ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติ

ผู้เชี่ยวชาญ	ผลการตรวจสอบข้อสอบ DIF ที่ถูกต้อง (4 ข้อ)		ผลการตรวจสอบข้อสอบ No DIF ที่ถูกต้อง (36 ข้อ)		สรุปผลการตรวจสอบข้อสอบทั้งหมด ที่ถูกต้อง (40 ข้อ)		สรุปผลการตรวจสอบข้อสอบทั้งหมด ที่ไม่ถูกต้อง (40 ข้อ)	
	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ
ฉันทามติ	0	0	30	83.33	30	75	10	15

จากตารางที่ 3 พบว่าอัตราความถูกต้องและอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจาก

กลุ่มผู้เชี่ยวชาญ มีอัตราความถูกต้องตามฉันทามติจำนวน 0 ข้อ (ร้อยละ 0) และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจำนวน 4 ข้อ (ร้อยละ 100)

ตารางที่ 4 อัตราความถูกต้องและอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์วาร์ด (TAPs) เปรียบเทียบกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติ

ผู้เชี่ยวชาญ	ผลการตรวจสอบข้อสอบ DIF ที่ถูกต้อง (4 ข้อ)		ผลการตรวจสอบข้อสอบ No DIF ที่ถูกต้อง (36 ข้อ)		สรุปผลการตรวจสอบข้อสอบทั้งหมดที่ถูกต้อง (40 ข้อ)		สรุปผลการตรวจสอบข้อสอบทั้งหมดที่ไม่ถูกต้อง (40 ข้อ)	
	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ
ท่านที่ 1	1	25	36	100	37	92.5	3	7.5
ท่านที่ 2	0	0	32	88.89	32	80	8	20
ท่านที่ 3	1	25	32	88.89	33	82.5	7	17.5
ท่านที่ 4	0	0	34	94.44	34	85	6	15
ท่านที่ 5	2	50	30	83.33	32	80	8	20
ท่านที่ 6	1	25	32	88.89	33	82.5	7	17.5
ท่านที่ 7	1	25	29	80.56	30	75	10	25
ค่าเฉลี่ย	1	25	32	88.89	33	82.5	7	17.5

จากตารางที่ 4 พบว่าอัตราความถูกต้องและอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์วาร์ด (TAPs) เปรียบเทียบกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติ มีอัตราความถูกต้องจากการตรวจสอบของค่าเฉลี่ยจากผู้เชี่ยวชาญทั้ง 7 คน พบว่ามีอัตราความถูกต้องในการตรวจสอบจำนวน 1 ข้อ (ร้อยละ 25) และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยจำนวน 3 ข้อ (ร้อยละ 75) ผู้เชี่ยวชาญท่าน

ที่ 1 ผู้เชี่ยวชาญคนที่ 3 ผู้เชี่ยวชาญคนที่ 6 และผู้เชี่ยวชาญคนที่ 7 เท่ากันจำนวน 1 ข้อ (ร้อยละ 25) และมีอัตราความคลาดเคลื่อนจำนวน 3 ข้อ (ร้อยละ 75) ผู้เชี่ยวชาญคนที่ 2 และผู้เชี่ยวชาญคนที่ 4 มีอัตราความถูกต้องจากการตรวจสอบจำนวน 0 ข้อ (ร้อยละ 0) และมีอัตราความคลาดเคลื่อนจำนวน 4 ข้อ (ร้อยละ 100) เท่ากัน ผู้เชี่ยวชาญคนที่ 5 มีอัตราความถูกต้องจากการตรวจสอบจำนวน 2 ข้อ (ร้อยละ 50) และมีอัตราความคลาดเคลื่อนจำนวน 2 ข้อ (ร้อยละ 50)

ตารางที่ 5 เปรียบเทียบอัตราความถูกต้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธี

ค่าเฉลี่ย	วิธีที่ 1		วิธีที่ 2		วิธีที่ 3	
	ผลการตรวจสอบ ข้อสอบ DIF ที่ถูกต้อง (4 ข้อ)		ผลการตรวจสอบ ข้อสอบ DIF ที่ถูกต้อง (4 ข้อ)		ผลการตรวจสอบ ข้อสอบ DIF ที่ถูกต้อง (4 ข้อ)	
	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ	จำนวน (ข้อ)	ร้อยละ
	2	50	0	0	1	25

จากตารางที่ 5 เป็นการเปรียบเทียบอัตราความถูกต้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธี พบว่าวิธีที่ 1 การวินิจฉัยโดยผู้เชี่ยวชาญ มีค่าเฉลี่ยของอัตราความถูกต้องในการตรวจสอบจำนวน 2 ข้อ (ร้อยละ 50) วิธีที่ 2 การประยุกต์ใช้การประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญมีฉันทามติตามเงื่อนไขจากการตรวจสอบ พบว่ามีอัตราความถูกต้องในการตรวจสอบจำนวน 0 ข้อ (ร้อยละ 0) วิธีที่ 3 การประยุกต์ใช้เทคนิคโพรโตคอลอะลาร์ด (TAPs) มีค่าเฉลี่ยของอัตราความถูกต้องในการตรวจสอบจำนวน 1 ข้อ (ร้อยละ 25)

การอภิปรายผล

จากผลการวิจัยครั้งนี้ มีประเด็นสำคัญหลัก ๆ ที่ควรนำมาอภิปรายดังนี้

ประเด็นแรก การพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธีนั้นมีความเหมือนและความแตกต่างกันในกระบวนการสำหรับการตรวจสอบ DIF ดังนี้ ในด้านความเหมือนกัน ได้แก่ ทั้ง 3 วิธีจะใช้ผู้เชี่ยวชาญกลุ่มละ 7 คน โดยไม่ซ้ำกันรวมเป็น 21 คน กระบวนการตรวจสอบในขั้นตอนแรกสำหรับทุกวิธีจะใช้เครื่องมือหลักในการตรวจสอบได้แก่ แบบสอบถามการเรียนรู้สู่ศึกษาและพลศึกษาที่ได้ทำการตรวจสอบ

DIF ด้วยวิธีทางสถิติแล้วและปรับปรุงคุณภาพเป็นเครื่องมือเป็นแบบวินิจฉัยการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ การตรวจสอบระยะแรกนี้ผู้เชี่ยวชาญทั้ง 21 คนจะทำการตรวจสอบ DIF และให้เหตุผลว่าข้อสอบ DIF เข้าข้างเพศใดหรือ No DIF และเพิ่มเติมทัศนะส่วนตนโดยการตรวจสอบมีความเป็นอิสระจากกัน ในด้านความแตกต่างกันนั้นกระบวนการในวิธีที่ 1 ผู้วิจัยได้ออกแบบให้ตรวจสอบเฉพาะระยะแรกเท่านั้น ส่วนวิธีที่ 2 การประยุกต์ใช้การจัดประชุมแบบเดลฟายนั้นผู้เชี่ยวชาญกลุ่มที่ 2 จะต้องมาประชุมร่วมกันเพื่อหาฉันทามติสำหรับการตรวจสอบโดยกลุ่มอีกครั้ง ซึ่งสอดคล้องกับผลการศึกษาของ Rowe and Wright (1999) ที่พบว่าวิธีการแบบเดลฟายรูปแบบดั้งเดิมจะไม่มี การเผชิญหน้ากันในกลุ่มผู้เชี่ยวชาญเพื่อการเก็บข้อมูลจากแบบสอบถามวิจัยและต้องใช้ระยะเวลาในการเก็บข้อมูลมาก แต่สำหรับเทคนิคเดลฟายแบบปรับปรุงซึ่งได้แก่ การจัดประชุมแบบเดลฟายเป็นการเก็บรวบรวมข้อมูลในลักษณะของการประชุมระหว่างการประชุมจะเก็บข้อมูลซ้ำด้วยแบบสอบถามและให้ผู้เชี่ยวชาญได้พิจารณาและตรวจสอบความคิดเห็นของตนเองอีกครั้งพร้อมกับการสนับสนุนให้เกิดการอภิปรายกันภายในกลุ่ม ผู้วิจัยสามารถสังเกตพฤติกรรมของผู้ให้ข้อมูล

ได้และได้ผลของฉันทามติจากกลุ่มได้ดีกว่า วิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด (TAPs) ซึ่งสอดคล้องกับผลการศึกษาของ Ercikan, Arim, Law, Domene, Gagnon and Lacroix (2010) พบว่า การตรวจสอบ DIF โดยให้ผู้เชี่ยวชาญใช้ข้อมูลจากนักเรียนเพื่อใช้ในการพิจารณาตรวจสอบ DIF นั้นจะส่งผลให้เกิดความแม่นยำในการตรวจสอบมากขึ้น

ประเด็นที่สอง จากผลการวิจัยจะพบว่าวิธีที่ 2 การประยุกต์ใช้การประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญตามฉันทามติ มีอัตราความถูกต้องจำนวน 0 ข้อ (ร้อยละ 0) แล้วไม่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้เลย นั้น เป็นเพราะข้อจำกัดตามเงื่อนไขของเกณฑ์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบซึ่งได้ระบุไว้ว่าข้อสอบข้อใดที่ผ่านการตรวจสอบจากกลุ่มผู้เชี่ยวชาญและจะถือว่าตรวจสอบพบการทำหน้าที่ต่างกันของข้อสอบได้นั้นต้องผ่านฉันทามติจากกลุ่มผู้เชี่ยวชาญด้วยคะแนนเสียง 4 ใน 7 เสียง ดังนั้น จึงถือได้ว่าวิธีที่ 2 นี้เกิดจุดอ่อนในเรื่องอิทธิพลของกลุ่มเข้ามาเกี่ยวข้องจากบริบทในการประชุมนั้นๆ

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1) ข้อควรระมัดระวังสำหรับผู้วิจัย โดยห้ามเปิดเผยข้อมูลเกี่ยวกับผลการตรวจสอบ DIF ทางสถิติให้กับกลุ่มผู้เชี่ยวชาญที่ต้องตรวจสอบ DIF จากวิธีการทั้ง 3 ดังกล่าวข้างต้นโดยเด็ดขาด เพื่อให้เครื่องมือที่ได้พัฒนาขึ้นนี้มีความน่าเชื่อถือในการทดลอง และเกิดผลการตรวจสอบในสภาพความเป็นจริงได้อย่างถูกต้องที่สุด

2) ควรนำวิธีที่ได้พัฒนาขึ้นทั้ง 3 วิธีนี้ไปใช้ในเชิงปฏิบัติโดยคำนึงถึงบริบทในการใช้เพื่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อมีกลุ่มผู้เชี่ยวชาญได้ออกข้อสอบมาแล้วก็ควรมีกกลุ่มผู้เชี่ยวชาญอีกกลุ่มทำการตรวจสอบเพื่อคัดกรองข้อสอบเหล่านั้นก่อนการนำไปใช้งานจริงกับผู้สอบเป็นการลดปัญหาตั้งแต่ต้นทาง ทั้งยังเป็นการตรวจสอบคุณภาพของเนื้อหา ภาษาที่ใช้ และเนื้อหาในหลักสูตรในการออกข้อสอบได้ ข้อสอบจะได้มีคุณภาพ มีความเป็นกลางสำหรับผู้สอบต่อไป

3) การนำข้อมูลเชิงคุณภาพจากผู้เชี่ยวชาญไปใช้ประโยชน์ในการปรับปรุงการออกข้อสอบให้เหมาะสมและมีมาตรฐาน

ข้อเสนอแนะในการศึกษาวิจัยครั้งต่อไป

1) การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญยังสามารถตรวจสอบสาเหตุของการทำหน้าที่ต่างกันของข้อสอบจากทักษะเพิ่มเติมของผู้เชี่ยวชาญ และหากนำไปใช้กับกลุ่มสาระการเรียนรู้อื่นๆ จะเกิดประโยชน์มาก เนื่องจากผู้สอบในสาระการเรียนรู้อื่นก็มีความหลากหลายทางด้านเพศ และบริบทอื่นๆ เช่น ควรมีการศึกษาเพิ่มเติมการศึกษาด้านบริบทของโรงเรียน เศรษฐฐานะของผู้สอบ เชื้อชาติ ภูมิภาคของผู้สอบ ขนาดโรงเรียน เข้าไปเพื่อศึกษาต่อไป

2) การศึกษาในครั้งนี้เป็นการวิจัยเชิงทดลองเพื่อพัฒนาวิธีการใหม่ๆ เพื่อบุกเบิกองค์ความรู้ใหม่ โดยมีการศึกษาจากข้อมูลจริงจากข้อสอบในกลุ่มสาระการเรียนรู้สุขศึกษาและพลศึกษา เท่านั้น ดังนั้น หากการศึกษาค้นคว้าในอนาคตควรศึกษาเพิ่มเติมในส่วนของกลุ่มสาระอื่นๆ จะเป็นการศึกษาที่มีประโยชน์และมีคุณค่าในทางปฏิบัติมากยิ่งขึ้น

รายการอ้างอิง

ภาษาไทย

- จิตติมา วรณศรี. (2539). การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมนเทิล-แฮนส์เซลกับวิธีชิปเทสต์ เมื่อความยาวของแบบสอบ ขนาดกลุ่มตัวอย่าง และ อัตราส่วนของกลุ่มอ้างอิง และกลุ่มเปรียบเทียบต่างกัน. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, ภาควิชาวิจัยการศึกษา บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.
- นิคม กิรติวรานุกร. (2542). การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด แมนเทิล-แฮนส์เซล และการตอบสนองข้อสอบ. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, ภาควิชาวิจัยการศึกษา บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.
- วิรัชชา ชะม้อย. (2550). การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่ตรวจให้คะแนนแบบพหุภาคระหว่างวิธีโครงสร้างความแปรปรวนร่วมและค่าเฉลี่ยกับวิธีการ วิเคราะห์ฟังก์ชันเชิงจำแนกแบบโลจิสติก. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, ภาควิชา วิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- วลีมาศ แซ่อึ้ง. (2543). การเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอนุกรมระหว่างวิธีชิปเทสต์ปรับปรุงใหม่ วิธี ชิปเทสต์ วิธีแมนเทิล-แฮนส์เซล และวิธีการถดถอยโลจิสติก. วิทยานิพนธ์ปริญญาครุศาสตรดุษฎี บัณฑิต, ภาควิชาวิจัยการศึกษา บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.
- ศิริชัย กาญจนวาสี. (2550). ทฤษฎีการทดสอบแนวใหม่. (พิมพ์ครั้งที่ 3). (ฉบับปรับปรุงเพิ่มเติม) กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- สุมาลี แก้วทรวงศ์. (2547). สาเหตุของการทำหน้าที่ต่างกันของข้อสอบสาระการเรียนรู้ภาษาไทย และ สาระการเรียนรู้สังคมศึกษา ศาสนาและวัฒนธรรม. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- สุวลี ทวีบุตร. (2540). การเปรียบเทียบผลการสร้างฉันทามติและระดับการให้ความร่วมมือของผู้เชี่ยวชาญ ระหว่างการใช้เทคนิคเดลฟายแบบเดิมและเทคนิคเดลฟายแบบปรับปรุงที่ใช้ในการประเมิน ความต้องการจำเป็น. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, ภาควิชาวิจัยการศึกษา บัณฑิต วิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

- Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues & Practice*, 29(2), 24-35.

- Flanders, L. (1989). Northern Ireland Hit-Squad Scandal. *Nation*, 249 (14), 491-494.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, 2, 313-334.
- Hüseyin Hüsni Yildim & Giray Berberoğlu. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, 9, 108-121.
- Murry, J. W., & Hammons, J. O. (1995). Delphi: A versatile methodology for conducting qualitative research. *The Review of Higher Education*, 18, (4), 423-436.
- Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous Items. *Journal of Educational Measurement*, 47. (2), 129-149.
- Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, 15, (4), 353-375.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

ผู้เขียน

อาจารย์ ดร.ชัยวัฒน์ หลุ่ยพันธ์ อาจารย์ประจำ สาขาการจัดการ คณะมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก จังหวัดชลบุรี 20110

อีเมล: roger_fox7@hotmail.com

ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี อาจารย์ประจำสาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10330 อีเมล: sirichai.k@chula.ac.th

รองศาสตราจารย์ ดร.โชติกา ภาษีผล อาจารย์ประจำสาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10330 อีเมล: aimornj@hotmail.com

หมายเหตุ: วิทยานิพนธ์นี้ได้รับการสนับสนุนทุนวิจัยจาก “ทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย” กองทุน รัชดาภิเษกสมโภช