2020

Relationship Between Formative Test Results and Final Exam Performance in First Year Medical School Course

Grace Owens
Wright State University - Main Campus, owens.211@wright.edu

Relationship Between Formative Test Results and Final Exam Performance in First Year Medical School Course

Grace Owens

Dr. Irina Overman, Assistant Professor of Geriatrics & Internal Medicine

Medical Education Research

Scholarship in Medicine Final Report

☒ By checking this box, I indicate that my mentor has read and reviewed my draft proposal prior to submission

Abstract

Objective: The purpose of the study was to evaluate whether formative testing from iRAT and MCQ data was predictive of final exam scores for the Staying Alive course at WSU BSOM. *Methods:* Data was collected from two consecutive classes of first-year medical students (n=234). Data included students' formative quiz scores (iRAT), formative exam scores (MCQ), final exam scores (NBME), race, and gender. Three regression models were created to analyze the relationship between formative and final scores. *Results:* The average iRAT score was not a significant predictor of NBME score. 53-56% of the variability in NBME score was attributed to iRAT, MCQ, race, and gender. However, the models lacked the accuracy to predict a score within one letter grade of the actual score. MCQ 5 and MCQ 1 were the strongest predictors of NBME score.

Key Words: frequent formative testing, linear regression model, medical student education

Introduction/Literature Review

In the first two years of medical school, students are asked to learn and remember a vast amount of content. Successful retention and assimilation of that knowledge is necessary for on-going success. Many methods have been used to predict student success in medical school and beyond. A study of medical students of Jefferson Medical College analyzed the correlation between Medical College Admission Test (MCAT) scores and performance in medical school, during residency, and on licensing examinations.[1] The study found that three previous versions of the MCAT predicted Step 1 scores with a validity coefficient in the mid 0.40s.[1] Others have studied whether academic performance during the first year of medical school can predict later performance on United States Medical Licensing Examination (USMLE) Step 1, and clinical abilities such as objective structured clinical exam performance.[2] The study found that low academic performance, measured by number of appearances in the bottom quartile of exam scores during the first year of medical school, is a meaningful risk factor for predicting low performance later in medical school.[2] West and colleagues focused on the effect of study skills on academic achievement during medical school. They found that time management and self-testing were generally stronger predictors of first-semester academic performance than aptitude.[3]

As mentioned by West, strong study skills are an asset to success. Frequent formative testing, a low-stakes testing strategy that occurs throughout the course, has been found as one successful method to enhance learning by improving motivation, study strategies, and spacing out study efforts.[4] Use of frequent formative testing such as weekly assessments and practice exams has a significant relationship with final exam performance.[5] Students taking cumulative assessments over the duration of the course spent more time studying than students who only

took an end-of-term assessment.[6] For medical students in an anatomy class, participation in frequent formative testing correlated to summative exam scores.[7]

A study of undergraduate students in low and high level biology courses demonstrated that frequent formative testing (multiple choice quizzes after every lecture) correlated with performance on the final exam. The study also showed that students with higher quiz scores were more likely to pass the final exam than those students with lower quiz scores. The authors advocated that students use the regression models developed in the study to predict course score in order to help students to self-motivate, adapt learning strategies, or seek additional resources, as needed.[8]

The benefits of frequent formative testing have not been adequately studied in a medical school setting. Formative testing applied to medical school courses has the potential to increase students' performance. Use of formative testing throughout the course can also help students identify gaps in learning, and low scores can act as a signal for students who may need additional support. This study evaluates whether the frequent formative testing in the educational model used in a first-year, 13-week, systems-based course at Wright State University Boonshoft School of Medicine (WSU BSOM) can also be used to predict final exam performance in the course.

Research Questions

Do quiz (iRAT) and formative exam (MCQ) results predict performance on course final examination (NMBE) for first year medical students at WSU BSOM in Staying Alive course?

Which MCQs are more associated with NBME score in the Staying Alive course?

Methods

Context/Protocol

The data reported here come from the performance of two consecutive years of first-year medical students' scores in a second semester course. The data was collected in the spring of 2018 and 2019 for the Classes of 2021 and 2022, respectively, at WSU BSOM. This study was deemed exempt by Wright State University's Institutional Review Board.

The course, titled Staying Alive, had a focus on physiology, pharmacology, and pathology of the renal, cardiovascular, and pulmonary systems. The primary teaching method for the 13-week course was Peer Instruction. Peer Instruction is an active learning strategy where students prepared for class by reviewing and studying the assigned reading. During class students answered multiple choice questions individually and then in small groups while taking notes on the content.

Seven times throughout the course, class time consisted of Team Based Learning (TBL) instead. TBL is an active learning strategy with three components. Students prepare for class by completing the assigned reading beforehand. Then each student takes an individual quiz of 10-15 multiple choice questions to demonstrate knowledge of the material – an individual Readiness Assurance Test (iRAT). Students then repeat the same question quiz in small groups (gRAT). The final component is application of concepts from the assigned reading by focused problem-solving exercises led by the professor and solved during class time in the same small groups. The class of 2021 had 7 TBL sessions and the class of 2022 had 6 sessions during the course.

Over the length of the course, students also took 5 formative multiple-choice question exams (MCQ). Each exam was 50 questions long, and students were given 60 minutes to complete the exam individually. Similar to a TBL, the students then completed the same 50 question exam together in small groups.

A passing score of at least 70% was required for each student to be eligible to sit for the final exam. Cumulatively, the PIs, TBLs, and a Problem-Based Learning component of the class (not further discussed in this paper) accounted for 30% of the points possible to earn toward a 70% overall passing score. The five MCQ exams accounted for the other 70%.

The final exam consisted of 150 questions from the (NBME) data bank selected by faculty to comprehensively cover the material taught throughout the course. Though the assessments (PI, TBL, MCQ, NBME final) were not identical between the two versions of the course, they were similar in content and difficulty.

Data Collection

Data consisted of student performance during the Staying Alive course for the Class of 2021 and Class of 2022 at WSU BSOM. Two students from the original Class of 2022 were excluded because they did not take the Staying Alive course. Nine students were excluded from analyses including race because they did not self-report their race. Data included individual scores for the 7 iRATs, 5 MCQs, and NBME Final Exam. Raw scores were converted to percentages. Demographic information on gender and race was also included.

Data Analysis

The Classes of 2023 and 2024 were analyzed together as a single cohort. Prior to the analysis, each iRAT, MCQ, and the NMBE were plotted as histograms and found to have an approximately normal distribution. A One-way analysis of variance (ANOVA) was used to compare NBME scores by gender and race. Bonferroni multiple comparisons tests were used for post hoc comparisons of the NBME scores among the four race categories. A multiple linear regression model was created with the independent variables of average iRAT and average MCQ scores to predict NBME score. A second model was created that included the average iRAT and

MCQ scores, gender, and race. Then a stepwise multiple linear regression was performed with the five individual MCQ scores, race, and gender as the independent variables predicting NBME scores, to determine whether different individual MCQ exams were greater predictors than others.

Results

A total of 234 students were included in the analysis from the Classes of 2021 and 2022 at Wright State University Boonshoft School of Medicine. A summary of the student participants is given in Table 1. The average Staying Alive iRAT score was 67.1%, the average Staying Alive MCQ score was 76.1%, and the average Staying Alive NBME score was 79.1%, see Table 2.

Table 3 shows the comparisons of NBME scores by gender and race. Male students scored significantly higher than female students ($F_{1, 232} = 23.3$, $p < 0.001$), and White, Non-Hispanic students scored significantly higher than Black, Non-Hispanic students ($F_{3, 221} = 4.31$, $p = 0.006$; Bonferroni $p < 0.05$). No other differences among the race categories were observed.

The regression model with the iRAT average and MCQ average as predictors for NBME score showed that iRAT average was not statistically significant ($p = 0.122$) as a predictor of NBME score, although the overall model was significant ($F_{2, 231} = 133.5$, $p < 0.001$). After controlling for iRAT average, a 1 point increase in MCQ average score resulted in a 0.762 point increase in NBME score (Table 4, model 1). The differences between the observed and predicted NBME scores in this model ranged from -18.7 points to 16.1 points. Adding gender and race to the model increased the adjusted R^2 by 0.026 (Table 4, model 2, $F_{6, 218} = 1555.9$, $p < 0.001$).

The stepwise regression model of individual MCQ scores, gender, and race against NBME score showed that the most significant predictors of NBME scores were MCQ 5 followed by MCQ 1 and Gender; all three together accounted for an adjusted R^2 of 0.51. With all seven variables, this model had an adjusted R^2 of 0.56. (Table 4, model 3, $F9, 215 = 1055.6$, $p < 0.001$). MCQ 4, Asian race, and Other race were not significant predictors of NBME score.

Discussion

The purpose of the study was to evaluate whether formative testing from iRAT and MCQ data was predictive of final exam scores for the Staying Alive course at WSU BSOM. iRAT scores did not significantly predict NBME score. iRAT scores only account for a small portion of the grade needed to reach a 70% to sit for the final exam. In addition, iRATs are usually only 10-15 questions in length. For these reasons, students may not take studying for an iRAT as seriously as studying for an MCQ, which accounts for a larger portion of the grade needed to sit for the final exam.

Race and gender were not the test variables of interest in the study, but they were included because they were found to have a significant effect on the outcome variable, NBME score. The results in this study align with the results found in other studies: males performed higher than females on a study of USMLE Step 1 scores; likewise for the differences in scores between White, Non-Hispanic and all other races.[9] However, adding these demographic variables to the model in this study had little effect on the adjusted R^2; controlling for race and gender had little effect.

Depending on the model used, 53 – 56% of the variability in the NBME score can be attributed to the independent variables in the study (iRAT, MCQ, gender, and race). However, these models do not account for enough of the variability in NBME score to be useful in

individually predicting final grades. Model 1 had a difference in observed and predicted scores ranging over 15 percentage points above or below the actual score. This model lacks the accuracy to predict a score within 1 letter grade of the actual score.

Model 3, the stepwise model, offers insight into the strength of individual independent variables' predictive value. The first and last MCQs the students took had the greatest ability to predict final scores. The model selected MCQ 5 as the most predictive of final MCQ score, followed by MCQ 1. But when using standardized coefficients, MCQ 1 had the greatest predictive validity. MCQ 4 was not a significant predictor of score. It may be useful to explore possible reasons for the differences such as considering if students try harder for the first and last tests of the semester. Or is there something about the 'newness' of a course's first exam and the proximity of the last exam and final that affects the predictive value of those two exams compared to the middle exams. It may be useful to evaluate if there is something about the content or structure of the material on the fourth test that brings down its predictive value or whether the content covered in each exam is tested equally on the NBME.

The study has many limitations. Comparing the study methods between this study and that reported by Wambuguh shows a number of opportunities for improvement. Data used by Wambuguh was collected over five years from multiple courses and included 1294 students.[8] This study was limited in scope to only include only one course taken by two classes of medical students. In addition, the quiz data used to build the model by Wambuguh was collected after every lecture rather than intermittently like the iRAT and MCQ data used in this study.[8]

In addition, the models developed in this study only predict around fifty percent of the variability in final exam scores, as seen by the R^2 values found in the regression models. The models may be improved by adding more classes, courses, or considering other variables. First

poll Peer Instruction results are one variable to consider including since Peer Instruction was the primary teaching method for this course and consisted of around 20 questions per class time.

Conclusion

The results of this study show that, overall, the formative test data collected were not adequate to predict final score. However, the first and last exam of the Staying Alive course were the most significant predictors of final NBME score of the variables considered. Course directors may consider using MCQ 1 scores to identify potential students who may be in need of additional support throughout the course. Another study with more data in the form of additional academic years, additional courses, or additional data points, such as first-poll Peer Instruction results, may be more useful in developing a model to predict final course scores.

References

1. Callahan CA, Hojat M, Veloski J, Erdmann JB, Gonnella JS. The Predictive Validity of Three Versions of the MCAT in Relation to Performance in Medical School, Residency, and Licensing Examinations: A Longitudinal Study of 36 Classes of Jefferson Medical College. *Acad Med*. 2010;85(6):980. http://ezproxy.libraries.wright.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edo&AN=51781375&site=eds-live

2. Krupat E, Pelletier SR, Dienstag JL. Academic Performance on First-Year Medical School Exams: How Well Does It Predict Later Performance on Knowledge-Based and Clinical Assessments? *Teach Learn Med*. 2017;29(2):181-187. doi:10.1080/10401334.2016.1259109

3. West C, Sadoski M. Do study strategies predict academic performance in medical school? *Med Educ*. 2011;45(7):696-703. doi:10.1111/j.1365-2923.2011.03929.x

4. Roediger HL, Karpicke JD, Roediger III HL, Karpicke JD. The Power of Testing Memory Basic Research and Implications for Educational Practice. *Perspect Psychol Sci*. 2006;1(3):181-210. doi:10.1111/j.1745-6916.2006.00012.x

5. Chang EK, Wimmers PF. Effect of Repeated/Spaced Formative Assessments on Medical School Final Exam Performance. *Heal Prof Educ*. 2017;3(1):32-37. http://10.0.3.248/j.hpe.2016.08.001

6. Kerdijk W, Cohen-Schotanus J, Mulder BF, Muntinghe FLH, Tio RA. Cumulative versus end-of-course assessment: effects on self-study time and test performance. *Med Educ*. 2015;49(7):709-716. doi:10.1111/medu.12756

7. Palmen LN, Vorstenbosch MATM, Tanck E, Kooloos JGM. What is more effective: a daily or a weekly formative test? *Perspect Med Educ*. 2015;4(2):73-78. doi:10.1007/s40037-015-0178-8

8. Wambuguh O, Yonn-Brown T. Regular Lecture Quizzes Scores as Predictors of Final Examination Performance: A Test of Hypothesis Using Logistic Regression Analysis. *Int J Scholarsh Teach Learn*. 2013;7(1). http://ezproxy.libraries.wright.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1135208&site=eds-live

9. Rubright JD, Jodoin M, Barone MA. Examining Demographics, Prior Academic Performance, and United States Medical Licensing Examination Scores. *Acad Med*. 2019;94(3):364. http://ezproxy.libraries.wright.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edo&AN=134933410&site=eds-live

Tables and Figures

Table 1. Descriptive statistics of class, race, and gender.

Demographic	N	Percent
Class		
2021	115	49.1
2022	119	50.9
Gender		
Male	105	44.9
Female	129	55.1
Race		
White, Non-Hispanic	148	65.8
Black, Non-Hispanic	30	13.3
Asian/ Pacific Islander	39	17.3
Other	8	3.6

Table 2. Average iRAT, MCQ, and NBME scores with standard deviation, minimum, and maximum.

Test	N	Mean	SD	Min	Max
iRAT	234	67.1	10.1	32.2	89.3
MCQ	234	76.1	7.4	56.0	93.0
NBME	234	79.1	8.5	50.0	98.0

Table 3. NBME scores (mean, SD) by gender and race.

Group	N	Mean	SD	P Value
Gender				<0.001
Female	129	76.8	8.4	
Male	105	81.9	7.8	
Race				0.006
White, Non-Hispanic	148	80.6	8.3	
Black, Non-Hispanic	30	75.4[a]	7.4	
Asian/Pacific Islander	39	77.0	9.1	
Other	8	79.3	9.1	

[a]P<0.05 vs White race

Table 4. Regression models

Model	Beta coefficient (95% confidence interval)	Standardized Beta coefficient	n	Adjusted R^2	ANOVA P value
Model 1			234	0.532	<0.001
iRAT average	0.083 (-0.022-0.189)	0.098			
MCQ average	0.762 (0.617-0.906)	0.659			
Model 2			225	0.558	<0.001
iRAT average	0.057 (-0.050-0.165)	0.067			
MCQ average	0.799 (0.652-0.947)	0.689			
Gender (reference = female)	2.871 (1.334-4.407)	0.167			
Race (reference = White, Non-Hispanic)					
Race = Black, Non-Hispanic	2.953 (0.490-5.416)	0.118			
Race = Asian/Pacific Islander	-0.603 (-2.690-1.485)	-0.027			
Race = Other	1.272 (-2.810-5.354)	0.028			
Model 3, Stepwise[a]			225		
MCQ 5	0.214 (0.094-0.334)	0.228		0.361	<0.001
MCQ 1	0.254 (0.164-0.334)	0.307		0.488	<0.001
Gender (reference = female)	2.854 (1.297—4.411)	0.166		0.513	0.001
MCQ 3	0.173 (0.060-0.285)	0.181		0.542	<0.001
Race = Black, Non-Hispanic	2.895 (0.399-5.392)	0.115		0.552	0.014
MCQ 2	0.115 (0.003-0.227)	0.119		0.560	0.032
MCQ 4	0.096 (-0.013-0.205)	0.112		0.564	0.064
Race = Asian/Pacific Islander	-0.959 (-0.3063-1.145)	-0.043		0.564	0.340
Race = Other	0.905 (-0.3171-4.981)	0.020		0.563	0.662

[a]Each subsequent adjusted R^2 represents the improvement from the previous R^2 as a result of the addition of each variable to the model.