

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 สามารถอธิบายคุณลักษณะของข้อสอบได้ละเอียดมากขึ้นนำไปใช้อย่างแพร่หลายในการวัดผลทางการศึกษาและการสอบมาตรฐานทั่วโลก สามารถอธิบายความสามารถแฝง (latent trait, θ) ของผู้เข้าสอบกับความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง โดยประมาณค่าคุณลักษณะของข้อสอบเป็นรายข้อได้อย่างอิสระจากกลุ่มตัวอย่างภายใต้เงื่อนไขของ model ทั้งยังให้ข้อมูลเชิงลึกของดัชนีความยาก ดัชนีอำนาจจำแนก และประสิทธิภาพตัวลอง 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** (unidimensionality) หมายถึง ข้อสอบทุกข้อนั้นวัดความสามารถหลักเดียวกัน **ความเป็นอิสระเฉพาะที่** (local independence) หมายถึง การตอบข้อสอบแต่ละข้อจะไม่ขึ้นต่อกัน และ**ความเป็นเอกภาพของความสัมพันธ์** (monotonicity) เมื่อระดับความสามารถของผู้เข้าสอบเพิ่มขึ้น ความน่าจะเป็นที่จะตอบถูกต้องจะไม่ลดลง แม้ว่าความสามารถแท้จริงของผู้เข้าสอบไม่สามารถวัดได้โดยตรง แต่ประมาณได้จากการตอบข้อสอบ คนเก่งมีโอกาสดตอบถูกเพิ่มขึ้นตาม model ทางคณิตศาสตร์ที่กำหนดไว้ใน CTT ความสามารถของผู้เข้าสอบจะดูจากคะแนนรวมตรงๆ ถ้าคะแนนเท่ากันก็เก่งเท่ากัน แต่ใน IRT จะมองลึกซึ้งไปกว่านั้นคือ ถ้าตอบข้อที่ยากได้ถูกก็จัดว่าเก่งกว่าคนที่ได้คะแนนรวมเท่ากัน แต่ตอบข้อยากไม่ถูกหรือทำถูกได้น้อยกว่า โปรแกรมจะทำการวิเคราะห์คะแนนกับความยากรายข้อโดยใช้เทคนิค maximum likelihood estimation หรือจะเป็น Bayesian estimation ก็ได้สรุปให้เข้าใจอย่างง่ายคือ latent trait เป็นการวิเคราะห์ย้อนกลับเริ่มจากคะแนน 0 แล้ว estimate คะแนนกับระดับความยากของข้อสอบในแต่ละข้อ โดยดู pattern ของการตอบถูกกับความยากของข้อสอบรายข้อ จากนั้น IRT จะทำการสร้าง **Item Characteristic Curve (ICC)** แสดงความสัมพันธ์ระหว่างระดับความสามารถบนแกนนอน (θ ซึ่งมีค่า $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติมักเห็นพิสัย -3 ถึง +3) กับโอกาสตอบถูกในแกนตั้งเป็นรายข้อ (item-level analysis) ข้อยากจะมีเส้นโค้งที่เลื่อนไปทางขวา ส่วน

ดัชนีการจำแนกดูที่ความชัน (slope) ของเส้นโค้งในบริเวณระดับความสามารถที่ต้องการ IRT model ซึ่งมีหลายระดับได้แก่ One-Parameter Logistic Model; 1PL หรือ Rasch model พิจารณาเฉพาะความยากของข้อสอบ 2PL เพิ่ม parameter ในการจำแนก และ 3PL เพิ่มโอกาสการเดาคำตอบในข้อสอบแบบเลือกตอบ

การแปลผลพารามิเตอร์และกราฟสำคัญใน IRT

1. ค่าความยาก (Difficulty parameter;

b) หมายถึง ระดับความสามารถ (θ) ที่ผู้เข้าสอบมีโอกาสดตอบข้อสอบถูก สะท้อน "ตำแหน่ง" ของข้อสอบบนแกนความสามารถ **b ต่ำ** (-2 ถึง -1) หมายถึงข้อสอบค่อนข้างง่าย เหมาะสำหรับผู้เรียนอ่อน **b ใกล้ศูนย์** หมายถึงความยากปานกลางเหมาะกับ ผู้เข้าสอบระดับเฉลี่ย 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 ต่ำมากหรือมีค่าเป็นลบ ควรตรวจสอบสาเหตุ เช่น โจทย์กำกวม เฉลยผิด หรือวัดคุณละมิตจากข้อสอบส่วนใหญ่

3. พารามิเตอร์การเดา (Guessing parameter; c)

ใช้ในข้อสอบแบบเลือกตอบ ด้วย 3PL เช่น ข้อสอบ 4 ตัวเลือก มีค่า c ประมาณ 0.20 - 0.25 หมายความว่า แม้ผู้เข้าสอบจะมีความสามารถต่ำมาก ก็ยังมีโอกาสตอบถูกร้อยละ 20 - 25 หากค่า c สูงผิดปกติ อาจสะท้อนว่าตัวลองไม่มีประสิทธิภาพ หรือมีค่าใบที่ช่วยให้เลือกคำตอบที่ถูกได้ง่าย

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

สามารถใช้ 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: lrt จะประมาณค่าพารามิเตอร์ a , b และ c เป็นรายข้อ รวมทั้งค่าความสามารถ (θ) ของผู้สอบรายคน

5. ประเมิน Model fit โดยตรวจสอบค่า parameter ผิดปกติหรือ ICC ไม่สอดคล้องกับข้อมูลจริง

6. แปลผลและนำไปใช้ร่วมกับทบทวนเนื้อหาข้อสอบ ไม่ควรตัดสินจากค่าทางสถิติเพียงอย่างเดียว ข้อสอบที่มีค่าอำนาจจำแนกต่ำ หรือค่า guessing สูง ผิดปกติ ควรตรวจสอบทั้งการเขียนโจทย์ ตัวลวง และความสอดคล้องกับวัตถุประสงค์การเรียนรู้ แม้ IRT จะให้ข้อมูลที่ละเอียดและมีความแม่นยำกว่า CTT แต่การวิเคราะห์จำเป็นต้องมีจำนวนผู้เข้าสอบที่เพียงพอและข้อมูลที่มีคุณภาพ

Stata commands

1. Import เลือก excel file นำเข้าโดยตรงจากคำสั่ง 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 (ภาพที่ 1)
`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 ในภาพนี้เห็นว่าข้อที่ง่ายที่สุดอยู่ทางซ้ายสุดซึ่งมีความชันมากที่สุดด้วยคนที่มีความสามารถในระดับกลางและคนเก่ง (θ ตั้งแต่ 0 ขึ้นไป) ทำได้ถูกหมดทุกคน

4. **irtgraph iif q1 - q5** Item Information Function (IIF) แสดงประสิทธิภาพหรือความแม่นยำรายข้อ ดังภาพที่ 3 จากกราฟนี้จะเห็นได้ว่า ข้อสอบทั้ง

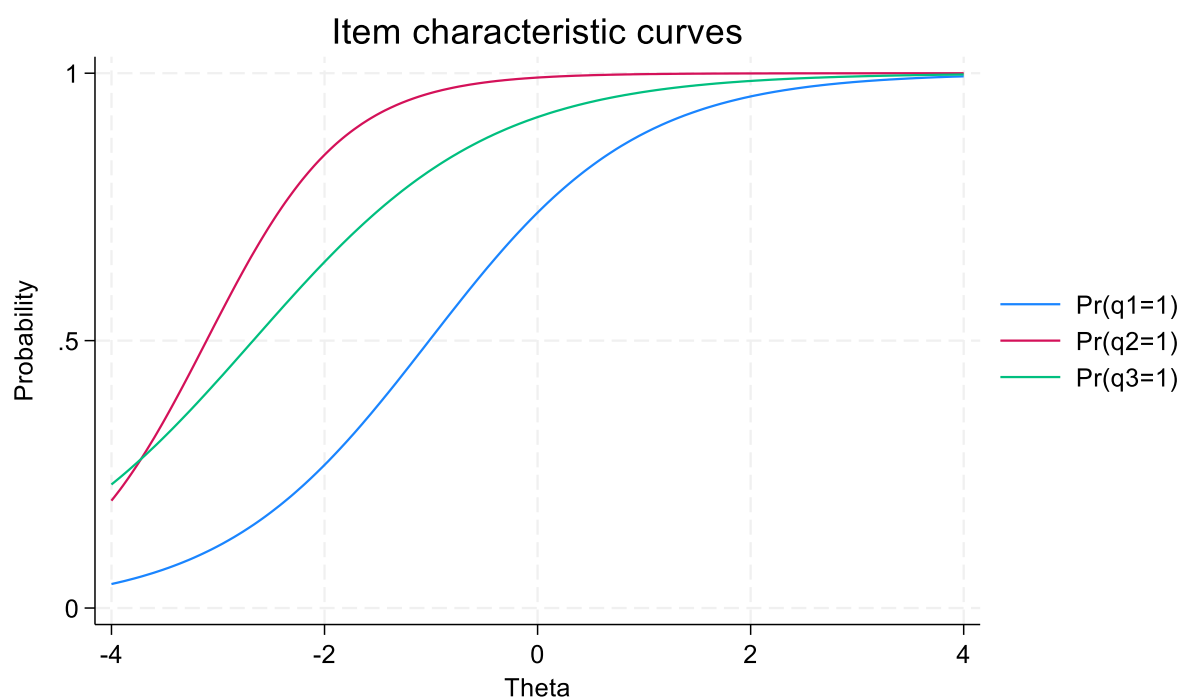
5 ข้อนั้นให้ information สูงสุดในช่วง θ -3 ถึง -1 และ information ลดลงอย่างรวดเร็วเมื่อ θ เข้าใกล้หรือมากกว่า 0 แสดงว่าทั้ง 5 ข้อนี้ให้ความแม่นยำในการวัดคนอ่อนที่ต่ำกว่าค่าเฉลี่ยเป็นหลัก

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

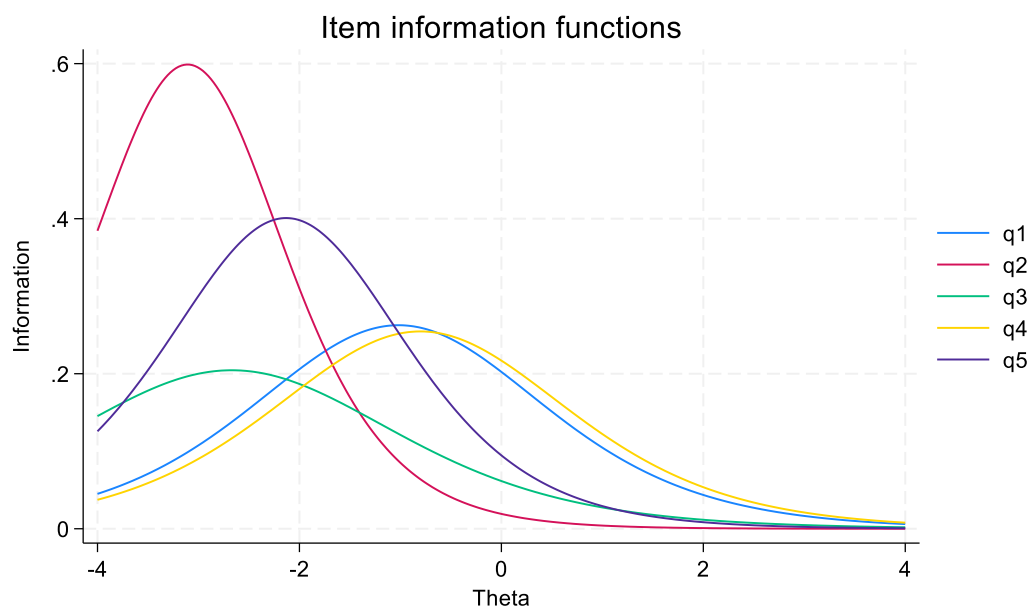
ภาพที่ 1 Stata output of irt

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
		-1.017409	.3523927	-2.89	0.004	-1.708086	-.3267319
q2	Discrim	1.547758	.6710877	2.31	0.021	.2324506	2.863066
		-3.107457	.7812146	-3.98	0.000	-4.638609	-1.576305
q3	Discrim	.9044026	.3207516	2.82	0.005	.275741	1.533064
		-2.6705	.814352	-3.28	0.001	-4.266601	-1.074399
q4	Discrim	1.009111	.2978247	3.39	0.001	.4253853	1.592837
		-.8027847	.323251	-2.48	0.013	-1.436345	-.1692243
q5	Discrim	1.266271	.3614336	3.50	0.000	.5578738	1.974668
		-2.131203	.4882656	-4.36	0.000	-3.088186	-1.17422

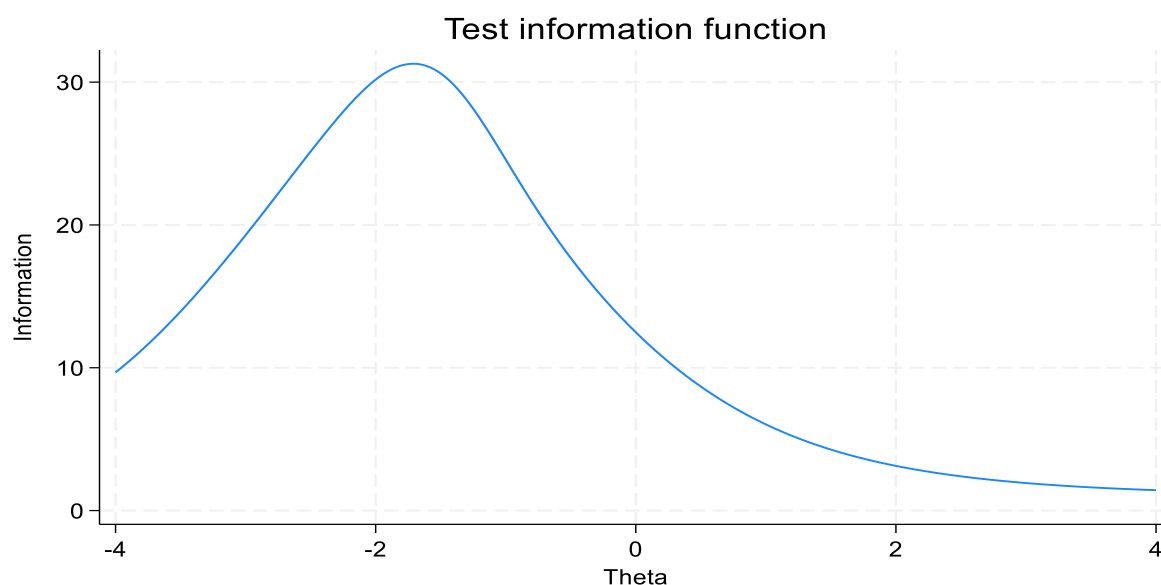
ภาพที่ 2 Item Characteristic Curve (ICC)



ภาพที่ 3 Item Information Function (IIF) ประสิทธิภาพของข้อสอบ



ภาพที่ 4 Test Information Function (TIF)



สรุปและคำแนะนำ

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

Disclosure

บทความนี้ใช้ AI ช่วยสืบค้นข้อมูลและร่างบทความ โดยทุกขั้นตอนถูกตรวจสอบความถูกต้องโดยผู้พิมพ์แล้ว

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://](https://www.academia.edu/30496631/Fundamentals_of_Item_Response_Theory)

www.academia.edu/30496631/Fundamentals_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/>