

Methodological Notes on Classroom (Action) Research: Rethinking the Role of Inferential Statistics and Individual Learning Evidence¹

Chayut Piromsombat²

Abstract

This article offers methodological reflections based on the author's experience reviewing classroom action research and educational theses. It points out common issues such as misuse of statistical tests, failure to question underlying research assumptions, and the neglect of individual-level data. The author proposes alternative approaches including the use of normalized gain scores, individual line graph analysis, and qualitative data collection and interpretation. These methods aim to capture meaningful change beyond simply reporting p-values. The article invites teachers, graduate students, researchers, and other stakeholders to reconsider prevailing research practices and adopt more appropriate, context-sensitive methods.

Keywords: Classroom Action Research, Gain Score, Individual Analysis, Normalized Gain Score

¹ Articles in the *Editor's Note* reflect the editor's own opinions and have not been peer-reviewed. Their primary purpose is to communicate with readers who are interested in sound and up-to-date research methodology.

² *Editor-in-Chief, Journal of Research Methodology*,

Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University
chayut.p@chula.ac.th

ข้อสังเกตเชิงวิธีวิทยาเกี่ยวกับการวิจัย(ปฏิบัติการ)ชั้นเรียน : ทบทวนบทบาทของสถิติเชิงอ้างอิงและหลักฐานการเรียนรู้

รายบุคคล¹

ชยุตม์ ภิรมย์สมบัติ²

บทคัดย่อ

บทความนี้เสนอข้อสังเกตเชิงวิธีวิทยาจากประสบการณ์ของผู้เขียนในการอ่านและประเมินงานวิจัยปฏิบัติการชั้นเรียนและวิทยานิพนธ์ทางการศึกษา โดยชี้ให้เห็นปัญหาที่พบบ่อย เช่น การใช้สถิติผิดวิธี การไม่ตั้งคำถามกับสมมติฐานพื้นฐานของปัญหาวิจัย และการมองข้ามข้อมูลรายบุคคล ผู้เขียนเสนอแนวทางวิเคราะห์ทางเลือก เช่น การใช้ normalized gain score การวิเคราะห์ด้วยกราฟรายบุคคล รวมถึงการเก็บและการวิเคราะห์ข้อมูลเชิงคุณภาพ เพื่อสะท้อนความเปลี่ยนแปลงอย่างมีความหมายมากกว่าการอ้างค่า p เพียงอย่างเดียว บทความนี้ต้องการเชิญชวนให้ครู นิสิตนักศึกษา นักวิจัย และผู้เกี่ยวข้องได้ทบทวนแนวปฏิบัติเกี่ยวกับวิจัยที่ทำกันตาม ๆ มา แล้วหันมาใช้วิธีที่เหมาะสมมากยิ่งขึ้น

คำสำคัญ : การวิจัยปฏิบัติการชั้นเรียน, คะแนนพัฒนาการ, การวิเคราะห์รายบุคคล, คะแนนพัฒนาการแบบนอร์มัลไลซ์

¹ บทความใน *บันทึกบรรณาธิการ* เป็นการนำเสนอความคิดเห็นของบรรณาธิการซึ่งไม่ได้ผ่านการประเมินโดยผู้ทรงคุณวุฒิ มีวัตถุประสงค์หลักเพื่อสื่อสารกับผู้อ่านที่สนใจวิธีวิทยาการวิจัยที่ถูกต้องและทันสมัย

² *บรรณาธิการบริหาร วารสารวิธีวิทยาการวิจัย*,
ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
chayut.p@chula.ac.th

บทความนี้ นำเสนอ มุมมองเชิงวิพากษ์วิทยาการวิจัย (research methodological perspective) จากประสบการณ์ของผู้เขียนในฐานะบรรณาธิการบริหารวารสารวิธีวิทยาการวิจัย (Journal of Research Methodology: JRM) และอาจารย์ประจำหลักสูตรวิธีวิทยาการวิจัย การศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย จุดประสงค์หลักมิใช่การสังเคราะห์ทฤษฎีหรือทบทวนเอกสารงานวิจัย แต่ต้องการชวนคิดผ่านข้อสังเกตจากประสบการณ์การประเมินบทความ การอ่านงานวิจัยและวิทยานิพนธ์ที่ทำในบริบทของห้องเรียน ตลอดจนงานที่นำเสนอในการประชุมวิชาการและช่องทางออนไลน์ต่าง ๆ โดยเฉพาะงานวิจัยปฏิบัติการชั้นเรียน เนื้อหาจึงสะท้อนความเห็นส่วนตัวเป็นหลัก พยายามเลี่ยงภาษาทางการหรือศัพท์เทคนิคหากไม่จำเป็น และบทความนี้ไม่ได้มุ่งอธิบายรายละเอียดเชิงเทคนิคเกี่ยวกับการวิจัยหรือการวิเคราะห์ข้อมูล โดยคาดหวังว่าผู้อ่านที่สนใจสามารถนำคำสำคัญหรือข้อแนวคิดต่าง ๆ ที่กล่าวถึงในบทความนี้ไปสืบค้นและศึกษาต่อเพิ่มเติมได้ด้วยตนเอง

งานวิจัยปฏิบัติการชั้นเรียนของไทย (ขอเรียกรวม ๆ กับงานวิจัยทางการศึกษาและวิทยานิพนธ์ทางการศึกษาที่พบเห็นทั่ว ๆ ไป) จำนวนมากที่ผู้เขียนได้อ่านมีโครงสร้างคล้ายกันคือ ครู/นักวิจัยมักเริ่มจากโจทย์กว้าง ๆ คล้าย ๆ กัน เช่น นักเรียนขาดแรงจูงใจ ผลสัมฤทธิ์/คะแนนต่ำ บางงานเท้าความไกลไปถึงผลประเมิน PISA และคะแนน O-NET ของประเทศไทยต่ำ แล้วออกแบบแผนการจัดการเรียนรู้ใหม่ในหัวข้อแคบ ๆ ที่ไม่ได้เกี่ยวข้องกับ PISA หรือ O-NET เลย (ซึ่งไม่ควรเป็นที่มาแบบอัตโนมัติของโจทย์วิจัยปฏิบัติการชั้นเรียนตั้งแต่แรก) จากนั้นใช้กิจกรรมที่ดูน่าสนใจขึ้น หรือเติมเทคโนโลยีเข้าไป แล้วปิดท้ายด้วยการเปรียบเทียบผลการเรียนก่อนและหลัง และ/หรือเปรียบเทียบคะแนนระหว่างกลุ่มควบคุมและกลุ่มทดลอง เพื่อสรุปว่าวิธีที่ใช้นั้น “ได้ผล” กระบวนการนี้ดูสมบูรณ์ในเชิงโครงสร้าง แต่เมื่อมองให้ลึกกลับพบปัญหาสำคัญที่มักถูกละเลย นั่นคือการไม่ตั้งคำถามกับความเข้าใจเบื้องต้นเกี่ยวกับปัญหาที่ต้องการจะแก้ไข

สิ่งที่มักหายไปจากกระบวนการคือการวิเคราะห์ต้นเหตุของปัญหาอย่างจริงจัง หลายงานข้ามขั้นตอนสำคัญ เช่น การตั้งคำถามวิจัยที่ตรงเป้า การกำหนดปัญหาอย่างมีมิติ การพิจารณาสมมติฐานเกี่ยวกับสาเหตุของปัญหาที่เป็นไปได้แล้วใช้ตรรกะคัดเลือกสาเหตุที่ควรใช้วิจัยเพื่อแก้ปัญหาและจะใช้แนวคิดอะไรมา “จัดการ” เหตุนั้น การเลือกหรือออกแบบนวัตกรรมที่มีตรรกะรองรับ การที่คะแนนต่ำอาจไม่ได้เกิดจากแผนการสอนที่น่าเบื่อเพียงอย่างเดียว แต่อาจมีสาเหตุอื่นที่ซับซ้อนกว่านั้น เช่น ความไม่เข้าใจภาษาในโจทย์ ความรู้พื้นฐานที่ไม่แม่นยำหรือมีมโนทัศน์พื้นฐานที่คลาดเคลื่อน ปัญหาทางจิตใจอารมณ์ความรู้สึก แรงกดดันจากครอบครัว หรือแม้แต่เครื่องมือวัดที่ไม่สอดคล้องกับสิ่งที่เรียน การสรุปปัญหาแบบผิวเผินแล้วรีบออกแบบกิจกรรมใหม่ทันที ทำให้การวิจัยปฏิบัติการชั้นเรียนจำนวนมากกลายเป็นการทำผิดตาม ๆ กันมาโดยไม่ตั้งคำถาม จนตกอยู่ในกับดักของ “research method cult” คือ การยึดติดกับรูปแบบหรือโครงสร้างที่เชื่อตาม ๆ กันว่าถูกต้อง มากกว่าจะใช้ตรรกะความรู้เชิงวิธีวิทยาการวิจัยที่ถูกต้องทันสมัยเพื่อทำความเข้าใจการเรียนรู้และแก้ปัญหาของผู้เรียนจริงจัง

หากจะยกระดับงานวิจัยปฏิบัติการชั้นเรียนให้ดีขึ้นจริง จุดเริ่มต้นคือต้องกล้าทบทวนโจทย์และตั้งคำถามใหม่ กล้าสงสัยว่าเราเข้าใจปัญหาถูกต้องแล้วหรือยัง และมีข้อมูลอะไรยืนยันความเข้าใจนั้น การใช้แนวคิดอย่าง design thinking หรือ design research (สுவิมล ว่องวานิช, 2563) หรือแนวคิดของ adaptive treatment ในการวิจัยเชิงทดลอง (Pistorello, 2017) อาจเป็น

ทางเลือกที่ช่วยให้นักวิจัยเข้าใจปัญหาได้ลึกซึ้ง ยืดหยุ่นและเปิดโอกาสให้ได้สะท้อนคิดและปรับ treatment (หรือ intervention หรือ innovation) ได้ต่อเนื่อง ไม่ใช่จบแค่ในรอบเดียวของการทดลองด้วย treatment แบบ fixed one-size-fits-all คือให้เหมือน ๆ กันทุกคนและใช้ตามที่ออกแบบตั้งแต่เริ่มการวิจัยโดยไม่มีการปรับปรุงพัฒนาระหว่างการดำเนินการวิจัย

ในบทความนี้ผู้เขียนจะไม่กล่าวถึงปัญหาทั้งหมดที่พบในงานวิจัยปฏิบัติการชั้นเรียน แต่อยากชวนโฟกัสเฉพาะจุดที่สำคัญมากอีกจุดหนึ่งแต่ถูกมองข้ามบ่อยเพราะทำตาม ๆ กันโดยไม่ยั้งคิด นั่นคือ การวิเคราะห์ข้อมูลจากงานวิจัยประเภทนี้ โดยเฉพาะกรณีที่ใช้แบบแผนการวิจัยอดนินยอมอย่าง pretest–posttest control group design หรือ PPCGD (ภาพ 1; Campbell & Stanley, 1963) หลายงานแปลผลข้อมูลแบบหยาบ ๆ ดีความผลการเปลี่ยนแปลงจากค่าเฉลี่ยคะแนนเพียงอย่างเดียว โดยไม่ย้อนกลับไปพิจารณาพฤติกรรม กระบวนการเรียนรู้ หรือบริบทของการสอนจริง การวิเคราะห์ข้อมูลที่ให้ insights ลึกพอจะช่วยให้ครูเข้าใจว่าการสอนเปลี่ยนพฤติกรรมใดของผู้เรียนได้บ้าง มากน้อยแค่ไหน และยังเหลือข้อจำกัดหรือสิ่งที่ควรพัฒนาในรอบถัดไปหรือไม่ ความสามารถในการอ่านข้อมูลอย่างมีความหมายจึงเป็นสิ่งที่แบ่งแยกระหว่างงานวิจัยที่ทำเพื่อให้ได้ “ส่ง” กับงานวิจัยที่มีคุณค่าจริงในการพัฒนาการเรียนรู้

ภาพ 1

แบบแผนการวิจัยแบบ Pretest–Posttest Control Group Design

C	R	O1		O2
E	R	O3	X	O4

งานวิจัยทางการศึกษาที่ใช้ PPCGD มักวิเคราะห์ข้อมูลโดยใช้ t-test หลายคู่ ซึ่งเป็นแนวปฏิบัติที่แพร่หลาย แต่ไม่เหมาะสมในเชิงวิธีวิทยา ตัวอย่างที่พบได้บ่อยคือการเปรียบเทียบ 4 คู่ ได้แก่ (1) O1 กับ O3 เพื่อดูว่าก่อนเรียนกลุ่มควบคุมกับกลุ่มทดลองมีค่าเฉลี่ยต่างกันหรือไม่ (2) O1 กับ O2 เพื่อดูการเปลี่ยนแปลงในกลุ่มควบคุม (3) O3 กับ O4 เพื่อดูการเปลี่ยนแปลงในกลุ่มทดลอง และ (4) O2 กับ O4 เพื่อเปรียบเทียบผลลัพธ์หลังเรียนระหว่างกลุ่มควบคุมและกลุ่มทดลอง

ในจำนวนนี้ การเปรียบเทียบ O1 กับ O3 มักถูกใช้โดยมีความเข้าใจคลาดเคลื่อนว่าถ้าค่าเฉลี่ยก่อนเรียนของกลุ่มควบคุมและกลุ่มทดลองแตกต่างกันตั้งแต่ต้น จะไม่สามารถเปรียบเทียบผลหลังเรียนได้ ซึ่งไม่เป็นความจริงในเชิงวิธีวิทยา และการใช้ t-test ถึง 4 ครั้งในลักษณะนี้ยังเพิ่มความเสี่ยงของ family-wise type I error ให้เกินระดับนัยสำคัญที่ตั้งไว้ (.05) โดยไม่มีการปรับค่า เช่นการใช้ Bonferroni correction หรือวิธีควบคุมข้อผิดพลาดอื่นที่เหมาะสม

แนวทางที่เหมาะสมกว่าคือการใช้การวิเคราะห์ที่สามารถจัดการความแตกต่างระหว่างกลุ่มก่อนเรียนอย่างเป็นระบบ และเปรียบเทียบผลลัพธ์หลังเรียนได้ในคราวเดียว เช่น ANCOVA โดยควบคุม pretest เป็น covariate การพยายามบังคับให้กลุ่มควบคุมและกลุ่มทดลอง “ไม่แตกต่างกัน” ตั้งแต่ต้นจึงไม่จำเป็น และในบริบทจริงก็ไม่สมเหตุผลผลด้วยซ้ำ เพราะเป้าหมายของ

การทดลองไม่ใช่การหากลุ่มที่เหมือนกันเป๊ะในทุกตัวแปร แต่คือการแยกผลของ intervention ออกจากปัจจัยอื่นให้ได้ชัดเจนที่สุด

นอกจากนี้ ยังมีการใช้ t-test ผิดแบบอื่น ๆ อีกมาก ที่พบได้บ่อย เช่น ใช้ independent-samples t-test เปรียบเทียบ pretest กับ posttest ของนักเรียนในกลุ่มเดียวกัน ทั้งที่ข้อมูลเป็นข้อมูลชุดเดียวกัน (dependent) ซึ่งควรใช้ dependent-samples หรือ paired-samples t-test แทน โปรดสังเกตข้อสถิติ ไม่ใช่ t-test independent เพราะผิดไวยากรณ์ และใช้ lowercase t ตาม Student (1908) ซึ่งเป็นนามแฝงของ William Sealy Gosset ผู้เสนอสถิติทดสอบนี้ ไม่ควรใช้ uppercase T ซึ่งอาจสับสนกับ Hotelling's T^2 ได้ ตลอดจนการละเลยข้อดกลงเบื้องต้น (assumption) ที่สำคัญ ๆ คือ ตัวแปรตามต้องมีการวัดแบบต่อเนื่อง (interval หรือ ratio) กลุ่มตัวอย่างสองกลุ่มต้องเป็นอิสระจากกัน (independent) กลุ่มแต่ละกลุ่มต้องมีการแจกแจงแบบปกติ (normal distribution) ความแปรปรวนของทั้งสองกลุ่มควรเท่ากัน (homogeneity of variances) ถ้าข้อมูลไม่เป็นไปตามสมมุติฐานเหล่านี้ นักวิจัยไม่ควรดันทุรังใช้ t-test ควรพิจารณาเลือกใช้สถิติอ้างอิง (inferential statistics) อื่นที่เหมาะสมกว่า เช่น Welch's t-test หรือ non-parametric test แทน หรือพิจารณาใช้วิธีการอื่นในการวิเคราะห์ เช่น กราฟเส้น และสถิติบรรยาย ซึ่งจะกล่าวถึงเพิ่มเติมต่อไป

แม้ว่าการวิจัยแบบ PPCGD จะเป็นงานวิจัยเชิงปริมาณโดยพื้นฐาน แต่ก็สามารถผสมการวิเคราะห์เชิงคุณภาพเข้าไปได้ หากนักวิจัยออกแบบอย่างตั้งใจ และเลิกทำตามแบบแผนแบบ cult ที่มุ่งแต่เปรียบเทียบคะแนนก่อน-หลังเท่านั้น การเก็บข้อมูลเชิงคุณภาพ เช่น บันทึกภาคสนาม การสังเกต การบันทึกการสนทนา การสัมภาษณ์ หรือการร่วมสะท้อนคิดกับผู้เรียน สามารถช่วยเปิดมุมมองใหม่ให้กับการตีความข้อมูลเชิงปริมาณได้มาก การวิเคราะห์เชิงคุณภาพเหล่านี้สามารถช่วยเสริมการอธิบายการเปลี่ยนแปลงที่เกิดขึ้น เช่น ด้านพฤติกรรม ทักษะคิด ความรู้ ความรู้สึก หรือความสัมพันธ์เชิงปฏิสัมพันธ์ในห้องเรียน ที่มักไม่ปรากฏในคะแนนสอบ การเข้าใจว่าการเปลี่ยนแปลงนั้นเกิดขึ้น “อย่างไร” และ “ทำไม” ช่วยเสริมคุณค่าให้กับข้อค้นพบ ไม่ใช่เพียงบอกแค่ “เกิดการเปลี่ยนแปลงหรือไม่” นอกจากนี้ การผสมผสานวิธีแบบนี้ยังตรงกับหลักวิธีวิทยาการวิจัยเชิงทดลองที่ดี ในมิติของ Fidelity of Treatment ด้าน Participant's Receipt คือ ผู้เรียนได้รับและตอบสนองต่อ treatment อย่างไรในความเป็นจริง ไม่ใช่แค่ในแบบที่นักวิจัยออกแบบไว้และเข้าใจเอาเองฝ่ายเดียว การมีข้อมูลเชิงคุณภาพจึงช่วยให้ประเมินความสมบูรณ์ของการนำ treatment ไปใช้ได้ลึกและแม่นยำยิ่งขึ้น และน่าจะมีประโยชน์มากกว่าการสอบถามความพึงพอใจของผู้เรียนแบบหยาบ ๆ อย่างที่เห็นกันทั่วไป

นอกจากปัญหาการใช้ t-test ที่กล่าวไปข้างต้น ยังมีอีกประเด็นหนึ่งที่สำคัญ และน่าจะเป็นจุดที่ช่วยให้ถูกคิด (ใจจริงอยากจะใช้คำว่า “เบกเนตร”) ให้นักวิจัยทบทวนว่าควรเลือกใช้ t-test แบบเดิม ๆ กับงานวิจัยประเภทนี้ได้หรือยัง ก่อนอื่นลองพิจารณาข้อมูลในตาราง 1 และตัวอย่างการสรุปผลที่มักพบในงานวิจัยแนวนี้นี้

หากนำข้อมูลในตารางไปวิเคราะห์ด้วย t-test ทั้ง 4 คู่ตามแนวทางผิด ๆ ที่หลายงานทำตาม research method cult กันมา อาจเขียนสรุปตามรูปแบบ APA 7th edition ได้ว่า “ผลการทดสอบ independent-samples t-test พบว่า คะแนนก่อนเรียน (หรือก่อนทดลอง) ของกลุ่มควบคุม ($M = 52.80$, $SD = 2.79$) และกลุ่มทดลอง ($M = 53.10$, $SD = 2.81$) แตกต่างกันแต่ไม่มี

นัยสำคัญทางสถิติ, $t(58) = -0.28, p = .78$ แสดงว่าทั้งสองกลุ่มมีความใกล้เคียงกันก่อนการทดลอง อย่างไรก็ตาม คะแนนหลังเรียนของกลุ่มทดลอง ($M = 63.90, SD = 4.82$) สูงกว่ากลุ่มควบคุม ($M = 60.03, SD = 4.16$) อย่างมีนัยสำคัญทางสถิติ, $t(58) = -2.26, p = .028$ โดยมีขนาดอิทธิพลในระดับปานกลาง (Cohen's $d = 0.58$) และการทดสอบ dependent-samples t พบว่า ทั้งสองกลุ่มมีคะแนนหลังเรียนสูงกว่าก่อนเรียนอย่างมีนัยสำคัญทางสถิติ กลุ่มควบคุมมีค่า $t(29) = 9.50, p < .001$ และกลุ่มทดลองมีค่า $t(29) = 8.01, p < .001$

ผลลัพธ์แบบนี้อาจจะดู “สวย” ทางสถิติ และครุณักวิจัยจำนวนไม่น้อยก็รู้สึกพอใจว่างานวิจัยสำเร็จเรียบร้อย แต่ถ้าเป็นนักวิจัยที่มีสติและรู้วิธีวิทยาการวิจัยที่ถูกต้องร่วมกับมุมมองของครูหรือจากจิตวิญญาณของผู้ที่ต้องการเข้าใจและแก้ปัญหาของผู้เรียนจริง ๆ ควรตั้งคำถามต่อ เช่น ทำไมนักเรียนบางคนมีคะแนนหลังเรียนน้อยกว่าก่อนเรียน ทำไมบางคนในกลุ่มทดลองทำคะแนนได้ต่ำกว่าค่าเฉลี่ยของกลุ่มควบคุม เกิดอะไรขึ้นกับผู้เรียนกลุ่มนี้ treatment ที่ใช้มีอะไรจุดอ่อนหรืออาจไม่ได้ออกแบบมาเพื่อตอบโจทย์ของเด็กกลุ่มนี้จริง ๆ และถ้าเกณฑ์ผ่านคือ 80 คะแนน ครุณักวิจัยยังยินดีกับผลเช่นนี้หรือไม่ คำถามเหล่านี้ชี้ให้เห็นว่าการตีความผลจาก t-test อย่างผิวเผิน โดยไม่กลับไปดูข้อมูลรายบุคคลหรือพฤติกรรมในชั้นเรียน อาจทำให้นักวิจัยพลาดประเด็นสำคัญ การวิเคราะห์แบบนี้บอกได้เพียงว่า “ผลต่างมีนัยสำคัญหรือไม่” แต่ไม่สามารถตอบคำถามว่า “เพราะอะไร” หรือ “อย่างไร” จึงเกิดผลลัพธ์เหล่านั้นขึ้น

ตาราง 1

ข้อมูลสมมุติจากการวิจัยเชิงทดลองแบบ Pretest-Posttest Control Group Design ($n_c = n_e = 30$)

Control Group						Experimental Group					
id	pretest	posttest	id	pretest	posttest	id	pretest	posttest	id	pretest	posttest
1	50	60	16	59	68	31	54	58	46	57	67
2	51	49	17	59	65	32	65	59	47	64	68
3	53	52	18	59	65	33	61	60	48	59	68
4	53	56	19	60	70	34	65	60	49	56	68
5	54	62	20	61	66	35	58	61	50	58	68
6	55	60	21	61	64	36	51	61	51	61	69
7	55	61	22	62	66	37	50	62	52	61	72
8	57	65	23	62	66	38	53	62	53	64	73
9	57	60	24	62	68	39	55	63	54	62	73
10	57	60	25	63	65	40	53	63	55	65	74
11	58	64	26	63	66	41	62	64	56	58	75
12	58	64	27	64	72	42	58	64	57	62	76
13	58	67	28	67	72	43	56	66	58	65	78
14	59	64	29	68	69	44	57	67	59	69	83
15	59	66	30	68	73	45	60	67	60	63	83

หมายเหตุ ตารางนี้ตั้งใจใช้หัวคอลัมน์เป็นภาษาอังกฤษเพื่อเน้นว่า pretest และ posttest ในแต่ละค่าไม่ต้องมีเครื่องหมายยัติภังค์ (- หรือ hyphen) ตามหลักการใช้ prefix และ hyphenation ของ APA 7th edition

นอกจากข้อสังเกตเรื่องการใช้ t-test ที่ผิดพลาดในเชิงวิวิธวิทยาแล้ว ผู้เขียนอยากชวนครุ
นิสิตนักศึกษา และนักวิจัยที่ทำวิจัยในลักษณะนี้ ลองพิจารณาทางเลือกอื่นที่ให้ข้อมูลเชิงลึกกว่า
การดูแค่ค่าเฉลี่ยแตกต่างกันอย่างมีนัยสำคัญทางสถิติหรือไม่ เพราะก่อนจะใช้สถิติอ้างอิง เรา
ควรตั้งสติถามตัวเองก่อนว่า งานวิจัยปฏิบัติการชั้นเรียนจำเป็นต้องอ้างอิงไปยังประชากรใด
หรือไม่ ต้องสุ่มตัวอย่าง (random sampling) หรือไม่ และถ้าเราทำวิจัยเพื่อแก้ปัญหาในชั้น
เรียนของตัวเองจริง ๆ ผลวิจัยนั้นควรถูกนำไปอ้างอิงกับประชากรอะไร เพื่ออะไร?! เคยอัปเดต
บ้างไหมว่างานวิจัยแนวนั้นในต่างประเทศเขาทำกันแบบนี้หรือไม่ และตรรกะเบื้องหลังมันคือ
อะไร งานลักษณะนี้หัวใจอยู่ที่การเข้าใจปัญหาและความเปลี่ยนแปลงของผู้เรียน โดยเฉพาะ
รายบุคคล (เพราะเราต้องการแก้ปัญหาให้เขา) ภายใต้อุปสรรคชั้นเรียนของนักวิจัยเอง สิ่งสำคัญคือ
causality ระหว่าง treatment ที่ออกแบบกับตัวแปรตาม กล่าวคือ treatment ที่ใช้ก่อให้เกิดผล
ที่ตั้งเป้าไว้ได้จริงหรือไม่ สุ่มสุ่มอะไรไม่เกี่ยวเลย ส่วนการนำผลวิจัยไปใช้ ก็ไม่จำเป็นต้อง infer
หรือ generalize กลับไปยังประชากรใด ผู้อ่านสามารถใช้ผลการวิจัยโดยพิจารณาเทียบเคียงกับ
บริบทของตนเอง สภาพปัญหาหรือสาเหตุรากของปัญหา (root cause) มีลักษณะคล้ายกันหรือไม่
นักเรียนในชั้นเรียนของตนมีบริบทที่ใกล้เคียงกับกรณีในงานวิจัยหรือไม่ แนวคิดนี้คล้ายกับการ
เทียบ case ในงานวิจัยเชิงคุณภาพหรือการใช้เหตุผลเชิงกรณี (case-based reasoning) เพื่อ
ตัดสินใจว่าจะนำ treatment มาปรับใช้กับชั้นเรียนตนเองได้หรือไม่ และน่าจะเกิดผลอย่างไร

หากครุ นิสิตนักศึกษา หรือนักวิจัยรู้สึกว่ายังไม่ถนัดการวิเคราะห์สถิติอ้างอิง ทางเลือกที่
ตรงไปตรงมาคือเรียนรู้เพิ่มเติมอย่างจริงจัง ไม่ต้องไปกลัวสถิติ หรืออย่างน้อยก็เลือกใช้วิธีการ
อื่นที่เหมาะสมกว่า โดยไม่จำเป็นต้องฟันใช้สถิติที่ตนเองไม่เข้าใจว่าใช้ไปทำไม หรือคัดลอก
แนวทางเดิม ๆ ที่ผิดมาตั้งแต่ต้นแล้วทำต่อกันมาอย่างไม่มีที่จบ ทบทวน ความไม่มั่นใจในสถิติเชิง
ซับซ้อน ไม่ใช่ข้ออ้างในการใช้สถิติผิด ๆ ซ้ำแล้วซ้ำอีก

วิธีที่เข้าใจง่ายและมีความหมายต่อการพัฒนาการเรียนรู้ของนักเรียนมากกว่า คือการ
โฟกัสที่ผลที่เกิดขึ้นกับผู้เรียนแต่ละคน การวิเคราะห์ด้วยวิธีง่าย ๆ เช่น (1) การใช้ตัวชี้วัดการ
เปลี่ยนแปลง เช่น relative gain score หรือ normalized gain score และ (2) การนำเสนอ
พัฒนาการรายบุคคลด้วยกราฟเส้น (line graph) ซึ่งคล้ายกับที่ใช้ใน single-subject design
หรือ single-case experimental design สามารถช่วยให้เข้าใจพัฒนาการของผู้เรียนแต่ละคนได้
ชัดเจนกว่าการดูแค่ค่าเฉลี่ยกลุ่มหรือความแตกต่างทางสถิติ

ตารางที่ 2 แสดงข้อมูลสมมุติของนักเรียน 5 คน ประกอบด้วยคะแนน pretest, posttest,
gain score, relative gain score (หรือ potential gain score) และ normalized gain score
โดยกำหนดให้คะแนนเต็มคือ 100 คะแนน การคำนวณ gain score ใช้วิธีตรงไปตรงมา คือ
posttest – pretest ส่วน relative gain score คำนวณโดยสูตร

$$\text{relative gain score} = \frac{\text{posttest} - \text{pretest}}{100 - \text{pretest}}$$

สูตรนี้แสดงสัดส่วนของคะแนนที่เพิ่มขึ้นจากโอกาสที่นักเรียนแต่ละคนจะพัฒนาได้ (ซึ่งไม่เท่ากัน
เพราะ pretest ต่างกัน) ส่วน normalized gain score หรือ <g> ซึ่งเสนอโดย Hake (1998) เป็น

ตัวชี้วัดการพัฒนาการเรียนรู้ที่ใช้กันแพร่หลายในงานวิจัยด้านฟิสิกส์ศึกษาและการเรียนรู้วิทยาศาสตร์ โดยคำนวณจากคะแนนที่แปลงเป็นเปอร์เซ็นต์แล้ว ดังนี้

$$\langle g \rangle = \frac{\% \langle post \rangle - \% \langle pre \rangle}{100 - \% \langle pre \rangle}$$

ค่า $\langle g \rangle$ แสดงให้เห็นว่าผู้เรียนพัฒนาได้มากน้อยเพียงใด (คิดเป็น %) เทียบกับโอกาสสูงสุดที่พัฒนาได้จากจุดเริ่มต้นเดิม

เมื่อพิจารณาค่าที่คำนวณได้จากตาราง 2 จะเห็นได้ชัดว่าการใช้ gain score แบบดิบ ๆ อาจทำให้เข้าใจผิด เพราะนักเรียนที่ได้คะแนนเพิ่มเท่ากันอาจมีระดับการพัฒนาที่ต่างกัน หากเขาเริ่มต้น (pretest) จากคนละจุด ตัวอย่างเช่น นักเรียน A ได้คะแนนเพิ่มขึ้น 10 คะแนนจาก pretest 30 ไป posttest 40 ขณะที่นักเรียน E ก็เพิ่ม 10 คะแนนจาก 80 ไป 90 ทั้งสองคนมี gain score เท่ากัน แต่ในแง่โอกาสในการพัฒนา นักเรียน A พัฒนาไปมากกว่านักเรียน E อย่างชัดเจน การใช้ relative gain score หรือ normalized gain score จึงช่วยให้เห็นรายละเอียดที่ gain score แบบดิบไม่สามารถอธิบายได้ และชี้เป้าได้ชัดเจนว่านักเรียนคนใดบ้างที่ยังต้องการการแก้ไขหรือช่วยเหลือให้สามารถพัฒนาเพิ่มขึ้นอีก นอกจากนี้ค่าสถิติทั้งสองนี้คำนวณง่าย ไม่ต้องใช้โปรแกรมขั้นสูง ติความได้ง่าย และสามารถนำเสนอควบคู่กับข้อมูลเชิงคุณภาพเพื่อเสริมความเข้าใจได้อีกชั้นหนึ่ง ถึงตอนนี้แล้วเห็นความแตกต่างจากการวิเคราะห์ด้วย t-test หรือยังว่าแบบไหนผิวนั้น แบบไหนให้รายละเอียดที่นำไปใช้ให้เป็นประโยชน์กับนักเรียนได้มากกว่า

ตาราง 2

ตัวอย่างการคำนวณคะแนนพัฒนาการรายบุคคลด้วยสูตรต่าง ๆ

Student	Pretest	Posttest	Gain Score	Relative Gain Score	Normalized Gain Score
A	30	40	10	.14	.14
B	40	50	10	.17	.17
C	50	60	10	.20	.20
D	60	70	10	.25	.25
E	80	90	10	.50	.50

ก่อนจะไปวิธีที่ 2 ผู้เขียนขอแทรกเกร็ดเล็กน้อยเกี่ยวกับการใช้ relative gain score หรือที่ในบางงานเรียกว่า potential gain score โดยทั่วไป บทความวิจัยทางการศึกษาของไทยที่กล่าวถึงหรือใช้การคำนวณในลักษณะนี้ มักอ้างอิงเอกสารหลักอยู่ 2 รายการ คือ

- (1) Kanjanawasee, S. (1989). *Alternative strategies for policy analysis: An assessment of school effects on students' cognitive and affective mathematics outcomes in lower secondary schools in Thailand* (Publication No. 8910299) [Doctoral dissertation, University of California, Los Angeles]. ProQuest Dissertations & Theses Global.
- (2) ศิริชัย กาญจนวาสี. (2557). การคำนวณคะแนนพัฒนาการ (Gain Scores). *วารสารสมาคมวิจัยสังคมศาสตร์แห่งประเทศไทย*, 1(1), 1–20.

เอกสารฉบับแรกเป็นวิทยานิพนธ์ระดับปริญญาเอกจาก UCLA นั้นเป็นแหล่งที่เข้าถึงได้ยากมาก ไม่มีฉบับตัวเล่มให้บริการในห้องสมุดของมหาวิทยาลัยไทย (เท่าที่ได้ลองสืบค้น) และไม่สามารถดาวน์โหลดจากฐานข้อมูลวิทยานิพนธ์ใด ๆ ได้ เว้นแต่จะสั่งซื้อผ่านฐานข้อมูลสากล เช่น ProQuest Dissertations & Theses Global เท่านั้น อดีตเคยมีสำเนาในหอสมุดคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อประมาณ 20–30 ปีก่อน แต่ในปัจจุบันไม่มีแล้ว หรือหากท่านได้ทราบแหล่งที่เข้าถึงได้โปรดกรุณาชี้แนะช่องทางเพื่อเป็นประโยชน์ทางวิชาการแก่ผู้อ่านท่านอื่น ๆ ด้วย (ส่วนผู้เขียนมีไฟล์ pdf ที่สั่งซื้อจาก ProQuest แล้ว)

สิ่งที่น่าสังเกตคือ บทความและวิทยานิพนธ์จำนวนมากอ้างอิงวิทยานิพนธ์ (1) เล่มนี้ โดยไม่ได้ระบุว่าได้อ่านจากเอกสารทุติยภูมิ (secondary source) ผู้เขียนจึงอดตั้งคำถามในใจไม่ได้ว่า นักวิจัยเหล่านี้ได้อ่านต้นฉบับจริงหรือไม่ ด้วยความคุ้นเคยกับแหล่งข้อมูลและระบบห้องสมุดระดับอุดมศึกษาของไทย ผู้เขียนมีความเชื่ออย่างสูงว่า นักวิจัยส่วนใหญ่ไม่ได้เข้าถึงต้นฉบับดังกล่าวโดยตรง แต่อ่านมาจากบทความหรือวิทยานิพนธ์ของคนอื่นที่อ้างไว้ก่อนแล้ว แล้วก็เอามาเขียนต่ออีกทีแต่อ้างเหมือนอ้างต้นฉบับเอง หากสมมุตินี้เป็นจริง ก็ถือว่าเป็นการละเมิดจรรยาบรรณทางวิชาการ เพราะเป็นการอ้างอิงในลักษณะที่ทำให้เข้าใจผิดว่าได้อ่านต้นฉบับแล้ว ทั้งที่จริงไม่ได้อ่าน แนวปฏิบัติที่ถูกต้องควรเป็นการอ้างอิงแบบ cited in หรือไม่ก็เลือกอ้างบทความปี 2557 ซึ่งผู้แต่งคนเดียวกันได้เรียบเรียงไว้ใหม่ และสามารถเข้าถึงได้ไม่ยาก การเลือกอ้างแหล่งที่เข้าถึงได้จริงและตรงตามที่ขอยืมแสดงถึงความรับผิดชอบในฐานะนักวิจัยที่ดี

เกร็ดเพิ่มเติมเกี่ยวกับ relative gain score ที่ผู้เขียนอยากบันทึกไว้ คือ Lord (1967) มักถูกอ้างถึงว่าเป็นผู้ชี้ให้เห็นข้อจำกัดของการใช้ gain score อย่างไรก็ตาม Lord ไม่ได้เสนอ relative gain score แต่เสนอให้ใช้วิธีวิเคราะห์อื่น เช่น ANCOVA โดยใช้ pretest เป็น covariate เพื่อลดอคติจากความไม่เท่ากันของค่าก่อนเรียนระหว่างกลุ่ม วิธีนี้สอดคล้องกับข้อเท็จจริงที่ว่า แม้แต่ผู้เรียนในกลุ่มเดียวกันก็แทบจะไม่มีทางได้คะแนน pretest เท่ากันทั้งหมด จึงควรปรับอิทธิพลของ pretest ระหว่างกลุ่มควบคุมและกลุ่มทดลอง แทนที่จะ assume หรือบังคับให้เท่ากัน ขณะเดียวกัน ผู้เขียนยังพบว่า มีงานวิจัยช่วง 1950–1980 ที่เสนอแนวคิดเกี่ยวกับ relative gain score จำนวนมาก อาทิ Zuckerman (1954) Osborn & Kraemer (1974) และ Hunt (1987) ซึ่งแสดงให้เห็นว่าการคิดเชิงสัดส่วนเพื่อสะท้อนพัฒนาการจากจุดเริ่มต้นของแต่ละบุคคลเป็นไอเดียที่มีรากฐานมานานมากแล้ว ดังแสดงในภาพ 2–4

การบันทึกเกร็ดเกี่ยวเรื่อง potential gain score นี้ไม่ได้มีเจตนาจะเสาะหาว่านักวิชาการท่านใดเป็นผู้คิดค้นสูตรเป็นคนแรก แต่ผู้เขียนต้องการบันทึกปรากฏการณ์ที่คิดว่า “มหัศจรรย์ทางวิชาการ” ในวิธีวิทยาการวิจัย โดยเฉพาะที่เกี่ยวกับการวัดผล สถิติ และการวิเคราะห์ทางจิตวิทยา มีปรากฏการณ์หนึ่งที่น่าสนใจ คือ การที่แนวคิดหรือโมเดลทางสถิติที่มีลักษณะคล้ายกันถูกพัฒนาขึ้นในช่วงเวลาที่ใกล้เคียงกัน โดยนักวิจัยจากต่างศาสตร์หรือต่างประเทศกัน ไม่ได้อ้างอิงงานของกันและกัน แต่เชื่อได้อย่างมั่นใจว่าไม่ได้ลอกเลียนกัน เพราะในยุคนั้นยังไม่สามารถสืบค้นผลงานของกันและกันได้ง่ายเหมือนปัจจุบัน และในทุกตัวอย่างที่จะนำเสนอต่อไปนี้ล้วนมีหลักฐานชัดเจนว่าเกิดจากการคิดหรือพัฒนาอย่างเป็นอิสระจากกัน (นั่นคือไม่มีใครลอกใคร)

ภาพ 2

แนวคิดเกี่ยวกับ *relative gain score* ในงานของ Zuckerman (1954)

ages for the two treatments. Several types of measures were used : (a) after-scores, (b) absolute-gain scores (after-minus-before percentages), (c) adjusted absolute-gain scores (after-minus-before differences), with the “before” base stabilized by lumping the before-test scores for both film and filmstrip group and basing the gains on the combined before percentages, and (d) adjusted “relative-gain” scores (obtained by dividing the difference between after-scores and combined before-values by the maximum change possible).⁵

Results. Table I lists the percentages of correct answers to

⁵Substantially the same pattern of results was obtained, however, when data for the discarded items were examined.

⁶Relative-gain scores take account of the fact that there was more room for the improvement on some items than others because of differences in before-level of correct responses. Positive differences in after-minus-before percentages were divided by 100 percent minus the before value to express the gains as a ratio of absolute gain to maximum possible gain. The few negative differences were expressed as a ratio of maximum room for negative change by dividing the after-minus-before differences by the before value.

ภาพ 3

แนวคิดเกี่ยวกับ *relative gain score* ในงานของ Osborn & Kraemer (1974)

An additional conversion was made in the basic data primarily to reduce the pre- vs. post-test scores to a single measure, but also to dispel concern over the small pre-test difference in the two final study groups. “Relative gain” scores were computed by dividing the difference in adjusted post- and pre-test scores by one of two quantities: (a) if post-test minus pre-test was positive, as in most cases, the difference was divided by the maximum possible gain that a man could have attained; (b) if the difference was negative, it was divided by a maximum possible loss. Relative gain scores indicate the amount of improvement shown by an individual as a proportion of what he might have achieved. These scores are used in the subsequent analysis.

ภาพ 4

แนวคิดเกี่ยวกับ *relative gain score* ในงานของ Hunt (1987)

academic gains. Correlations are often computed with teaching behavior rating scores and student raw gain scores (pretest scores subtracted from posttest scores). When criterion measures vary in difficulty, item type, and length and are used in the same study with different students to measure student academic gain, a percentage gain score:

$$\frac{\text{raw gain (posttest - pretest)}}{\text{possible score - pretest}} \times 100$$

is often computed to adjust for the differences in the tests being used. Correlations are then computed using teaching behavior rating scores and percentage gains (Ellett et al., 1980; McGarity, 1981; Warren, 1985).

ปรากฏการณ์เช่นนี้มักเรียกว่า multiple discovery หรือ parallel development ส่วนกรณีที่ห่างกันมากแต่บังเอิญคิดซ้ำในภายหลัง อาจเรียกว่า reinvention หรือ rediscovery (โดยบริสุทธิ์ใจ ไม่ได้ลอกแล้วมาอ้างว่าเป็นต้นฉบับหรือคิดพัฒนาเองเป็นคนแรก) เช่น

- **Rasch model** โดย Rasch (1960) กับ **1-parameter logistic model** ในสาย IRT โดย Lord (1968) ที่พัฒนาเกือบสิบปีให้หลังจากคนละศาสตร์ แต่มีการ parameterization คล้ายกันมาก
- **Confirmatory Factor Analysis** โดย Jöreskog (1969) ในสาย psychometrics และ **Structural Equation Modeling** โดย Duncan (1975) ในสายสังคมวิทยา ต่างฝ่ายต่างใช้กรอบแนวคิด covariance structure modeling โดยไม่ได้อ้างอิงซึ่งกันและกัน
- **Principal Component Analysis** จาก Pearson (1901) และ Hotelling (1933) กับ **Exploratory Factor Analysis** จาก Spearman (1904) และ Thurstone (1947) ที่แม้เป้าหมายต่างกัน แต่ใช้วิธีการคณิตศาสตร์ใกล้เคียงกันมาก

ตัวอย่างเหล่านี้สะท้อนว่า นักวิจัยในสาขาต่าง ๆ อาจเริ่มต้นจากโจทย์หรือบริบทที่ต่างกัน แต่เมื่อเจอข้อจำกัดของข้อมูลหรือแนวคิดทางคณิตศาสตร์เหมือนกัน แนวทางการวิเคราะห์หรือการสร้างโมเดลก็อาจคล้ายกันมากจนดูเหมือนเป็นการคิดซ้ำ ทั้งที่เกิดจากการคิดค้นอิสระอย่างบริสุทธิ์ใจและตรวจสอบย้อนกลับได้ เป็นเรื่องที่คุณเขียนคิดว่าน่าฟังดี

อีกหนึ่งทางเลือกในการวิเคราะห์ข้อมูลที่คุณเขียนขอเสนอ คือ การใช้กราฟเส้นรายบุคคล ลักษณะเดียวกับที่ใช้ใน single-case design โดยคำว่า single ในที่นี้ไม่ได้หมายความว่าต้องเป็นงานวิจัยที่มีผู้เข้าร่วมเพียงหนึ่งคนเท่านั้น (ซึ่งกรณีนั้นจะเรียกเจาะจงว่า N-of-1 design) แต่หมายถึงการวิเคราะห์ข้อมูลรายบุคคลเป็นหลัก แม้ว่าการศึกษาอาจมีผู้เข้าร่วมหลายคนที่มีลักษณะหรือปัญหาใกล้เคียงกันก็ตาม จุดเน้นของวิธีการนี้คือการติดตามการเปลี่ยนแปลงของแต่ละคน โดยไม่พึ่งการวิเคราะห์เชิงสถิติอ้างอิงเป็นหลัก แต่ใช้การตรวจสอบด้วยสายตา (visual inspection) เพื่อทำความเข้าใจลักษณะและแนวโน้มของพฤติกรรมหรือผลลัพธ์ที่เกิดขึ้นในแต่ละช่วงเวลา (Lane & Gast, 2014)

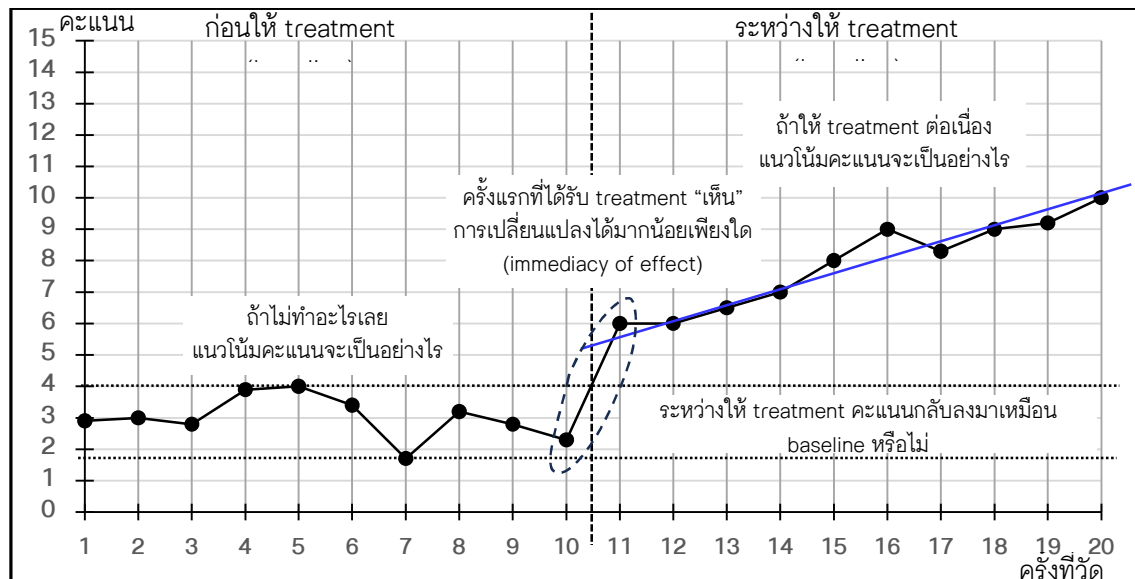
ผู้เขียนเตรียมกราฟตัวอย่างไว้ในภาพ 5 ซึ่งสร้างจากโปรแกรม Excel อย่างง่าย ๆ โดยมีแกนนอนแสดงลำดับเวลา (ครั้งที่เก็บข้อมูลหรือจะเป็นคาบ เป็นวัน เป็นสัปดาห์ แล้วแต่ผู้วิจัยออกแบบ) ตั้งแต่ครั้งที่ 1 ถึง 20 โดยช่วงที่ 1–10 เป็นช่วงก่อนเริ่มใช้ treatment (baseline) และช่วงที่ 11–20 เป็นช่วงที่มีการให้ treatment ส่วนแกนตั้งแสดงคะแนนของตัวแปรตาม เช่น คะแนน quiz จำนวนพฤติกรรมที่สังเกตได้ หรือจำนวนคำที่อ่านถูกต้อง โดยในกราฟตัวอย่างสมมติว่าเป็นคะแนน quiz มีคะแนนเต็ม 15 คะแนน

แม้กราฟลักษณะนี้จะไม่มีการทดสอบทางสถิติใด ๆ แต่สามารถให้ข้อมูลที่เปี่ยมประโยชน์ในเชิงลึกได้ เช่น คะแนนต่ำสุด-สูงสุด ความผันแปรของคะแนน ความเปลี่ยนแปลงของคะแนนทันทีเมื่อเริ่มให้ treatment (จุดที่ 10–11) ที่เรียกว่า immediacy of effect หรือ onset of change ความคงเส้นคงวาของผลระหว่างช่วงที่มี treatment มีแนวโน้มเพิ่มขึ้น สม่ำเสมอ หรือแปรปรวนมากน้อยเพียงใด คะแนนกลับไปต่ำเท่ากับช่วง baseline หรือไม่ ฯลฯ เหล่านี้เป็นข้อมูลที่สะท้อนผลของ treatment ได้ละเอียดชัดเจนยิ่งกว่าการทดสอบค่าเฉลี่ยของกลุ่ม

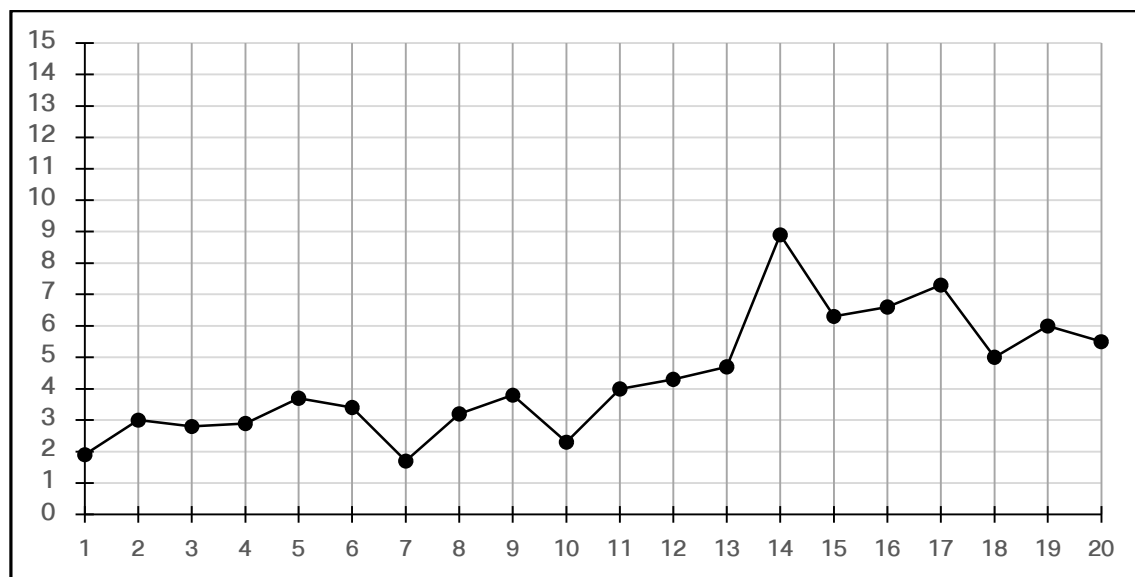
เมื่อมีข้อมูลรายบุคคลหลายคน ก็สามารถเปรียบเทียบกราฟของแต่ละคนเพื่อสังเกตรูปแบบการตอบสนองต่อ treatment เช่น นักเรียนคนใดตอบสนองในทางบวกและมีลักษณะร่วมกันอย่างไร นักเรียนคนใดไม่ตอบสนองหรือมีแนวโน้มเชิงลบและอาจมีอุปสรรคคล้ายคลึงกันอย่างไรบ้าง สิ่งเหล่านี้จะเป็นประโยชน์ต่อการปรับปรุง treatment ให้เหมาะสมและแม่นยำยิ่งขึ้น การใช้กราฟเพื่อแสดงพัฒนาการรายบุคคล เป็นแนวทางวิเคราะห์ที่เหมาะสมกับงานวิจัยแนวปฏิบัติการชั้นเรียน โดยเฉพาะกรณีที่ต้องการสะท้อนความเปลี่ยนแปลงของผู้เรียนแต่ละคนมากกว่าการวิเคราะห์ผลของกลุ่มในภาพรวม

ภาพ 5

กราฟแสดงการเปลี่ยนแปลงคะแนนของนักเรียน 2 คนและตัวอย่างคำถามการวิเคราะห์อย่างง่าย



นักเรียนคนที่ 1



นักเรียนคนที่ 2 (ลอง visual inspection เทียบกับของนักเรียนคนที่ 1 ว่าเหมือน/ต่างกันอย่างไร)

การวิเคราะห์แบบที่กล่าวมาทั้งหมดข้างต้นนี้ ไม่ได้ลดทอนความเป็นวิชาการแต่อย่างใด ตรงกันข้าม มันคือการก้าวออกจากกรอบสถิติที่ผิด ๆ ซ้ำ ๆ ไปสู่การมองข้อมูลอย่างมีเหตุผล ละเอียดยิ่งขึ้น ใครที่ถนัดสถิติก็ใช้เครื่องมือให้เหมาะกับคำถามวิจัย ใครที่ยังไม่ถนัด ก็สามารถเลือกวิธีที่ช่วยให้เข้าใจปัญหาในห้องเรียนและพฤติกรรมของผู้เรียนได้จริง โดยไม่ต้องอ้าง $p < .05$ ไปเรื่อย โดยไม่รู้ด้วยซ้ำว่าเกิดอะไรขึ้นระหว่างการสอน

สำหรับผู้สนใจงานสถิติเชิงลึกและมีภูมิด้านทฤษฎีความยากของเทคนิคขั้นสูง ปัจจุบันมีเครื่องมือทางสถิติและวิทยาการข้อมูล (data science) มากมายที่สามารถนำมาใช้กับการวิเคราะห์ข้อมูลในงานวิจัยปฏิบัติการชั้นเรียน หรือการวิจัยที่พัฒนา treatment ได้อย่างมีประสิทธิภาพ เช่น การใช้ propensity score matching, causal inference, causal forest, หรือ double machine learning หรือในกรณีที่มีการวัดซ้ำ เช่น latent growth curve modeling, time series modeling หรือ latent change modeling รวมถึงกลุ่มที่ผสมผสานระหว่างโมเดลสมการเชิงโครงสร้าง (SEM) กับ machine learning และ neural network หรือที่เรียกว่า hybrid SEM-ML/NN ที่ตรวจสอบทั้งความสัมพันธ์เชิงทฤษฎีระหว่างตัวแปรแฝงและสังเกตได้ และความไม่แน่นอนของค่าประมาณและผลการวิเคราะห์ที่เป็นจุดอ่อนของ SEM แบบดั้งเดิม

แม้เทคนิคเหล่านี้จะดูเหมือน “ซีโดโนเสาร์จับดักแดน” เมื่อใช้กับงานวิจัยขนาดเล็กหรือข้อมูลในห้องเรียน แต่ถ้าทำได้ก็ถือว่าคุ้มค่า เพราะช่วยแก้ปัญหาก็การใช้สถิติอ้างอิงแบบเดิมไม่สามารถจัดการได้ เช่น การที่กลุ่มทดลองและกลุ่มควบคุมมีความแตกต่างกันก่อนการทดลอง หรือมีปัจจัยแทรกแซงที่ไม่สามารถควบคุมได้โดยตรง ช่วยให้เห็นการเปลี่ยนแปลงในมิติเวลา มองเห็นรูปแบบของพัฒนาการที่ไม่เป็นเส้นตรง หรือไม่เกิดขึ้นพร้อมกันในทุกคนได้

สิ่งที่ได้จึงไม่ใช่แค่คำตอบว่า “ได้ผลหรือไม่” แต่ยังรวมถึงคำอธิบายว่า “ได้ผลกับใคร อย่างไร เมื่อไหร่ และภายใต้เงื่อนไขแบบไหน” ซึ่งมีประโยชน์ต่อการออกแบบการเรียนรู้ให้แม่นยำมากขึ้น และปรับใช้กับกลุ่มผู้เรียนที่มีลักษณะแตกต่างกันในอนาคต การวิเคราะห์แบบนี้ยังช่วยให้เห็น “กลุ่มย่อย” ของผู้เรียนที่ตอบสนองต่างจากภาพรวม ซึ่งมักถูกกลบด้วยค่าเฉลี่ยในวิธีวิเคราะห์แบบเดิม นอกจากนี้ยังเปิดโอกาสให้นักวิจัยตั้งคำถามใหม่ที่เฉียบคมขึ้นจากสิ่งที่พบในข้อมูล ไม่ใช่แค่ทดสอบสมมติฐานแบบตายตัวแล้วจบ พูดอีกแบบคือ หากมีเครื่องมือที่แม่นยำ ชัดกว่า และลึกกว่า แล้วใช้เป็น ก็จะได้ insights มากกว่าผลลัพธ์เชิงปริมาณที่ “นัยสำคัญหรือไม่” เท่านั้น

อย่างไรก็ตาม บทความนี้ไม่ได้ต้องการขึ้นว่านักวิจัยทุกคน โดยเฉพาะครูหรือนิสิตที่ไม่ได้เรียนสถิติเข้มข้น ต้องหันมาใช้เทคนิคขั้นสูงเหล่านี้ทันที จุดประสงค์คืออยากให้รู้ว่ามีเครื่องมือที่ก้าวหน้าและหลากหลายมากกว่าที่เคยใช้กันมา และการรู้ว่ามีตัวเลือกเหล่านี้อยู่ ก็อาจช่วยเปิดมุมมองใหม่ในการออกแบบงานวิจัยให้เหมาะกับปัญหามากขึ้น รวมถึงขยับไปสู่การวิเคราะห์ที่ลึกและมีเหตุผลมากกว่าในอนาคต หากอัปเดตความรู้และทักษะเหล่านี้ได้ ก็ถือเป็นอาวุธทางวิชาชีพที่มีคุณค่า แต่หากยังไม่พร้อม อย่างน้อยก็ควรเลี่ยงการใช้ t-test แบบผิด ๆ เช่น ใช้ผิดประเภท ไม่ตรวจสอบข้อตกลงเบื้องต้น หรือสนใจแค่ p-value ว่ามีนัยสำคัญหรือไม่ โดยไม่สนใจพฤติกรรมจริงของผู้เรียนในชั้นเรียน ข้อมูลของนักเรียนเพียงคนเดียว อาจให้ insight ที่สำคัญกว่าค่าเฉลี่ยของกลุ่มทั้งกลุ่ม การใช้กราฟเส้นง่าย ๆ หรือการวิเคราะห์เชิงพรรณนาอย่างมีเหตุผล อาจให้คำตอบที่สัมพันธ์กับบริบทของการสอนได้มากกว่า และมีคุณค่าทางวิชาการมากกว่าการสรุป

เพียงว่า “ค่าเฉลี่ยของกลุ่มทดลองสูงกว่ากลุ่มควบคุม” หรือ “คะแนนหลังเรียนสูงกว่าก่อนเรียน อย่างมีนัยสำคัญทางสถิติที่ระดับ .05”

บทสรุปทิ้งท้าย

บทความนี้ไม่ได้มีเป้าหมายเพื่อโจมตีสถิติหรือการวิจัยรูปแบบใดรูปแบบหนึ่ง แต่ต้องการชวนให้ทุกคนที่มีส่วนเกี่ยวข้องกับการวิจัยทางการศึกษา โดยเฉพาะการวิจัยปฏิบัติการชั้นเรียน หยุดแล้วคิด ทบทวน และตั้งคำถามกับแนวปฏิบัติที่เราเคยทำตาม ๆ กันมาโดยไม่ทันได้พิจารณาว่ายังเหมาะสมอยู่หรือไม่ในวันนี้

ขอเชิญชวน

- ครูและนักวิจัยในชั้นเรียน ใช้การสังเกตและติดตามพัฒนาการของผู้เรียนเป็นรายบุคคลมากกว่าการอ้าง p -value ว่า “มีนัยสำคัญ” หรือไม่
- อาจารย์ที่ปรึกษาและผู้ให้ความรู้ด้านวิจัย ชี้แนะให้นักศึกษาเลือกแนวทางการวิจัยที่เหมาะสมกับคำถามที่แท้จริงของงาน ไม่ติดกับดักสถิติที่ไม่สอดคล้องกับธรรมชาติของข้อมูล
- ผู้ประเมินงานวิชาการและบรรณาธิการวารสาร หยุดยึดติดกับข้อกำหนดทางสถิติที่ไม่จำเป็น และเปิดรับการวิเคราะห์ที่มีเหตุผล แม้ไม่ใช่สถิติขั้นสูง
- ผู้บริหารและคณะกรรมการประเมินวิทยฐานะ พิจารณางานวิจัยในชั้นเรียน (CAR) ให้สอดคล้องกับคุณภาพของการเรียนการสอนจริง ๆ ไม่ใช่แค่การผ่านเกณฑ์ทางสถิติ

ท้ายที่สุด ผู้เขียนคาดหวังและขอเรียกร้องไปยัง นักวิจัย นิสิตนักศึกษาระดับบัณฑิตศึกษา อาจารย์ที่ปรึกษาวิทยานิพนธ์ กรรมการสอบวิทยานิพนธ์ อาจารย์ผู้สอนวิชาวิจัย/สถิติ/วัดประเมิน ผู้ประเมินงานวิจัย/บทความของครูผู้ขอ ค.ศ. หน่วยงานที่เกี่ยวข้องที่มีอำนาจหน้าที่สามารถชี้ถูกผิดได้ เช่น สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) ศุภสภา คณะครุศาสตร์/ศึกษาศาสตร์ บรรณาธิการวารสารทางการศึกษา และผู้ประเมินบทความวิจัยให้วารสารต่าง ๆ ร่วมกันยกระดับการใช้วิธีวิทยาการวิจัยทางการศึกษาของไทยให้ดีขึ้น ถูกต้อง และทันสมัยเป็นสากล หลุดพ้นจาก cult ที่มีความเชื่องมภายใน “รูปแบบที่ดูวิชาการ” แต่ทั้งล้าสมัยและไม่ถูกต้องตามหลักวิธีวิทยา รวมถึงไม่ตอบโจทย์การพัฒนาการเรียนรู้ของผู้เรียนจริง ๆ

หากผู้อ่านท่านใดมีประสบการณ์เกี่ยวกับการใช้หรือการตั้งคำถามเกี่ยวกับการใช้สถิติในงานวิจัยปฏิบัติการชั้นเรียน ไม่ว่าจะในฐานะนักวิจัย ผู้สอน ผู้ประเมิน หรือผู้ถูกประเมิน ผู้เขียนขอเชิญร่วมสนทนาโดยส่งความเห็นมายังอีเมลของวารสารหรือโดยตรงถึงผู้เขียน ความคิดเห็นข้อเสนอแนะ และตัวอย่างจากประสบการณ์ที่เป็นประโยชน์อาจได้รับพิจารณาให้ตีพิมพ์เผยแพร่ใน JRM ต่อไป

รายการอ้างอิง

- ศิริชัย กาญจนวาสี. (2557). การคำนวณคะแนนพัฒนาการ (Gain Scores). *วารสารสมาคมวิจัยสังคมศาสตร์แห่งประเทศไทย*, 1(1), 1-20.
- สุวิมล ว่องวานิช. (2563). *การวิจัยการออกแบบทางการศึกษา*. สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally College Publishing.
Retrieved from <https://www.jameslindlibrary.org/campbell-dt-stanley-jc-1963/>
- Coletta, V. P., & Steinert, J. J. (2020). Why normalized gain should continue to be used in analyzing preinstruction and postinstruction scores on concept inventories. *Physical Review Physics Education Research*, 16(1), 010108.
<https://doi.org/10.1103/PhysRevPhysEducRes.16.010108>
- Duncan, O. D. (1975). *Introduction to structural equation models*. Academic Press.
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74.
<https://doi.org/10.1119/1.18809>
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441.
<https://doi.org/10.1037/h0071325>
- Hunt, C. (1987). *Predicting student academic gain through a competency-based teacher assessment instrument: A validation study* (Publication No. 8727198) [Doctoral dissertation, Georgia State University – College of Education]. ProQuest Dissertations Publishing.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202.
<https://doi.org/10.1007/BF02289343>
- Kanjanawasee, S. (1989). *Alternative strategies for policy analysis: An assessment of school effects on students' cognitive and affective mathematics outcomes in lower secondary schools in Thailand* (Publication No. 8910299) [Doctoral dissertation, University of California, Los Angeles]. ProQuest Dissertations & Theses Global.
- Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, 24(3–4), 445–463. <https://doi.org/10.1080/09602011.2013.815636>
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68(5), 304–305. <https://doi.org/10.1037/h0025105>

- Osborn, W. C., & Kraemer, R. E. (1974). *Development and evaluation of a pre-school study manual for drill sergeant candidates* (Professional Paper No. 5-74). Human Resources Research Organization. <https://eric.ed.gov/?id=ED095352>
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 559–572. <https://doi.org/10.1080/14786440109462720>
- Pistorello, J., Jobes, D. A., Compton, S. N., Locey, N. S., Walloch, J. C., Gallop, R., Au, J. S., Noose, S. K., Young, M., Johnson, J., Dickens, Y., Chatham, P., Jeffcoat, T., Dalto, G., & Goswami, S. (2018). Developing adaptive treatment strategies to address suicidal risk in college students: A pilot sequential, multiple assignment, randomized trial (SMART). *Archives of Suicide Research*, 22(4), 644–664. <https://doi.org/10.1080/13811118.2017.1392915>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>
- Student. (1908). The probable error of a mean. *Biometrika*, 5(1), 1–25. <https://doi.org/10.1093/biomet/6.1.1>
- Thurstone, L. L. (1947). *Multiple-factor analysis*. University of Chicago Press.
- Zuckerman, J. V. (1954). Predicting film learning by pre-release testing. *Audio-Visual Communication Review*, 2(1), 49–56. <https://doi.org/10.2307/30216707>

Translated Thai References

- Kanjanawasee, S. (2014). Calculation of gain scores. *Journal of the Social Science Research Association of Thailand*, 1(1), 1–20.
- Wongwanich, S. (2020). *Educational design research*. Chulalongkorn University Press.