Analysis Report: Differences in Medical Exam Scores between Underepresented in Medicine Students and Non-Underepresented in Medicine Students

Marin Haluza

2025-11-15

Background

The purpose of this study is to investigate differences in medical school exam scores between underrepresented (URM) medical students and students who are not underrepresented in medicine (non-URM). Understanding these differences may facilitate understanding medical student populations and bridge any biases amongst exam scores. Because the data was collected as assessments were completed, this is a cross sectional study design. The primary outcome of interest is OSCE scores, while USMLE and clerkship scores are secondary, more exploratory, outcomes of interest.

Methods

Using data from an observational, cross sectional study collected from medical students attending the University of Colorado Denver at the Anschutz Medical School, an independent samples t-tests were used to investigate differences in mean OSCE scores in three different domains. All observations having less than 60% of the reported OSCE scores were omitted from analysis. Additionally, clerkship grades, USMLE Step Scores, and OSCE scores that were missing more than 60% of observations were dropped from analysis. Lastly, any observations missing URM status were dropped from analysis due to URM status being our primary predictor. The sample size is comprised of 42 URM students and 112 non-URM students. Table 1 represents the baseline characteristics (measured in sample counts), mean OSCE scores, mean USMLE scores, and clerkship grades. Clerkship grades were grouped into passing and high passing. High passing grouping combines students who passed with honors with students who passed with high honors.

Mean OSCE scores (graded on a continuous scale of 1 - 100) were aggregated across three different domains: physical exam scores, medical documentation scores, and communication scores. Because the data was collected over two years and domains were tested multiple times, each domain score was averaged into one mean domain score for each student. Oral presentation exam scores were only collected once and were thus omitted from analysis. Normality was not tested for the purpose of this analysis because the sample sizes were both greater than 30 in each group. In order to adjust for multiple comparisons, a Bonferroni Correction was applied to significance levels for this analysis with three comparison groups ($p = 0.017$) An independent samples t-test was conducted for the three domains and estimate statistics with p-values are reported.

In order to investigate differences amongst mean USMLE Step 1 and 2 Scores (graded on a continuous scale of 1 - 300) between URM status we will use an independent sample t-test to see if there is a difference amongst the two groups. This is an exploratory component of our analysis so reported p-values will not be adjusted and will be provided not as a conclusive interpretation, but rather as an overview of the sample. Significance levels were not adjusted and set to alpha = 0.05.

For clerkship grades, we will use a Chi Squared analysis to see if there is any difference in scores between URM status. A chi Square analysis was chosen for the purposes of comparing categorical scores, rather than continous scores as seen in OSCE and Step scores. Since there was a substantial amount of missingness in some of the clerkship grades, we will only compare the final clerkship scores for variables that are not

missing more than 60% of the observations. We will compare proportions between passing and high passing (a combined variable for passing with honors and passing with high honors) for the final clerkship scores. EC and MSK Final Clerkship grades were dropped from analysis because every student received the same score. Because of the amount of comparisons being made, these results will be included only as an exploratory component to provide an overview of the sample. Significance levels were not adjusted and set to alpha = 0.05.

Setting Up Project Directory: CIDATools library was used to create the folders of this project. All documentation was done through R Markdown.

GitHub Repository Link: https://github.com/marinhaluza-web/BIOS6621-Final-Project-.git

Table 1: Participant Characteristics by URM Status

Characteristic	**Overall** N = 154[1]	**Not Underepresented in Medicine** N = 112[1]	**Underepresented in Medicine** N = 42[1]
Gender			
Male	76 (49%)	54 (48%)	22 (52%)
Female	78 (51%)	58 (52%)	20 (48%)
Race			
American Indian or Alaska Native	7 (4.9%)	0 (0%)	7 (21%)
Asian	25 (17%)	21 (19%)	4 (12%)
Black or African American	11 (7.6%)	0 (0%)	11 (32%)
Other	5 (3.5%)	4 (3.6%)	1 (2.9%)
White	96 (67%)	85 (77%)	11 (32%)
(Missing)	10	2	8
Mean Physical Exam OSCE Scores	89.9 (4.5)	90.4 (4.1)	88.7 (5.4)
Mean Communication Exam OSCE Scores	92.4 (4.6)	92.6 (4.7)	91.9 (4.4)
Mean Medical Documentation Exam OSCE Scores	91.9 (5.1)	92.5 (4.5)	90.2 (6.0)
Oral Presentation OSCE Scores	96.9 (5.5)	96.9 (5.7)	96.9 (4.8)
(Missing)	2	1	1
Step 1 Exam Scores	226 (21)	229 (21)	220 (19)
Step 2 Exam Scores	244 (14)	246 (13)	239 (13)
(Missing)	1	0	1
CPC Final Clerkship Grade			
Passing	42 (33%)	30 (31%)	12 (38%)
High Passing	86 (67%)	66 (69%)	20 (63%)
(Missing)	26	16	10
EC Final Clerkship Grade			

Table 1: Participant Characteristics by URM Status

Characteristic	**Overall** N = 154[1]	**Not Underepresented in Medicine** N = 112[1]	**Underepresented in Medicine** N = 42[1]
Passing	154 (100%)	112 (100%)	42 (100%)
High Passing	0 (0%)	0 (0%)	0 (0%)
HAC Final Clerkship Grade			
Passing	36 (26%)	20 (20%)	16 (42%)
High Passing	101 (74%)	79 (80%)	22 (58%)
(Missing)	17	13	4
ICAC Final Clerkship Grade			
Passing	31 (24%)	18 (19%)	13 (34%)
High Passing	100 (76%)	75 (81%)	25 (66%)
(Missing)	23	19	4
MSK Final Clerkship Grade			
Passing	121 (100%)	88 (100%)	33 (100%)
High Passing	0 (0%)	0 (0%)	0 (0%)
(Missing)	33	24	9
NC Final Clerkship Grade			
Passing	52 (35%)	33 (31%)	19 (46%)
High Passing	95 (65%)	73 (69%)	22 (54%)
(Missing)	7	6	1
OBG Final Clerkship Grade			
Passing	36 (27%)	19 (20%)	17 (47%)
High Passing	95 (73%)	76 (80%)	19 (53%)
(Missing)	23	17	6
OPC Final Clerkship Grade			
Passing	26 (19%)	18 (18%)	8 (23%)
High Passing	110 (81%)	83 (82%)	27 (77%)
(Missing)	18	11	7
PC Final Clerkship Grade			
Passing	41 (32%)	28 (29%)	13 (37%)
High Passing	89 (68%)	67 (71%)	22 (63%)
(Missing)	24	17	7

[1]n (%); Mean (SD)

Results for Independent Sample T-Tests for USMLE Scores Independent sample t-tests for USMLE Step 1 and Step 2 scores between URM and non-URM students was conducted as an exploratory analysis of the sample (Table 2; Figure 1). Mean Step 1 Scores demonstrated statistical significance when comparing non-URM students to URM students. On average, non-URM student's Step 1 scores were 8.67 points higher than URM students ($p = 0.0158$). Similarly, Step 2 Exam Scores in non-URM students had a higher mean score by 6.93 points ($p = 0.00541$) on average. These results suggest a statistical difference in USMLE Step 1 and Step 2 scores; however, they are not part of the primary study aims and included to provide insight on the sample. Further research investigating differences in medical USMLE exam scores may be of interest.

Figure 1. USMLE Step Exam Scores by URM Status

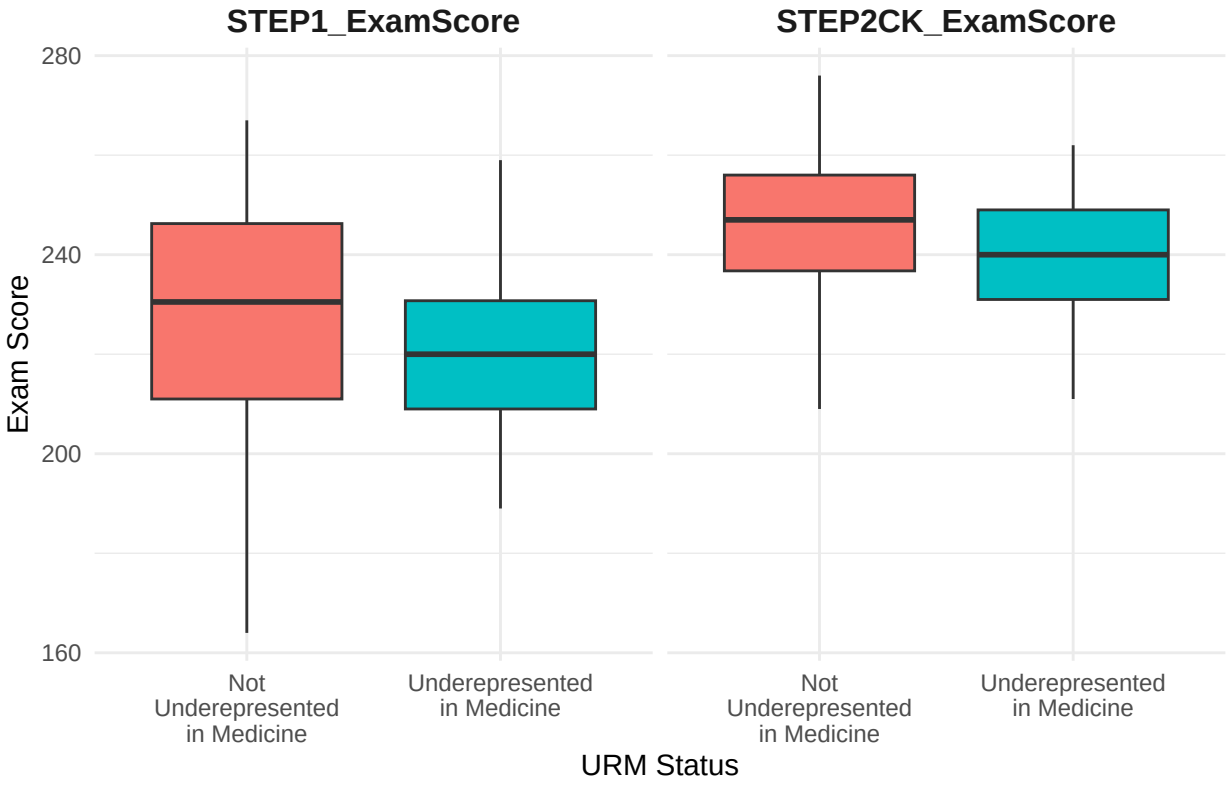


Table 2: T-test Comparison of STEP Exam Scores by URM Status

Test	Mean Difference	t statistic	P-value	95% CI Lower	95% CI Upper
STEP 1 Exam Score	8.667	2.465	0.016	1.671	15.662
STEP 2 Exam Score	6.929	2.867	0.005	2.112	11.745

Results for Chi Square Analysis of Clerkship Scores The results comparing clerkship score differences aim to explore the study sample and provide insight rather than making conclusive statements about differences in clerkship scores between URM an non-URM students.

Using a Chi-Square Analysis, these results demonstrate differences in final clerkship course grade proportions (passing versus high passing) by URM status (Table 3). There was an observed statistically significant difference in OBG ($p=0.002$) and HAC ($p=0.009$) final clerkship grade proportions. These results indicate that final course grades differed by URM status in these clerkship for this sample. No statistical significant differences were found for CPC($p = 0.5$), ICAC($p=0.069$), NC($p=0.084$), OPC($p=0.5$), and PC ($p=0.4$) clerkship scores, suggesting that this sample did not have uniform differences across clerkship grades. The exploratory nature and differing results for this analysis indicates that further research may provide further insight on clerkship grade differences.

Table 3: Chi Square Analysis: Final Clerkship Grades by URM Status

Characteristic	**Not Underepresented in Medicine** N = 112[1]	**Underepresented in Medicine** N = 42[1]	**p-value**[2]
CPC Final Clerkship Grade			0.5
Passing	30 (31%)	12 (38%)	
High Passing	66 (69%)	20 (63%)	
(Missing)	16	10	
HAC Final Clerkship Grade			0.009
Passing	20 (20%)	16 (42%)	
High Passing	79 (80%)	22 (58%)	
(Missing)	13	4	
ICAC Final Clerkship Grade			0.069
Passing	18 (19%)	13 (34%)	
High Passing	75 (81%)	25 (66%)	
(Missing)	19	4	
NC Final Clerkship Grade			0.084
Passing	33 (31%)	19 (46%)	
High Passing	73 (69%)	22 (54%)	
(Missing)	6	1	
OBG Final Clerkship Grade			0.002
Passing	19 (20%)	17 (47%)	
High Passing	76 (80%)	19 (53%)	
(Missing)	17	6	
OPC Final Clerkship Grade			0.5
Passing	18 (18%)	8 (23%)	
High Passing	83 (82%)	27 (77%)	
(Missing)	11	7	
PC Final Clerkship Grade			0.4
Passing	28 (29%)	13 (37%)	
High Passing	67 (71%)	22 (63%)	
(Missing)	17	7	

Table 3: Chi Square Analysis: Final Clerkship Grades by URM Status

Characteristic	**Not Underepresented in Medicine** N = 112[1]	**Underepresented in Medicine** N = 42[1]	**p-value**[2]

[1]n (%)

[2]Pearson's Chi-squared test

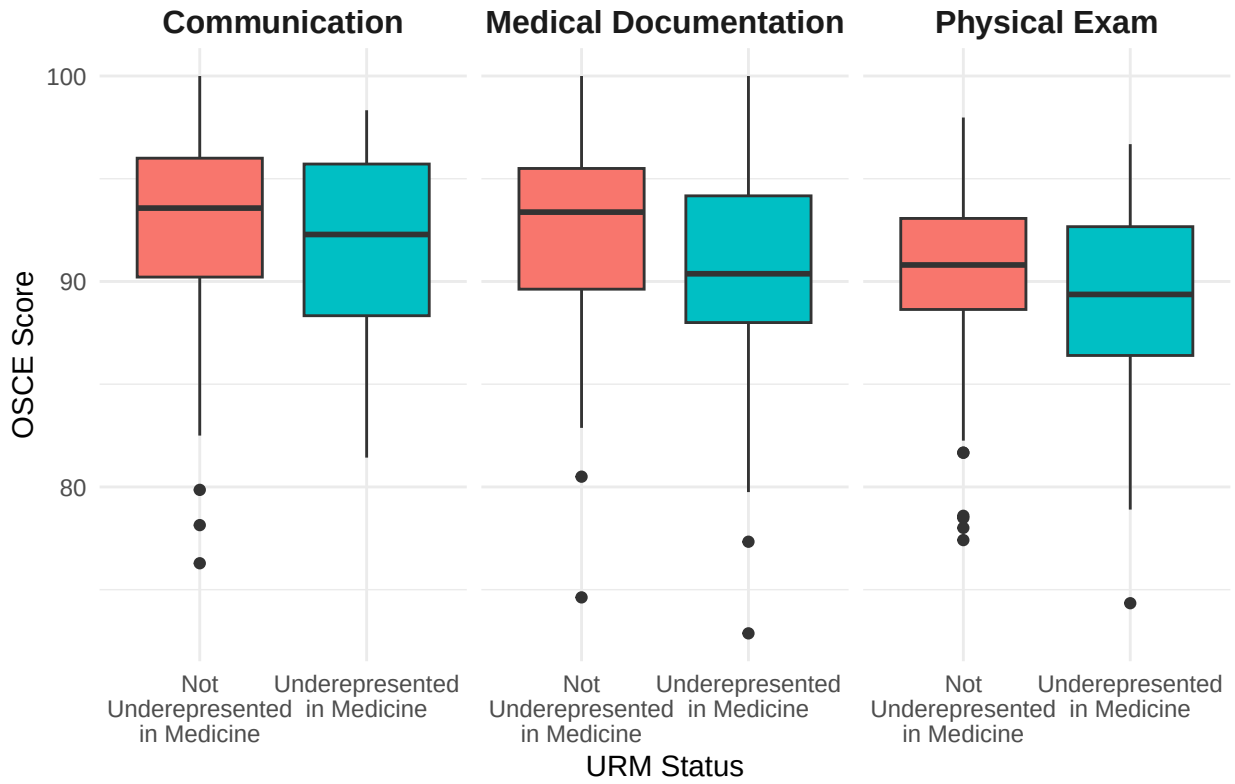
Results for Independent Sample T-Test of Mean OSCE Scores For our primary analysis, we investigated differences in mean OSCE scores. Among the three domains that were analyzed (Physical Exam Scores, Communication Exam Scores, and Medical Documentation Exam Scores), we see no statistically significant differences between URM and non-URM students (Table 4; Figure 2).Mean Physical Exam scores in non-URM students were 1.72 points higher than URM students (p = 0.066). Mean Communication Exam scores differed by 0.78 points(p = 0.342), and mean Medical Documentation Exam scores differed by points 2.33 on average (p = 0.0261).

Overall, these not statistically significant results indicate that there is no differences in OSCE scores and further research may be needed to provide more insight on differences between URM and non-URM students.

Table 4: T-test Comparison of OSCE Scores by URM Status

Test	Mean Difference	t statistic	P-value	95% CI Lower	95% CI Upper
Mean Physical Exam Score	1.718	1.873	0.066	-0.117	3.552
Mean Communication Exam Score	0.778	0.957	0.342	-0.841	2.398
Mean Medical Documentation Exam Score	2.329	2.282	0.026	0.287	4.371

Figure 2.OSCE Scores by URM Status Across Three Domains



Discussion and Limitations The primary objective of this study was to evaluate whether OSCE performance differed between URM and non-URM medical students across three core domains: Physical Exam, Communication, and Medical Documentation. Across all three domains, mean scores did not differ significantly by URM status, indicating that students achieved comparable performance levels on structured clinical assessments early in the medical curriculum.

In contrast to the OSCE results, statistically significant differences were observed in USMLE Step 1 and Step 2 examination scores and some clerkship grades, with non-URM students achieving higher average scores.These results are not the primary focus of this study and are only representative of the sample at hand, but may provide meaningful insight for future research.

Several limitations should be considered when interpreting these results. First, the study sample was observational and based on data from a single institution, limiting generalizability to broader medical student populations. OSCE scores were graded by several individuals, so inconsistencies in grading methods may be of interest for further research. Furthermore, the sample size—particularly for the URM group—differed in size from non-URM students, which may reduce the ability to detect statistical differences. Additionally, differences in missing data across variables and domains may introduce bias if scores were not missing at random.