

ความสัมพันธ์ระหว่างคะแนนเฉลี่ยและความยากของข้อสอบแบบเลือกตอบ

ปัทมา ล้วนิช¹ วีรวรรณ เล็กสกุลไชย²

¹ภาควิชาเภสัชวิทยา คณะแพทยศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

²ภาควิชาพยาธิวิทยา คณะแพทยศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

Received: October 31, 2018

Revised: September 25, 2019

Accepted: November 26, 2019

บทคัดย่อ

โดยทั่วไปการตัดเกรดจะใช้วิธีอิงกลุ่มตามคะแนนที่ผู้เรียนทำได้ แต่บางครั้งเมื่อคะแนนที่ได้สูงหรือต่ำกว่าที่ผู้สอนคาดหวัง จะพิจารณาความง่ายของรายวิชาหรือความเก่งของผู้เรียนตามมุมมองของผู้สอนแต่ละคนร่วมด้วย ทำให้อาจเกิดปัญหาความเที่ยงตรงและความเชื่อมั่นในการตัดเกรด การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างคะแนนเฉลี่ย (mean) ที่ผู้เรียนทั้งกลุ่มทำได้ และดัชนีความยาก (difficulty index) หรือค่า p ของชุดข้อสอบแบบเลือกตอบ (MCQ) โดยทำการเก็บข้อมูลแบบย้อนกลับของข้อสอบ 2 รายวิชาที่กลุ่มผู้เรียนแตกต่างกัน คือ วิชา พพ201 และ ภส203 โดยข้อสอบทั้งหมดถูกตรวจด้วยเครื่องตรวจข้อสอบและวิเคราะห์คุณภาพข้อสอบด้วยโปรแกรมคอมพิวเตอร์ ผลการศึกษาพบว่า คะแนนเฉลี่ยที่แปลงเป็นร้อยละ (%) มีความสัมพันธ์อย่างสมบูรณ์ ($R = 1.000$) กับค่า p ของชุดข้อสอบที่มีค่าระหว่าง 0.0-1.0 โดยไม่ขึ้นอยู่กับคะแนนที่ผู้เรียนทำได้และสามารถเขียนเป็นสมการได้ว่า คะแนนเฉลี่ย (%) = ค่าดัชนีความยาก $\times 100$ ดังนั้น ในการจัดการเรียนการสอนรายวิชาใดๆ จึงสามารถใช้คะแนนเฉลี่ยที่กลุ่มผู้เรียนทำได้บอกถึงระดับความง่ายของรายวิชานั้น ในทำนองเดียวกันกับการใช้ค่าดัชนีความยากบอกถึงความยากของข้อสอบ

คำสำคัญ: ข้อสอบแบบเลือกตอบ ดัชนีความยาก คะแนนเฉลี่ย

ผู้พิมพ์และประสานงาน:

ปัทมา ล้วนิช

ภาควิชาเภสัชวิทยา คณะแพทยศาสตร์

มหาวิทยาลัยศรีนครินทรวิโรฒ

114 ถนนสุขุมวิท ซอยสุขุมวิท 23

แขวงคลองเตยเหนือ เขตวัฒนา กรุงเทพฯ 10110

อีเมล: pathama@gs.wu.ac.th

The relationship between mean and difficulty index of multiple choice questions test

Pathama Leewanich¹, Veeravan Lekskulchai²

¹Department of Pharmacology, Faculty of Medicine, Srinakharinwirot University

²Department of Pathology, Faculty of Medicine, Srinakharinwirot University

Abstract

In general, we decide students grades based on the average scores that students have done. When the average score is sometimes higher or lower than the instructors expecting, grades will be also based on the difficulty of the course or the skill of the students in the perspective of individual instructor. This may cause a problem of the accuracy and reliability in grading. This study aims to determine the relationship between average scores obtained by groups of students or mean values and difficulty index or p values of multiple choice question (MCQ) test. A retrospective study has been done by collecting data from 2 different subjects: PR201 and PM 203. The test items were examined and analyzed using a computer program. It was found that the percentage of means and the p-value that has a value between 0.00-1.00 of the tests showed absolutely and statistically significant relationship ($R = 1.000$) by mean (%) = $p \times 100$. Therefore, the mean value of the total scores can be used to indicate the difficulty of the course. Similarity, p-value can be used to indicate the difficulty of the test.

Keywords: MCQ test, difficulty index, mean

Corresponding Author:

Pathama Leewanich

Department of Pharmacology, Faculty of Medicine,
Srinakharinwirot University

114 Sukhumvit Road, Sukhumvit 23,
Wattana, Bangkok 10110 Thailand

E-mail: pathama@g.swu.ac.th

บทนำ

การประเมินความรู้ความสามารถของผู้เรียนในแต่ละรายวิชาที่เป็นวิทยาศาสตร์การแพทย์พื้นฐาน ในระดับปริคณิการของคณะแพทยศาสตร์มหาวิทยาลัยศรีนครินทรวิโรฒ แบ่งออกเป็น 2 ส่วน คือ ด้านความรู้ (knowledge) และด้านพฤติกรรม (performance) การประเมินด้านความรู้ใช้ข้อสอบแบบเลือกตอบ (multiple choice questions, MCQ) การประเมินด้านพฤติกรรมมีรูปแบบที่หลากหลายขึ้นอยู่กับธรรมชาติของแต่ละรายวิชา เมื่อมีการตัดเกรดจะนำคะแนนทั้งหมดมารวมกันตามสัดส่วนที่กำหนดด้วยวิธีการแบบอิงกลุ่ม และใช้ค่าเบี่ยงเบนมาตรฐานกำหนดช่วงของคะแนนแต่ละเกรดซึ่งเป็นวิธีที่เหมาะสมสำหรับวิชาที่มีผู้เรียนจำนวนมากและมีการประเมินด้วยข้อสอบแบบเลือกตอบ^{1,2} ถึงแม้จะมีทั้งข้อดีและข้อเสีย^{2,3,4} ปัญหาที่พบคือ บางรายวิชาผู้เรียนไม่สามารถทำคะแนนได้ตามที่ผู้สอนคาดหวัง ซึ่งอาจเกิดจากข้อสอบหรือวิธีการวัดผลยากเกินไป หรือผู้เรียนเป็นกลุ่มที่เรียนไม่ค่อยเก่ง การตัดเกรดก็จะนำเหตุผลเหล่านี้มาพิจารณาด้วย การระบุว่าวิชาใดวิชาหนึ่งง่ายหรือยากนั้นผู้สอนมักเป็นผู้กำหนด เป็นการใช้อำนาจดุลยพินิจและอาจทำให้การพิจารณาเกิดเกิดความคลาดเคลื่อน หรือไม่เที่ยงตรงได้ ถ้าสามารถกำหนดความยากง่ายของข้อสอบที่อยู่บนพื้นฐานของข้อมูลทางสถิติจะทำให้การพิจารณาเกรดมีความแม่นยำและเที่ยงตรงมากขึ้น ค่า p เป็นสัญลักษณ์ของค่าดัชนีความยาก (difficulty index) ของข้อสอบและชุดข้อสอบที่ได้จากการตรวจข้อสอบด้วยเครื่องตรวจและทำการวิเคราะห์ข้อสอบด้วยโปรแกรมคอมพิวเตอร์ เป็นค่าที่บ่งชี้ถึงคุณภาพของข้อสอบค่าหนึ่ง มีที่มาจากสูตร $p = (H + L)/N$ เมื่อ N คือ จำนวนผู้เรียนทั้งหมด H และ L คือ จำนวนผู้เรียนที่ตอบถูกในกลุ่มคนเก่งและคนอ่อนตามลำดับ^{5,6} แสดงถึงสัดส่วนระหว่างจำนวนคนที่ตอบข้อสอบข้อนั้นถูกกับจำนวนคนที่ตอบข้อนั้นทั้งหมด ดัชนีความยากจะมีค่าอยู่ระหว่าง 0 ถึง 1 ถ้า

มีค่า 0 หมายความว่าไม่มีใครตอบข้อนั้นถูกเลย ถ้ามีค่า 1 หมายความว่าทุกคนตอบข้อนั้นถูก ดัชนีความยากที่ยอมรับได้ควรอยู่ระหว่าง 0.2- 0.8^{7,8} ถ้าดัชนีความยากน้อยกว่า 0.2 หรือมากกว่า 0.8 จัดเป็นข้อสอบที่ยากหรือง่ายเกินไปตามลำดับ เป็นข้อสอบที่ต้องปรับปรุง ด้วยเหตุนี้คณะผู้วิจัยจึงต้องการหาความสัมพันธ์ระหว่างคะแนนเฉลี่ย (mean) และดัชนีความยากของชุดข้อสอบหนึ่งๆ โดยคาดหวังว่าผลที่ได้จากการศึกษาครั้งนี้จะสามารถนำไปประยุกต์ใช้ในการตัดเกรด โดยใช้คะแนนเฉลี่ยของคะแนนทั้งหมดที่ผู้เรียนทำได้ในรายวิชาใดวิชาหนึ่งเป็นตัวบ่งชี้ความยากง่ายในการทำคะแนนของรายวิชานั้น แทนการใช้ความคิดเห็นของผู้สอนเพียงอย่างเดียว เพื่อป้องกันอคติของผู้สอน

วัตถุประสงค์

เพื่อศึกษาความสัมพันธ์ระหว่างคะแนนเฉลี่ย และดัชนีความยากของข้อสอบ โดยคาดหวังว่าจะสามารถนำคะแนนเฉลี่ย มาใช้บอกระดับความยากง่ายของรายวิชาในลักษณะเดียวกันกับการใช้ค่า p บอกความยากง่ายของชุดข้อสอบ

วิธีการศึกษา

1) เลือกศึกษารายวิชา พพ201 และ ภส 203 เพราะเป็นรายวิชาที่ผู้เรียนเป็นคนละกลุ่มสาขา และระดับคะแนนที่ทำได้มีความแตกต่างกัน ภายใต้วัตถุประสงค์ที่ต้องการศึกษาความสัมพันธ์ระหว่างคะแนนเฉลี่ย และดัชนีความยากของข้อสอบ คณะผู้วิจัยต้องการศึกษาว่าความสัมพันธ์นั้นขึ้นอยู่กับระดับคะแนนสูงต่ำที่ผู้เรียนทำได้หรือไม่

2) เนื่องจากรายวิชา พพ201 มี 1 หน่วยกิต ทำการสอบด้วยข้อสอบแบบเลือกตอบเพียง 1 ครั้ง และรายวิชา ภส203 มี 3 หน่วยกิต ทำการสอบด้วยข้อสอบแบบเลือกตอบ 3 ครั้ง เพื่อให้ได้จำนวนข้อมูลที่มากพอจึงใช้ข้อมูลย้อนหลัง 3 ปีการศึกษา (ปีการศึกษา 2556-2558) และ 2 ปีการศึกษา (ปีการศึกษา

2556-2557) สำหรับวิชา พพ201 และ ภส203 ตามลำดับ จำนวนข้อมูลที่เกิดขึ้นในการศึกษาครั้งนี้จึงประกอบด้วยข้อมูลของข้อสอบรายวิชา พพ201 จำนวน 3 ชุด และรายวิชา ภส203 จำนวน 6 ชุด

3) ทำการรวบรวมข้อมูลผลการตรวจและวิเคราะห์ข้อสอบในการสอบแต่ละครั้งของรายวิชา พพ201 และ ภส203 ที่ตรวจด้วยเครื่องตรวจข้อสอบ OMR ยี่ห้อ SCANTRON รุ่น OpScan 4ES และโปรแกรมวิเคราะห์ข้อสอบ Scan Tool Plus ตามทฤษฎีการทดสอบแบบดั้งเดิม (classical test theory) โดยทำการวิเคราะห์ข้อสอบเป็นรายข้อ (item analysis) และการวิเคราะห์ข้อสอบโดยรวมทั้งฉบับ (test analysis)⁹

4) เนื่องจากคะแนนเฉลี่ยที่ได้จากการตรวจด้วยเครื่องเป็นคะแนนเฉลี่ยที่ผู้เข้าสอบทำได้จากการสอบแต่ละครั้งที่มีการจำนวนคำถามไม่เท่ากัน จึงแปลงคะแนนเฉลี่ยให้อยู่ในรูปร้อยละ (%)

5) เปรียบเทียบความแตกต่างของค่าดัชนีความยากของชุดข้อสอบของ 2 รายวิชาโดยใช้ t-test ชนิด two-sample unequal variance และหาความสัมพันธ์ระหว่างคะแนนเฉลี่ยและดัชนีความยากของข้อสอบแบบ regression ด้วยโปรแกรม Microsoft Excel 2013 บน Window version 10

ผลการศึกษา

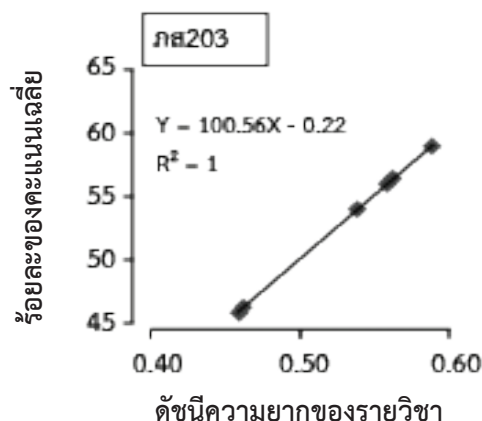
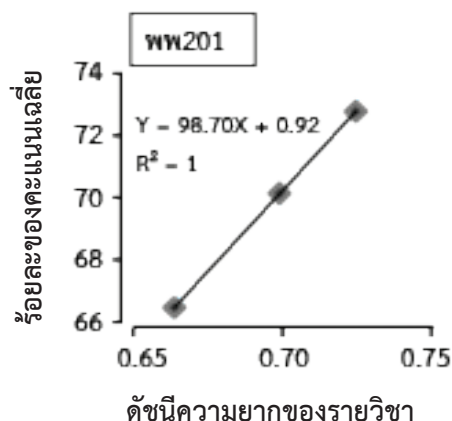
1. ความสัมพันธ์ระหว่างคะแนนเฉลี่ยและดัชนีความยากของชุดข้อสอบ

หลังจากรวบรวมข้อมูลคะแนนเฉลี่ยและค่าดัชนีความยากในการสอบแต่ละครั้งของแต่ละรายวิชา และได้เปลี่ยนคะแนนเฉลี่ยที่นิสิตทำได้เป็นค่าร้อยละ (ตารางที่ 1) พบว่าคะแนนเฉลี่ยมีความสัมพันธ์กับค่าดัชนีความยากของชุดข้อสอบด้วยค่า $R^2 = 1$ (รูปที่ 1) ทั้ง 2 รายวิชา

ตารางที่ 1 สถิติของการสอบแต่ละรายวิชา

รายวิชา	ปีการศึกษา	ครั้งที่สอบ	จำนวนผู้สอบ	จำนวนข้อสอบ	คะแนนเฉลี่ย	ร้อยละของคะแนนเฉลี่ย	ดัชนีความยาก
พพ201 (PR201)	2556	1	160	64	46.6	72.8	0.728
	2557	1	184	60	39.9	66.5	0.664
	2558	1	180	60	42.1	70.1	0.701
					Mean	69.8	0.698*
					SD	2.59	0.030
ภส203 (PM203)	2556	1	102	75	42.1	56.1	0.560
	2556	2	102	60	27.8	46.3	0.463
	2556	3	102	75	40.5	54.1	0.540
	2557	1	83	60	35.4	59.0	0.589
	2557	2	83	75	34.4	45.9	0.458
	2557	3	83	75	42.4	56.5	0.564
					Mean	53.0	0.529
					SD	5.56	0.060

* $p = 0.005$ เปรียบเทียบกับรายวิชา ภส203 (PM203)



รูปที่ 1 ความสัมพันธ์ระหว่างร้อยละของคะแนนเฉลี่ยและดัชนีความยากของรายวิชา PPW201 และ RS203 ตามลำดับ

2. เปรียบเทียบความยากของรายวิชา

ดัชนีความยากของชุดข้อสอบรายวิชา PPW201 เท่ากับ 0.698 และ รายวิชา RS203 เท่ากับ 0.529 (ตารางที่ 1) เมื่อเปรียบเทียบความยากของข้อสอบทั้ง 2 รายวิชา พบว่ามีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p < 0.005$)

อภิปรายผล

การศึกษาค้นคว้าครั้งนี้แสดงให้เห็นว่าร้อยละของคะแนนเฉลี่ยมีความสัมพันธ์อย่างสมบูรณ์กับค่าดัชนีความยากของชุดข้อสอบ โดยไม่ขึ้นกับคะแนนที่ผู้เรียนทำได้ ซึ่งความสัมพันธ์นี้ไม่มีใครเคยรายงานมาก่อน อาจเนื่องจากการตรวจสอบด้วยเครื่องตรวจจะแสดงผลเป็นคะแนนที่ผู้เข้าสอบแต่ละคนทำได้ และคิดค่าสถิติคะแนนเฉลี่ยของการสอบในรูปของ “คะแนนดิบ” ไม่ใช่ร้อยละ แต่ค่าดัชนีความยากของข้อสอบเป็นการแสดงสัดส่วนของจำนวน “คน” ที่ตอบถูกในข้อนั้นๆ มีค่าระหว่าง 0.0-1.0 ถ้าดัชนี

ความยากมีค่าเท่ากับ 1.0 หมายความว่า ผู้เข้าสอบตอบข้อนั้นถูกทุกคน ดัชนีความยากของชุดข้อสอบเป็นค่าเฉลี่ยดัชนีความยากของข้อสอบแต่ละข้อในชุดข้อสอบนั้น เมื่อแปลง “คะแนนดิบ” ของคะแนนเฉลี่ยที่ผู้เข้าสอบทั้งหมดทำได้เป็นร้อยละก็จะได้เห็นความสัมพันธ์ระหว่างคะแนนเฉลี่ยและค่าดัชนีความยากของชุดข้อสอบชัดเจน เช่น ข้อสอบชุดหนึ่งมีข้อสอบทั้งหมด 100 ข้อ มีผู้เข้าสอบทั้งหมด 100 คน ทำคะแนนเฉลี่ยได้ 70 คะแนน จะได้ค่าดัชนีความยากของชุดข้อสอบนี้เท่ากับ 0.7 หมายความว่าผู้เข้าสอบแต่ละคนตอบถูกเฉลี่ย 70 ข้อ และข้อสอบแต่ละข้อมีคนเลือกข้อถูกเฉลี่ย 70 คน จะเห็นว่าค่าทั้งสองมีที่มาต่างกันแต่มีความสัมพันธ์ไปในทิศทางเดียวกัน ดังนั้น เมื่อเปรียบเทียบกับเกณฑ์การแปลความหมายค่าดัชนีความยากของข้อสอบ¹⁰ จะสามารถใช้ค่าเฉลี่ยของคะแนนทั้งหมดที่ผู้เรียนทำได้ บอกระดับความยากของรายวิชาได้ดังนี้

ดัชนีความยาก (p)	ร้อยละของ คะแนนเฉลี่ย	ความหมาย
0.80 – 1.00	80 – 100	ง่ายมาก
0.60 – 0.79	60 – 79	ค่อนข้างง่าย
0.40 – 0.59	40 – 59	ยากปานกลาง
0.20 – 0.39	20 – 39	ค่อนข้างยาก
0.00 – 0.19	0 – 19	ยากมาก

สิ่งที่ได้จากการศึกษาครั้งนี้คือ ถ้าใช้คะแนนเฉลี่ยที่ผู้เรียนทั้งกลุ่มทำได้เป็นตัวบ่งชี้ความยากง่ายของรายวิชา แทนการใช้จรรยาบรรณของผู้สอนเพียงอย่างเดียว จะทำให้การตัดเกรดแบบอิงกลุ่มมีความเที่ยงตรงและน่าเชื่อถือยิ่งขึ้น

สรุปผล

คะแนนรวมทั้งหมดของรายวิชาใดๆ เมื่อนำมาหาค่าเฉลี่ยในรูปร้อยละ จะสะท้อนถึงความง่ายหรือยากในการเรียนวิชานั้นได้โดยตรง เหมือนกับการแปลความหมายของค่าดัชนีความยากที่ได้จากการวิเคราะห์ชุดข้อสอบมีปัจจัยหลายอย่างเข้ามาเกี่ยวข้องกับค่าดัชนีความยาก ซึ่งจะต้องทำการศึกษาต่อไป

References

1. Chan WS. A better norm-referenced grading using the standard deviation criterion. Teach Learn Med [Internet]. 2014 Oct [cited 2017 Aug 11];26(4):364-365. Available from: <https://www.tandfonline.com/doi/full/10.1080/10401334.2014.945031?scroll=top&needAccess=true>
2. The glossary of education reform. [Internet]. Portland: Great Schools Partnership; 2015. Norm-referenced test; [updated 2015 Jul 22; cited 2017 Aug 11]; [about 5 screens]. Available from: <https://www.edglossary.org/norm-referenced-test/>.
3. Hussain S, Tadesse T. Norm-referenced and criterion-referenced test in EFL classroom. Int J Soc Sci Humanit Invent [Internet]. 2015 Oct [cited 2017 Sept 19];4(10):24-30. Available from: [http://www.ijhssi.org/papers/v4\(10\)/Version-3/D04103024030.pdf](http://www.ijhssi.org/papers/v4(10)/Version-3/D04103024030.pdf).
4. Hall M. To curve or not to curve revisited. 2016 Sept 16 [cited 2018 Aug12]. In: The innovative instructor blog [Internet]. Baltimore: Johns Hopkins University Teaching and Learning Centers. 2012-[about 2 screens]. Available from: <http://ii.library.jhu.edu/tag/norm-referenced-grading/>.
5. Boopathiraj C, Chellamani K. Analysis of test items on difficulty level and discrimination index in the test for research in education. Int J Soc Sci Int Res [Internet]. 2013 Feb [cited 2017 Aug 11];2():189-93. Available from: <http://indianresearchjournals.com/pdf/IJSSIR/2013/February/15.pdf>.

6. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. J Pak Med Assoc [Internet]. 2012 Feb [cited 2017 Aug 11];62(2):142-7. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/22755376>.
7. Nauman M. Factors affecting difficulty & discrimination Indices. 2016 Sept 19 [cited 2018 Aug 12]. In: drnaumanshad [Internet]. Unknown: A great WordPress. com. 2016-[about 3 screens]. Available from: <https://drnaumanshad.wordpress.com/2016/09/19/factors-affecting-difficulty-discrimination-indices/>.
8. Mitra N K, Nagaraja H S, Ponnudurai G, et al. The Levels Of Difficulty And Discrimination Indices In Type A Multiple Choice Questions Of Pre-clinical Semester 1 Multidisciplinary Summative Tests. IeJSME 2009;3(1):2-7.
9. Shultz KS, Whitney DJ. Classical Test Theory Item Analysis. [Internet]. Thousand Oaks: SAGE Publications, Inc. 2005 [cited 2015 Aug 2]. p. 191-214. Available from: <http://sk.sagepub.com/books/measurement-theory-in-action/n12.xml>.
10. Lord, F.M. The Relationship of the reliability of multiple-choice test to the distribution of item difficulties. Psychometrika 1952;18:181-94.