Vol. 2 No. 2 (May - August 2026)
ปีที่ 2 ฉบับที่ 2 (พฤษภาคม - สิงหาคม 2569)
ISSN 3088-2893 (Online)

สำนักส่งเสริม สนับสนุนการผลิตพัฒนาแพทย์และบุคลากรทางการแพทย์



Medical and Health Professions Education

วารสารแพทยศาสตรศึกษาสำหรับบุคลากรทางการแพทย์

02-590-1980
https://so11.tci-thaijo.org
cpird_mhpe@cpird.in.th

Medical and Health Professions Education

วารสารแพทยศาสตรศึกษาสำหรับบุคลากรทางการแพทย์

Vol2 No. 2 (May - August 2026)

ปีที่ 2 ฉบับที่ 2 (พฤษภาคม - สิงหาคม 2569)

ISSN : 3088-2893 (Online)

วัตถุประสงค์

วารสารแพทยศาสตรศึกษาสำหรับบุคลากรทางการแพทย์ เป็นวารสารของสำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์ มีวัตถุประสงค์เพื่อการยกระดับคุณภาพการศึกษาของบุคลากรทางการแพทย์ โดยการส่งเสริมการศึกษาค้นคว้า การเผยแพร่บทความวิจัยและบทความวิชาการทางการศึกษาที่มีคุณภาพสูง และเข้าถึงได้ (open access) วารสารนี้มุ่งเน้นบทความทางการศึกษาในหัวข้อด้านวิชาชีพสุขภาพที่หลากหลาย เช่น วิธีการสอน และการเรียนรู้ การประเมิน และนวัตกรรมทางการศึกษา สำหรับบุคลากรทางการแพทย์ นักวิจัย และผู้นำทางการศึกษาด้านการดูแลสุขภาพ ทั้งนี้บทความที่ตีพิมพ์ลงวารสารจะได้รับการประเมินแบบ Double blinded peer - reviewed process (การตรวจสอบแบบปกปิดรายชื่อ) จากผู้ทรงคุณวุฒิที่มีความเชี่ยวชาญจากหลากหลายสถาบัน จำนวนอย่างน้อย 2 ท่าน ผ่านระบบ ThaiJO

บรรณาธิการอำนวยการ (Editor In Chief)

พญ.กนกวรรณ ศรีรักษา	ผู้ช่วยผู้อำนวยการสำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์

ที่ปรึกษากองบรรณาธิการ (Editorial Advisor)

นพ.ชิษณุพงศ์ ตั้งอดุลย์รัตน์	ผู้อำนวยการสำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์
ผศ.(พิเศษ) พญ.วาสนา หงษ์กัน	รองผู้อำนวยการสำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์
ผศ.(พิเศษ) นพ.ไพโรจน์ บุญลักษณ์ศิริ	ที่ปรึกษาผู้อำนวยการสำนักส่งเสริมสนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์
นพ.ปริญญา ชำนาญ	จุฬาลงกรณ์มหาวิทยาลัย

ผู้จัดการวารสาร (Journal Manager)

พญ.ทิพาพร ทองมาก	ผู้ช่วยผู้อำนวยการสำนักส่งเสริมสนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์
นพ.วิชช ธรรมปัญญา	โรงพยาบาลเชียงรายประชานุเคราะห์

กองบรรณาธิการ (Editorial Board Members)

พญ.ศรัญญา ประทัยเทพ	ผู้ช่วยผู้อำนวยการสำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์
พญ.อำนวยพร อภิรักษากร	ผู้ช่วยผู้อำนวยการสำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์
พญ.วรนาฏ รัตนากร	โรงพยาบาลชลบุรี
นพ.ธรรมสรณ์ จีรอำพรวัฒน์	โรงพยาบาลขอนแก่น
ดร.ภญ.เบญจพร ศิลารักษ์	โรงพยาบาลขอนแก่น
พญ.อรุณี ทิพย์วงศ์	โรงพยาบาลสุราษฎร์ธานี
พญ.วชิราภรณ์ อรุโณทอง	โรงพยาบาลลำปาง
พญ.ปิยาภรณ์ ศิริจันทร์ชื่น	โรงพยาบาลเชียงรายประชานุเคราะห์

ทีมงานวารสาร

นางสาวดลชนก ศรีตพงษ์	Medical Educator Support of MHPD
นางสาวดวงตะวัน ชิณวงศ์	Medical Educator Support of MHPD
นางสาวสิราวรรณ อุดมกิจทวี	Medical Educator Support of MHPD
นางสาวพชรพร หฤษฎ์นิธิพร	Medical Educator Support of MHPD
นางสาวขนิษฐา สินสมุทร	Medical Educator Support of MHPD
นายศุภกร จุลสำรวล	Medical Educator Support of MHPD

เจ้าของ

Medical and Health Professions Education วารสารแพทยศาสตรศึกษาสำหรับบุคลากรทางการแพทย์

สำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์

88/20 ถนนติวานนท์ ตำบลตลาดขวัญ จังหวัดนนทบุรี 11000, ประเทศไทย

E-mail : cpird_mhpe@cpird.in.th

Telephone : 02-590-1980

Website : https://so11.tci-thaijo.org

กำหนดออก

การเผยแพร่ 3 ฉบับต่อปี (ราย 4 เดือน) ดังนี้

ฉบับที่ 1 กันยายน - ธันวาคม

ฉบับที่ 2 มกราคม - เมษายน

ฉบับที่ 3 พฤษภาคม - สิงหาคม

รูปแบบการพิมพ์

จัดพิมพ์เป็นรูปแบบออนไลน์และเผยแพร่ผ่านเว็ปไซต์วารสารวารสารแพทยศาสตรศึกษาสำหรับบุคลากรทางการแพทย์ (Thaijo)

- บทความที่ได้รับการตีพิมพ์เป็นลิขสิทธิ์ของสำนักส่งเสริม สนับสนุนการผลิต พัฒนาแพทย์ และบุคลากรทางการแพทย์ กระทรวงสาธารณสุข
- ข้อความที่ปรากฏในบทความแต่ละเรื่องในวารสารวิชาการเล่มนี้เป็นความคิดเห็นส่วนตัวของผู้เขียนแต่ละท่าน ความรับผิดชอบ องค์ประกอบทั้งหมดของบทความแต่ละเรื่องเป็นของผู้เขียนแต่ละท่าน

บทบรรณาธิการ

วารสารแพทยศาสตรศึกษาสำหรับบุคลากรทางการแพทย์ หรือชื่อภาษาอังกฤษว่า Medical and Health Profession Education (MHPE) ฉบับนี้เป็นฉบับที่ 2 ของปี พ.ศ. 2569 ภายใต้การสนับสนุนของสำนักส่งเสริมสนับสนุนการผลิต พัฒนาแพทย์และบุคลากรทางการแพทย์ (สพพ.) ในฉบับนี้มีบทความปริทรรศน์ที่น่าสนใจจำนวน 1 เรื่อง ได้แก่ “การวิเคราะห์คุณภาพข้อสอบปรนัยด้วย Item Response Theory (IRT)” ซึ่งเป็นหนึ่งในกระบวนการที่ใช้ประเมินคุณสมบัติด้าน psychometric ของข้อสอบแบบปรนัย ซึ่งน่าจะเป็นประโยชน์กับอาจารย์ที่ต้องติดตามกำกับคุณภาพของข้อสอบที่ใช้ในการประเมินผลที่สำคัญ นอกจากการประเมินความรู้ด้วยข้อสอบปรนัยแล้ว การประเมินสมรรถนะทางคลินิก (clinical competency) ด้วยการสอบ OSCE ก็ยังเป็นเรื่องที่สำคัญ และวารสารฉบับนี้ได้ตีพิมพ์งานวิจัย 3 เรื่อง ได้แก่ “ความสามารถในการใส่ท่อหายใจของนักศึกษาแพทย์ประเมินจากการสอบลงกองในรายวิชาวิสัญญีวิทยา การสอบใบประกอบวิชาชีพขั้นตอนที่ 3 (OSCE) และประเมินตามการรับรู้สมรรถนะตนเอง หลังจากผ่านกระบวนการฝึกฝนอย่างตั้งใจ” เรื่องที่สองคือ “High-Fidelity Simulation for Emergency and Difficult Airway Management Training in Novice Learners” ที่เปรียบเทียบการประเมินสมรรถนะการใส่ท่อช่วยหายใจของนักศึกษาแพทย์ซึ่งเป็นหัตถการที่เป็น life-saving และเป็นหัตถการที่บัณฑิตแพทย์ต้องทำได้ นอกจากนี้ในยุคที่ AI ได้เข้ามามีบทบาทมากขึ้นในการศึกษางานวิจัยเรื่องที่สาม “ปัจจัยที่มีผลต่อคะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉิน โดยใช้ ChatGPT” ผู้วิจัยได้ศึกษาเกี่ยวกับการรับรู้และสมรรถนะของนักศึกษาแพทย์ในการใช้ AI เข้ามาช่วยในการค้นหาข้อมูลทางการแพทย์เพื่อช่วยแก้ปัญหาผู้ป่วย ซึ้งน่าจะเป็นข้อมูลเบื้องต้นสำหรับผู้อ่านที่สนใจในหัวข้อ AI literacy ในนักศึกษาแพทย์

วารสาร MHPE มีกำหนดการเผยแพร่ทุก 3 ฉบับต่อปี จึงขอเชิญชวนผู้ที่สนใจงานด้านแพทยศาสตร์ศึกษาหรืองานที่เกี่ยวข้องกับการศึกษาของสาขาวิทยาศาสตร์สุขภาพส่งบทความและงานวิจัยมาตีพิมพ์เพื่อเป็นการเผยแพร่องค์ความรู้ด้านนี้ในวงกว้างมากขึ้น ทั้งนี้สามารถส่งผลงานตีพิมพ์ได้ทั้งภาษาไทยและภาษาอังกฤษ

กองบรรณาธิการ

สารบัญ

เรื่อง **หน้า**

บทความปริทัศน์

1. การวิเคราะห์คุณภาพข้อสอบปรนัยด้วย Item Response Theory (IRT) 1
Using Item Response Theory to Evaluate the Quality of Multiple-Choice Questions

บทความวิจัย

2. ความสามารถในการใส่ท่อหายใจของนักศึกษาแพทย์ประเมินจากการสอบลงกองในรายวิชาวิสัญญี วิทยาการสอบใบประกอบวิชาชีพขั้นตอนที่ 3 (NL3) และประเมินตามการรับรู้สมรรถนะตนเอง หลังจากผ่านกระบวนการฝึกฝนอย่างตั้งใจ 8
Tracheal intubation skills in medical students: A comparison between course examinations, the national licensing examination step 3 (NL3), and self-efficacy assessment following deliberate practice

3. High-Fidelity Simulation for Emergency and Difficult Airway Management Training in Novice Learners 20

4. ปัจจัยที่มีผลต่อคะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยผู้ป่วยภาวะฉุกเฉินโดยใช้ ChatGPT 33
Factors Influencing Competency Scores of Medical Students in Researching Emergency Care Principles Using ChatGPT

บทความปริทัศน์

การวิเคราะห์คุณภาพข้อสอบปรนัยด้วย Item Response Theory (IRT)

ไพโรจน์ บุญลักษณ์ศิริ

รับบทความ: 14 กรกฎาคม 2569
ปรับแก้ไขบทความ: 14 กรกฎาคม 2569
ตอบรับตีพิมพ์: 14 กรกฎาคม 2569

บทนำ

การวิเคราะห์คุณภาพข้อสอบ multiple-choice questions (MCQs) ส่วนใหญ่อาศัยแนวคิดของ Classical Test Theory (CTT) ใช้ดัชนีพื้นฐาน เช่น ค่าความยาก (difficulty index) ค่าอำนาจจำแนก (discrimination index) และค่าความเชื่อมั่น (reliability coefficient) แม้ว่าวิธีนี้จะเข้าใจง่าย คำนวณไม่ซับซ้อน และทำได้ด้วยโปรแกรมสถิติทั่วไป แต่มีข้อจำกัดโดยเฉพาะเรื่องคุณลักษณะข้อสอบที่ขึ้นกับกลุ่มผู้เข้าสอบ (sample dependent) และความสามารถของผู้เข้าสอบก็ขึ้นกับชุดข้อสอบนั้นๆ (test dependent) ทำให้เปรียบเทียบคุณภาพข้อสอบต่างชุดหรือคนต่างรุ่นทำได้จำกัด นอกจากนี้ CTT ยังให้ข้อมูลเพียงค่าเดียวตลอดช่วงความสามารถของผู้เข้าสอบ จึงไม่อธิบายว่าข้อสอบแต่ละข้อทำงานได้ดีในผู้เข้าสอบที่มีความสามารถแตกต่างกันอย่างไร

แนวคิดของ Item Response Theory (IRT) มีการพัฒนาจากหลายคน เริ่มตั้งแต่รากฐานแนวคิดแรกในช่วงปี 1925 - 1927 โดย L.L. Thurstone จนในปี 1950 Frederic M. Lord ได้นำมาพัฒนาและใช้ต่อจนได้รับยกย่องเป็นบิดาแห่ง IRT และต่อมาในปี 1960 Georg Rasch ได้เสนอ Rasch model (1PL) มี 1 parameter และในช่วงปลายทศวรรษ 1960 Allan Birnbaum ได้ขยาย model เป็น 2PL (2 parameters) และ 3PL (3 parameters) ทำให้ IRT สามารถอธิบายคุณลักษณะของข้อสอบได้ละเอียดมากขึ้น นำไปใช้อย่างแพร่หลายในการวัดผลทางการศึกษาและการสอบมาตรฐานทั่วโลก สามารถอธิบายความสามารถแฝง (latent trait, θ) ของผู้เข้าสอบกับความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง โดยประมาณค่าคุณลักษณะของข้อสอบเป็นรายข้อได้อย่างอิสระจากกลุ่มตัวอย่างภายใต้เงื่อนไขของ model ทั้งยังให้ข้อมูลเชิงลึกของดัชนีความยาก ดัชนีจำแนก และประสิทธิภาพตัวลวง หรือโอกาสการเดาในข้อสอบ MCQ นอกจากนี้ยังเป็นพื้นฐานสำคัญของการสร้างคลังข้อสอบ การสอบแบบปรับตามความสามารถของผู้สอบ (computerized adaptive testing) และการเทียบเคียงคะแนนระหว่างข้อสอบต่างชุด (test equating) IRT เป็นวิธีซับซ้อนกว่า CTT แต่ปัจจุบันโปรแกรมต่างๆ เช่น Stata, R และโปรแกรมด้าน psychometrics ได้พัฒนาให้การวิเคราะห์ IRT ทำได้สะดวกขึ้นอย่างมาก

บทความนี้มีวัตถุประสงค์เพื่อสรุปแนวคิดพื้นฐานของ IRT อธิบายความหมายและการแปลผลพารามิเตอร์ที่สำคัญ ตลอดจนการประยุกต์ใช้ IRT วิเคราะห์คุณภาพ MCQs ได้อย่างเหมาะสม

Review

Using Item Response Theory to Evaluate the Quality of Multiple-Choice Questions

Pairoj Boonluksiri

Received: 14 July 2026
Revised: 14 July 2026
Accepted: 14 July 2026

Background

The quality of multiple-choice questions (MCQs) has traditionally been evaluated using Classical Test Theory (CTT), which relies on indices such as item difficulty, item discrimination, and test reliability. Although CTT is simple, widely used, and easily implemented with conventional statistical software, it has important limitations. Item characteristics are dependent on the examinee sample, whereas estimates of examinee ability depend on the specific test administered, making comparisons across different cohorts or test forms difficult. Furthermore, CTT provides only a single summary statistic for each item and does not describe how item performance varies across different ability levels. Item Response Theory (IRT) was developed to address these limitations by modeling the relationship between an examinee's latent ability (θ) and the probability of answering an item correctly. Since its development by Frederic M. Lord, Georg Rasch, and Allan Birnbaum, IRT has become the dominant psychometric framework for educational and professional testing. Unlike CTT, IRT estimates item parameters—including difficulty, discrimination, and guessing—independently of the examinee sample, provided that model assumptions are satisfied. It also provides detailed information on how individual items function across the ability continuum. These advantages have made IRT the foundation of modern assessment practices, including item banking, computerized adaptive testing, and test equating. Although IRT is more mathematically sophisticated than CTT, recent advances in statistical software such as Stata and R have made its application increasingly accessible.

This review summarizes the fundamental concepts of IRT, explains the interpretation of its key parameters, and discusses its practical application in evaluating the quality of MCQs in medical education.

หลักการพื้นฐานของ IRT

IRT เป็นทฤษฎีใช้อธิบายความสัมพันธ์ระหว่างความสามารถแฝง (latent trait หรือ latent ability; θ theta) ของผู้ที่เข้าสอบกับโอกาส (probability) เลือกตอบข้อสอบได้ถูกต้อง อาศัยสมมติฐานสำคัญหลายประการ ได้แก่ **การวัดเพียงมิติเดียว** (uni-dimensionality) หมายถึง ข้อสอบทุกข้อนั้นวัดความสามารถหลักเดียวกัน **ความเป็นอิสระเฉพาะที่** (local independence) หมายถึง การตอบข้อสอบแต่ละข้อจะไม่ขึ้นต่อกัน และ**ความเป็นเอกภาพของความสัมพันธ์** (monotonicity) เมื่อระดับความสามารถของผู้เข้าสอบเพิ่มขึ้น ความน่าจะเป็นที่จะตอบถูกต้องจะไม่ลดลง แม้ว่าความสามารถแท้จริงของผู้เข้าสอบไม่สามารถวัดได้โดยตรง แต่ประมาณได้จากการตอบข้อสอบ คนเก่งมีโอกาสตอบถูกเพิ่มขึ้นตาม model ทางคณิตศาสตร์ที่กำหนดไว้ใน CTT ความสามารถของผู้เข้าสอบจะดูจากคะแนนรวมตรงๆ ถ้าคะแนนเท่ากันก็เก่งเท่ากัน แต่ใน IRT จะมองลึกซึ้งไปกว่านั้นคือ ถ้าตอบข้อที่ยากได้ถูกก็จัดว่าเก่งกว่าคนที่ได้คะแนนรวมเท่ากัน แต่ตอบข้อยากไม่ถูกหรือทำถูกได้น้อยกว่า โปรแกรมจะทำการวิเคราะห์คะแนนกับความยากรายข้อโดยใช้เทคนิค maximum likelihood estimation หรือจะเป็น Bayesian estimation ก็ได้ สรุปให้เข้าใจอย่างง่ายคือ latent trait เป็นการวิเคราะห์ย้อนกลับเริ่มจากคะแนน 0 แล้ว estimate คะแนนกับระดับความยากของข้อสอบในแต่ละข้อ โดยดู pattern ของการตอบถูกกับความยากของข้อสอบรายข้อ จากนั้น IRT จะทำการสร้าง **Item Characteristic Curve (ICC)** แสดงความสัมพันธ์ระหว่างระดับความสามารถบนแกนนอน (θ ซึ่งมีค่า −∞ ถึง +∞ แต่ในทางปฏิบัติมักเห็นพิสัย -3 ถึง +3) กับโอกาสตอบถูกในแกนตั้งเป็นรายข้อ (item-level analysis) ข้อยากจะมีเส้นโค้งที่เลื่อนไปทางขวา ส่วนดัชนีการจำแนกดูที่ความชัน (slope) ของเส้นโค้งในบริเวณระดับความสามารถที่ต้องการ IRT model ซึ่งมีหลายระดับได้แก่ One-Parameter Logisti Model; 1PL หรือ Rasch model พิจารณาเฉพาะความยากของข้อสอบ 2PL เพิ่ม parameter ในการจำแนก และ3PL เพิ่มโอกาสการเดาคำตอบในข้อสอบแบบเลือกตอบ

การแปลผลพารามิเตอร์และกราฟสำคัญใน IRT

1. ค่าความยาก (Difficulty parameter; b) หมายถึง ระดับความสามารถ (θ) ที่ผู้เข้าสอบมีโอกาสตอบข้อสอบถูก สะท้อน "ตำแหน่ง" ของข้อสอบบนแกนความสามารถ **b ต่ำ** (−2 ถึง −1) หมายถึงข้อสอบค่อนข้างง่าย เหมาะสำหรับคนอ่อน **b ใกล้ศูนย์** หมายถึงความยากปานกลางเหมาะกับผู้เข้าสอบระดับเฉลี่ย **และ b สูง** (+1 ถึง +2) หมายถึงค่อนข้างยากซึ่งมักจำแนกคนเก่งได้ดี โดยทั่วไปข้อสอบควรที่จะมีค่าของความยากกระจายครอบคลุมทุกช่วงความสามารถของผู้เข้าสอบ

2. ค่าอำนาจจำแนก (Discrimination parameter; a) แยกคนเก่งและคนอ่อน โดยดูจากความชันของเส้นโค้ง Item Characteristic Curve (ICC) ค่า **a สูง** เส้นโค้งชันจำแนกได้ดี แต่ถ้า **a ต่ำ** เส้นโค้งค่อนข้างราบจำแนกยาก ข้อสอบที่มี a ต่ำมากหรือมีค่าเป็นลบควรตรวจสอบสาเหตุ เช่น โจทย์กำกวม เฉลยผิด หรือวัดคนละมิติจากข้อสอบส่วนใหญ่

3. พารามิเตอร์การเดา (Guessing parameter; c) ใช้ในข้อสอบแบบเลือกตอบ ด้วย 3PL เช่น ข้อสอบ 4 ตัวเลือก มีค่า c ประมาณ 0.20 - 0.25 หมายความว่า แม้ผู้เข้าสอบจะมีความสามารถต่ำมากก็ยังมีโอกาสตอบถูกร้อยละ 20 - 25 หากค่า c สูงผิดปกติ อาจสะท้อนว่าตัวลวงไม่มีประสิทธิภาพ หรือมีคำใบ้ที่ช่วยให้เลือกคำตอบที่ถูกได้ง่าย

4. Item Characteristic Curve (ICC) เป็นกราฟพื้นฐานที่แสดงถึงความสัมพันธ์ระหว่างความสามารถของผู้เข้าสอบ (θ) กับความน่าจะเป็นในการตอบถูก ได้แก่ (1) ตำแหน่งกราฟตามแนวนอนสะท้อนค่าความยาก (b), (2) ความชันของกราฟสะท้อนค่าอำนาจจำแนก (a), (3) จุดเริ่มต้นของกราฟด้านซ้ายสะท้อนค่าโอกาสการเดา (c)

5. Item Information Function (IIF) เป็นกราฟแสดง information ข้อสอบรายข้อ เช่น ค่า a สูงมักให้ IFF สูง ในขณะที่ตำแหน่งของยอดกราฟจะอยู่ใกล้ค่าความยาก (b) ของข้อสอบ หากต้องการประเมินเลือกคนเก่ง ควรเลือกข้อที่มีค่า b สูง (กราฟเลื่อนมาทางขวา) และให้ IFF สูง (slope ชัน)

6. Test Information Function (TIF) เป็นผลรวมของ IFF ทั้งฉบับ ซึ่งหากยอดของ TIF อยู่บริเวณ $\theta = 0$ แสดงว่าแบบทดสอบมีความแม่นยำสูงในการวัดผู้เข้าสอบระดับปานกลาง แต่หากต้องการใช้คัดเลือกคนเก่ง เช่น การสอบคัดเลือก ควรเลือก TIF ที่มีค่าสูงในช่วง θ มากกว่า 1 ถึง 2 ทั้งนี้กราฟ TIF จะผกผันกับ Standard Error of Measurement (SEM) TIF สูงซึ่งจะสื่อถึงความสามารถที่แม่นยำ

ขั้นตอนการวิเคราะห์ IRT ด้วยโปรแกรม Stata

สามารถที่จะใช้ Stata v.14 ขึ้นไป คำสั่ง IRT ครอบคลุม 3 models และสร้างกราฟสำคัญ ทำได้ดังนี้

1. เตรียมข้อมูลผลสอบ MCQ ในรูปแบบ row เป็นลำดับผู้เข้าสอบ (id) และข้อสอบแต่ละข้อเป็นหนึ่งคอลัมน์ (column) กำหนด 1 คือตอบถูก 0 คือตอบผิด

2. ตรวจสอบสมมติฐานของ IRT โดยก่อนเลือก model ควรประเมินว่าข้อสอบวัดความสามารถเพียงมิติเดียว (unidimensionality) และมีความเป็นอิสระเฉพาะที่ (local independence) โดยสามารถใช้การวิเคราะห์ factor analysis รวมทั้งตรวจสอบ residual correlation ระหว่างข้อสอบ

3. เลือก model ที่เหมาะสม ได้แก่ Rasch model (1PL), 2PL หรือ 3PL ถ้าลองทำหลาย models สามารถพิจารณาเลือก model ที่ดีที่สุดได้จากดัชนีความเหมาะสมได้ เช่น Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), likelihood-ratio test และเปรียบเทียบเชิงเนื้อหา

4. ใช้ **Stata command: Irt** จะประมาณค่าพารามิเตอร์ a, b และ c เป็นรายข้อ รวมทั้งค่าความสามารถ (θ) ของผู้สอบรายคน

5. ประเมิน Model fit โดยตรวจสอบค่า parameter ผิดปกติหรือ ICC ไม่สอดคล้องกับข้อมูลจริง

6. แปลผลและนำไปใช้ร่วมกับทบทวนเนื้อหาของข้อสอบ ไม่ควรตัดสินจากค่าทางสถิติเพียงอย่างเดียว ข้อสอบที่มีค่าจำแนกต่ำ หรือค่า guessing สูงผิดปกติ ควรตรวจสอบทั้งการเขียนโจทย์ ตัวลวง และความสอดคล้องกับวัตถุประสงค์การเรียนรู้ แม้ IRT จะให้ข้อมูลที่ละเอียดและมีความแม่นยำกว่า CTT แต่การวิเคราะห์จำเป็นที่จะต้องมีจำนวนผู้เข้าสอบที่เพียงพอและข้อมูลที่มีคุณภาพ

Stata commands

1. **Import** เลือก excel file นำเข้าโดยตรง จากคำสั่ง สมมุติให้เป็นผล MCQ 100 ข้อ ผู้เข้าสอบ 100 คน **egen totalscore = rowtotal (q1 - q100)** สร้าง column คะแนนรวม (option ถ้าต้องการ)

2. **irt 2pl q1 - q100** ในที่นี้เลือกใช้ 2PL model จะให้ค่า b และ a ตามรายข้อ (ภาพที่1)

predict theta สร้าง column แสดงค่า probability of theta (option ถ้าต้องการ)

predict setheta, stdp สร้าง column แสดงค่า SEM (option ถ้าต้องการ)

3. **irtgraph icc q1 - q5** แสดงกราฟของข้อสอบรายข้อ ในที่นี้เลือก q1 - q5 ดังภาพที่ 2 โอกาสที่ตอบถูก (probability of θ) ตามสมการ $P(\theta) = 1/[(1+\exp[-a(\theta-b)])]$ สำหรับ 2P model ในการแปลผลเส้นแต่ละเส้นคือข้อสอบแต่ละข้อ เส้นทางซ้ายสุดคือข้อที่ง่ายที่สุดหรือค่า b น้อยๆ และกราฟขยับไปทางขวามากขึ้น ค่า b มากขึ้นก็จะยากขึ้น ส่วน slope ยิ่งชันแสดงถึงยิ่งจำแนกได้สูง ค่า a มาก ในภาพนี้เห็นว่าข้อที่ง่ายที่สุดอยู่ทางซ้ายสุดซึ่งมีความชันมากที่สุดด้วย คนที่มีความสามารถในระดับกลางและคนเก่ง (θ ตั้งแต่ 0 ขึ้นไป) ทำได้ถูกหมดทุกคน

4. **irtgraph iif q1 - g5** Item Information Function (IIF) แสดงประสิทธิภาพหรือความแม่นยำรายข้อ ดังภาพที่ 3 จากกราฟนี้จะเห็นได้ว่า ข้อสอบทั้ง 5 ข้อนั้นให้ information สูงสุดในช่วง θ −3 ถึง −1 และ information ลดลงอย่างรวดเร็วเมื่อ θ เข้าใกล้หรือมากกว่า 0 แสดงว่าทั้ง 5 ข้อนี้ให้ความแม่นยำในการวัดคนอ่อนที่ต่ำกว่าค่าเฉลี่ยเป็นหลัก

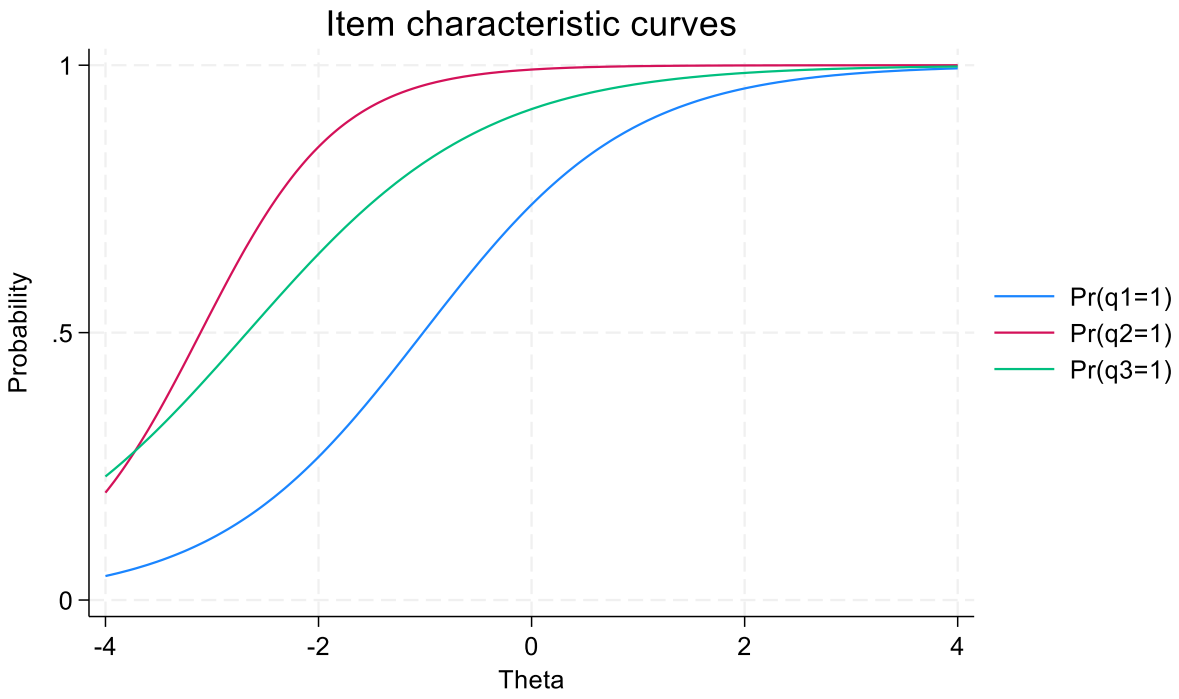
5. **irtgraph tif** Test Information Function (TIF) ผลรวม IFF ของข้อสอบทั้งฉบับ ดังภาพที่ 4 แปลผล จุดสูงสุดของกราฟอยู่ที่ $\theta \approx -1.7$ แสดงว่าทั้ง 5 ข้อ โดยรวมนี้แม่นยำสูงในการวัดคนอ่อนที่ต่ำกว่าค่าเฉลี่ยประมาณ -1.7 ของส่วนเบี่ยงเบนมาตรฐาน

ภาพที่ 1 Stata output of irt

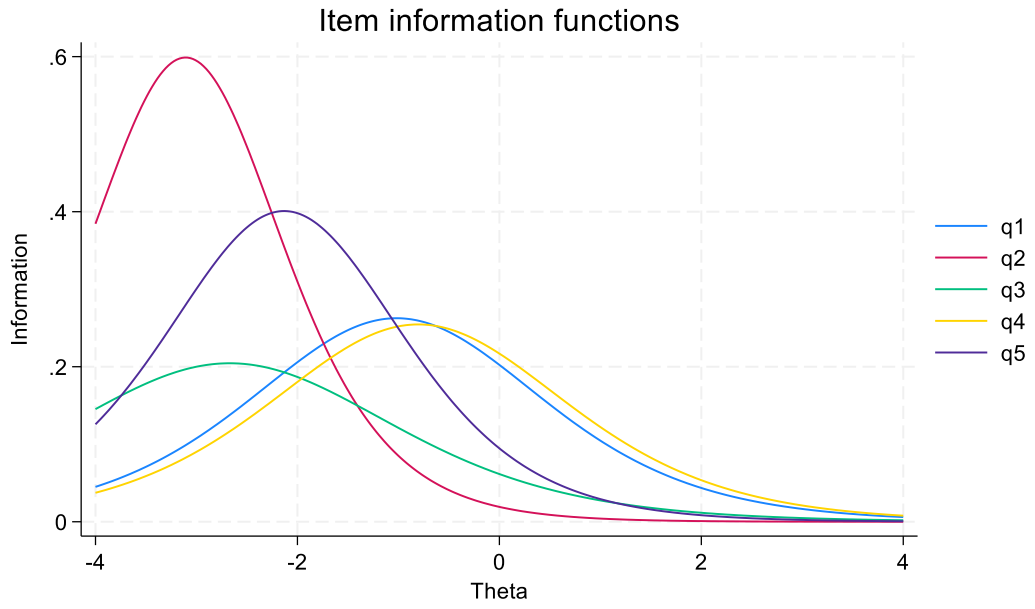
Two-parameter logistic model — Number of obs = 100
Log likelihood = -4009.8529

		Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
q1							
	Discrim	1.024878	.297381	3.45	0.001	.4420221	1.607734
	Diff	-1.017409	.3523927	-2.89	0.004	-1.708086	-.3267319
q2							
	Discrim	1.547758	.6710877	2.31	0.021	.2324506	2.863066
	Diff	-3.107457	.7812146	-3.98	0.000	-4.638609	-1.576305
q3							
	Discrim	.9044026	.3207516	2.82	0.005	.275741	1.533064
	Diff	-2.6705	.814352	-3.28	0.001	-4.266601	-1.074399
q4							
	Discrim	1.009111	.2978247	3.39	0.001	.4253853	1.592837
	Diff	-.8027847	.323251	-2.48	0.013	-1.436345	-.1692243
q5							
	Discrim	1.266271	.3614336	3.50	0.000	.5578738	1.974668
	Diff	-2.131203	.4882656	-4.36	0.000	-3.088186	-1.17422

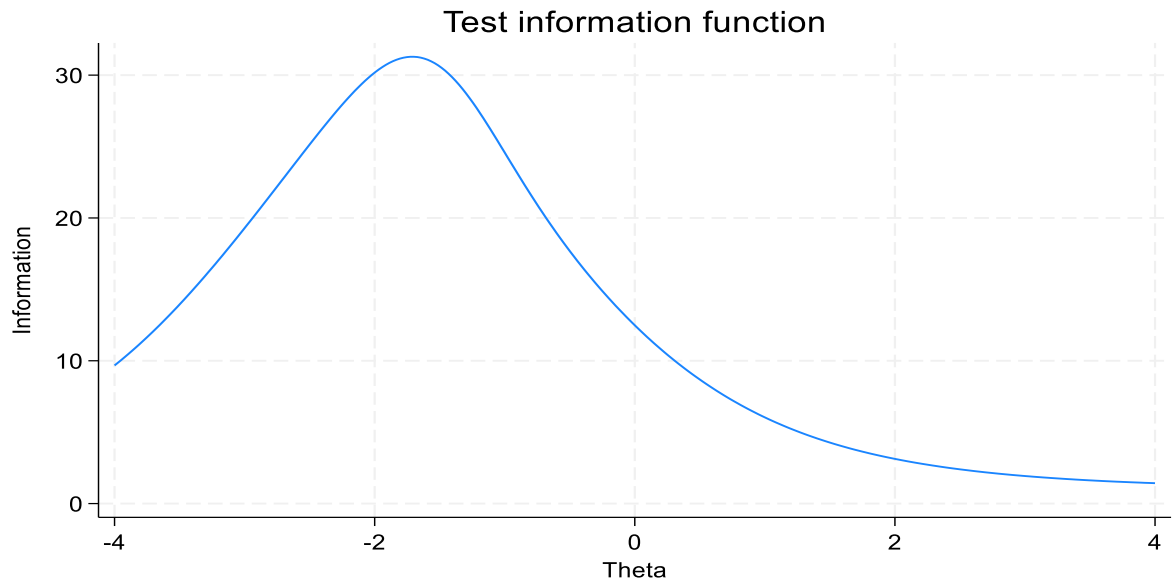
ภาพที่ 2 Item Characteristic Curve (ICC)



ภาพที่ 3 Item Information Function (IIF) ประสิทธิภาพของข้อสอบ



ภาพที่ 4 Test Information Function (TIF)



สรุปและคำแนะนำ

คำแนะนำสำหรับผู้เริ่มต้นใช้ IRT ให้วิเคราะห์ข้อสอบด้วย CTT ก่อน แล้วตรวจสมมติฐานของ IRT โดยเฉพาะ unidimensionality เลือกใช้ 2PL or 3PL ที่เหมาะสม ตรวจสอบพารามิเตอร์ a, b, c, ICC, IIF และ TIF พิจารณาปรับปรุงรายข้อที่มีปัญหาและเลือกข้อที่คุณภาพเหมาะสมเข้า item bank

Disclosure

บทความนี้ใช้ AI ช่วยสืบค้นข้อมูลและร่างบทความ โดยทุกขั้นตอนถูกตรวจสอบความถูกต้องโดยผู้นิพนธ์แล้ว

Conflict of interest

ไม่มี

References

1. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park (CA): Sage Publications; 1991. Available at:https://www. academia.edu/30496631/Fundame tals_of_Item_Response_Theory
2. Kean J, Brodke DS, Biber J, Gross P. An introduction to Item Response Theory and Rasch Analysis of the Eating Assessment Tool (EAT-10). Brain Impair. 2018; 19: 91–102. doi:10.1017/BrImp.2017.31.
3. Kumar D., Jaipurkar R., Shekhar A., Sikri G., V Srinivas. Item analysis of multiple-choice questions: A quality assurance test for an assessment tool. Med J Armed Forces India. 2021;77(Suppl1): S85-S89.doi: 10.1016/j.mjafi.2020.11.007
4. StataCorp LLC. IRT (item response theory) [Internet]. College Station (TX): StataCorp LLC; [cited 2026 Jul 11]. Available from:https://www.stata.com/features/overview/irt/

ORIGINAL ARTICLE

Tracheal intubation skills in medical students: A comparison between course examinations, the national licensing examination step 3 (NL3), and self-efficacy assessment following deliberate practice

Chompoonik Jiemjitpolchai, Darika Thanbuasawan

Sawanpracharak Medical Education Center

Received: 15 March 2026

Revised: 14 May 2026

Accepted: 2 June 2026

Background: Tracheal intubation is a critical life-saving competency in medical practice. Proficiency in this skill requires not only technical precision but also psychological readiness. This study aimed to identify clinical performance gaps and evaluate the correlation between measured competence and perceived self-efficacy among medical students after undergoing a structured deliberate practice curriculum.

Summary of Work: A retrospective observational cohort study was conducted involving 35 sixth-year medical students at the Sawanpracharak Medical Education Center. The study compared performance scores from the internal fifth-year anesthesiology course examination (OSCE) with results from the National Licensing Examination Step 3 (NL3) administered by the Center for Medical Competency Assessment (CMA). Participants also completed a self-efficacy assessment questionnaire one day prior to the national exam. Data were analyzed to identify statistical differences in three domains: preparation, procedural execution, and tube confirmation.

Summary of Results: The students demonstrated significantly higher scores in the preparation domain during their fifth-year OSCE examination compared to the NL3 exam (31.1 ± 3.3 vs. 28.0 ± 5.6; $p < 0.05$). Conversely, performance in the actual intubation and tube confirmation phases showed statistically significant improvement in the NL3 exam (37.1 ± 5.4 vs. 35.7 ± 5.2 and 23.3 ± 4.3 vs. 16.9 ± 8.0, respectively; $p < 0.05$). Interestingly, students' self-efficacy scores were significantly lower than their actual performance levels across nearly all procedural steps ($p < 0.05$).

Conclusion: Following deliberate practice, students scored significantly higher on the NL3 than on the fifth-year course examination in the intubation and tube confirmation domains but significantly lower in the preparation domain, and their self-efficacy was significantly below their actual performance in almost every sub-behavior.

บทความวิจัย

ความสามารถในการใส่ท่อหายใจของนักศึกษาแพทย์ประเมินจากการสอบลงกองในรายวิชาวิสัญญีวิทยา การสอบใบประกอบวิชาชีพขั้นตอนที่ 3 (NL3) และประเมินตามการรับรู้สมรรถนะตนเอง หลังจากผ่านกระบวนการฝึกฝนอย่างตั้งใจ

ชมพูณิก เจียมจิตพลชัย , ดาริกา ธารบัวสวรรค์

ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสวรรค์ประชารักษ์

รับบทความ: 15 มีนาคม 2569

ปรับแก้ไขบทความ: 14 พฤษภาคม 2569

ตอบรับตีพิมพ์: 2 มิถุนายน 2569

ภูมิหลัง: การใส่ท่อช่วยหายใจเป็นทักษะวิชาชีพที่สำคัญและจำเป็นอย่างยิ่งสำหรับแพทย์ ทักษะนี้ไม่เพียงแต่ต้องการความแม่นยำทางเทคนิค แต่ยังรวมถึงความพร้อมทางจิตวิทยา การเข้าใจถึงข้อบกพร่องและระดับการรับรู้สมรรถนะตนเองของผู้เรียนภายหลังกระบวนการฝึกฝนอย่างตั้งใจ จะช่วยสนับสนุนการสร้างกลยุทธ์การสอนที่มีประสิทธิภาพเพื่อพัฒนาขีดความสามารถของนักศึกษาแพทย์

สรุปการดำเนินงาน: การวิจัยเชิงสังเกตแบบย้อนหลัง (Retrospective Observational Cohort Study) นี้ศึกษาในนักศึกษาแพทย์ชั้นปีที่ 6 จำนวน 35 คน ณ ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสวรรค์ประชารักษ์ โดยเปรียบเทียบผลคะแนนการใส่ท่อช่วยหายใจระหว่างการสอบลงกองในรายวิชาวิสัญญีวิทยาเมื่อครั้งอยู่ชั้นปีที่ 5 กับผลการสอบใบประกอบวิชาชีพเวชกรรมขั้นตอนที่ 3 (**NL3**) โดยศูนย์ประเมินและรับรองความรู้ความสามารถตามมาตรฐานวิชาชีพเวชกรรม (ศรว.) นอกจากนี้ นักศึกษาได้ประเมินสมรรถนะแห่งตนผ่านแบบสอบถามหนึ่งวันก่อนการสอบระดับชาติ ข้อมูลถูกนำมาวิเคราะห์เปรียบเทียบใน 3 ด้านหลัก ได้แก่ ขั้นการเตรียมการ ขั้นการปฏิบัติการใส่ท่อ และขั้นการยืนยันตำแหน่งท่อ

สรุปผลการวิจัย: ผลการศึกษาพบว่า นักศึกษาแพทย์ทำคะแนนในส่วนการเตรียมความพร้อมในการสอบชั้นปีที่ 5 ได้สูงกว่าการสอบ ศรว. อย่างมีนัยสำคัญ (31.1 ± 3.3 เทียบกับ 28.0 ± 5.6) ในทางตรงกันข้ามคะแนนในส่วนการปฏิบัติการใส่ท่อและการยืนยันตำแหน่งท่อในการสอบ ศรว. กลับสูงกว่าการสอบในรายวิชาอย่างมีนัยสำคัญทางสถิติ (37.1 ± 5.4 เทียบกับ 35.7 ± 5.2 และ 23.3 ± 4.3 เทียบกับ 16.9 ± 8.0 ตามลำดับ; p-value < 0.05) นอกจากนี้ยังพบประเด็นสำคัญคือ นักศึกษาประเมินการรับรู้สมรรถนะแห่งตนต่ำกว่าระดับความสามารถที่ปฏิบัติได้จริงอย่างมีนัยสำคัญในเกือบทุกขั้นตอน (p-value < 0.05)

บทสรุป: ภายหลังกระบวนการฝึกฝนอย่างตั้งใจ นักศึกษาแพทย์มีคะแนนขั้นการปฏิบัติการใส่ท่อและขั้นการยืนยันตำแหน่งท่อในการสอบ ศรว. ขั้นตอนที่ 3 สูงกว่าการสอบลงกองอย่างมีนัยสำคัญ แต่คะแนนขั้นการเตรียมการกลับลดลง และการรับรู้สมรรถนะตนเองนั้นต่ำกว่าสมรรถนะจริงอย่างมีนัยสำคัญในเกือบทุกพฤติกรรมย่อย การฝึกฝนอย่างตั้งใจจึงพัฒนาทักษะทางเทคนิคที่วิกฤตได้จริง แต่ยังคงมีช่องว่างที่ขั้นการเตรียมการและการปรับความมั่นใจให้ตรงกับความสามารถ ซึ่งต้องแก้ไขด้วยการพัฒนาเกณฑ์การประเมินภายในให้สอดคล้องกับมาตรฐานระดับชาติ (ศรว.)

บทนำ

ทักษะการใส่ท่อช่วยหายใจ (Tracheal Intubation) เป็นสมรรถนะหลักที่สำคัญและเป็นหัตถการช่วยชีวิตขั้นสูง (Advanced Life Support) ที่แพทย์ทุกคนจำเป็นต้องปฏิบัติได้อย่างถูกต้องแม่นยำและมั่นใจภายหลังสำเร็จการศึกษาตามหลักสูตรแพทยศาสตรบัณฑิต[1-2] อย่างไรก็ตาม ในบริบทของการศึกษาแพทยศาสตร์ยุคปัจจุบัน การเรียนการสอนทักษะนี้เผชิญกับความท้าทายหลายประการ โดยเฉพาะประเด็นเรื่องสิทธิผู้ป่วยและความปลอดภัย (Patient Safety) ซึ่งทำให้นักศึกษาแพทย์มีโอกาสฝึกปฏิบัติหัตถการวิกฤตกับผู้ป่วยจริงลดลงอย่างมีนัยสำคัญ[3-4] ส่งผลทำให้บัณฑิตแพทย์จำนวนหนึ่งขาดความมั่นใจเมื่อต้องเผชิญกับสถานการณ์ฉุกเฉินในเวชปฏิบัติจริง[5]

เพื่อให้มีการก้าวข้ามข้อจำกัดดังกล่าวนั้น ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสวรรค์ประชารักษ์ จึงได้นำแนวคิด "การฝึกฝนอย่างตั้งใจ" (Deliberate Practice) ตามโมเดลของ Ericsson[6] มาประยุกต์ใช้ในการจัดการเรียนการสอนรายวิชาวิสัญญีวิทยาสำหรับนักศึกษาแพทย์ชั้นปีที่ ๕ แนวคิดนี้มุ่งเน้นการแบ่งทักษะที่ซับซ้อนออกเป็น "พฤติกรรมย่อย" (Microskills) หรือพฤติกรรมย่อยทางคลินิก (Sub-behaviors) เพื่อให้ผู้เรียนได้ฝึกฝนกับหุ่นจำลองอย่างจดจ่อ มีเป้าหมายที่ชัดเจน และมีการทำซ้ำภายใต้การดูแลของอาจารย์แพทย์ที่ให้ข้อมูลป้อนกลับ (Feedback) ทันที[7-8] กระบวนการนี้ไม่เพียงแต่ช่วยพัฒนาทักษะทางกายภาพ (Motor Skills) แต่ยังช่วยสร้างสะพานเชื่อมระหว่างความรู้เชิงทฤษฎีกับการปฏิบัติการในสถานการณ์วิกฤต[9]

นอกเหนือจากทักษะทางเทคนิค (Technical Skills) ปัจจัยทางจิตวิทยาโดยเฉพาะ "การรับรู้สมรรถนะแห่งตน" (Self-Efficacy) ตามทฤษฎีปัญญาทางสังคมของ Bandura ซึ่งถือเป็นตัวแปรสำคัญที่กำหนดความสำเร็จในการปฏิบัติงานทางคลินิก[10-11] การรับรู้สมรรถนะแห่งตนไม่ใช่เพียงความเชื่อมั่นทั่วไป แต่เป็นการตัดสินความสามารถของตนเองในการจัดการและดำเนินการปฏิบัติเพื่อบรรลุผลสำเร็จในสถานการณ์ที่กำหนด[12] วรรณกรรมทางการแพทย์ระบุว่านักศึกษาที่ประเมินสมรรถนะตนเองได้สอดคล้องกับความสามารถจริงจะมีความผิดพลาดทางการแพทย์น้อยกว่ากลุ่มที่มีความมั่นใจสูงเกินความเป็นจริง (Over-confidence) หรือในกลุ่มที่ประเมินตนเองต่ำกว่าความเป็นจริง (Under-confidence) ซึ่งมักพบในผู้เรียนระดับต้น (Novice) ที่ได้เริ่มมีการตระหนักถึงความเสี่ยงและภาวะแทรกซ้อน[13-14]

ในประเทศไทย มาตรฐานความรู้ความสามารถวิชาชีพเวชกรรมถูกกำหนดและควบคุมโดยแพทยสภาผ่านการประเมินโดยศูนย์ประเมินและรับรองความรู้ความสามารถตามมาตรฐานวิชาชีพเวชกรรม (ศรว.) หรือ Center for Medical Competency Assessment (CMA)[15] ซึ่งการสอบใบประกอบวิชาชีพเวชกรรมขั้นตอนที่ 3 ในรูปแบบการสอบทักษะทางคลินิก (Objective Structured Clinical Examination: OSCE) ถือเป็นเกณฑ์ตัดสินสำคัญก่อนการสำเร็จการศึกษา[16] อย่างไรก็ตาม ความแตกต่างระหว่างเกณฑ์การประเมินภายในสถาบัน (การสอบลงกอง) และเกณฑ์มาตรฐานระดับชาติ รวมไปถึงช่องว่างที่เกิดขึ้นระหว่างความมั่นใจและความสามารถจริงของนักศึกษา ยังเป็นประเด็นที่ต้องศึกษาอย่างเป็นระบบ

ผู้วิจัยในฐานะอาจารย์แพทย์ผู้รับผิดชอบการจัดการเรียนการสอน จึงสนใจที่จะทำการศึกษาและประเมินทักษะการใส่ท่อหายใจของนักศึกษาแพทย์ชั้นปีที่ 6 โดยแบ่งวัตถุประสงค์เฉพาะออกเป็น 3 ประการ ดังนี้

1. วัตถุประสงค์หลัก (Primary objective) เพื่อทำการเปรียบเทียบคะแนนสมรรถนะทางคลินิก

ระหว่างการสอบภายในรายวิชาวิสัญญีวิทยา (ปี 5) และการสอบใบประกอบวิชาชีพเวชกรรมขั้นตอนที่ 3 (NL3) (ปี 6)

2. วัตถุประสงค์รอง (Secondary objective) เพื่อประเมินความแตกต่าง (discrepancy) ระหว่างการรับรู้สมรรถนะตนเอง (self-efficacy) และคะแนนความสามารถที่ทำได้จริง (actual perfor- mance) และ

3. วัตถุประสงค์เชิงสำรวจ (Exploratory objective) เพื่อที่จะทำการระบุช่องว่างของสมรรถนะในแต่ละโดเมนย่อย โดยเปรียบเทียบระหว่างขั้นตอนการเตรียมการ (preparation) และขั้นตอนการปฏิบัติการ (intubation และ endotracheal tube confirmation) เพื่อนำข้อมูลเชิงประจักษ์มาใช้ในการปรับปรุงกลยุทธ์การสอนและเสริมสร้างสมรรถนะของบัณฑิตแพทย์ให้พร้อมสำหรับการปฏิบัติหน้าที่ในฐานะแพทย์เพิ่มพูนทักษะอย่างมีประสิทธิภาพและปลอดภัยต่อไป

วิธีการศึกษา (Methods)

รูปแบบการวิจัย (Study design)

การศึกษานี้เป็นการวิจัยเชิงสังเกตแบบย้อนหลัง (Retrospective observational cohort study)

กลุ่มตัวอย่าง (Participants)

ประชากรที่ศึกษาคือนักศึกษาแพทย์ชั้นปีที่ 6 ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสวรรค์ประชารักษ์ ที่เข้าสอบประเมินทักษะทางคลินิก (OSCE) ขั้นตอนที่ 3 (NL3) ในปีการศึกษา 2566 จำนวน 35 คน โดยมีเกณฑ์การคัดเข้า (Inclusion criteria) คือ 1) เป็นนักศึกษาที่ผ่านการฝึกอบรมและสอบวัดผลในรายวิชาวิสัญญีวิทยาเมื่อครั้งอยู่ชั้นปีที่ 5 และ 2) มีผลการสอบ NL3 ในสถานีการใส่ท่อช่วยหายใจ (Tracheal intubation) ครบถ้วน โดยนักศึกษาทุกรายให้ความยินยอมเข้าร่วมการวิจัย (Response rate 100%) นักศึกษาแพทย์กลุ่มนี้ได้รับการฝึกอบรมทักษะการใส่ท่อช่วยหายใจผ่านกระบวนการ "ฝึกฝนอย่างตั้งใจ" (Deliberate practice) เริ่มต้นในรายวิชาวิสัญญีวิทยาชั้นปีที่ 5 โดยแบ่งทักษะออกเป็นพฤติกรรมย่อย ฝึกปฏิบัติกับหุ่นจำลองความแม่นยำสูง และได้รับข้อมูลป้อนกลับ (Feedback) ทันทีจากอาจารย์แพทย์จนเกิดความชำนาญก่อนเข้าสู่การสอบวัดผลในรายวิชา หลังจากนั้นนักศึกษาแพทย์ที่เข้าร่วมการศึกษาทุกคนได้ผ่านกระบวนการเรียนรู้แบบการฝึกฝนอย่างตั้งใจกับผู้ป่วยจริงอย่างต่อเนื่องตลอดระยะเวลา 1 ปีการศึกษาในชั้นคลินิก ก่อนการเข้าทดสอบระดับชาติโดย ศรว. ขั้นตอนที่ 3 ทั้งนี้นักศึกษาแพทย์ชั้นปีที่ 6 ทุกรายได้เข้าทำแบบประเมินการรับรู้สมรรถนะตนเอง (Self-efficacy assessment) ล่วงหน้า 1 วันก่อนการเข้าสอบประเมินทักษะทางคลินิก (OSCE)

เครื่องมือที่ใช้ในการวิจัย (Research Instruments)

เครื่องมือที่ใช้ประกอบด้วย 2 ส่วนหลัก:

1. **เกณฑ์การประเมินทักษะ OSCE:** ใช้ใบให้คะแนน (Checklist) มาตรฐานของศูนย์ประเมินและรับรองความรู้ความสามารถตามมาตรฐานวิชาชีพเวชกรรม (ศรว.) หรือ Center for Medical Competency Assessment (CMA) ซึ่งระบุรายละเอียดพฤติกรรมย่อยและเกณฑ์การให้คะแนนไว้อย่างชัดเจน รวมคะแนนเต็ม 100 คะแนน แบ่งเป็น 3 โดเมนหลัก ได้แก่ ขั้นในการเตรียมผู้ป่วยและอุปกรณ์ (Preparation) ขั้นการปฏิบัติการใส่ท่อ (Intubation) และขั้นการยืนยันตำแหน่งท่อ (Tube confirmation) โดยการประเมินทั้งในการสอบลงกอง (ปี 5) และการสอบระดับชาติ (ปี 6) ซึ่งดำเนินการโดยอาจารย์แพทย์ผู้เชี่ยวชาญ

ท่านเดียวกัน (Same evaluator) เพื่อรักษามาตรฐานความเชื่อมั่นในการประเมิน (Reliability)

2. **แบบประเมินการรับรู้สมรรถนะตนเอง (Self-efficacy Questionnaire):** ได้พัฒนาขึ้นเพื่อประเมินความเชื่อมั่นต่อขั้นตอนการปฏิบัติ 10 ขั้นตอนย่อย ตามมาตรวัดระดับความมั่นใจ 0–10 คะแนน เครื่องมือผ่านการตรวจสอบความตรงในเชิงเนื้อหา (Content Validity) โดยวิสัญญีแพทย์ 3 ท่าน พบค่า IOC เท่ากับ 0.8 และได้นำไปทดสอบเบื้องต้น (Pilot test) กับนักศึกษาแพทย์ชั้นปีที่ 5 จำนวน 10 คน เพื่อตรวจสอบความชัดเจนของภาษา ซึ่งพบว่าแบบประเมินมีค่าความสอดคล้องภายใน (Cronbach's alpha coefficient) เท่ากับ 0.82

การวิเคราะห์ทางสถิติ

วิเคราะห์ข้อมูลด้วยโปรแกรมสำเร็จรูปทางสถิติ (SPSS version 26.0) โดยใช้สถิติเชิงพรรณนา ได้แก่ ค่าเฉลี่ย (mean) และส่วนเบี่ยงเบนมาตรฐาน (SD) ส่วนการเปรียบเทียบความแตกต่างของคะแนนประเมินพฤติกรรมการปฏิบัติในระหว่างช่วงเวลาและการเปรียบเทียบระหว่างสมรรถนะจริงกับสมรรถนะตามการรับรู้ ใช้สถิติ Paired sample t-test สำหรับข้อมูลที่ได้มีการแจกแจงแบบปกติ และ Wilcoxon signed-rank test สำหรับข้อมูลที่มีการแจกแจงแบบไม่ปกติ โดยกำหนดระดับนัยสำคัญทางสถิติที่ $p < 0.05$

จริยธรรมการวิจัย (Research ethics)

โครงการวิจัยได้รับการรับรองจากคณะกรรมการจริยธรรมการวิจัยในคน โรงพยาบาลสวรรค์ประชารักษ์ หมายเลข COA.10/2566

ผลการศึกษา (Results)

1. ข้อมูลทั่วไปของกลุ่มตัวอย่าง (Participants Characteristics)

กลุ่มตัวอย่างคือนักศึกษาแพทย์ชั้นปีที่ 6 จำนวน 35 คน เป็นเพศชายร้อยละ 51.4 (n=18) และเพศหญิงร้อยละ 48.6 (n=17) มีอายุเฉลี่ย 23.69 ± 0.72 ปี เมื่อพิจารณาประสบการณ์ปฏิบัติหัตถการในชั้นคลินิก พบว่านักศึกษามีประสบการณ์ใส่ท่อช่วยหายใจในผู้ป่วยมารับการผ่าตัดเฉลี่ย 12.40 ± 7.52 ครั้ง และในผู้ป่วยภาวะวิกฤตเฉลี่ย 8.17 ± 6.55 ครั้ง

2. การเปรียบเทียบสมรรถนะการใส่ท่อช่วยหายใจระหว่างช่วงเวลาการเรียนและการสอบระดับชาติ

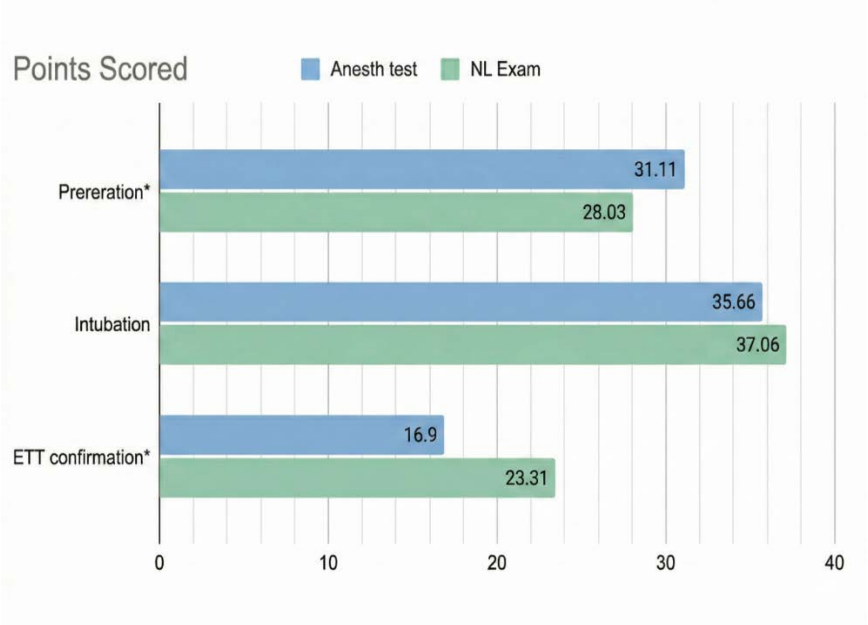
เมื่อเปรียบเทียบคะแนนเฉลี่ยใน 3 โดเมนหลักของการใส่ท่อช่วยหายใจ พบการเปลี่ยนแปลงที่มีนัยสำคัญทางสถิติ ($p < 0.05$) โดยในโดเมนการเตรียมผู้ป่วยและอุปกรณ์ (Preparation) นักศึกษามีคะแนนสอบในระดับชาติ (NL3) ลดลงอย่างชัดเจนเมื่อเทียบกับการสอบในรายวิชาวิสัญญีวิทยา (28.03 เทียบกับ 31.11 คะแนน; mean difference −3.08 คะแนน) อย่างไรก็ตาม ในโดเมนการปฏิบัติการใส่ท่อ (Intubation) และโดเมนการยืนยันตำแหน่งท่อ (ETT Confirmation) นักศึกษากลับมีคะแนนสูงขึ้นอย่างมีนัยสำคัญในการสอบระดับชาติ (37.06 และ 23.31 คะแนน ตามลำดับ; mean difference +1.40 และ +6.41 คะแนน ตามลำดับ) โดยมีรายละเอียดแสดงในตารางที่ 1 และภาพที่ 1

ตารางที่ 1 การเปรียบเทียบคะแนนประเมินพฤติกรรมใส่ท่อหายใจจำแนกตามโดเมนและหัวข้อย่อย โดยรายงานเป็นผลต่างค่าเฉลี่ย (mean difference) ระหว่างการสอบ ศรว. ปี 6 กับการสอบรายวิชา ปี 5

ขั้นตอน (คะแนนเต็ม)	สอบรายวิชา (ปี 5) (n = 35)	สอบ ศรว. (ปี 6) (n = 35)	Mean difference (ปี 6 – ปี 5)	95% CI Lower	95% CI Upper	p-value
1. การเตรียมผู้ป่วยและอุปกรณ์ (35) Mean ± SD	31.1 ± 3.3	28.0 ± 5.6	-3.08	-5.36	-0.81	<0.01*
1.1 การเลือก mask (5)	4.7 ± 0.7	5	0.30	0.04	0.53	0.02*
1.2 การหนุนผ้าที่ท้ายทอย (occiput) (5)	2.6 ± 2.4	3.3 ± 2.4	0.70	-0.37	1.69	0.2
1.3 การครอบหน้ากากให้กระชับ (5)	4.8 ± 0.9	3.9 ± 2.1	-0.90	-1.78	-0.11	<0.03*
1.4 การดันศีรษะและเชยคาง (10)	9.7 ± 1.1	8.2 ± 2.0	-1.50	-2.30	-0.66	<0.01*
1.5 การช่วยหายใจให้ทรวงอกขยับ (10)	9.3 ± 1.5	7.7 ± 2.8	-1.60	-2.68	-0.52	<0.01*
2. การปฏิบัติการใส่ท่อ (40) Mean ± SD	35.7 ± 5.2	37.1 ± 5.4	1.40	-1.42	4.22	0.32
2.1 การเลือกขนาดท่อช่วยหายใจ (5)	5	5	0.00	-	-	-
2.2 การสอด blade โดยไม่งัดฟัน (10)	9.4 ± 1.4	9.9 ± 0.7	0.50	-0.10	1.01	0.1
2.3 การยก laryngoscope ทำมุม 45 องศา (10)	8.97 ± 1.77	9.03 ± 2.89	0.06	-1.21	0.50	0.93
2.4 การสอดท่อช่วยหายใจ (5)	4.6 ± 1.0	4.4 ± 1.6	-0.20	-0.90	1.32	0.56
2.5 การยึดท่อไว้ขณะดึง stylet ออก (5)	4.6 ± 1.2	4.0 ± 2.0	-0.60	-1.37	0.17	0.12
2.6 การเติมลมใน Pilot ballloon (5)	3.0 ± 2.4	4.7 ± 1.2	1.70	0.69	2.68	<0.01*

ขั้นตอน (คะแนนเต็ม)	สอบรายวิชา (ปี 5) (n = 35)	สอบ ศรว. (ปี 6) (n = 35)	Mean difference (ปี 6 – ปี 5)	95% CI Lower	95% CI Upper	p-value
3. การยืนยันตำแหน่งท่อ (25) Mean ± SD	16.9 ± 8.0	23.3 ± 4.6	6.41	3.31	9.52	<0.01*
3.1 การฟังเสียงปอดด้วย stethoscope (10)	8.5 ± 3.3	9.6 ± 1.8	1.10	-0.15	2.33	0.08
3.2 การตรวจเช็คความลึกของท่อ (10)	5.7 ± 4.8	9.4 ± 2.0	3.70	1.89	5.65	<0.01*
3.3 การยึดท่อช่วยหายใจ (5)	2.7 ± 2.3	4.3 ± 1.8	1.60	0.67	2.44	<0.01*

ภาพที่ 1 แผนภูมิแท่งแสดงการเปรียบเทียบคะแนนเฉลี่ย (mean score) ± ส่วนเบี่ยงเบน มาตรฐานของการประเมินพฤติกรรมใส่ท่อหายใจจำแนกตามโดเมน ระหว่างการสอบรายวิชาวิสัญญีวิทยา (ปี 5) และการสอบ NL3 (ปี 6) (n = 35)



ภาพที่ 1 เปรียบเทียบคะแนนเฉลี่ยใน 3 โดเมนหลักระหว่างการสอบรายวิชาวิสัญญีวิทยาและการสอบระดับชาติ (NL3) แผนภูมิแสดงให้ถึงเห็นแนวโน้มการพัฒนาทักษะในทางเทคนิค (Technical skills) ที่สูงขึ้นตามประสบการณ์ในชั้นปีสุดท้าย โดยเฉพาะในส่วนการยืนยันตำแหน่งท่อที่มีการพัฒนาอย่างก้าวกระโดด ขณะที่ทักษะในด้านการเตรียมตัวเบื้องต้นมีคะแนนลดลงในการสอบระดับชาติ

3. ความสอดคล้องระหว่างการรับรู้สมรรถนะตนเองและสมรรถนะที่ปฏิบัติได้จริง

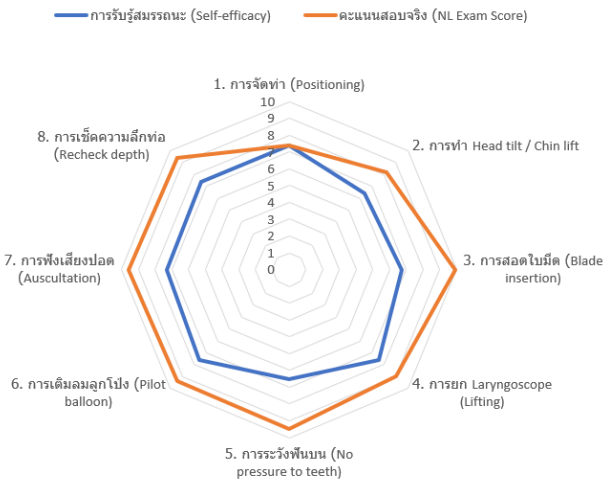
ผลการวิเคราะห์พบสภาวะ "Under-confidence" อย่างมีนัยสำคัญทางสถิติ ($p < 0.05$) โดย

นักศึกษาประเมินระดับความมั่นใจในตนเองเฉลี่ยอยู่ในระดับความมั่นใจปานกลาง (6.29 - 7.64 คะแนน) ซึ่งระดับต่ำกว่าคะแนนที่ปฏิบัติได้จริงในการสอบ NL3 ในเกือบทุกหัวข้อพฤติกรรมย่อย โดยเฉพาะในขั้นตอนที่วิกฤต เช่น การใส่ Blade และการยืนยันตำแหน่งท่อผ่านการฟัง (Auscultation) ซึ่งนักศึกษามีความเชื่อมั่นเฉลี่ยเพียง 6.71 และ 7.64 คะแนน ตามลำดับ ในขณะที่คะแนนสอบจริงเฉลี่ยสูงถึง 9.89 และ 9.33 คะแนน (mean difference - 3.18 และ - 1.69 คะแนน ตามลำดับ) โดยมีรายละเอียดแสดงในตารางที่ 2 และภาพที่ 2

ตารางที่ 2 เปรียบเทียบคะแนนการใส่ท่อหายใจระหว่างนักศึกษาประเมินตนเอง และผลการสอบ NL3 โดยรายงานเป็นผลต่างค่าเฉลี่ย (mean difference)

ขั้นตอน	นักศึกษาประเมินตนเอง (n = 35)	สอบ NL3 (n = 35)	Mean difference (Self-efficacy - NL3)	95% CI Lower	95% CI Upper	p-value
การจัดท่า Mean ± SD	7.36 ± 0.9	7.4 ± 1.2	-0.04	-0.33	0.41	0.78
การทำ head tilt chin lift Mean ± SD	6.45 ± 1.7	8.17 ± 2.02	-1.72	-1.71	0.08	0.11
การสอด blade Mean ± SD	6.71 ± 2.3	9.89 ± 0.7	-3.18	0.40	-3.98	<0.01*
การยก laryngoscope Mean ± SD	7.36 ± 1.9	9.03 ± 2.9	-1.67	-3.85	-1.49	<0.01*
การระวังไม่งัดฟันบน Mean ± SD	6.5 ± 1.6	9.03 ± 2.8	-2.53	-3.65	-1.41	<0.01*
การเติมลมใน pilot balloon Mean ± SD	7.36 ± 1.8	9.43 ± 2.0	-2.07	-2.99	-1.16	<0.01*
การฟังเสียงปอด Mean ± SD	7.64 ± 2.0	9.33 ± 2.4	-1.69	-2.83	-0.74	<0.01*
การเช็คความลึกท่อ Mean ± SD	7.28 ± 1.9	9.6 ± 1.8	-2.32	-3.21	-1.42	<0.01*

ภาพที่ 2 แผนภูมิใยแมงมุมแสดงความสัมพันธ์ระหว่างการรับรู้สมรรถนะตนเองและคะแนนการสอบจริงรายพฤติกรรมย่อย



ภาพที่ 2 แผนภูมิใยแมงมุมแสดงความสัมพันธ์ระหว่างการรับรู้สมรรถนะตนเองและคะแนนการสอบจริงรายพฤติกรรมย่อย พื้นที่ว่าง (Gap) ระหว่างเส้นคะแนนความมั่นใจ (เส้นใน) และเส้นคะแนนที่ปฏิบัติได้จริง (เส้นนอก) สะท้อนให้เห็นว่านักศึกษาแพทย์ขาดความเชื่อมั่นในทักษะของตนเอง แม้จะมีความสามารถทางการปฏิบัติอยู่ในเกณฑ์ดีเยี่ยมก็ตาม

วิจารณ์ (Discussion)

การศึกษาครั้งนี้แสดงให้เห็นถึงพัฒนาการของทักษะการใส่ท่อช่วยหายใจในนักศึกษาแพทย์ชั้นปีที่ 6 เมื่อเปรียบเทียบกับการประเมินในรายวิชาเมื่อครั้งอยู่ชั้นปีที่ 5 โดยมีประเด็นที่น่าสนใจ 3 มิติหลัก คือ พัฒนาการของทักษะทางเทคนิค การลดลงของความใส่ใจในรายละเอียดการเตรียมการ และช่องว่างระหว่างสมรรถนะจริงกับการรับรู้สมรรถนะตนเอง

1. พัฒนาการของทักษะในทางเทคนิคและประสบการณ์สะสม ผลการศึกษาพบว่าคะแนนในโดเมนการปฏิบัติการใส่ท่อ (Intubation) และการยืนยันตำแหน่งท่อ (ETT Confirmation) เพิ่มสูงขึ้นอย่างมีนัยสำคัญในการสอบระดับชาติ (NL3) สอดคล้องกับภาพที่ 1 ซึ่งแสดงให้เห็นถึงผลสำเร็จของ "การฝึกฝนอย่างตั้งใจ" (Deliberate Practice) และประสบการณ์ทางคลินิกที่สะสมในช่วงชั้นปีสุดท้าย หรือระยะ Externship พัฒนาการที่โดดเด่นในขั้นตอนการยืนยันตำแหน่งท่อ (คะแนนเพิ่มจาก 16.90 เป็น 23.31) สะท้อนถึงความเข้าใจในความปลอดภัยของผู้ป่วย (Patient safety) ที่ลึกซึ้งขึ้นซึ่งเป็นผลจากการได้ปฏิบัติจริงในห้องผ่าตัดและหอผู้ป่วยวิกฤต โดยงานวิจัยในบริบทของประเทศไทยระบุว่านักศึกษาแพทย์ที่ไม่มีประสบการณ์มาก่อนจำเป็นต้องทำหัตถการนี้อย่างน้อย 5 - 10 ครั้งจึงจะเข้าสู่ช่วงของการเรียนรู้ที่ชัดขึ้น (Learning curve) และความสำเร็จในการใส่ท่อในครั้งแรก (First-pass success) จะเพิ่มขึ้นตามจำนวนครั้งที่ปฏิบัติจริง[3,14]

2. การลดลงของคะแนนในส่วนการเตรียมการ (Preparation decay) ในทางตรงกันข้าม คะแนนในโดเมนการเตรียมผู้ป่วยและอุปกรณ์กลับลดลงอย่างมีนัยสำคัญในการสอบ NL3 (28.03 เทียบกับ 31.11 คะแนน) ปรากฏการณ์นี้อาจจะเกิดจากปัจจัยทางจิตวิทยาของการสอบระดับชาติที่มีสภาวะกดดันสูง (High-stakes assessment) ทำให้นักศึกษาพยายามมุ่งเน้นไปที่ขั้นตอนที่วิกฤตกว่าคือการสอดท่อให้สำเร็จจนละเลยรายละเอียดปลีกย่อยในขั้นเตรียมตัว เช่น

การครอบหน้ากากให้กระชับ หรือการจัดท่าศีรษะที่ถูกต้อง นอกจากนี้ยังอาจสะท้อนถึงความแตกต่างของเกณฑ์การประเมินและการเตรียมผู้ประเมิน (Inter-rater standardization) แม้จะเป็นผู้ประเมินคนเดียวกัน แต่บริบทสนามสอบที่ต่างกันอาจส่งผลต่อความเข้มงวดในการให้คะแนนพฤติกรรมย่อย

3. ช่องว่างระหว่างสมรรถนะและความมั่นใจ (Confidence-Competence Gap) ประเด็นที่สำคัญที่สุดคือสภาวะ "Under-confidence" ที่ปรากฏในภาพที่ 2 โดยนักศึกษาประเมินการรับรู้สมรรถนะตนเอง (Self-efficacy) ต่ำกว่าคะแนนที่ปฏิบัติได้จริงอย่างมากในทุกหัวข้อ สภาวะนี้นั้นสามารถที่จะอธิบายได้ด้วยปรากฏการณ์ Dunning-Kruger ในส่วนผู้ที่มีสมรรถนะสูง (High performers) ซึ่งเมื่อเริ่มมีความรู้และตระหนักถึงความซับซ้อนของหัตถการ รวมถึงภาวะแทรกซ้อนที่อาจเกิดขึ้น จะเริ่มประเมินความสามารถของตนเองอย่างระมัดระวังและถ่อมตัวมากขึ้น[17] การประเมินตนเองต่ำกว่าความเป็นจริงนี้แม้จะช่วยลดความประมาท แต่อาจส่งผลให้เกิดความลังเลใจ (Decision-making hesitation) ในสถานการณ์วิกฤตจริง อาจารย์แพทย์จึงควรนำรูปแบบการให้ข้อมูลป้อนกลับแบบทันที (Rapid cycle deliberate practice: RCDP) มาใช้เพื่อเสริมสร้างความเชื่อมั่นให้สอดคล้องกับระดับสมรรถนะที่แท้จริง

ข้อเสนอแนะเชิงการจัดการเรียนการสอน

อาจารย์แพทย์ควรให้ความสำคัญกับการชี้แนะ "พฤติกรรมย่อย" (Sub-behaviors) ในช่วงการฝึกฝนอย่างตั้งใจอย่างต่อเนื่อง และควรมีการพัฒนาเกณฑ์การประเมินภายใน (In-course assessment) ให้มีความสอดคล้องกับเกณฑ์มาตรฐานระดับชาติ (CMA) มากขึ้น เพื่อให้นักศึกษาแพทย์มีความพร้อมทั้งในด้านทักษะทางเทคนิคและความมั่นใจที่มีความสอดคล้องกับขีดความสามารถจริง นำไปสู่การประกอบวิชาชีพได้อย่างปลอดภัยและมีประสิทธิภาพต่อไป

ข้อจำกัดของการศึกษา

การศึกษานี้ทำในกลุ่มประชากรขนาดเล็ก (n = 35) ภายในสถาบันเดียว (single center) ซึ่งส่งผลให้งานวิจัยนี้นั้นมีลักษณะเป็นการศึกษาเชิงสำรวจ (exploratory study) และเป็นการศึกษาแบบย้อนหลัง ข้อมูลบางส่วนอาจได้รับอิทธิพลจากความแตกต่างของอุปกรณ์ในห้องจำลองและสนามสอบจริง นอกจากนี้ยังอาจจะเกิดอคติจากประสบการณ์การเรียนรู้ทางคลินิกที่มีความแตกต่างกัน (learning exposure bias) เนื่องจากความแตกต่างของจำนวนและประเภทของผู้ป่วยที่นักศึกษาแต่ละรายได้มีการฝึกปฏิบัติจริง (variability in clinical exposure) เช่น ในบางรายที่มีประสบการณ์ 8 ครั้ง ในขณะที่บางรายอาจได้ทำถึง 12 ครั้ง ซึ่งอาจจะส่งผลโดยตรงต่อระดับสมรรถนะและความมั่นใจที่วัดได้ ทั้งนี้การแปลผลควรทำด้วยความระมัดระวังเนื่องจากข้อจำกัดด้านขนาดตัวอย่าง และควรมีการศึกษาในอนาคตที่เป็นลักษณะหลายสถาบันเพื่อเพิ่มอำนาจทางสถิติ (statistical power) และยืนยันผลการศึกษานี้ให้มีความน่าเชื่อถือยิ่งขึ้น อย่างไรก็ตาม จุดแข็งคือการเปรียบเทียบข้อมูลรายบุคคลในช่วงเวลาที่ต่างกันและใช้ผู้ประเมินคนเดียวกัน ซึ่งช่วยลดอคติจากการวัดได้ในระดับหนึ่ง

สรุปผลการศึกษา

นักศึกษาแพทย์ชั้นปีที่ 6 ที่ผ่านกระบวนการฝึกฝนอย่างตั้งใจมีคะแนนขั้นการปฏิบัติการใส่ท่อและขั้นการยืนยันตำแหน่งท่อในการสอบ ศรว. ขั้นตอนที่ 3 สูงกว่าการสอบลงกองในรายวิชาวิสัญญีวิทยาอย่างมีนัยสำคัญ แต่คะแนนขั้นการเตรียมการกลับลดลงอย่างมีนัยสำคัญ และประเมินการรับรู้สมรรถนะ

ตนเองต่ำกว่าสมรรถนะจริงในเกือบทุกพฤติกรรมย่อย ช่องว่างของสมรรถนะจึงอยู่ที่ขั้นของการเตรียมการและการปรับความมั่นใจให้ตรงกับความสามารถจริง

Reference

1. ปาริชาต อภิเดชากุล,ปฏิภาณ ตุ่มทอง, ธัชวรรณ จิระติวานนท์. การสังเกตความสามารถในการใส่ท่อหายใจของนักศึกษาแพทย์ ชั้นปีที่5: ความเชื่อมโยงจากผู้เรียนระดับต้นสู่ผู้เชี่ยวชาญ. วิสัญญีสาร. 2560;43(1):19-26.
2. สุพรรณา ครองแถว, สุภามาศ ผาติประจักษ์, อรุณวรรณ อุดมเกษมทรัพย์. ผลของโปรแกรมฝึกอบรมการจัดการทางเดินหายใจ และการช่วยหายใจ สำหรับนักศึกษาพยาบาลระดับปริญญาตรี. Rama Nurs J. 2559;22(1):93-105.
3. รัชยา วีระยุทธวัฒนะ, ธนาวดี ธีรัชฌานันท์, สุจารีย์ ภู่พิพัฒน์ภาพ, บุษบา ศุภวัฒน์ธนบดี. การศึกษาหาจำนวนผู้ป่วยขั้นต่ำในการฝึกทักษะหัตถการใส่ท่อช่วยหายใจในนักศึกษาแพทย์. เชียงใหม่เวชสาร. 2564;60(3):317-23.
4. Ovatakanont P, Kovindha A, Wittayanin W, Mongkol S. Mobility and self-care activities related to the ability to drive and control a modified motorcycle: a pilot study in paraplegic persons. J Thai Rehabil Med. 2006;16:85-97. (Reference for context of clinical skill autonomy in Thai settings).
5. Sripiman P, Thiangchanya P, Purinthrapibal S. The effectiveness of a self-efficacy promoting program on knowledge, self-confidence, and caring behaviors. CMU Journal of Nursing. 2020;47 (1):1-15.
6. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med. 2008;15(11):988-94.
7. Petrosoniak A, Lu M, Gray S, Monteiro S. Deliberate practice model for improved task completion in medical education. ResearchGate [Internet]. 2024 [cited 2024 Mar 2]. Available from:https://www.researchgate.net/figure/Deliberate-practice-model_fig2_332239055.
8. Kim Sawtelle KR. Rapid Cycle Deliberate Practice (RCDP) to Master Airway Intubation. S D Med. 2024; 77(9):408.
9. Remmen R, Scherpbier A, Van der Vleuten C. Effectiveness of basic clinical skills training programs: A cross-sectional comparison. Med Educ. 2001; 35:121-8.
10. Bandura A. Self-efficacy: The exercise of control. New York: Freeman; 1997.
11. ภาสิต ศิริเทศ, ณพวิทย์ ธรรมสีหา. ทฤษฎีการรับรู้ความสามารถของตนเองกับพฤติกรรมการดูแลสุขภาพของผู้สูงอายุ. วารสารพยาบาลทหารบก 2562;20(2):58-65.
12. Sacaja N, Romic Z, Prka Z. Self-perceived versus objectively measured competence in performing clinical practical procedures by final year me-

dical students. Int J Med Educ. 2016; 7:122-9.

13. Barbosa J, Santos A, Castro-Vale I. Poorer correlation between self-perceived and objectively measured competence in practical skills. BMC Med Educ. 2016;16:1-8.
14. Bismilla Z, Lopushinsky SR, Ratnapalan S. Evaluation of a pediatric intubation checklist for residents. Pediatr Emerg Care. 2010;26(8):545-9.
15. ศูนย์ประเมินและรับรองความรู้ความสามารถตามมาตรฐานวิชีพเวชกรรม (ศรว.). ประวัติศูนย์ประเมินและรับรองความรู้ความสามารถตามมาตรฐานวิชาชีพเวชกรรมของแพทยสภา [อินเทอร์เน็ต]. 2565 [สืบค้นเมื่อวันที่ 2 มี.ค. 2569]. จาก https://cmathai.org/about.
16. Kittrakulrat J, Jongjatuporn W, Jurjai R, Jarupanich N. The ASEAN economic community and medical qualification. Chula Med J. 2014;58(4):415-28
17. Sanchez-Mendiola M, Durante-Montiel I, Morales-Lopez S, Lozano-Sanchez R, Martinez-Gonzalez A. Self-assessment and external assessment: A study of the Dunning-Kruger effect in medical students. BMC Med Educ. 2020; 20(1):77.

ORIGINAL ARTICLE

High-Fidelity Simulation for Emergency and Difficult Airway Management Training in Novice Learners

Chompoonik Jiemjitpolchai, Darika Thanbuasawan

Sawanpracharak Medical Education Center

Received: 28 February 2026

Revised: 7 July 2026

Accepted: 26 July 2026

Introduction: Emergency airway management presents a significant challenge for medical students, demanding critical skills with limited opportunities for real-world practice on live patients. Simulation-based medical education offers a valuable platform for acquiring knowledge, skills, and appropriate attitudes across all healthcare professions. This study aimed to evaluate the efficacy of a high-fidelity simulator-based training program in enhancing knowledge acquisition and learners' perceptions regarding emergency and difficult airway management among novice learners.

Methods: A quasi-experimental, single-group pretest-posttest design was employed. Twenty-nine fifth-year medical students enrolled in an anesthesiology course participated in this study. All students were novices in high-fidelity simulation-based learning (SBL) and had no prior experience with intubation in real-life emergency airway situations. Prior to SBL, students received lectures and basic intubation training with part-task simulation under instructor guidance. A pre-test was administered one day before the SBL session. Two scenarios were designed: emergency airway management and difficult airway management. All students were divided into 16 leaders and 13 observers. Upon completion of the SBL program, all students underwent a post-test to assess their knowledge improvement and completed a survey to evaluate their perception of the SBL experience.

Results: The results demonstrated a statistically significant increase in the overall knowledge scores post-intervention compared to the pre-test ($p < 0.05$). Both Leaders and Observers showed comparable improvements in knowledge acquisition. Furthermore, the participants reported high levels of satisfaction and increased self-confidence in managing emergency airways.

Conclusion: The integration of SBL significantly improved the learners' knowledge and perceived confidence in handling emergency and difficult airway scenarios. Knowledge improvement and learners' perceptions were comparable between leaders and observers.

Introduction

Emergency airway management is a critical, life-saving intervention in clinical practice. Failure to secure an airway promptly can lead to severe and catastrophic complications, including profound hypoxia, permanent neurological damage, and death[1-2]. For novice learners, such as medical students, acquiring these essential skills through the traditional "see one, do one" model on live patients is increasingly limited by ethical concerns and patient safety priorities[3]. Consequently, there is an urgent need for effective educational strategies that provide a safe yet realistic environment for training in high-stakes airway scenarios. Simulation-based medical education (SBME) offers a valuable platform to bridge this gap, providing a safe environment for fostering multidisciplinary communication and structured feedback without compromising patient safety[4].

Simulation-based medical education (SBME) has emerged as a cornerstone in medical training, bridging the gap between theoretical knowledge and clinical practice[5]. A systematic review and meta-analysis have demonstrated that technology-enhanced simulation is consistently associated with significant improvements in knowledge and skill acquisition compared to no intervention[6]. Specifically, high-fidelity simulators (HFS) provide a realistic physiological response to interventions, allowing learners to experience the stress and complexity of emergency and difficult airway management without risking patient harm[7-8].

In the context of Thailand, the national competency standards for medical graduates (B.E. 2566/2023) strictly mandate proficiency in basic and emergency life-saving procedures prior to entering clinical internship[9]. However, integrating high-fidelity simulation into the undergraduate curriculum presents significant logistical challenges, particularly for regional medical education centers. High-fidelity simulators are resource-intensive and expensive, often resulting in high student-to-simulator ratios[8]. To maximize the utility of these limited resources, simulation sessions are frequently designed as group activities where a few students act as active participants or "Leaders," while the rest act as "Observers."

While the general efficacy of simulation is well-documented, prior research has largely focused on experienced residents or specialized practitioners, leaving a paucity of evidence regarding its immediate impact on novice learners facing emergency intubation for the first time. Furthermore, the practical necessity of utilizing observer roles raises a critical gap in knowledge: Does the role of a directed observer provide equivalent know- ledge acquisition and perceived learning value compared to the active leader role in high-

fidelity airway scenarios?[10-11].

Understanding this "role equivalence" is vital for optimizing medical curricula and resource management in resource-limited settings. Consequently, this study sought to address the existing knowledge gaps through a structured simulation intervention. The primary objective of this study was to evaluate the effectiveness of a high-fidelity simulation program in enhancing knowledge acquisition and learners' perceptions regarding emergency and difficult airway management among fifth-year medical students. The secondary objective was to compare these educational outcomes—specifically knowledge improvements and satisfaction levels—between students acting in active participant roles (Leaders) and those in directed observation roles (Observers).

Methods

Study Design

This cross-sectional, quasi-experimental study utilized a single-group pretest-posttest design to assess knowledge acquisition and learner perceptions among novice medical students.

Study Population

The study population consisted of fifth-year medical students from a six-year MD program during their anesthesiology clerkship. Since the simulation was conducted as part of the mandatory, routine educational curriculum, all students actively participated, and no explicit informed consent was required for the educational activity itself.

The inclusion criteria required that all students were from the same academic year, shared an identical curriculum during their anesthesiology rotation, and were novices in high-fidelity simulation-based learning (SBL) without prior experience in real-life emergency endotracheal intubation. This selection process ensured a homogeneous baseline across all learners.

The exclusion criteria were defined as: (1) absence from the simulation session, (2) incomplete attendance from the briefing to the debriefing phase, and (3) failure to complete either the pre-test or the post-test and survey. No participants met the exclusion criteria.

Simulation Intervention and Role Allocation

One week prior to the session, students were provided with self-study materials, foundational lectures, and basic intubation training using low-fidelity part-task trainers. The primary SBL intervention utilized the CAE Apollo high-fidelity patient simulator (CAE Healthcare, Sarasota, FL, USA). The simulation was conducted in a dedicated laboratory equipped to mimic a standard general ward and an Intensive Care Unit (ICU) environment.

Participants were organized into small groups of 3–4 members. Role allocation

within each group was determined through voluntary self-selection. To ensure a comprehensive learning experience, a role rotation strategy was implemented between different scenarios so that no participant performed the same role twice. For comparative analysis, participants were categorized based on their primary active role during the assessed scenarios into a "Leader" group (responsible for decision-making and interventions) and an "Observer" group (responsible for observing the scenario participants, recording key points to provide constructive feedback, and contributing to the shared reflection during the debriefing).

Two core scenarios were designed based on ASA 2022 guidelines: emergency airway and unanticipated difficult airway management. To enhance psychological fidelity, a failure to secure the airway within two minutes triggered physiological deterioration, including a SpO2 drop to 70% and bradycardia. Each SBL session followed a standardized structure: a 5-minute pre-briefing, a 10–15 minute scenario, and a 15–20 minute debriefing using the Gather-Analyze-Summarize model.

Outcome Measurements and Data Collection

The educational outcomes were evaluated using two primary instruments designed to assess knowledge acquisition and learners' perceptions.

1. Knowledge Assessment: A 10-item multiple-choice questionnaire (MCQ) was administered to evaluate cognitive knowledge. The questions covered critical areas such as indications for intubation, rapid airway assessment, confirmation of endotracheal tube placement, alternative airway devices, and team communication. The tool demonstrated excellent content validity (Index of Item Objective Congruence [IOC] = 0.67–1.0). Each correct answer was awarded 1 point. This assessment was conducted as a pre-test one day prior to the session and as a post-test immediately following completion.

2. Learners' Perception Survey: Following the training, a self-reported survey was administered to evaluate perceptions using a 5-point Likert scale (Cronbach’s alpha = 0.85). The survey captured perceptions across the briefing, scenario, and debriefing phases, and evaluated the overall impact of the high-fidelity manikin on learning stimulation, enjoyment, and critical thinking.

Regarding skill evaluation, direct observation of clinical and non-technical skills was conducted by instructors during the scenarios. However, this was utilized exclusively for formative feedback during the debriefing phase and was not recorded for summative quantitative analysis.

Statistical Analysis

Statistical analysis was performed using SPSS. The normality of data distribution for continuous variables was initially assessed using the Shapiro-Wilk test. For normally distributed data, such as participants' age, descriptive statistics were reported as mean ± standard deviation (SD), and comparisons between the Leader and Observer groups were conducted using an independent t-test. Categorical data, such as gender distribution, were analyzed using the Chi-square test.

Given the non-normal distribution of the knowledge scores, nonparametric statistics were utilized for these outcomes. Knowledge scores and Likert-scale perception ratings were reported as median and interquartile range (IQR). Within-group differences between pre-test and post-test knowledge scores were analyzed using the Wilcoxon signed-rank test. Between-group comparisons (Leader vs. Observer) for both knowledge gains and perception ratings were evaluated using the Mann-Whitney U test. In addition, the magnitude of the change from pre-test to post-test was quantified as the mean difference with its corresponding 95% confidence interval (95% CI). A p-value of < 0.05 was considered statistically significant.

Ethical Considerations

Ethical approval was granted by the Institutional Review Board of Sawanpracharak Hospital (COE No. 32/2568).

Results

A total of 29 fifth-year medical students (Leader group, n=16; Observer group, n=13) completed the simulation session and all associated assessments. Table 1 presents the baseline characteristics of the participants. The mean age of the cohort was 22.24±0.5 years. There were no statistically significant differences between the Leader and Observer groups regarding age, gender distribution, or baseline knowledge scores (all $p > 0.05$), indicating that the groups were well-matched prior to the intervention.

Table 1 Baseline characteristics and initial knowledge scores of the participants (N=29)

Characteristics	Total (n=29)	Leader (n=16)	Observer (n=13)	p-value
Age (years), Mean ± SD	22.24 ± 0.5	22.31 ± 0.6	22.15 ± 0.5	0.471
Gender, n (%)				
Male	13 (44.8)	6 (46.20)	7 (53.8)	0.379
Female	16 (55.2)	10 (62.5)	6 (37.5)	
Pre-test Score (Max 10), Median (IQR)	10 (9-10)	10 (9-10)	10 (9-10)	0.918
Pre-test Score (Max 10), Mean ± SD	9.28 ± 0.9	9.31 ± 0.9	9.23 ± 1	0.831

Knowledge Acquisition Overall knowledge scores significantly increased from a pre-test median of 10 (IQR 9–10) to a post-test median of 10 (IQR 10–10) (mean difference 0.586, 95% CI 0.227–0.946; p = 0.003). When comparing the two roles, both the Leader and Observer groups demonstrated significant and nearly identical improvements in knowledge. There was no statistically significant difference in the post-test scores between Leaders (Median 10, IQR 9-10) and Observers (Median 10, IQR 9–10) (p = 0.83). These findings suggest equivalent knowledge acquisition regardless of whether the student assumed an active or observational role (Table 2)

Table 2 Comparison of Mean Knowledge Scores by Role (n=29)

Roles	Statistic	Pre-Test	Post-Test	Mean difference (95% CI)	p-value
Leader (n=16)	Mean ± SD Median (IQR)	9.31 ± 0.94 10 (9-10)	9.88 ± 0.34 10 (10-10)	0.562 (0.049–1.076)	0.034 0.034
Observer (n=13)	Mean ± SD Median (IQR)	9.23 ± 1.09 10 (9-10)	9.85 ± 0.37 10 (10-10)	0.615 (0.035–1.196)	0.040 0.039
Total (n=29)	Mean ± SD Median (IQR)	9.28 ± 0.99 10 (9-10)	9.86 ± 0.35 10 (10-10)	0.586 (0.227–0.946)	0.002 0.003

Learners' Perceptions and Attitudes The simulation program received consistently high ratings across all perception domains. As shown in Table 3, the mean scores for the brie- fing, scenario participation, and debriefing phases were uniformly high for both groups. Notably, students in the Observer role reported substantial levels of satisfaction, particularly regarding the simulation's ability to stimulate learning and enhance critical thinking through directed observation. Comparisons between roles revealed no significant differences in any perception category (all $p > 0.05$).

Table 3 Comparative Learner Perceptions of Simulation-Based Learning Phases (Leader vs. Observer)

Issues	Leader (Mean ± SD)	Observer (Mean ± SD)	p-value
Pre-Simulation (Preparation)			
Difficulty of pre-test	3.69 ± 0.946	3.92 ± 0.862	0.494
Assigned self-study materials	3.88 ± 1.258	4.00 ± 0.707	0.752
Scenario Execution			
Realism, excitement, participation	4.56 ± 0.512	4.31 ± 0.480	0.182
Scenario difficulty	4.19 ± 0.981	4.00 ± 0.577	0.548
Debriefing (Reflection)			
Suitable atmosphere (no stress)	4.13 ± 1.310	4.38 ± 0.650	0.521
Debriefing covered subject matter	4.31 ± 0.793	4.54 ± 0.519	0.385

Strategic Sentiment Toward Simula- tion-Based Learning When evaluating the strategic sentiment regarding simulation as a teaching tool, a distinct "perception gap" was observed, although it did not reach statistical significance (Figure 1). While 100% of Observers strongly agreed or agreed that SBL stimulates learning and promotes comprehensive thinking, a slightly lower proportion of Leaders (87.5%) shared this sentiment. Most notably, 92.3% of Observers advocated for the expansion of SBL to other clinical subjects, compared to only 62.5% of Leaders ($p = 0.141$). Furthermore, a substantial majority of the overall cohort (89.7%) indicated that simulation training should be introduced earlier in the medical curriculum.

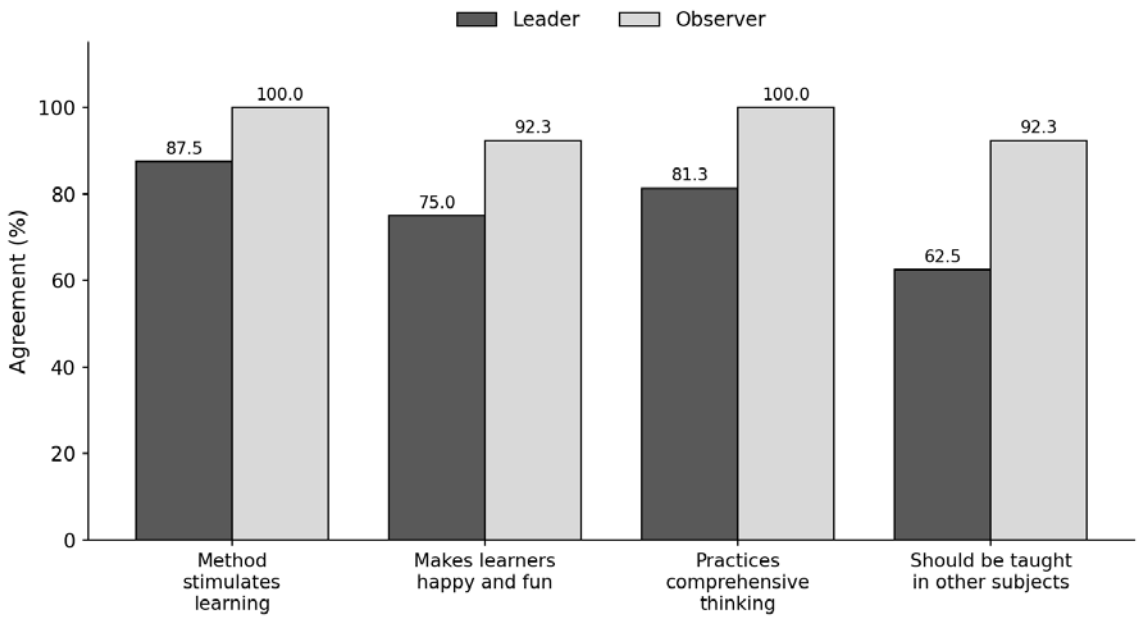


Figure 1 Learner Agreement on Educational Benefits of High-Fidelity Simulation-Based Learning (Leader vs. Observer)

Discussion

The primary findings of this study underscore the effectiveness of high-fidelity simulation in accelerating the cognitive preparedness of novice medical students for high-stakes airway crises. The significant improvement in knowledge scores-and the notable reduction in score variance suggest that simulation-based learning (SBL) effectively standardizes knowledge levels toward near-mastery across a diverse student cohort. These findings align with current evidence suggesting that technology-enhanced simulation consistently yields superior knowledge acquisition compared to traditional didactic methods[-6,12].

Role Equivalence and Learning Theories

A critical observation of this research is the parity in knowledge acquisition between students in the Leader role and those in the Observer role. This "Role Equivalence" suggests that learners can acquire complex behaviors and knowledge vicariously through directed observation, a concept deeply rooted in Social Learning Theory[13-14]. In our study, Observers were not passive; they were cognitively engaged through the pre-briefing and the shared debriefing phase. This engagement facilitates the reflective observation and abstract conceptualization stages of the experiential learning cycle[15], which has been shown

to provide significant educational value in medical school settings[16]. Our results are consistent with recent literature suggesting that structured observation is a potent pedagogical tool that allows students to learn without the immediate performance anxiety often experienced by active participants[10-11,17].

The "Perception Gap" and Psychological Fidelity

Interestingly, our data revealed a "Perception Gap" regarding the strategic value of simulation. While both groups were highly satisfied, Observers were significantly more likely than Leaders to advocate for the expansion of SBL to other clinical subjects. This may be attributed to the lower cognitive load experienced by Observers, allowing them to step back from the technical pressures of intubation and view the broader pedagogical framework[11]. Conversely, Leaders likely focused more intensely on technical mastery under stress, which may narrow their immediate perspective on curriculum-wide benefits. However, educators must ensure that such high satisfaction does not lead to overconfidence, which can occasionally occur in high-fidelity simulation environments[18].

The utilization of high-fidelity simulators played a cornerstone role in providing psychological fidelity. By simulating physiological triggers such as rapid desaturation and bradycardia, students were forced to navigate difficult airway algorithms under realistic pressure[19]. This immersion is critical for developing non-technical skills, including situational awareness and team communication[20-21]. Furthermore, the use of a structured debriefing model successfully facilitated the transition from individual experience to shared mental models, as emphasized in modern reflective learning frameworks[22-24].

Clinical Implications and Limitations

In the context of Thailand, this study demonstrates a successful implementation of a simulation curriculum that aligns with national competency standards for medical graduates[9] and supports digital literacy initiatives in higher education[25]. Despite these positive educational outcomes, this study has several limitations that should be acknowledged.

First, the primary limitation is the relatively small sample size (n=29) drawn from a single institution. Although this number reflects the total enrollment of a single academic year cohort at our center—thereby minimizing sampling bias—it may limit the statistical power to detect subtle differences between the leader and observer groups, as well as the overall generalizability of the findings. Second, the pre-test and post-test scores exhibited a skewed distribution, reflecting the high baseline knowledge of the participants and a potential ceiling effect. To address this, non-parametric statistical methods were strictly employed to

ensure the robustness of the data analysis.

Third, a structured instrument for quantitative clinical skill assessment was not utilized. Because this research was integrated into a time-constrained SBL curriculum, instructors focused on formative feedback across multiple holistic competencies beyond just technical skills, including teamwork, decision-making, and crisis resource management. In our setting, detailed technical proficiency is typically evaluated later through formal Objective Structured Clinical Examinations (OSCE).

Finally, this study assessed immediate knowledge gain but did not evaluate long-term retention or the direct transfer of these skills to live patients. Future multicenter, longitudinal research over multiple academic years is recommended to assess skill decay over time[26] and to formally incorporate OSCEs for quantitative skill measurement.

Conclusion

High-fidelity simulation-based training significantly improved knowledge of emergency and difficult airway management among novice fifth-year medical students, and learners reported high levels of satisfaction and self-confidence across the briefing, scenario, and debriefing phases. Knowledge improvement and learners' perceptions did not differ between students who took the active Leader role and those in the directed Observer role, indicating that the two roles produced comparable educational outcomes in this setting.

References

1. Bielka K, Kuchyn I, Fomina H, et al. Difficult airway simulation-based training for anaesthesiologists: efficacy and skills retention within six months. BMC Anesthesiol. 2024;24(1):44.
2. Cook TM, Woodall N, Frerk C, Fourth National Audit Project. Major complications of airway management in the UK: results of the Fourth National Audit Project of the Royal College of Anaesthetists and the Difficult Airway Society. Part 1: anaesthesia. Br J Anaesth. 2011;106(5):617-31.
3. Baker PA, Dua G, Iliff HA. Education in airway management: a narrative review. Br J Anaesth. 2026;136(1):308-16.
4. Grande B, Kolbe M, Biro P. Difficult airway management and training: simulation, commu nication, and feedback. Curr Opin Anaesthesiol. 2017;30(6):743-747.
5. Pareek S, Singh G. Pre-experimental study to assess effectiveness of airway management training among nurses and resident doctors working at Goyal Hospital, Jodhpur. Int J Med Sci Clin Res Rev. 2025;8(6):1336-40.
6. Kennedy CC, Cannon EK, Warner DO, Cook DA. Advanced airway management simulation training in medical education: a systematic review and meta-analysis. Crit Care Med.

2014;42(1):169-78.
7. Komasawa N, Berg BW. Simulation-based airway management training for anesthesiologists - a brief review of its essential role in skills training for clinical competency. J Educ Perioper Med. 2017;19(4):E612.
8. Maurya I, Ahmed SM, Garg R. Simulation in airway management teaching and training. Indian J Anaesth. 2024;68(1):52-57.
9. The Medical Council of Thailand. National Competency Standards for Medical Graduates 2023 (B.E. 2566). Nonthaburi: The Medical Council of Thailand; 2023.
10. O'Regan S, Molloy E, Watterson L, Nestel D. Observer roles that optimize learning in healthcare simulation education: a systematic review. Adv Simul (Lond). 2016;1:4.
11. Inam M, Tariq U, Durrani R, Karim MM, Surani Z, Ismail FW. Comparison of the observer role vs. active participant role in interprofessional simulation-based education for affective skills: a randomized-controlled trial. BMC Med Educ. 2025;26(1):134.
12. Sun Y, Pan C, Li T, Gan TJ. Airway management education: simulation based training versus non-simulation based training-A systematic review and meta-analyses. BMC Anesthesiol. 2017;17(1):17.
13. Bandura A. Social foundations of thought and action: a social cognitive theory. Englewood Cliffs (NJ): Prentice-Hall; 1986.
14. Bethards ML. Applying social learning theory to the observer role in simulation. Clin Simul Nurs. 2014;10(2):e65-e69.
15. Kolb DA. Experiential learning: experience as the source of learning and development. 2nd ed. Upper Saddle River (NJ): Pearson Education; 2014.
16. Ahmad A, Abd Rahim AF, Wan Mohammad WMZ, Mat Nor MZ. Experiential learning values based on Kolb theory: a qualitative output in the School of Medical Sciences, Universiti Sains Malaysia. Educ Med J. 2024;16(Suppl 1):11-22.
17. Moraes EMB, Carvalho DRDS, Sandars J, Ozahata TM, Patel R, Cecilio-Fernandes D, et al. Unravelling the differences between observation and active participation in simulation-based education. Med Teach. 2025;47(6):991-996.
18. Massoth C, Röder H, Ohlenburg H, Hessler M, Zarbock A, Pöpping DM, et al. High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. BMC Med Educ. 2019;19(1):29. doi: 10.1186/s12909-019-1464-7.
19. Apfelbaum JL, Hagberg CA, Connis RT, Abdelmalak BB, Agarkar M, Dutton RP, et al. 2022 American Society of Anesthesiologists practice guidelines for management of the difficult airway. Anesthesiology. 2022;136(1):31-81.

20. Lei C, Palm K. Crisis resource management training in medical simulation. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023.
21. Schram A, Bonne NL, Henriksen TB, Paltved C, Hertel NT, Lindhard MS. Simulation-based team training for healthcare professionals in pediatric departments: study protocol for a nonrandomized controlled trial. BMC Med Educ. 2024;24(1):607.
22. Dreifuerst KT, Decker SI. Debriefing: an essential component for learning in simulation pedagogy. In: Jeffries PR, National League for Nursing, editors. Simulation in nursing education: from conceptualization to evaluation. 2nd ed. New York: National League for Nursing; 2012. p. 105-130.
23. Eppich W, Cheng A. Promoting Excellence and Reflective Learning in Simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. Simul Healthc. 2015;10(2):106-115.
24. Şahin G, Başak T. Debriefing methods in simulation-based education. J Educ Res Nurs. 2021;18(3):341-346.
25. Ministry of Higher Education, Science, Research and Innovation. (2024). Digital literacy policy for undergraduate students. Bangkok: MHESI.

ORIGINAL ARTICLE

Factors Influencing Competency Scores of Medical Students in Researching Emergency Care Principles Using ChatGPT

Sathaya Buri

Yala Hospital Medical Education Center

Received: 13 February 2026

Revised: 8 June 2526

Accepted: 29 August 2026

Abstract

Background: Artificial intelligence (AI), particularly ChatGPT, offers significant potential for enhancing medical education and clinical decision-making, especially in high-pressure environments like emergency departments. This study investigated the factors influencing medical students' competency and satisfaction when using ChatGPT to research emergency care principles.

Methods: This descriptive-analytical study was conducted on 55 medical students from years 4–6 using total population sampling. Outcome measurements comprised two components: 1) competency in retrieving emergency care principles, assessed via three clinical scenario tests, and 2) user satisfaction, evaluated using a 4-point Likert scale questionnaire. Data were analyzed using descriptive statistics, independent t-tests, and ANOVA.

Results: The quantity and completeness of information provided in the clinical scenarios significantly impacted retrieval competency scores. Scenarios with limited information yielded the lowest mean score (34.9 ± 19.3), which was significantly lower than scenarios with comprehensive information (mean score 52.8 ± 13.5; MD = 17.9, $p < 0.001$) and moderate information (mean score 54.1 ± 19.8; MD = 19.2, $p < 0.001$). This indicates that scenarios with comprehensive data lead to more accurate outcomes. Additionally, participants reported the highest satisfaction regarding ChatGPT's utility as a medical education tool (85.4% rated satisfied or highly satisfied; mean = 3.3 ± 0.7 out of 4). In contrast, personal factors, including gender, academic year, GPA, and prior experience with ChatGPT, did not significantly affect competency scores.

Conclusion: The quantity and quality of input data are critical factors determining medical students' competency scores in retrieving emergency medicine information using ChatGPT, whereas personal factors and prior experience have no impact on this performance. Furthermore, students exhibited positive attitudes and satisfaction toward utilizing ChatGPT for clinical decision support and medical education.

Key words: ChatGPT, AI, Medical education, Decision making, Problem solving

บทความวิจัย

ปัจจัยที่มีผลต่อคะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉินโดยใช้ ChatGPT

สาธญา บุรี

ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลยะลา

รับบทความ: 13 กุมภาพันธ์ 2569

ปรับแก้ไขบทความ: 8 มิถุนายน 2569

ตอบรับตีพิมพ์: 29 สิงหาคม 2569

บทคัดย่อ

หลักการและเหตุผล: ปัจจุบันปัญญาประดิษฐ์ (AI) เข้ามามีบทบาทสำคัญในการเพิ่มประสิทธิภาพการศึกษาแพทยศาสตร์และการตัดสินใจทางคลินิก โดยเฉพาะในงานเวชศาสตร์ฉุกเฉินที่ต้องการความรวดเร็วและแม่นยำ การศึกษานี้จึงมีวัตถุประสงค์เพื่อทำการศึกษาปัจจัยที่มีผลต่อสมรรถนะของนักศึกษาแพทย์ในการใช้ ChatGPT เพื่อสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉิน

วิธีการศึกษา: การวิจัยเชิงพรรณนาเชิงวิเคราะห์ กลุ่มตัวอย่างคือนักศึกษาแพทย์ชั้นปีที่ 4-6 จำนวน 55 คน โดยใช้การเลือกตัวอย่างแบบครอบคลุมประชากรทั้งหมด การศึกษาตัวแปรผลลัพธ์ ประกอบด้วย 2 ส่วน ได้แก่ 1) สมรรถนะในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉิน วัดผลโดยใช้แบบทดสอบสถานการณ์จำลองทางคลินิก (clinical scenarios) จำนวน 3 ข้อ และ 2) ความพึงพอใจต่อการใช้งาน วัดผลโดยใช้แบบสอบถามชนิดมาตราส่วนประมาณค่า (Likert scale) 4 ระดับ วิเคราะห์ข้อมูลด้วยสถิติเชิงพรรณนา การทดสอบ t-test และ ANOVA

ผลการศึกษา: ปริมาณและความสมบูรณ์ของข้อมูลในโจทย์สถานการณ์จำลองมีผลต่อคะแนนสมรรถนะในการสืบค้นอย่างมีนัยสำคัญทางสถิติ โดยสถานการณ์ที่มีข้อมูลน้อยมีคะแนนเฉลี่ยต่ำที่สุด (34.9 ± 19.3 คะแนน) ซึ่งต่ำกว่าสถานการณ์ที่มีข้อมูลมาก (52.8 ± 13.5 คะแนน; MD = 17.9, $p < 0.001$) และสถานการณ์ที่มีข้อมูลปานกลาง (54.1 ± 19.8 คะแนน; MD = 19.2, $p < 0.001$) อย่างมีนัยสำคัญ กล่าวคือโจทย์ที่มีข้อมูลครบถ้วนนำไปสู่ผลลัพธ์ที่ถูกต้องมากกว่า นอกจากนี้กลุ่มตัวอย่างมีความพึงพอใจว่า ChatGPT มีประโยชน์ในด้านสื่อการเรียนการสอนทางการแพทย์สูงที่สุด (ร้อยละ 85.4 ประเมินระดับมากถึงมากที่สุด; ค่าเฉลี่ย = 3.3 ± 0.7 จากคะแนนเต็ม 4) ในขณะที่ปัจจัยส่วนบุคคล ได้แก่ เพศ ระดับชั้นปี เกรดเฉลี่ย และประสบการณ์การใช้ ChatGPT ไม่มีผลต่อคะแนนสมรรถนะอย่างมีนัยสำคัญทางสถิติ

สรุปผล: ปริมาณและคุณภาพของข้อมูลนำเข้า (Input Data) เป็นปัจจัยสำคัญที่มีผลต่อคะแนนสมรรถนะของนักศึกษาแพทย์ในการใช้ ChatGPT เพื่อสืบค้นข้อมูลเวชศาสตร์ฉุกเฉิน ในขณะที่ปัจจัยส่วนบุคคลและประสบการณ์การใช้ ChatGPT ไม่มีผลต่อสมรรถนะดังกล่าว นอกจากนี้ นักศึกษาแพทย์มีทัศนคติเชิงบวกและพึงพอใจต่อการใช้ ChatGPT ในการสนับสนุนการตัดสินใจทางคลินิกและการเรียนการสอน

คำสำคัญ: ChatGPT, AI, Medical education, dicison making, problem solving

บทนำ

ในยุคปัจจุบันเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence: AI) โดยเฉพาะโมเดลภาษาขนาดใหญ่ (Large Language Models: LLMs) เช่น ChatGPT ได้เข้ามามีบทบาทเปลี่ยนแปลงพื้นที่การศึกษาแพทยศาสตร์และวิทยาศาสตร์สุขภาพในช่วงไม่กี่ปีที่ผ่านมา จากการศึกษางานวิจัยทั่วโลกพบหลักฐานที่สะท้อนให้เห็นถึงแนวโน้มศักยภาพของ AI ในฐานะเครื่องมือการสอนและสนับสนุนการตัดสินใจทางคลินิก โดยงานวิจัยของ Kung และคณะ[1] พบว่า ChatGPT มีฐานข้อมูลความรู้พื้นฐานที่ครอบคลุมจนสามารถทำข้อสอบใบประกอบวิชาชีพแพทย์ของสหรัฐอเมริกา (USMLE) ผ่านเกณฑ์มาตรฐานได้ ในขณะเดียวกัน Gilson และคณะ[2] เน้นว่า AI สามารถตีความเชิงตรรกะของการให้เหตุผลในการวินิจฉัย (Rationales) ได้ ซึ่งช่วยส่งเสริมความเข้าใจเชิงลึกให้นักศึกษาแพทย์ได้เป็นอย่างดี

สาขาเวชศาสตร์ฉุกเฉิน (Emergency Medicine) ซึ่งเป็นสาขาที่มีความกดดัน ท้าทายสูง ต้องอาศัยการตัดสินใจที่รวดเร็วและแม่นยำ เนื่องจากมีบริบทการเรียนการสอนที่แตกต่างจากหอผู้ป่วยทั่วไป (Ward) อย่างสิ้นเชิง นักศึกษาแพทย์ในห้องฉุกเฉินต้องเผชิญกับผู้ป่วยที่ยังไม่ได้รับการวินิจฉัย ภายใต้ข้อมูลที่จำกัดเวลาที่กดดัน ทำให้ต้องอาศัยการตัดสินใจและรักษาผู้ป่วยที่คัดกรองตามระดับความเสี่ยง (Triage) สูงก่อนอย่างรวดเร็ว ต้องฝึกคิดวินิจฉัยแยกโรคที่อันตรายถึงชีวิตออกไปก่อน ซึ่งการปฏิบัติงานจริงต้องอาศัยการเรียนรู้จากผู้ป่วยหน้างานเป็นหลัก ทำให้มีข้อจำกัดด้านความหลากหลายของโรค อาจทำให้นักศึกษาที่เพิ่งเริ่มขึ้นชั้นคลินิกเกิดความสับสนหรือจับประเด็นสำคัญไม่ได้ แตกต่างจากการปฏิบัติงานในหอผู้ป่วยทั่วไป ผู้ป่วยมักได้รับการวินิจฉัยในเบื้องต้นมาแล้วระดับหนึ่ง ทำให้นักศึกษาสามารถที่จะโฟกัสไปที่การเจาะลึกพยาธิสภาพ และการวางแผนการรักษาแบบองค์รวมได้ ขณะเดียวกันในการเรียนการสอนและบทบาทของอาจารย์แพทย์จะสอนจากสถานการณ์จริงตรงหน้า นักศึกษาจะได้รับ Feedback ที่ทันทีจากการตัดสินใจของนักศึกษา อาจารย์สามารถประเมินทักษะการแก้ปัญหาเฉพาะหน้าของนักศึกษาได้ชัดเจน แต่ด้วยภาระงานและความเร่งด่วน อาจารย์มักไม่มีเวลาสำหรับการอภิปรายเชิงลึกหรือไม่สามารถทำ Bedside teaching แบบค่อยเป็นค่อยไปได้ ทำให้บางครั้งนักศึกษาอาจแค่ทำตามคำสั่งโดยไม่ทันได้ทำความเข้าใจเหตุผลทั้งหมด ด้านการทำหัตถการและการลงมือปฏิบัติ (Procedural Skills) เป็นแหล่งรวมหัตถการฉุกเฉินที่หลากหลายที่สุด เช่น CPR, Intubation, Suturing, Splinting, FAST scan นักศึกษามีโอกาสได้ลงมือปฏิบัติจริงบ่อยครั้ง แต่ภายใต้ความกดดันของเวลาและชีวิตผู้ป่วย หากนักศึกษายังไม่มีความมั่นใจเพียงพออาจจะถูกลดบทบาทให้ไปเป็นเพียงผู้สังเกตการณ์เพื่อความปลอดภัยของผู้ป่วย งานวิจัยของ Strong และคณะ[3] และ Rao และคณะ[4] พบว่า AI สามารถใช้สร้างสถานการณ์จำลอง (Simulated Scenarios) เพื่อช่วยในการฝึกทักษะในการให้เหตุผลทางคลินิกได้อย่างมีประสิทธิภาพแบบโต้ตอบในทีละขั้นตอนและให้มีแนวทางการตัดสินใจสั่งการรักษาภายในข้อมูลจำกัด การนำ ChatGPT เข้ามาประยุกต์ใช้จึงมีส่วนสำคัญอย่างยิ่งในการทำหน้าที่เป็นกระดานสะท้อนความคิดให้นักศึกษาได้เรียนรู้แนวทางในการวินิจฉัยโรค การบริหารจัดการผู้ป่วย และชดเชยประสบการณ์การตัดสินใจที่อาจขาดหายไปจากข้อจำกัดหน้างาน อย่างไรก็ตามประสิทธิภาพของ AI ยังขึ้นอยู่กับทักษะของผู้ใช้ โดยเฉพาะการกำหนดคำสั่ง (Prompt Engineering) ซึ่ง Hosseini และคณะ[5] ระบุว่าเป็นปัจจัยสำคัญที่มากำหนดความถูกต้องแม่นยำของคำตอบซึ่งบริบทของเวชศาสตร์ฉุกเฉิน ข้อมูลนำเข้าที่นักศึกษาจะนำมาป้อนให้กับ AI นั้น ได้มาจากความสามารถในการประเมินผู้ป่วยแรกรับจากสัญญาณชีพ การซัก

ประวัติและการตรวจร่างกาย การให้การวินิจฉัยโรคภายใต้ที่ข้อมูลยังไม่สมบูรณ์ หากนักศึกษาสามารถรวบรวมและคัดกรองเฉพาะข้อมูลทางคลินิกที่สำคัญป้อนให้ AI ได้อย่างครบถ้วน ย่อมส่งผลให้ AI สามารถประมวลผลและสนับสนุนการตัดสินใจได้แม่นยำขึ้น

ด้วยเหตุนี้ผู้วิจัยจึงเห็นความสำคัญของศึกษาปัจจัยที่มีผลต่อสมรรถนะของนักศึกษาแพทย์ ในการใช้ ChatGPT เพื่อสืบค้นและแก้ไขปัญหาทางเวชศาสตร์ฉุกเฉิน การศึกษานี้ต้องการเปรียบเทียบปัจจัยด้านพื้นฐานทางวิชาการ (GPA) หรือปัจจัยด้านเทคนิค เช่น ความละเอียดของข้อมูลในโจทย์ มีอิทธิพลอย่างมีนัยสำคัญต่อประสิทธิภาพการเรียนรู้

วัตถุประสงค์ของการวิจัย

วัตถุประสงค์หลัก

เพื่อศึกษาปัจจัยที่มีผลต่อคะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉิน โดยใช้ ChatGPT

วัตถุประสงค์รอง

เพื่อสำรวจความคิดเห็นของนักศึกษาแพทย์ต่อการใช้ ChatGPT ในการสนับสนุนการตัดสินใจทางคลินิกและการเรียนการสอน

วิธีการศึกษา

เป็นการวิจัยเชิงพรรณนาเชิงวิเคราะห์ (descriptive analytic study) ในนักศึกษาแพทย์ชั้นปีที่ 4 - 6 จำนวน 55 คน ที่ศึกษาอยู่ปีการศึกษา พ.ศ. 2567 ตั้งแต่วันที่ 28 พฤษภาคม 2567 ถึงวันที่ 28 มิถุนายน 2567 และปฏิบัติงานในศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลยะลา

การกำหนดขนาดตัวอย่าง ผู้วิจัยใช้วิธีการเลือกตัวอย่างแบบครอบคลุมประชากรทั้งหมด (Total population sampling) จากนักศึกษาที่เข้าเกณฑ์และยินดีที่จะเข้าร่วมการศึกษา (Inclusion criteria) ในช่วงเวลาที่กำหนด ทำให้ได้กลุ่มตัวอย่างรวมทั้งสิ้นจำนวน 55 คน ซึ่งประกอบด้วยชั้นปีที่ 4 จำนวน 19 คน ชั้นปีที่ 5 จำนวน 18 คน และชั้นปีที่ 6 จำนวน 18 คน ซึ่งเป็นนักศึกษาแพทย์ทั้งหมดของศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลยะลา

การดำเนินการวิจัย

นักศึกษาแพทย์ที่ปฏิบัติงานอยู่ในศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลยะลา ได้รับเชิญให้เข้าร่วมศึกษาโดยเน้นย้ำว่ามีสิทธิปฏิเสธและไม่มีผลต่อการเรียนการสอน การทดสอบจัดขึ้นในช่วงประมาณ 1 เดือนแรกหลังจากเปิดปีการศึกษาของแต่ละระดับชั้นปี (ซึ่งในขณะนั้นมีเพียงกลุ่มตัวอย่างนักศึกษาแพทย์ชั้นปีที่ 6 เท่านั้นที่เคยผ่านการฝึกปฏิบัติงานในรายวิชาเวชศาสตร์ฉุกเฉิน) นักศึกษาทำแบบทดสอบให้เสร็จภายในระยะเวลา 1 ชั่วโมง โดยให้ใช้เพียง ChatGPT Version 3.5 (Free version) ในการค้นคว้าเท่านั้น ไม่อนุญาตให้ใช้ตำราเรียน แอปพลิเคชันทางการแพทย์หรือสืบค้นผ่าน Search Engine

การให้คะแนนโดยแพทย์เวชศาสตร์ฉุกเฉิน จำนวน 1 ท่าน ประเมินความถูกต้องของคำตอบเทียบกับแบบตรวจสอบรายการ (Checklist rubrics) ที่อิงตามแนวทางมาตรฐานการรักษา โดยให้เกณฑ์การผ่านที่ร้อยละ 60 ซึ่งเป็นระดับสมรรถนะขั้นต่ำที่ปลอดภัยสำหรับการดูแลผู้ป่วย

เครื่องมือที่ใช้ในการวิจัย

1. แบบทดสอบสถานการณ์จำลองทางคลินิก (Clinical Scenarios) เป็นข้อสอบอัตนัยจำนวน 3 ข้อ สถานกาณ์จำลองที่ 1 เป็นโจทย์ที่มีปริมาณข้อมูลน้อย สถานการณ์จำลองที่ 2 เป็นโจทย์ที่มีปริมาณข้อมูลมากและสถานการณ์จำลองที่ 3 เป็นโจทย์ที่มีปริมาณข้อมูลปานกลาง ที่สร้างขึ้นโดยผู้วิจัยซึ่งเป็นแพทย์

เวชศาสตร์ฉุกเฉิน โดยอ้างอิงเนื้อหาและเกณฑ์การรักษาจากแนวทางมาตรฐานสากล ได้แก่ ACLS, ATLS และ PALS ซึ่งใช้วัดความถูกต้องของคำตอบในการแก้โจทย์สถานการณ์จำลองทางคลินิก โดยให้นักศึกษาตอบการวินิจฉัยโรคที่เป็นไปได้มากที่สุด วางแผนการรักษา ระบุชื่อยา ขนาดยาและคำแนะนำ เพื่อให้มีความสอดคล้องกับสถานการณ์จริงทางคลินิก แบบทดสอบมีคะแนนเต็มข้อละ 100 คะแนน โดยกำหนดน้ำหนักคะแนนทั้ง 3 ข้อเท่ากัน

2. แบบสอบถามความพึงพอใจ เป็นแบบสอบถามชนิดมาตราส่วนประมาณค่า (Likert scale) 4 ระดับ เพื่อวัดความพึงพอใจ 6 ด้าน ได้แก่ ความน่าเชื่อถือทางการแพทย์, ความถูกต้องทางการแพทย์, การสนับสนุนการตัดสินใจทางการแพทย์, ประโยชน์ในด้านสื่อการเรียนการสอนทางการแพทย์, ความพึงพอใจที่เกิดจากการใช้งาน ChatGPT ในด้านการปฏิบัติงานทางคลินิก และความพึงพอใจต่อการใช้งาน ChatGPT โดยภาพรวมทางการแพทย์ ซึ่งเป็นแบบสอบถามที่ผู้วิจัยอ้างอิงจากการทบทวนวรรณกรรมที่เกี่ยวข้องกับการใช้ปัญญาประดิษฐ์ในแพทยศาสตรศึกษา[4,6,7,8]

การวิเคราะห์ข้อมูล:

การศึกษานี้ใช้สถิติเชิงพรรณนา (ค่าเฉลี่ย ร้อยละ ส่วนเบี่ยงเบนมาตรฐาน) **ใช้ Independent t-test** และ ANOVA เปรียบเทียบคะแนนตามปัจจัย กำหนดระดับนัยสำคัญทางสถิติที่ $P < 0.05$ และความคิดเห็นด้วยแบบสอบถามใช้ **Likert scale**

ผลการศึกษา

ตารางที่ 1 ข้อมูลลักษณะของนักศึกษาแพทย์และปัจจัยทางวิชาการของนักศึกษาแพทย์

ปัจจัยพื้นฐาน	จำนวน N=55	ร้อยละ
เพศ		
เพศชาย	13	23.6
เพศหญิง	42	76.4
ระดับชั้นปีการศึกษา		
ชั้นปีที่ 4	19	34.5
ชั้นปีที่ 5	18	32.7
ชั้นปีที่ 6	18	32.7
เกรดเฉลี่ย		
2.5-2.99	11	20.0
3.0-3.49	27	49.1
3.5-4.00	17	30.9
ประสบการณ์ใช้งาน ChatGPT		
ไม่เคยมีประสบการณ์	21	38.2
เคยมีประสบการณ์	34	61.8

ตารางที่ 1 แสดงถึงการกระจายตัวข้อมูลลักษณะของนักศึกษาแพทย์และปัจจัยทางวิชาการของผู้เข้าร่วมการศึกษา พบว่าผู้เข้าร่วมส่วนใหญ่เป็นเพศหญิงร้อยละ 76.4 การกระจายตามชั้นปีพบว่ามีสัดส่วนใกล้เคียงกันในนักศึกษา ชั้นปีที่ 4 ร้อยละ 34.5 ชั้นปีที่ 5 ร้อยละ 32.7 และชั้นปีที่ 6 ร้อยละ 32.7 ในด้านผลการเรียนพบว่านักศึกษามีคะแนนเฉลี่ยสะสม (GPA) อยู่ในช่วง 3.0 - 3.49 มากที่สุด คิดเป็นร้อยละ 49.1 รองลงมาคือช่วง 3.5 - 4.00 คิดเป็นร้อยละ 30.9 และช่วง 2.5 - 2.99 คิดเป็นร้อยละ 20.0 นอกจากนี้ นักศึกษาส่วนใหญ่ร้อยละ 61.8 เคยมีประสบการณ์ใช้งาน ChatGPT มาก่อน

ตารางที่ 2: คะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉินโดยใช้ ChatGPT

โจทย์	**คะแนนดิบ** **Mean ± SD**	**ความถี่** **เกณฑ์การผ่านร้อยละ 60**
สถานการณ์จำลองที่ 1 (ข้อมูลน้อย)	34.9 ± 19.3	6 (10.9)
สถานการณ์จำลองที่ 2 (ข้อมูลมาก)	52.8 ± 13.5	16 (29.1)
สถานการณ์จำลองที่ 3 (ข้อมูลปานกลาง)	54.1 ± 19.8	27 (49.1)

ตารางที่ 2 แสดงคะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉินโดยใช้ ChatGPT พบว่า สถานการณ์จำลองที่ 1 ซึ่งเป็นโจทย์ที่มีปริมาณข้อมูลน้อย นักศึกษามีคะแนนเฉลี่ยต่ำที่สุดเท่ากับ 34.9 ± 19.3 คะแนน และมีนักศึกษาผ่านเกณฑ์เพียง 6 คน คิดเป็นร้อยละ 10.9 เท่านั้น สำหรับ สถานการณ์จำลองที่ 2 ซึ่งเป็นโจทย์ที่มีปริมาณข้อมูลมาก นักศึกษามีคะแนนเฉลี่ยเพิ่มขึ้นเป็น 52.8 ± 13.5 คะแนน โดยมีนักศึกษาผ่านเกณฑ์จำนวน 16 คน คิดเป็นร้อยละ 29.1 และในสถานการณ์จำลองที่ 3 ซึ่งเป็นโจทย์ที่มีปริมาณข้อมูลปานกลาง นักศึกษามีคะแนนเฉลี่ยสูงที่สุดเท่ากับ 54.1 ± 19.8 คะแนน โดยมีนักศึกษาผ่านเกณฑ์จำนวน 27 คน คิดเป็นร้อยละ 49.1 เมื่อพิจารณาในภาพรวม พบว่านักศึกษาแพทย์ส่วนใหญ่มีคะแนนสมรรถนะต่ำกว่าเกณฑ์ผ่านร้อยละ 60 ในทุกสถานการณ์ โดยสถานการณ์จำลองที่ 1 มีสัดส่วนผู้ผ่านเกณฑ์น้อยที่สุด รองลงมาคือสถานการณ์จำลองที่ 2 และสถานการณ์จำลองที่ 3 ตามลำดับ นอกจากนี้ยังพบว่าค่าเบี่ยงเบนมาตรฐานในสถานการณ์จำลองที่ 1 และสถานการณ์จำลองที่ 3 มีค่าค่อนข้างสูง ร้อยละ 19.3 และร้อยละ 19.8 ตามลำดับ บ่งชี้ถึงความแตกต่างของคะแนนระหว่างนักศึกษาแต่ละคนในระดับสูง ในขณะที่สถานการณ์จำลองที่ 2 มีค่าเบี่ยงเบนมาตรฐานต่ำกว่าร้อยละ 13.5 แสดงให้เห็นว่าการกระจายของคะแนนมีความสม่ำเสมอกว่า

ตารางที่ 3: ความสัมพันธ์ระหว่างปัจจัยต่าง ๆ กับคะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉินโดยใช้ ChatGPT

ปัจจัย	สถานการณ์จำลองที่ 1 Mean ± SD	P-Value	สถานการณ์จำลองที่ 2 Mean ± SD	P-Value	สถานการณ์จำลองที่ 3 Mean ± SD	P-Value
เพศ		0.549		0.752		0.404
เพศชาย	37.7 ± 21.4		53.9 ± 14.7		50.1 ± 23.0	
ระดับชั้นปีการศึกษา		0.501		0.102		0.937
ชั้นปีที่ 4	31.1 ± 14.7		50.5 ± 18.5		52.8 ± 23.8	
ชั้นปีที่ 5	38.7 ± 20.0		58.3 ± 10.4		54.4 ± 17.7	
ชั้นปีที่ 6	35.0 ± 23.0		49.7 ± 7.5		55.2 ± 8.2	
เกรดเฉลี่ย		0.343		0.255		0.462
2.5-2.99	27.4 ± 19.0		48.1 ± 15.3		48.9 ± 20.0	
3.0-3.49	36.0 ± 21.2		52.3 ± 13.5		53.6 ± 20.0	
3.5-4.00	38.0 ± 16.0		56.7 ± 11.8		58.4 ± 19.6	
ประสบการณ์ใช้งาน ChatGPT		0.461		0.290		0.245
เคยมีประสบการณ์	36.4 ± 19.4		54.3 ± 13.7		56.6 ± 16.6	

ตารางที่ 3 แสดงความสัมพันธ์ระหว่างปัจจัยต่าง ๆ กับคะแนนสมรรถนะของนักศึกษาแพทย์ในการสืบค้นหลักการดูแลรักษาผู้ป่วยภาวะฉุกเฉินโดยใช้ ChatGPT ผลการศึกษาพบว่า ไม่พบความแตกต่างอย่างนัยสำคัญทางสถิติระหว่างนักศึกษาเพศชายและเพศหญิงในทุกข้อคำถาม เช่นเดียวกับการเปรียบเทียบระหว่างชั้นปีที่ 4, 5 และ 6 ที่ไม่พบความแตกต่างอย่างมีนัยสำคัญทางสถิติ นักศึกษาที่มีคะแนนเฉลี่ยสะสม (GPA) สูงกว่า

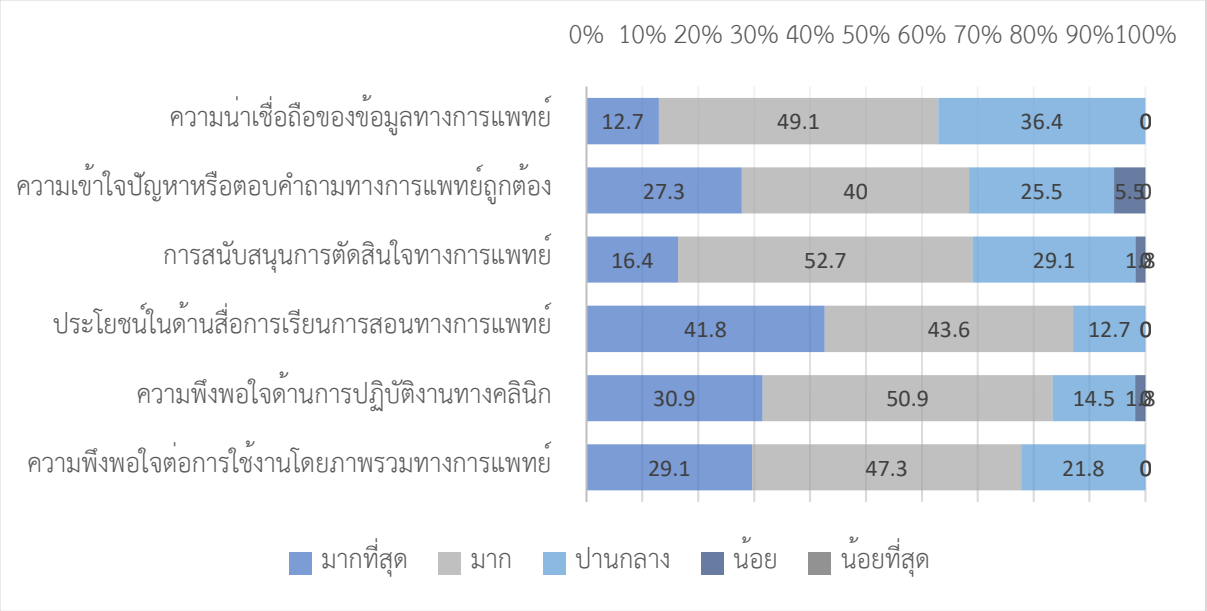
มีแนวโน้มที่จะได้คะแนนสูงกว่า แต่ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ ขณะที่นักศึกษาที่เคยมีประสบการณ์การใช้ ChatGPT มีคะแนนเฉลี่ยสูงกว่าผู้ที่ไม่เคยใช้งานเล็กน้อย แต่ก็ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ โดยสรุปปัจจัยที่ทำการศึกษาทั้งหมดไม่พบว่ามีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติกับคะแนนสมรรถนะ

ตารางที่ 4 การเปรียบเทียบคะแนนการตอบคำถามตามปริมาณข้อมูลในสถานการณ์จำลอง

สถานการณ์จำลอง		Mean Difference	p-value
สถานการณ์จำลองที่ 1 (ข้อมูลน้อย)	สถานการณ์จำลองที่ 2 (ข้อมูลมาก)	17.9	p<0.001
สถานการณ์จำลองที่ 1 (ข้อมูลน้อย)	สถานการณ์จำลองที่ 3 (ข้อมูลปานกลาง)	19.2	p<0.001
สถานการณ์จำลองที่ 2 (ข้อมูลมาก)	สถานการณ์จำลองที่ 3 (ข้อมูลปานกลาง)	1.3	0.620

ตารางที่ 4 การเปรียบเทียบคะแนนการตอบคำถามตามปริมาณข้อมูลในสถานการณ์จำลอง ผลการวิเคราะห์จากการวิเคราะห์ความแปรปรวน พบว่าปริมาณข้อมูลตั้งต้นในสถานการณ์จำลองมีผลต่อคะแนนสมรรถนะของนักศึกษาอย่างมีนัยสำคัญทางสถิติ ($p < 0.001$) เมื่อพิจารณาคะแนนเฉลี่ยในแต่ละสถานการณ์จำลอง พบว่าสถานการณ์จำลองที่ 1 ซึ่งมีปริมาณข้อมูลน้อย มีคะแนนเฉลี่ยต่ำที่สุดเท่ากับ 34.9 ± 19.3 คะแนน รองลงมาคือสถานการณ์จำลองที่ 2 ซึ่งมีปริมาณข้อมูลมากได้คะแนนเฉลี่ย 52.8 ± 13.5 คะแนน และสถานการณ์จำลองที่ 3 ซึ่งมีปริมาณข้อมูลปานกลาง มีคะแนนเฉลี่ยสูงที่สุดเท่ากับ 54.1 ± 19.8 คะแนน เมื่อทำการเปรียบเทียบรายคู่ พบว่าสถานการณ์จำลองที่ 1 มีคะแนนเฉลี่ยต่ำกว่าสถานการณ์จำลองที่ 2 อย่างมีนัยสำคัญทางสถิติ (Mean difference = 17.9, $p < 0.001$) และมีคะแนนต่ำกว่าสถานการณ์จำลองที่ 3 อย่างมีนัยสำคัญทางสถิติเช่นกัน (Mean difference = 19.2, $p < 0.001$) อย่างไรก็ตาม เมื่อเปรียบเทียบระหว่างสถานการณ์จำลองที่ 2 และสถานการณ์จำลองที่ 3 พบว่าคะแนนเฉลี่ยของทั้งสองสถานการณ์ไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ (Mean difference = 1.3, $p = 0.620$) แม้ว่าสถานการณ์จำลองที่ 3 จะมีคะแนนเฉลี่ยสูงกว่าเล็กน้อยก็ตาม

กราฟที่ 1: แสดงสัดส่วนร้อยละความพึงพอใจแต่ละด้านต่อการใช้งาน ChatGPT ของนักศึกษาแพทย์



กราฟที่ 1 แสดงสัดส่วนร้อยละความพึงพอใจแต่ละด้านต่อการใช้งาน ChatGPT ของนักศึกษาแพทย์ พบว่ากลุ่มตัวอย่างมีความพึงพอใจต่อการใช้งาน ChatGPT ระดับมากถึงมากที่สุด ในด้านประโยชน์ในสื่อการเรียนการสอนทางการแพทย์สูงที่สุด รวมคิดเป็นร้อยละ 85.4 รองลงมาคือความพึงพอใจในด้านการปฏิบัติงานทางคลินิก รวมคิดเป็นร้อยละ 81.8 ในขณะที่ด้านความน่าเชื่อถือของข้อมูลทางการแพทย์ กลุ่มตัวอย่างส่วนใหญ่ประเมินให้อยู่ในระดับมาก คิดเป็นร้อยละ 49.1 และปานกลาง คิดเป็นร้อยละ 36.4 ซึ่งเป็นด้านที่มีคะแนนเฉลี่ยต่ำที่สุดจากการประเมินทั้งหมด

อภิปรายผลการศึกษา

จากผลการศึกษาภาพรวมที่พบว่า นักศึกษาแพทย์ส่วนใหญ่มีคะแนนสมรรถนะในการสืบค้นข้อมูลต่ำกว่าเกณฑ์ผ่านร้อยละ 60 ในทุกสถานการณ์จำลองนั้น อาจสามารถอธิบายได้จากการประยุกต์ใช้ปัญญาประดิษฐ์ในบริบททางการแพทย์ โดยถึงแม้ว่านักศึกษาจะคุ้นเคยกับการใช้งาน AI ในชีวิตประจำวัน (ร้อยละ 61.8 เคยใช้งานมาก่อน) แต่การใช้ ChatGPT เพื่อที่จะแก้ปัญหาทางคลินิกที่ซับซ้อนนั้น จำเป็นต้องอาศัยในทักษะของความฉลาดทางปัญญาประดิษฐ์ ซึ่งนักศึกษาในปัจจุบันอาจจะยังไม่ได้รับการฝึกฝนอย่างเป็นระบบ การป้อนคำถามในลักษณะการคัดลอกข้อมูลทั้งหมด โดยไม่สกัดเฉพาะประเด็นทางคลินิกที่สำคัญ มักนำไปสู่ผลลัพธ์ที่กว้างเกินไปหรือไม่ตรงประเด็น ส่งผลให้คะแนนสมรรถนะความถูกต้องโดยรวมนั้นยังไม่ถึงเกณฑ์มาตรฐานที่กำหนดไว้

นอกจากนี้ถึงแม้มีเพียงนักศึกษาชั้นปีที่ 6 เท่านั้นที่ผ่านรายวิชาเวชศาสตร์ฉุกเฉิน แต่กลับไม่พบความแตกต่างอย่างมีนัยสำคัญของคะแนนสมรรถนะระหว่างชั้นปีในทุกสถานการณ์ ชี้ให้เห็นถึงศักยภาพของเทคโนโลยีปัญญาประดิษฐ์ในการทำหน้าที่ช่วยลดความเหลื่อมล้ำทางวิชาการและอุดช่องโหว่ทางด้านประสบการณ์ เป็นเครื่องมือช่วยสนับสนุนการตัดสินใจที่มีประสิทธิภาพมากขึ้น ทักษะในเรื่องของการตั้งคำถามและการมีข้อมูลนำเข้าที่ครบถ้วน อาจมีความสำคัญมากกว่าความจำจากความรู้พื้นฐานเดิม ทำให้ผู้เรียนสามารถเข้าถึงและประยุกต์ใช้แนวทางการดูแลผู้ป่วย

ฉุกเฉินที่ซับซ้อนได้อย่างเท่าเทียมกัน อย่างไรก็ตาม ประเด็นที่จะต้องพิจารณาด้วยความระมัดระวังคือ การที่นักศึกษาแพทย์ชั้นปีที่ 4 และ 5 ได้คะแนนใกล้เคียงกับชั้นปีที่ 6 อาจไม่ได้สะท้อนสมรรถนะที่แท้จริง เนื่องจากนักศึกษาที่ขาดความรู้พื้นฐานอาจนำคำตอบของ ChatGPT มาใช้โดยไม่สามารถตรวจสอบความถูกต้องได้ ซึ่งเป็นความเสี่ยงที่ซ่อนอยู่และต้องการการศึกษาเพิ่มเติมในอนาคต

นักศึกษาแพทย์ส่วนใหญ่มีทัศนคติเชิงบวกและมีความพึงพอใจระดับสูงในการค้นหาหลักการดูแลผู้ป่วยฉุกเฉินผ่าน ChatGPT ซึ่งสอดคล้องกับงานวิจัยในต่างประเทศของ Ali และคณะ[6] Peralta และคณะ[7] ที่ระบุว่าการใช้เทคโนโลยีปัญญาประดิษฐ์สามารถช่วยสร้างสภาพแวดล้อมการเรียนรู้ที่ปลอดภัย ส่งผลให้เกิดการพัฒนาความมั่นใจและการปรับปรุงกระบวนการเรียนรู้ที่มีประสิทธิภาพของนักศึกษาแพทย์และบุคลากรทางการแพทย์ โดยเฉพาะอย่างยิ่งขีดความสามารถของ AI ในการเป็นเครื่องมือช่วยคิดที่ตอบสนองต่อความจำเป็นในการตัดสินใจที่รวดเร็วและซับซ้อน เช่น ในบริบทของการแพทย์ฉุกเฉินและการดูแลผู้ป่วยวิกฤต ไม่มีความแตกต่างอย่างนัยสำคัญระหว่างประเด็นเรื่องเพศและปีการศึกษา อย่างไรก็ตามระดับเกรดเฉลี่ย (GPA) มีความสัมพันธ์อย่างมีนัยสำคัญกับคะแนนความพึงพอใจในมิติความถูกต้องทางการแพทย์ และประโยชน์ในด้านสื่อการเรียนการสอน ($p < 0.05$) ซึ่งบ่งชี้ว่านักศึกษาที่ระดับเกรดเฉลี่ยที่สูงกว่า อาจสามารถประเมินและใช้ประโยชน์จากเครื่องมือดังกล่าวอย่างมีประสิทธิภาพมากกว่า แม้ในงานวิจัยของ Roganovic[8] ระบุว่า ความคุ้นเคยกับเทคโนโลยี AI นั้นมีผลเชิงบวกต่อประสิทธิภาพในการประมวลผลและการทำโจทย์ แต่การศึกษาครั้งนี้กลับไม่พบความแตกต่างอย่างมีนัยสำคัญทางสถิติระหว่างกลุ่มนักศึกษาที่เคยและไม่เคยใช้งานมาก่อน อาจจะอธิบายได้ว่าประสบการณ์การใช้งานทั่วไปเพียงอย่างเดียว ไม่เพียงพอต่อการดึงศักยภาพสูงสุดของ AI มาใช้ในการแก้ปัญหาสถานการณ์ทางคลินิกที่ซับซ้อน สอดคล้องกับแนวคิดของ Knoth และคณะ[9] ที่ระบุว่าความฉลาดทางปัญญาประดิษฐ์ และทักษะการออกแบบคำสั่ง ในบริบททางการแพทย์นั้น ถือเป็น 'ทักษะอิสระ' ไม่ใช่ทักษะที่เกิดขึ้นเองตามธรรมชาติจากการเป็นผู้ใช้งานทั่วไป แต่นั้นเป็นทักษะเฉพาะที่ต้องได้รับการฝึกฝนอย่างเป็นระบบ โดยทักษะที่ส่งผลต่อความแม่นยำในการสืบค้นอย่างแท้จริงนั้นคือความสามารถในการสกัดและเรียบเรียงข้อมูลอย่างเป็นระบบ ดังนั้นนักศึกษาที่ได้เคยได้งาน AI ในชีวิตประจำวัน แต่อาจขาดทักษะการตั้งคำถามเชิงลึกทางการแพทย์ จึงมีคะแนนสมรรถนะที่ไม่แตกต่างจากผู้ที่ไม่เคยใช้งานมาก่อน

นอกจากนี้ปริมาณและคุณภาพของข้อมูลในคำถามสถานการณ์มีผลโดยตรงต่อคะแนนสมรรถนะการสืบค้นของนักศึกษา โดยพบว่า โจทย์สถานการณ์จำลองที่มีข้อมูลครบถ้วนเหมาะสมให้ผลลัพธ์คะแนนที่สูงกว่าสถานการณ์จำลองที่มีข้อมูลน้อย ซึ่งสะท้อนให้เห็นข้อจำกัดของ AI เมื่อต้องเจอกับภาวะความคลุมเครือของข้อมูลหรือโจทย์ที่ขาดรายละเอียดที่จำเป็น (AI จะไม่สามารถเชื่อมโยงตรรกะหรือสร้างการให้เหตุผล ที่ถูกต้องได้ ส่งผลทำให้เกิดการตอบแบบกว้าง หรือการสร้างข้อมูลเท็จ เพื่อเติมเต็มช่องว่างของข้อมูลที่ขาดหายไป สอดคล้องกับการศึกษาของ Kim และคณะ[10] ที่ตรวจสอบข้อจำกัดของโมเดลภาษาขนาดใหญ่ทางการแพทย์ โดยพบว่าเมื่อข้อมูลตั้งต้นมีความคลุมเครือ AI จะเกิดความล้มเหลวในการให้เหตุผล และมักจะทำการสร้างหรือจำลองประวัติผู้ป่วยขึ้นมาเอง เพื่อพยายามตอบคำถามให้สำเร็จ ขณะเดียวกันปริมาณข้อมูลที่มากที่สุด ไม่ได้นำไปสู่ความแม่นยำสูงสุดเสมอไป เนื่องจากผลการศึกษานี้พบว่าผลลัพธ์ที่ได้มีความขัดแย้งกับสมมติฐานเบื้องต้นที่ผู้วิจัยได้ตั้งไว้ โดยก่อนการศึกษานั้นผู้วิจัยคาดว่า

ยิ่งโจทย์สถานการณ์มีปริมาณข้อมูลทางคลินิกที่มากเท่าใด จะยิ่งช่วยให้ปัญญาประดิษฐ์ประมวลผลคำตอบได้ถูกต้องและทำให้นักศึกษามีคะแนนสมรรถนะสูงขึ้นตามไปด้วย แต่ผลการวิเคราะห์เปรียบเทียบแสดงให้เห็นว่าเมื่อข้อมูลตั้งต้นมีความสมบูรณ์ถึงระดับปานกลางแล้ว (สถานการณ์จำลองที่ 3) การเพิ่มปริมาณข้อมูลเข้าไปอีก (สถานการณ์จำลองที่ 2) ไม่ได้ทำให้คะแนนสมรรถนะดีขึ้นอย่างมีนัยสำคัญ อาจอธิบายได้ว่ากลไกการประมวลผลของโมเดลภาษาขนาดใหญ่ การใส่ข้อมูลที่มากเกินความจำเป็นในกลุ่มข้อมูลมาก รวมถึงข้อมูลที่ไม่เกี่ยวข้องในระบบจะทำหน้าที่เป็นสัญญาณรบกวน ที่เบี่ยงเบนความสนใจของ AI ออกจากประเด็นสำคัญ อ้างอิงจากการศึกษาของ Cai และคณะ[11] ด้านข้อจำกัดของ Medical LLMs ในปี 2025 ซึ่งพบว่าปัญญาประดิษฐ์มักถูกทำให้ไขว้เขวได้ง่าย จากข้อมูลส่วนเกิน โดยการแทรกข้อมูลทางการแพทย์ที่ไม่เกี่ยวข้องเข้าไปในโจทย์จะทำให้ความแม่นยำในการวินิจฉัยโรคของ AI ลดลงถึงร้อยละ 9.4 เนื่องจาก AI นั้นขาดกลไกเชิงตรรกะในการคัดกรองข้อมูลขยะด้วยตนเอง แม้โจทย์ในสถานการณ์ที่ 2 (กลุ่มข้อมูลมาก) จะให้ข้อมูลทางคลินิกที่ครบถ้วนและจำเป็นต่อการประเมินผู้ป่วย (เช่น อาการที่แสดง สัญญาณชีพ และน้ำหนักตัว) แต่กลับพบว่าคะแนนเฉลี่ยไม่ได้สูงกว่ากลุ่มที่มีข้อมูลปานกลาง อาจอธิบายได้จากความซับซ้อนของการประมวลผลคำสั่ง ซึ่งสอดคล้องกับผลการศึกษาของ Vishwanath และคณะ[12] ที่ได้พบว่า Medical LLMs มีความอ่อนแอต่อ distracting information อย่างมีนัยสำคัญ โดยเฉพาะในหมวด Patient Care: Diagnosis ที่พบความแม่นยำลดลงถึงร้อยละ 10.1 และหมวดระบบทางเดินหายใจได้รับผลกระทบมากที่สุดที่ร้อยละ 11.0 และจากแนวคิดของ Zheng และคณะ[13] กล่าวว่าความขัดแย้งที่เกิดขึ้นระหว่าง clinical presentation กับ vital signs ซึ่งอาจจะทำให้ attention mechanism ของ LLM ถ่วงน้ำหนักผิด และนำไปสู่การวางแผนการรักษาที่คลาดเคลื่อน ไม่สอดคล้องกับความรุนแรงที่แท้จริงของผู้ป่วย สภาวะที่ข้อมูลสองชุดส่งสัญญาณในทิศทางที่แตกต่างกันนี้ อาจก่อให้เกิด conflicting signals ต่อ attention mechanism ของโมเดล ดังนั้นแม้ว่าเทคโนโลยีโมเดลภาษาขนาดใหญ่จะมีศักยภาพสูงในการเป็นเครื่องมือช่วยคิด แต่ผลการศึกษานั้นชี้ให้เห็นถึงข้อจำกัดของ AI เมื่อต้องเผชิญกับข้อมูลที่คลุมเครือหรือมีข้อมูลส่วนเกินนั้น ยิ่งทำให้สะท้อนให้เห็นถึงความสำคัญของทักษะการให้เหตุผลทางคลินิก (clinical reasoning) ในตัวนักศึกษาแพทย์ โดยเฉพาะอย่างยิ่งในบริบทของเวชศาสตร์ฉุกเฉิน ซึ่งมักต้องเผชิญกับผู้ป่วยที่ยังไม่ได้รับการวินิจฉัยภายใต้ระยะเวลาที่จำกัดและสถานการณ์วิกฤต เช่น ภาวะช็อกหรืออุบัติเหตุบาดเจ็บรุนแรงที่ต้องการการคัดแยก (Triage) และกู้ชีพอย่างทันท่วงที ในสถานการณ์เช่นนี้ ChatGPT ไม่สามารถทดแทนวิจารณญาณทางคลินิกของมนุษย์ได้ แต่ทำหน้าที่เป็นเพียงส่วนสนับสนุนข้อมูลเท่านั้น ความสำเร็จในการแก้ปัญหานั้นจึงไม่ได้ขึ้นอยู่กับความสามารถในการสืบค้นเพียงอย่างเดียว แต่ขึ้นอยู่กับความสามารถของผู้เรียนในการตีความและบูรณาการข้อมูลเหล่านั้นเข้ากับบริบทของผู้ป่วยจริง

การศึกษานี้แสดงให้เห็นว่าการใช้ ChatGPT หรือ AI สำหรับการศึกษาด้านการแพทย์ฉุกเฉินจำเป็นต้องออกแบบสถานการณ์ที่เหมาะสม การฝึกฝนรวบรวมประวัติผู้ป่วยที่ดีและทำการตรวจร่างกายในชีวิตจริง เพื่อให้ได้เฉพาะข้อมูลทางคลินิกที่สำคัญ (prompt Engineering) มีความสำคัญอย่างยิ่งต่อการใช้งานเทคโนโลยีปัญญาประดิษฐ์ให้เกิดประสิทธิผลสูงสุด และนักศึกษาแพทย์จำเป็นต้องมีรากฐาน clinical reasoning ที่ดี เพื่อใช้ในการประเมินและคัดกรองความน่าเชื่อถือของผลลัพธ์ที่ได้จากปัญญาประดิษฐ์ เริ่มตั้งแต่การประเมินผู้ป่วยด้วยตนเองอย่างเป็นระบบ การวิเคราะห์เพื่อสกัดเฉพาะประเด็นทางคลินิกที่สำคัญ

และการนำคำตอบที่ได้รับจาก AI มาสังเคราะห์ร่วมกับหลักฐานเชิงประจักษ์และเหตุผลทางพยาธิสรีรวิทยาเพื่อวางแผนการดูแลรักษาผู้ป่วยอย่างปลอดภัยที่สุด[5] อย่างไรก็ตามข้อจำกัดของการศึกษานี้เป็นการจำลองสถานการณ์ซึ่งไม่ใช่การประเมินการปฏิบัติจริงและขึ้นอยู่กับสิ่งที่นักศึกษาป้อนข้อมูลให้กับ ChatGPT ซึ่งอาจส่งผลต่อความถูกต้องของคำตอบและสะท้อนถึงทักษะในการเขียนคำสั่ง (prompting skills) ของนักศึกษาในแต่ละบุคคลมากกว่าที่จะสะท้อนความรู้ทางคลินิกที่แท้จริงของผู้เรียน ดังนั้นอาจารย์แพทย์อาจบรรจุหลักสูตร prompt engineering ให้กับนักศึกษาแพทย์ พิมพ์เฉพาะ key clinical features ป้อนให้ AI ไม่ใช่การ copy/paste ข้อมูลคนไข้ทั้งหมด[9] และบูรณาการปัญญาประดิษฐ์เข้ากับการศึกษาแพทยศาสตร์จึงไม่ควรหยุดอยู่เพียงการสอนทักษะการเขียนคำสั่ง แต่จะต้องมุ่งเน้นให้นักศึกษาสามารถอธิบายเหตุผลเบื้องหลังการตัดสินใจและรู้เท่าทันข้อจำกัดของเครื่องมือ เพื่อยกระดับให้ AI เป็นเครื่องมือที่ช่วยความคิดทางคลินิกมากกว่าจะเป็นสิ่งที่ผู้เรียนพึ่งพาโดยปราศจากการคิดวิเคราะห์อย่างรอบด้าน

การศึกษานี้นั้นมีข้อจำกัดอยู่หลายประการ ประการแรก คือขนาดของกลุ่มตัวอย่างที่มีจำนวนน้อยเพียง 55 คน ซึ่งอาจทำให้ผลการศึกษาวิเคราะห์ได้จำกัดในบางประเด็น เช่น ความรู้พื้นฐานของนักศึกษาแพทย์โดยรวม กลุ่มตัวอย่างเป็นนักศึกษาแพทย์จากโรงพยาบาลเดียว ซึ่งอาจจะยังไม่ครอบคลุมถึงบริบทและประสบการณ์ที่หลากหลายจากสถานศึกษาอื่น ประการที่สอง ข้อจำกัดทางด้านเครื่องมือและการประเมินผล กล่าวคือ แบบทดสอบสถานการณ์จำลองและแบบสอบถามความพึงพอใจที่ได้สร้างขึ้น แม้อ้างอิงจากมาตรฐานวิชาชีพและวรรณกรรมที่เกี่ยวข้อง แต่ยังไม่ได้ผ่านกระบวนการทดสอบความตรง (Validity) และความเที่ยง (Reliability) ทางสถิติอย่างเป็นทางการก่อนที่จะนำมาใช้ ประการที่สาม การตรวจให้คะแนนแบบทดสอบอัตนัยดำเนินการโดยแพทย์เฉพาะทางเพียง 1 ท่าน ซึ่งแม้จะใช้เกณฑ์มาตรฐาน (Checklist rubric) แต่อาจขาดการประเมินความสอดคล้องระหว่างผู้ประเมิน (Inter-rater reliability) ซึ่งควรได้รับการพัฒนาและปรับปรุงในการวิจัยครั้งต่อไป และประการที่สี่ การศึกษานี้เป็นการจำลองสถานการณ์ไม่ใช่การประเมินการปฏิบัติจริงและขึ้นอยู่กับสิ่งที่นักศึกษาป้อนข้อมูลให้กับ ChatGPT ซึ่งอาจส่งผลต่อความถูกต้องของคำตอบและอาจสะท้อนถึงทักษะที่ใช้ในการเขียนคำสั่ง (prompting skills) ของนักศึกษาแต่ละบุคคลมากกว่าที่จะสะท้อนถึงความรู้ทางคลินิกที่แท้จริงของผู้เรียน

ข้อเสนอแนะสำหรับการวิจัยในอนาคต การศึกษาควรขยายขอบเขตการวิจัยไปยังนักศึกษาแพทย์จากโรงพยาบาลหรือสถาบันการศึกษาอื่น ๆ เพื่อเพิ่มขนาดกลุ่มตัวอย่างที่มีให้ได้ผลลัพธ์ที่แม่นยำและมีความเชื่อถือมากขึ้น ใช้การออกแบบแบบ Randomized Controlled Trial เพื่อทดสอบประสิทธิผลการฝึก Medical Prompt Engineering และเพิ่มการวัดสมรรถนะทางด้าน Critical Appraisal of AI Output เพื่อให้ได้ภาพรวมของสมรรถนะที่ครบถ้วนยิ่งขึ้น

สรุปผลการศึกษา

การศึกษานี้แสดงให้เห็นว่านักศึกษาแพทย์ส่วนใหญ่ยังมีสมรรถนะในการสืบค้นข้อมูลเวชศาสตร์ฉุกเฉินผ่าน ChatGPT ต่ำกว่าเกณฑ์มาตรฐาน โดยปริมาณอย่างมีนัยสำคัญทางสถิติ กล่าวคือโจทย์ที่มีข้อมูลน้อยส่งผลให้คะแนนต่ำอย่างมีนัยสำคัญ ข้อมูลตั้งต้นในโจทย์เป็นปัจจัยสำคัญที่มีผลต่อสมรรถนะ ในขณะที่เมื่อข้อมูลถึงระดับปานกลางแล้ว การเพิ่มข้อมูลอีกไม่ได้ทำให้สมรรถนะดีขึ้น ปัจจัยทางด้านเพศ ชั้นปี ผลการเรียน และประสบการณ์การใช้ ChatGPT ไม่มีความสัมพันธ์อย่างมีนัยสำคัญกับสมรรถนะ

ผลประโยชน์ทับซ้อน: ไม่มี

แหล่งเงินทุนสนับสนุน: ไม่มี

เอกสารอ้างอิง

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
2. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications for Medical Education. JMIR Med Educ. 2023;9:e45312.
3. Strong E, DiGiammarino A, Weng Y, et al. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. JAMA Intern Med. 2023;183(11): 1228-30
4. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res. 2023;25:e48659.
5. Hosseini M, Gao CA, Liebovitz DM, et al. An Exploratory Study on the Performance of ChatGPT in Answering Clinical Questions: The Importance of Prompt Engineering. J Med Syst. 2023;47(1):116.
6. Ali A, et al. Artificial intelligence in medical education: a narrative review. Med Sci Forum. 2025;6(12):322.
7. Peralta Ramirez AA, Trujillo López S, Navarro Armendariz GA, De la Torre Othón SA, Sierra Cervantes MR, Medina Aguirre JA. Clinical simulation with ChatGpt: a revolution in medical education? J CME. 2025;14(1):2525615.
8. Roganović J. Familiarity with ChatGPT Features Modifies Expectations and Learning Outcomes of Dental Students. Int Dent J. 2024 Dec;74(6):1456-1462. doi: 10.1016/ j.identj.2024.04.012. Epub 2024 Apr 26. PMID: 38677973; PMCID: PMC11551562.
9. Knoth N, Tolzin A, Janson A, Leimeister JM. AI literacy and its implications for prompt engineering strategies. Comput Educ Artif Intell. 2024;6:100225. doi: 10.1016/ j.caeai.2024.100225
10. Kim Y, Jeong H, Chen S, Li SS, Park C, Lu M, et al. Medical hallucination in foundation models and their impact on healthcare. medRxiv [Preprint]. 2025.
11. Cai B, Wang X, Ratan U, Shah Y. Understanding the limitations of medical reasoning in large language models. In: Machine Learning for Health (ML4H) 2025 Symposium; 2025.
12. Vishwanath K, Agarwal R, Singhal K, Tu T, Azizi S, Natarajan V, et al. UltraMedical: Specialized Large Language Models for Medical Knowledge and Reasoning.

arXiv [Preprint]. 2025 [cited 2026 Apr 14]: arXiv:2504.01201.

13. Zheng Z, Wang Y, Huang Y, Song S, Yang M, Tang B, Xiong F, Li Z. Attention heads of large language models. Patterns (N Y). 2025 Feb 6;6(2):101176. doi: 10.1016/j.patter.2025.101176. PMID: 40041856; PMCID: PMC11873009.