

การวิเคราะห์ข้อสอบ: เชื่อมโยงความเชื่อมั่น ความยาก และอำนาจจำแนก  
ให้ก่อเกิดความน่าเชื่อถือ

Analyzing the Interrelations of Reliability, Item Difficulty, and Item Discrimination in Test Construction

ผู้ช่วยศาสตราจารย์ ดร.ปรเมตต์ปัญญปรัชญ์ ต้องประสงค์

ดร.กิริติ วัชรสินธุ์

มหาวิทยาลัยสวนดุสิต

การสอบคัดเลือกเข้าศึกษาหรือเข้าทำงาน สามารถพิจารณาค่าความเชื่อมั่น (Reliability) ค่าความยาก (p) และค่าอำนาจจำแนก (r) จากการวิเคราะห์ข้อสอบ เพื่อให้มีความสำคัญต่อการประเมินคุณภาพของแบบทดสอบ กล่าวได้ว่าทั้งสามค่านี้มีความสัมพันธ์และส่งผลต่อประสิทธิภาพของการวัดผล การทดสอบ จึงสามารถสรุปสาระโดยสำคัญได้ดังนี้

### 1. ค่าความเชื่อมั่น (Reliability)

ค่าความเชื่อมั่น หมายถึง ความสามารถของเครื่องมือทดสอบที่ให้ผลได้อย่างสม่ำเสมอเมื่อทำการทดสอบซ้ำ ๆ การคำนวณค่าความเชื่อมั่นในแบบทดสอบสามารถทำได้โดยวิธีต่าง ๆ เช่น Cronbach's  $\alpha$  วิธี KR-20 วิธี test-retest วิธี split-half โดยการใช้เครื่องมือเหล่านี้ช่วยให้ผู้พัฒนาแบบทดสอบสามารถประเมินความเสถียรของผลลัพธ์ที่ได้จากการทดสอบซ้ำ ๆ ได้ นั่นคือค่า Cronbach's  $\alpha$  หรือ KR 20 (สำหรับข้อปรนัย) ที่มีค่าความเชื่อมั่น  $\geq 0.7$  ถือว่าใช้ได้  $\geq 0.8$  ดี

เมื่อวิเคราะห์ความเชื่อมโยงระหว่างค่า p และ r พบว่า ค่าความยาก (p) อยู่ในช่วงที่เหมาะสม คือ 0.20–0.80 จะช่วยให้สามารถแยกแยะผู้เรียนที่มีผลการเรียนดีและไม่ดีได้ ส่งผลให้ ความเชื่อมั่น (reliability) ของแบบทดสอบสูงขึ้น นอกจากนี้ ค่าอำนาจจำแนก (r) ที่สูงช่วยให้การแยกแยะผู้เรียนมีความแม่นยำ ทำให้ผลการทดสอบมีความเชื่อมั่นสูงขึ้น (Cohen, Swerdlik, & Sturman, 2021) ดังตาราง 1

ตาราง 1 การแปลความหมายค่า Cronbach's  $\alpha$

| ค่า Reliability ( $\alpha$ ) | การแปลความหมาย     | คำแนะนำในการใช้งาน                                                          |
|------------------------------|--------------------|-----------------------------------------------------------------------------|
| 0.90–1.00                    | สูงมาก (Excellent) | เหมาะสำหรับการประเมินที่มีผลกระทบสูง เช่น การคัดเลือกเข้าศึกษาหรือเข้าทำงาน |
| 0.80–0.89                    | สูง (Good)         | เหมาะกับการประเมินทั่วไปในระดับหลักสูตรหรือรายวิชา                          |

| ค่า Reliability ( $\alpha$ ) | การแปล<br>ความหมาย | คำแนะนำในการใช้งาน                         |
|------------------------------|--------------------|--------------------------------------------|
| 0.70–0.79                    | พอใช้ (Acceptable) | ใช้ได้เ็นบริบททั่วไป แต่ควรพิจารณาพัฒนาต่อ |
| 0.60–0.69                    | ต่ำ (Questionable) | ควรพิจารณาทบทวนเครื่องมือหรือวิธีประเมิน   |
| 0.00–0.59                    | ต่ำมาก (Poor)      | ไม่ควรใช้ในการตัดสินใจที่สำคัญ             |

ที่มา: Nunnally & Bernstein (1994); Cohen, Swerdlik, & Sturman (2013)

หรืออาจใช้ลักษณะอนุรักษนิยม โดยระบุ “ในการตีความเบื้องต้น”

| ค่า Reliability ( $\alpha$ ) | การแปลความหมาย       |
|------------------------------|----------------------|
| 0.91–1.00                    | สูงมาก (Excellent)   |
| 0.71–0.90                    | สูง (Good)           |
| 0.41–0.70                    | ปานกลาง (Acceptable) |
| 0.21–0.40                    | ต่ำ (Questionable)   |
| 0.00–0.20                    | ต่ำมาก (Poor)        |

ที่มา: George & Mallery (2003); DeVellis (2016)

## 2. ค่าความยาก (p)

ค่าความยากของข้อสอบ (Item Difficulty; p-value) คือสัดส่วนของผู้สอบที่สามารถตอบข้อสอบข้อนั้นได้ถูกต้อง ซึ่งเป็นตัวชี้วัดว่าแต่ละข้อมีระดับความง่ายหรือความยากมากน้อยเพียงใด หากค่า p สูง ใกล้ 1 หมายถึงข้อสอบง่ายเกินไป ในขณะที่ค่าต่ำใกล้ 0 หมายถึงข้อสอบยากเกินไป (Nitko & Brookhart, 2013)

ค่าความยาก (p) วัดจากสัดส่วนของผู้สอบที่ตอบได้ถูกต้องจากข้อสอบ ซึ่งสะท้อนถึงความง่ายหรือยากของข้อสอบ หาก p สูง (ใกล้ 1) หมายถึงข้อสอบง่าย แต่หาก p ต่ำ (ใกล้ 0) หมายถึงข้อสอบยากเกินไป

การเลือกใช้ข้อสอบที่มีค่า p อยู่ในช่วง .20–.80 จะช่วยให้สามารถแยกความสามารถของผู้สอบได้ดี ส่งผลให้ค่าความเชื่อมั่น (reliability) ของแบบทดสอบสูงขึ้น (Embretson & Reise, 2013; Kline, 2016) ทั้งนี้ ข้อสอบที่ง่ายหรือยากเกินไป จึงไม่สามารถจำแนกระดับของผู้สอบได้อย่างแม่นยำ (Crocker & Algina, 2006)

เมื่อวิเคราะห์ความเชื่อมโยงกับ Reliability และ  $r$  ค่า  $p$  ที่มีความยากที่เหมาะสมจะช่วยให้การทดสอบสามารถ จำแนกผู้สอบ ระดับต่าง ๆ ได้ดี ส่งผลให้  $r$  สูงขึ้น ซึ่งจะช่วยให้เพิ่มความแม่นยำและความน่าเชื่อถือของผลการทดสอบ (Embretson & Reise, 2013; Kline, 2016)

ตัวอย่างการคำนวณ ผู้ที่ตอบถูก จำนวน 45 คนจาก 50 คนจะได้  $45/50 \rightarrow p = 0.90$

เกณฑ์มาตรฐานในการใช้งาน ข้อสอบที่ดีควรมีค่า  $p$  ระหว่าง 0.20–0.80 เพื่อให้ทั้ง แยกผู้สอบได้ และไม่ยากเกินไป การตีความค่า  $p$  ดังตาราง 2

ตาราง 2 การแปลความหมายค่า  $p$

| ค่า $p$  | ความยากของข้อสอบ        | ข้อเสนอแนะ                            |
|----------|-------------------------|---------------------------------------|
| .00–.19  | ยากมาก (Very Difficult) | ควรปรับหรือยกเลิก                     |
| .20–.39  | ค่อนข้างยาก (Difficult) | ใช้ได้หากต้องการวัดผู้มีความสามารถสูง |
| .40–.60  | ปานกลาง (Moderate)      | เหมาะสมที่สุด                         |
| .61–.80  | ค่อนข้างง่าย (Easy)     | ใช้ได้เพื่อรักษาคุณภาพในข้อสอบ        |
| .81–1.00 | ง่ายมาก (Very Easy)     | ควรปรับหรือตัดออกจากชุดข้อสอบ         |

ที่มา: Crocker & Algina (2006)

### 3. ค่าอำนาจจำแนก ( $r$ )

ค่าอำนาจจำแนก ( $r$ ) หมายถึง ดัชนีที่ใช้วัดความสามารถของข้อสอบในการแยกแยะระหว่างผู้สอบที่มีผลการเรียนดีและไม่ดี โดยสามารถคำนวณได้จากการเปรียบเทียบคะแนนของผู้สอบกลุ่มที่มีผลคะแนนสูงกับกลุ่มที่มีผลคะแนนต่ำ

เมื่อวิเคราะห์ความเชื่อมโยงกับ Reliability และ  $p$  ค่า  $r$  ที่สูงช่วยให้การแยกแยะกลุ่มผู้เรียนทำได้ดี ทำให้ ค่าความเชื่อมั่น (reliability) เพิ่มขึ้น เพราะข้อสอบสามารถแยกแยะผู้เรียนได้ชัดเจนและแม่นยำ (Thompson, 2017)

เกณฑ์การแปลความค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ระหว่างคะแนนรายข้อกับคะแนนรวม เพื่อวัดความสามารถของข้อสอบในการจำแนกผู้สอบ (Item Discrimination Index) ดังตาราง 3

ตาราง 3 การแปลความหมายค่า  $r$

| ค่า $r$   | การแปลความหมาย    |
|-----------|-------------------|
| 0.60–1.00 | ดีมาก (Excellent) |
| 0.40–0.59 | ดี (Good)         |

|           |                                  |
|-----------|----------------------------------|
| 0.20–0.39 | พอใช้ (Fair)                     |
| 0.00–0.19 | ต่ำ ควรปรับปรุงหรือตัดออก (Poor) |

ที่มา: Crocker & Algina (2006); Nitko & Brookhart (2013); Kline (2016)

### การวิเคราะห์ข้อสอบอย่างมีประสิทธิภาพ

การพิจารณาค่าทั้งสามนี้ร่วมกันช่วยให้การวิเคราะห์ข้อสอบมีความ **น่าเชื่อถือ** และ **แม่นยำ** โดยสามารถสรุปได้ดังนี้

1. **ค่าความเชื่อมั่น (Reliability)** ช่วยประเมินความเสถียรของการทดสอบ การตรวจสอบค่าความเชื่อมั่นช่วยให้มั่นใจได้ว่าเครื่องมือทดสอบจะให้ผลที่คงที่ไม่เปลี่ยนแปลงจากการทดสอบซ้ำ ๆ

2. **ค่าความยาก (p)** ช่วยตรวจสอบว่าแต่ละข้อสอบอยู่ในระดับที่เหมาะสมไม่ยากหรือไม่ง่ายเกินไป เพื่อให้สามารถแยกแยะผู้สอบได้ดี

3. **ค่าอำนาจจำแนก (r)** ช่วยในการแยกแยะผู้สอบที่มีผลการเรียนดีจากผู้สอบที่มีผลการเรียนต่ำได้อย่างมีประสิทธิภาพ

การใช้ทั้งสามค่านี้ร่วมกันจะช่วยให้การออกแบบข้อสอบมีความแม่นยำและสามารถสะท้อนความสามารถที่แท้จริงของผู้สอบได้อย่างถูกต้อง จึงสรุปความสัมพันธ์ระหว่าง **p** ค่า **r** และ **ความเชื่อมั่น**

- ข้อสอบที่มี  $p$  อยู่ในช่วง 0.2–0.8 และ  $r$  มากกว่า 0.2 จึงมีคุณภาพดี อีกทั้งสามารถแยกผู้สอบและสร้างความเชื่อมั่นให้ทั้งฉบับ
- รูปแบบ KR-20 จะรวมค่า  $p$  และ  $q$  ( $1-p$ ) เข้าคำนวณความเชื่อมั่นของแบบทดสอบ
- สรุปเชิงประจักษ์ ดังตาราง 4

ตาราง 4 สรุปเกณฑ์เชิงประจักษ์

| ตัวชี้วัด                       | เกณฑ์มาตรฐาน      |
|---------------------------------|-------------------|
| $p$                             | 0.20–0.80         |
| $r$                             | $\geq 0.20$       |
| reliability ( $\alpha$ / KR-20) | $\geq 0.70$ –0.80 |

เมื่อดำเนินการจัดทำข้อสอบ (ออกข้อสอบ) ควรพิจารณาดังนี้

- พัฒนาและวิเคราะห์ **p** และ **r** รายข้อ

2. ตรวจสอบ  $\alpha$  หรือ KR-20 เพื่อประเมินความเชื่อมั่นของฉบับแบบทดสอบ
3. ปรับปรุงข้อที่ p หรือ r ต่ำ และตรวจสอบรูปแบบการเขียนเพื่อปรับปรุงคุณภาพเครื่องมือ

### รายการอ้างอิง

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2021). *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (9th ed.). McGraw-Hill Education.
- Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Cengage Learning.
- DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). Sage Publications.
- Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory for Psychologists*. Psychology Press.
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn & Bacon.
- Kline, T. J. B. (2016). *Psychometrics: An Introduction* (2nd ed.). SAGE Publications.
- Kuder, G. F., & Richardson, M. W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, 2(3), 151–160.
- Nitko, A. J., & Brookhart, S. M. (2013). *Educational assessment of students* (7th ed.). Pearson.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Thompson, B. (2017). *A Primer on the Validity of Psychological Tests: A Guide for Students and Practitioners*. Springer.