Section 1: USMLE Step 1 and Step 2 CK Score Uses and Interpretations (General Studies)

Section Overview

This section includes summaries of articles published between 2009 and 2018. The articles examine aspects of Step 1 and Step 2 CK scores, including their relationships to external measures of performance and clinical practice, their use in making residency selection decisions, and score differences by student and school characteristics.

Research Summaries and Abstract Links

Andriole DA, Jeffe DB. A national cohort study of U.S. medical school students who initially failed Step 1 of the United States Medical Licensing Examination. Academic Medicine. 2012;87(4): 529–536.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3315604/

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study sought to describe educational outcomes for a cohort of U.S. medical students who initially failed Step 1, with particular attention on attempting and passing Step 2 CK.

How was the study conducted (i.e., what data and methodologies were used)?
Data from 6,594 students who failed Step 1 between 1993 and 2000 were examined using chi-square analyses and multivariate logistic regression techniques.

What were the primary results of the study?
Approximately 91% of students who failed Step 1 on the first attempt sat for Step 2 CK.
Students who attempted Step 2 CK after failing Step 1 on the first attempt were more likely to be women and of Asian/Pacific Islander descent. Students in the cohort with lower Step 1 scores and more recent matriculates were less likely to attempt Step 2 CK. Approximately 70% of students who initially failed Step 1 eventually passed Step 2 CK. Students who passed Step 2 CK were more likely to be women, have recently matriculated, and have higher Step 1 scores. Students of Asian/Pacific Islander descent, underrepresented minorities, older students, and students with lower failing Step 1 scores were less likely to pass Step 2 CK.

What are the implications of the findings?
Findings may help to inform medical school educational interventions and support programs aimed at identifying and helping students at risk for not attempting or not passing Step 2 CK. Also, the identification of variables associated with students who have difficulties with USMLE may have implications for understanding subsequent physician workforce patterns.

What are the limitations of the study?
The study did not include school-specific information, including Step 1 score requirements for graduation or advancement and existing academic support programs. Additionally, issues of selection bias may be present as average Step 1 scores for the failing students included in the sample were higher than the average Step 1 scores from students who were excluded. Lastly, a general theoretical assumption is made that all students who initially fail Step 1 should be encouraged and supported to pursue Step 2 CK, although this may not always be appropriate.

Cuddy MM, Young A, Gelman A, Swanson DB, Johnson DA, Dillon GF, and Clauser B. Exploring the relationships between USMLE performance and disciplinary action in practice: A validity study of score inferences from a licensure examination. Academic Medicine. 2017;92(12):1780–85.
https://www.ncbi.nlm.nih.gov/pubmed/28562454

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study assessed the relationship between performance on USMLE Step 1 and Step 2 CK examinations and subsequent disciplinary actions by state medical boards.

How was the study conducted (i.e., what data and methodologies were used)?
The study combined Step 1 and Step 2 CK scores from passing examinees with state medical board disciplinary action data. Logistic multi-level modeling was used to assess the relationships between Step 1 and Step 2 CK scores and the likelihood of receiving a disciplinary action from a state medical board, accounting for state and specialty differences. Gender and number of years since graduation were treated as covariates.

What were the primary results of the study?
A small percent of physicians face disciplinary actions; Step 2 CK scores are related to the likelihood of receiving an action. As stated in the paper, "1-SD increase in Step 2 CK scores (approximately 23 score points) corresponds to a decrease in the chance of disciplinary action by about 25% (odds ratio of 0.75)," (p. 4). While the Step 1 score had an independent effect on the likelihood of receiving an action, when included in the model with Step 2 CK score, the effect of Step 1 score was no longer significant.

What are the implications of the findings?
As lower Step 2 CK scores are related to higher likelihoods of disciplinary action by a state medical board, Step 2 CK scores may be a useful tool for state boards to create a composite picture of a licensure candidate. As Step 1 scores were unrelated to disciplinary actions in practice, after accounting for Step 2 CK scores, it may be that Step 2 CK scores reflect skills more congruent with characteristics that later result in actionable behavior.

What are the limitations of the study?
This study looked only at whether a disciplinary action was taken or not, and did not examine the type of offense or the severity of the action. Many claims brought to medical boards do not end in a formal action, and so looking at lower thresholds of complaints could be another fruitful avenue of exploration. Only Step 1 and Step 2 CK scores were under study. Step 2 CS and Step 3 scores may better reflect what is expected of physicians in practice. Lastly, given that physicians may practice and be licensed in various states, the issue of data quality arises. This is also a concern with the quality of the data available for a physician's specialty.

Gauer JL, Jackson JB. The association of USMLE Step 1 and Step 2 CK scores with residency match specialty and location. Medical Education. 2017;22:1-7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5653932/

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined two questions: 1) How are USMLE Step 1 and 2 CK scores related to matching into various residency specialties, and 2) do USMLE Step 1 and Step 2 CK scores predict whether students remained in-state for their residency training?

How was the study conducted (i.e., what data and methodologies were used)?
The sample consisted of 1,054 students from the University of Minnesota who graduated between 2011 and 2015 (five graduating classes), and who matched into a residency program. A one-way MANOVA (multivariate analysis of variance) was used to examine the impact of Step 1 and Step 2 CK scores on matching, where scores served as the dependent variables and residency specialty functioned as the independent variable (20 specialties were included). A t-test was used to examine score differences for students who stayed in- vs out-of-state.

What were the primary results of the study?
Scores from both examinations significantly differed across matched residency specialty. Dermatology matches performed the highest on both examinations (Step 1 = 244 and Step 2 CK = 251) and family medicine matches performed the lowest (Step 1 = 213 and Step 2 CK = 229). The variability of other specialty average scores was moderately high, with specialties falling across the range of scores between Dermatology and Family Medicine. For Step 1, students who matched in-state (Mean = 225, SD = 19) performed statistically below those who matched out-of-state (Mean = 228, SD = 19), though the difference was small. In contrast, there was no statistical difference between in-state (Mean = 238, SD = 17) and out-of-state (Mean = 239, SD = 16) performance on Step 2.

What are the implications of the findings?
Step 1 and 2 CK performance plays an important role in residency selection. Students may be self-selecting into residency specialty areas based on their Step scores, with high performers applying for more competitive specialties. Students aiming to match with a high average score specialty may dedicate more time to studying for USMLE. Findings showing students with higher scores are more likely to leave the state probably reflect the additional opportunities available to these high performers.

What are the limitations of the study?
All data come from a single school and therefore results may not generalize to the national landscape. Given the small sample size (1,054) relative to number of specialties, some specialties were unable to be represented in the analysis (< 5 students matched). Therefore, the findings may be excluding some nuance regarding more specialized residencies.

Green M, Jones P, Thomas, JX Jr. Selection criteria for residency: Results of a national program directors survey. Academic Medicine. 2009;84(3):362-367.
https://www.ncbi.nlm.nih.gov/pubmed/19240447

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined the relative importance of different criteria used for residency selection across medical specialties.

How was the study conducted (i.e., what data and methodologies were used)?
A survey was administered in 2006 to a sample of residency program directors (N = 1,201; 49% response rate) from 21 medical specialties. Program directors were asked to rate the relative importance of various selection criteria (e.g., measures of academic achievement, extracurricular activities, supporting information, issues of concern) using a five-point Likert scale (5 = critical; 1 = unimportant). The mean values for all selection criteria were computed, ranked order, and compared across and within specialties.

What were the primary results of the study?
Across all specialties, program directors rated grades in required clerkships, grades in specialty-specific electives, number of honors grades, Step 1 score, and Step 2 CK score as the five most important residency selection criteria. In contrast, program directors rated medical school academic awards, grades in other senior electives, grades in preclinical courses, published medical school research, and research experience as the five least important residency selection criteria.

What are the implications of the findings?
Program directors placed more importance on academic criteria generally reflective of clinical performance than non-academic criteria when making residency decisions. Program directors ranked grades in preclinical courses as one of the five least important selection criteria. This may be due to the considerable variability in these courses across schools. Program directors may view USMLE Step 1 scores as an alternative for preclinical grades, which may be one reason why Step 1 score was highly ranked. Program directors also ranked Step 2 CK score as a top selection criterion.

What are the limitations of the study?
Program directors varied in their response rate by specialty. Thus, program directors who responded to the survey may not be representative of program directors generally.

Green M, Angoff N, Encandela J. Test anxiety and United States Medical Licensing Examination scores. The Clinical Teacher. 2016;13(2):142-146.
https://www.ncbi.nlm.nih.gov/pubmed/26037042

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study investigated the relationship between anxiety and USMLE Step 1 performance. It also examined whether a test-taking strategy course intended to lower anxiety affected this relationship.

How was the study conducted (i.e., what data and methodologies were used)?
The sample included 93 second-year students from a private medical school who were preparing to take Step 1. Students volunteered to participate in a test-taking strategy course. Twenty-five volunteers were randomly selected to attend the course, which lasted 6 days and consisted of 6-hour lectures on answer strategies, study techniques/preparation, and relaxation techniques. The remaining students, including those who did not volunteer, were used as a control group. Students in the treatment group completed the Westside Test Anxiety Scale at three time points: baseline, after completing the course (4 weeks after baseline), and after taking Step 1 (10 weeks after baseline). Control students completed the scale at baseline and after Step 1. The scale ranges from 1 to 5, where higher scores indicate increased anxiety. Multiple regression techniques were used to determine the relationship between Step 1 scores and test anxiety after adjusting for MCAT scores. T-tests were used to compare pre-post data for the test-anxiety treatment and control groups. Correlations between course exposure and scores were computed to determine see if attendance improved performance.

What were the primary results of the study?
The baseline test anxiety scores indicated that 22% of students were moderately to highly anxious about Step 1 (score above 3.0). Anxiety was moderately inversely related to Step 1 scores after accounting for MCAT scores. Students completing the course showed a small decrease in anxiety after completing the course and after taking Step 1, whereas students in the control showed a small increase in anxiety. The mean USMLE score in the course group was 234 ± 18 vs 243 ± 16 in the controls ($p = 0.05$). No correlation existed between the exposure group and USMLE scores, adjusting for baseline anxiety and MCAT scores.

What are the implications of the findings?
Some students have moderate to high anxiety regarding Step 1. This anxiety is inversely related to scores. However, a test strategy course that lowered anxiety did not increase these students' performance. It may be that a stronger treatment is needed to see an effect on scores, or that test anxiety is a proxy for another variable that has more direct influence on scores.

What are the limitations of the study?
The study was conducted with a relatively small number of students in a single institution and results may not generalize to students at other institutions. The assignment to the test anxiety course was not randomized, as only students who volunteered were selected. There may be confounding difference between students who did and did not volunteer for the course.

Greenburg DL, Durning SJ, Cruess DL, Cohen DM, Jackson JL. The prevalence, causes, and consequences of experiencing a life crisis during medical school. Teaching and Learning in Medicine. 2010;22(2):85-92.
https://www.ncbi.nlm.nih.gov/pubmed/20614371

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study addressed the following question: What are the prevalence, causes, and consequences of experiencing a major life crisis while enrolled as a medical student at Uniformed Services University (USU)?

How was the study conducted (i.e., what data and methodologies were used)?
This was a retrospective review of USU graduate performance based on the following measures:

- Long Term Career Outcomes Survey (LTCOS) delivered after graduation, which asked (yes/no) if they experienced a "major life crisis" while enrolled at USU.
- Demographic data, including age, sex, race, marital status, and number of children at both matriculation and graduation.
- Academic data prior to medical school, including undergraduate GPA, MCAT scores, and USU admission interview scores
- Medical school performance data, including USMLE Step 1 and Step 2 CK scores

What were the primary results of the study?
The LTCOS had a 67% response rate (1,807 graduates). 22% of USU medical students experienced a major life crisis during medical school, which had a significantly negative effect on student performance, including lower USMLE Step 1 scores (202 vs. 207, $p < .001$) and lower USMLE Step 2 CK scores (198 vs. 204, $p < .001$). Students experiencing a crisis were more likely to be older, female, married at matriculation, and have children at matriculation to medical school. Students most likely to suffer adversely after a crisis had lower college GPAs, lower MCAT scores, and were less likely to have children at matriculation.

What are the implications of the findings?
Personal crises were common in medical school and had impacts on academic and licensure exam performance. Students with lower academic achievement prior to medical school had the most significant impact on their USMLE scores. This raises the question whether certain groups of students will suffer the most significant decreases on USMLE performance after a life crisis.

What are the limitations of the study?
The study design was retrospective, so students may have perceived situations to be crises after noting a decline in their performance. The population is predominantly white and male, so findings may not be reflective of the experiences of non-majority students. Students from other medical schools with tuition policies different than those at USU (which has free tuition) may experience different types of financial crises and thus findings may not generalize to them. Neither the timing nor the duration of crises was collected, so whether a crisis had a lasting impact or a crisis persisted for several years (e.g. divorce) is unknown.

Hecker K, Violato C. How much do differences in medical schools influence student performance? A longitudinal study employing hierarchical linear modeling. Teaching and Learning in Medicine. 2008;20(2):104-113.
https://www.ncbi.nlm.nih.gov/pubmed/18444195

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study investigated the effects of medical school variables related to curricula and educational policies on USMLE Step 1, Step 2 CK, and Step 3 scores from 1994-2002.

How was the study conducted (i.e., what data and methodologies were used)?
Longitudinal data for 116 medical schools was examined using hierarchical linear modeling techniques to partition the total variation in USMLE scores into within- and between-school components and to study the effects of school variables on Step 1, Step 2 CK, and Step 3 scores.

What were the primary results of the study?
The between-school variance for Step 1, Step 2 CK and Step 3 schools was approximately 15%, 10%, and 11%, respectively. Between-school variation decreased to less than 5% when student factors were considered. School variables accounted for approximately 28% to 36% of the variation in Step 1 scores, approximately 16% to 24% of the variation in Step 2 CK scores, and approximately 19% to 25% of the variation in Step 3 scores. Student variables accounted for approximately 81% to 88% of the variation in Step 1 scores, approximately 48% to 80% of the variation in Step 2 CK scores, and 68% to 81% of the variance in Step 3 scores. School-level variables did not consistently predict adjusted mean school Step performance.

What are the implications of the findings?
Individual student differences account for most of the variation in USMLE Step performance, with small differences between schools. The vast majority of within- and between-school variation in scores can be accounted for by incoming student differences, mostly MCAT scores. Curricula and educational policies do not appear to be particularly influential. Findings suggest that school differences may not matter as much in terms of USMLE performance, compared to student differences, with implications for efforts geared toward curricular reform.

What are the limitations of the study?
USMLE scores are one measure of student achievement, and curricular differences may be related to other types of assessments used for different purposes and/or focused on different knowledge areas or skill sets. The measurement of curricular type was categorical and self-reported; therefore certain nuances of how curricula are implemented in practice may have been lost in the analysis. As this study is now about 10 years old, findings also may not fully represent current patterns, especially given recent changes in the organization of medical school curricula surrounding integrated curricular approaches and USMLE timing.

Jones MD, Yamashita T, Ross RG, Gong J. Positive predictive value of medical student specialty choices. BMC Medical Education. 2018;18(33):1-7.
https://www.ncbi.nlm.nih.gov/pubmed/29523127

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined two questions: (1) to what extent do students' specialty preferences and their confidence in those preferences at different time points during medical school predict their eventual postgraduate training specialty area, and (2) what is the relationship between intended and actual specialty area discrepancies and Step 1 scores?

How was the study conducted (i.e., what data and methodologies were used)?
Data from 634 medical students from the University of Colorado School of Medicine who trained in 23 residencies from 2011-2015 were analyzed. Students were asked to rank their first, second, and third specialty preferences and confidence in them at the end of their first, second, and third year of medical school. Positive predictive values (PPV; % of students who pursued postgraduate residency training in the desired specialty area) and negative predictive values (NPV; % of students who did not pursue postgraduate residency training in the desired specialty area) were calculated and compared across 8 specialties with at least 45 trainees at each year. Step 1 scores were compared across students who trained in a less, equally, and more competitive specialty than the first-choice specialty identified at the end of their second year of medical school.

What were the primary results of the study?
PPVs improved by medical school year (year 1 = ~40%; year 2 = ~55%; year 3 = ~85%). Students' confidence ratings, however, did not generally improve PPVs across specialties. NVPs remained constant at about 90% in all years. Step 1 scores were higher for those who eventually trained in specialties more competitive than their desired specialty at end of year 2.

What are the implications of the findings?
The results suggest that the career plans of first- and second-year students with respect to specialty areas may not reflect actual specialty area placements/choices. This information may be useful for medical schools who offer specialty-specific program for medical students. Given the relationship between Step 1 scores and changes in intended and eventual specialties, it is plausible that students may have altered their specialty choices based on their Step 1 performance.

What are the limitations of the study?
As this study was done at a single institution, results may not generalize. Also, there may be some student or curricula characteristics that were not examined that may influence the relationships examined. The PPVs of the rankings may have been higher if the school had offered specialty-specific training and/or included other student characteristics. Lastly, the number of students within certain specialties was small.

Jurich D, Daniel M, Paniagua M, Fleming A, Harnik V, Pock A, Swan-Sein A, Barone MA, Santen SA. Moving the United States Medical Licensing Examination Step 1 after core clerkships: An outcomes analysis. [published online ahead of print, September 11, 2018]. Academic Medicine. https://www.ncbi.nlm.nih.gov/pubmed/30211755

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study provides an investigation of how school policies changing the timing of Step 1 from after completion of the basic science curricula to after core clerkships impacts USMLE Step 1 performance.

How was the study conducted (i.e., what data and methodologies were used)?
Performance data from four medical schools that changed the timing of Step 1 to after core clerkships were analyzed. Step 1 scores from three years before and three years after the change were analyzed in a pre-post study design. Across schools, the pre-change sample included 1,668 students and the post-change sample included 1,529 students. All students took Step 1 between 2008 and 2016. Several confounders were addressed, including rising national scores and potential differences in cohort abilities using deviation scores and ANCOVA. A resampling procedure compared study schools' score changes to similar schools' in the same time period.

What were the primary results of the study?
The ANCOVA indicated that post-change Step 1 scores were higher compared to pre-change, after adjusting for MCAT scores and rising national averages (adjusted difference = 2.67, 95% confidence interval [CI]: 1.50–3.83, P < .001; effect size = 0.14). The average score increase in the study schools was larger than changes seen in similar schools. Step 1 failure rates also decreased from 2.87% (n = 48) pre-change to 0.39% (n = 6) post-change (P < .001).

What are the implications of the findings?
Results suggest that moving the timing of Step 1 to after core clerkships yielded a small increase in Step 1 scores and a reduction in failure rates. These findings suggest that such a shift is no worse than when Step 1 is taken prior to clerkships. For institutions interested in curricular reform related to basic science instruction and Step 1 timing, findings may provide guidance for both faculty and administrators.

What are the limitations of the study?
The change in Step 1 timing may be difficult to disentangle from other curricular changes implemented during the same timeframe, including varying Step 1 preparation practices at the selected schools. In addition, three of the four schools examined in this study tended to score historically above national averages on Step 1 relative to other schools, and thus findings may not generalize. Lastly, the small benefit gained in Step 1 scores when the timing of the exam is moved to after clerkships likely does not reflect meaningful student knowledge gains; however, it may be that the placement of Step 1 has other learning benefits not explored in this study.

Kenny S, McInnes M, Singh V. Associations between residency selection strategies and doctor performance: A meta-analysis. Medical Education, 2013;47(8):790-800. https://www.ncbi.nlm.nih.gov/pubmed/23837425

What was primary purpose of the study or the research question(s) that the study sought to answer?
This meta-analysis examined the types of information available to resident selection committees that are associated with subsequent professional practice (in residency and beyond).

How was the study conducted (i.e., what data and methodologies were used)?
An extensive literature search up to 2012 was conducted, which identified 1,877 studies, 80 of which met inclusion criteria for a meta-analysis. These 80 studies included 41,704 participants. To conduct the meta-analysis, random-effects models without pooling were used for each selection strategy and sensitivity analysis with pooled effect sizes was used for each strategy-outcome combination.

What were the primary results of the study?
Seventeen selection strategies and 17 resident or physician practice outcomes were identified and examined across the studies included in the sample. The strongest positive associations were between examination-based selection strategies, such as USMLE Step 1, and examination-based outcomes, such as scores on in-training examinations. Medical school grades were moderately positively related to both examination-based and more subjective practice outcomes. Minimal or no associations were found for the selection tools involving interviews, reference letters, and deans' letters.

What are the implications of the findings?
Standardized examination performance (e.g., Step 1 scores) and medical school grades show the strongest associations with measures of resident and physician performance in practice, suggesting that they may provide some utility for making selection decisions about entry into residency programs. Despite the value placed on them for making residency program admission decisions, deans' letters, reference letters, and interviews all demonstrated a lower than expected association with measures of resident and physician performance in practice, thus raising questions about their use for evaluating candidates for residency positions.

What are the limitations of the study?
Several of the publications selected for the meta-analysis stem from a single institution and involved relatively large sample sizes. Thus, their inclusion may have biased the results toward a reflection of one institution. Also, many of the sensitivity analyses by outcome had considerable or substantial levels of heterogeneity, and either the participants or residency training may have contributed to heterogeneity. Lastly, interpretation of results may be limited by the lack of reliable outcome measures reflecting resident and physician performance in practice. The measures used in the studies analyzed in the meta-analysis typically included residency in-training examination scores and licensing examination scores. These measures were used as a proxy of long-term physician performance, yet their value for predicting such performance is unknown.

Kumar AD, Shah MK, Maley J H, Evron J, Gyftopoulos A, Miller C. Preparing to take the USMLE Step 1: A survey on medical students' self-reported study habits. Postgraduate Medical Journal. 2015;91(1075):257-261.
https://www.ncbi.nlm.nih.gov/pubmed/25910497

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined how students' self-reported study habits correlate with Step 1 scores.

How was the study conducted (i.e., what data and methodologies were used)?
A survey regarding Step 1 study habits was administered to third-year medical students at Tulane University School of Medicine from 2009-2011 (N = 256). Students were eligible for the study if they had recently completed Step 1 (within the last three months) and had already received their scores. Students were asked to self-report their Step 1 score, the number of days they prepared for the exam, the number of practice problems they completed, the percentage of study time they spent doing practice questions, the hours they spent studying per day, and the percentage of study time they spent in a group setting. Responses were categorized and a one-way ANOVA was conducted to evaluate the effects of different study habits on Step 1 scores.

What were the primary results of the study?
Hours spent studying per day, the number of days spent preparing, and the number of practice problems completed all had a significant impact on Step 1 scores. Specifically, students who reported studying for 8 or more hours, preparing for 40 days or less, and completing 2000 or more practice problems had significantly higher Step 1 scores than students who reported studying less than 8 hours, preparing for 40 or more days, and completing fewer than 2000 practice problems. Studying in a group setting had no statistically significant impact on Step 1 scores.

What are the implications of the findings?
This study examined different study habits that had not been previously examined by other studies, such as practice problem utilization, hours spent studying per day, and group studying. Findings may help to guide student study practices at both the student and school levels. Of interest is the somewhat counterintuitive finding that preparing for more than 40 days has a detrimental effect on performance.

What are the limitations of the study?
Data was collected from students at a single institution, thus findings may not be representative of the entire medical student population. The survey required that students retrospectively estimate their study habits, which may have resulted in some misremembered and poorly estimated responses.

Lee M, Vermillion M. Comparative values of medical school assessments in the prediction of internship performance. *Medical Teacher.* 2018; 1:1-6. https://www.ncbi.nlm.nih.gov/pubmed/29390938

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study investigated the ability of various assessments used in medical school to predict performance during internship.

How was the study conducted (i.e., what data and methodologies were used)?
The initial sample included archival data for 637 David Geffen School of Medicine students from the classes of 2011 to 2014 whose residency program directors were sent a survey regarding their internship performance. Hierarchical multiple regression techniques were used to compare the relationships between each medical school assessment and subsequent internship performance. Independent predictors included: USMLE Step 1 score, USMLE Step 2 CK score, NBME Medicine Subject Examination score, inpatient internal medicine clerkship performance ratings, and scores from an eight-case OSCE taken at end of the third year. Internship performance ratings, measured on a 1-3 scale, served as the dependent variable. MCAT scores were treated as a covariate to account for initial differences in academic achievement.

What were the primary results of the study?
The program director survey response rate was 75%. Step 1 score, NBME Medicine Subject Examination score, and Step 2 CK score showed the highest correlations with graduate internship ratings. In terms of predictive utility, medicine subject examination scores, OSCE scores, and Step 2 CK scores explained a statistically significant amount of variance in internship performance, above and beyond MCAT scores. Step 1 score and medicine clerkship ratings did not statistically aid predictive value when included in the model with measures of clinical sciences knowledge.

What are the implications of the findings?
Findings suggest that success in internships requires proficiency across multiple domains/skills. It seems that clinical measures of student ability best predict internship success. Step 2 CK may be a better indicator of internship performance than Step 1, given its intended purpose and focus and its closer proximity to when learners begin clinical internships. However, the final regression model accounted for only 16% of the variation in internship ratings, indicating that many other factors influence internship ratings.

What are the limitations of the study?
More subjective judgments, such as those captured by program director ratings of learners, tend to have high measurement error, making it difficult to account for variance in these types of measures. Internship ratings also appeared to have low variance and a potential ceiling effect. These factors may have contributed to the low overall amount of variance explained by the model. Also, the study was conducted with students in a single institution and results may not generalize to students at other institutions who completed a different curriculum with different assessment methods.

McGaghie WC, Cohen ER, Wayne DB. Are United States Medical Licensing Exam Step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? Academic Medicine. 2011;86(1):48–52.
https://www.ncbi.nlm.nih.gov/pubmed/21099388

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study asked the following research questions: (1) is there strong validity evidence for the use of Step 1 and Step 2 CK score for residency selection, and (2) is the validity argument for the use of these scores for residency selection structured, coherent, and evidence-based?

How was the study conducted (i.e., what data and methodologies were used)?
Using a critical review approach, this study synthesizes nine reports involving data from 393 medical students and residents across a five-year time span (2005-2010). Eight clinical skills were assessed via the measures of clinical performance (e.g., cardiac auscultation, central venous catheter insertion, communication). Correlations between Step 1 and Step 2 CK scores and reliable measures of clinical performance from the nine research reports were extracted, tabulated, and compared.

What were the primary results of the study?
Correlations between Step 1 scores and the eight assessments of clinical skills ranged from -0.05 to 0.29, with none being statistically significant. Correlations between Step 2 CK scores and the eight assessments of clinical skills ranged from -0.16 to 0.24; only one relationship (Step 2 CK and advanced cardiac life support scenarios) yielded a statistically significant result that accounts for just 5% of the variance in scores from an assessment of advanced cardiac life support scenarios. When corrected for attenuation, none of the correlations are statistically significant.

What are the implications of the findings?
This study reveals a lack of evidence to support the extrapolation and interpretation of components of a validity argument for the use of Step 1 and Step 2 CK scores for residency selection purposes. This is because scores were not found to relate to some measures of clinical skills important among advanced medical students, residents, and subspecialty fellows. The absence of such an empirical link suggests that use of USMLE Step 1 and 2 CK scores for postgraduate medical residency selection decisions may be unwarranted.

What are the limitations of the study?
Because the number of studies reviewed in this study is small and primarily from trainees at one institution, generalization of results is not guaranteed. Moreover, the number of students/residents included in the calculation of each correlation was relatively small. In addition, although a range of skills among medical trainees were reviewed, many of them were technical and procedural in nature, raising questions about the expected relationships between scores from a general licensing examination and such skills in clinical practice. The selected clinical skills measures and the assessments used to do so may have produced a selection effect that cannot be teased apart from the current study.

Marcus-Blank B, Dahlke JA, Braman JP, Borman-Shoap E, Tiryaki E, Chipman J, Andrews J, Sacket P, Cullen MJ. Predicting performance of first-year residents: Correlations between structured interview, licensure exam, and competency scores in a multi-institutional study. [EPub ahead of print August 28, 2018]. Academic Medicine.
https://www.ncbi.nlm.nih.gov/pubmed/30157088

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined the relationship between ACGME milestone ratings and scores from a structured interview (SI) of non-cognitive competencies, and whether the SI scores add incremental validity evidence above USMLE Step 1 and Step 2 CK scores.

How was the study conducted (i.e., what data and methodologies were used)?
A scorable SI for entry into residency programs was developed. Correlations were examined between these scores and faculty ratings of student milestones, along with changes in milestone variance explained by adding Step 1 and Step 2 CK scores via sequential block regression models. Data were obtained from 164 students who were interviewed using the SI, subsequently matched in a residency program, and rated using ACGME milestones.

What were the primary results of the study?
SI scores were modestly related to first-year and end-of-year milestone ratings on noncognitive competencies, including interpersonal and communication skills. Step 1 and Step 2 CK scores were modestly related to first-year medical knowledge and end-of-year patient care competencies. Step 1 scores, Step 2 CK scores, and SI scores added incremental predictive value for understanding different milestone competencies.

What are the implications of the findings?
Both cognitive (Step 1 and Step 2 CK scores) and non-cognitive (SI scores) contribute value to the prediction of resident performance as measured by milestone ratings. Depending on the milestone of focus, cognitive or non-cognitive measures may be a better predictor of future performance.

What are the limitations of the study?
Milestone ratings lack reported reliability evidence, which makes interpretation of results difficult. Additionally, programs utilized different questions from the SI and performed ratings using different sets of milestones. Although composites were created and statistically centered, the constructs across institutions may be different, perhaps meaningfully so. It is not reported how well-aligned SI raters were in their ratings, as inter-rater reliabilities are absent. Small sample sizes prevent more accurate estimates of relationships between the variables under study. Lastly, the outcomes relate non-cognitive measures to faculty ratings; if available, patient outcome data would be valuable to study as well.

Margolis MJ, Clauser BE, Winward M, Dillon GF. Validity evidence for USMLE examination cut scores: Results of a large-scale survey. Academic Medicine. 2010; 85(10): S93-97. https://www.ncbi.nlm.nih.gov/pubmed/20881714

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined whether the cut scores used to make pass/fail decisions for USMLE are appropriate.

How was the study conducted (i.e., what data and methodologies were used)?
Nine key stakeholder groups with knowledge of USMLE examinee skills were sent a survey. These stakeholders included NBME and FSMB board members, USMLE committee members, state medical board presidents and executive directors, medical school associate deans, basic science course directors, clinical clerkship directors, residency program directors, chief residents, and recent USMLE examinees. Approximately 15,000 respondents provided their opinions about initial and ultimate fail rates for Step 1, Step 2 CK, and Step 3, as well as the highest acceptable, lowest acceptable, and optimal initial and ultimate fail rates for each exam.

What were the primary results of the study?
Respondents generally viewed initial fail rates as appropriate; more variation in opinions was found for ultimate fail rates, although actual fail rates for each examination fell within the respondent-identified acceptable range. Some differences were identified between respondent groups. Most notably, residency directors identified higher initial fail rates as appropriate, whereas Step 1 examinees reported lower initial fail rates as appropriate.

What are the implications of the findings?
Findings provide some validity support for the use of current cut scores for each Step, as key stakeholder groups converge on the appropriateness of the resulting fail rates. The opinions of residency program directors and Step 1 examinees may reflect the current use of Step 1 and Step 2 CK scores for determining entry into residency programs.

What are the limitations of the study?
The survey response rate was modest (28%), with some groups having lower response rates than others.

Norcini JJ, Boulet JR, Opalek A, Dauphinee WD. The relationship between licensing examination performance and the outcomes of care by international medical school graduates. Academic Medicine. 2014; 89(8): 1157–62.
https://www.ncbi.nlm.nih.gov/pubmed/24853199

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined the relationship between USMLE Step 2 CK scores and patient outcomes (in-hospital mortality) for graduates of international medical schools (IMGs).

How was the study conducted (i.e., what data and methodologies were used)?
Using a retrospective observational study design, 60,958 hospitalizations from 2003-2009 in a single state were analyzed. The principle diagnosis in these cases was restricted to acute myocardial infarction or congestive heart failure, and the physician had to have attended a medical school outside of the US and have taken Step 1 CK (N = 2,525 physicians). Multivariate regression models were estimated where the dependent variable was in-hospital mortality and the primary independent variable was Step 2 CK score.

What were the primary results of the study?
Among IMGs, Step 2 CK scores were inversely related to in-hospital mortality. In addition, board-certification and self-designated specialty areas of family or internal medicine were associated with a lower relative risk of patient mortality.

What are the implications of the findings?
The results of this study suggest that Step 2 CK examinations have a positive relationship with patient outcomes. They also provide some validity evidence for primary and secondary uses of Step 2 CK scores.

What are the limitations of the study?
The relationships between Step 2 CK scores and patient outcomes are possibly understated – meaning that the relationships described in this study are subject to restriction of range issues. Also, this study examined only two inpatient conditions in one state, so the generalizability of the results may be limited.

Rubright JD, Jodoin M, Barone MA. Examining demographics, prior academic performance, and United States Medical Licensing Examination Scores. [Published online ahead of print July 17, 2018]. Academic Medicine.
https://www.ncbi.nlm.nih.gov/pubmed/30024473

What was primary purpose of the study or the research question(s) that the study sought to answer?
This paper sought to explore whether USMLE Step 1, Step 2 CK, and Step 3 performance varied by examinee demographics, including sex, race, citizenship, English as a second language, and age. Additionally, MCAT scores and undergraduate GPA were included as covariates to see if they altered any of the relationships.

How was the study conducted (i.e., what data and methodologies were used)?
The study used hierarchical linear models: first utilizing demographic variables only, followed by the addition of pre-medical school covariates. Data from 45,154 examinees from U.S. and Canadian allopathic and osteopathic medical were used to model Step 1, Step 2 CK, and Step 3 scores separately.

What were the primary results of the study?
Scores from all Steps showed demographic differences, although specific results varied. On average, men outperformed women on Step 1 and women outperformed men on Step 2 CK. Self-identified non-white examinees scored lower than self-identified white examinees. Adding MCAT scores and GPA to the models both explained more variance in the outcomes and either reduced the size of or eliminated the originally observed differences.

What are the implications of the findings?
Differences in USMLE scores by examinee demographic characteristics that mirror differences found in past research are present, yet attenuated by previous academic performance. Additional factors for explaining the remaining differences need to be researched. Differences in Step 1 and Step 2 CK scores may not reflect differences in Step 1 and Step 2 CK pass rates, thus results should be interpreted with respect to specific uses and interpretations of Step 1 and Step 2 CK performance.

What are the limitations of the study?
Although findings show score differences by important examinee characteristics (e.g., gender, race), the reasons for these differences are not knowable from the current analysis. Moreover, variables related to other examinee factors known or suspected to impact performance, such as study habits and ability to afford external study resources, were not included.

Sandella JM, Gimpel JR, Smith LL, Boulet JR. The use of COMLEX-USA and USMLE for residency applicant selection. Journal of Graduate Medical Education. 2016;8(3):358-63. https://www.ncbi.nlm.nih.gov/pubmed/27413438

What was primary purpose of the study or the research question(s) that the study sought to answer?
This study examined the relationship between performance, both continuous scores and categorical passing status, on COMLEX Level 1 and USMLE Step 1.

How was the study conducted (i.e., what data and methodologies were used)?
The data consisted of COMLEX and Step 1 scores for 914 students from 3 colleges of osteopathic medicine who matriculated in 2010 or 2011. Of the 914 students, 795 completed both examinations and thus were used in the analysis. Pearson correlations were used to describe the linear relationship between COMLEX and Step 1 scores. The phi coefficient was used to quantify the relationship between passing status on the two examinations. The 2x2 contingency table of passing status for both exams was also examined.

What were the primary results of the study?
COMLEX and Step 1 scores were highly related, with a correlation of 0.84 across the three schools. A scatterplot of scores showed that the linear trend held except at the high end of COMLEX scores. Examinees with high COMLEX scores did not perform as well on Step 1 as predicted by the correlation. The relationship between passing status on the two exams was moderate, with a phi correlation of 0.39. Most students passed both examinations (90%), but more students who passed COMPLEX failed Step 1 (7%) than the converse (1%).

What are the implications of the findings?
The strong association across multiple schools suggests the two exams may measure similar constructs and may both provide reasonably comparable estimates of examinee ability. However, the lack of a near perfect correlation and differences in passing status suggest that the examination scores are not equivalent. Score users should take the high relationship between scores into account, but use caution when interpreting scores from either examination for secondary uses, such as residency selection.

What are the limitations of the study?
Although the study included data from multiple schools, only three of the 29 colleges of osteopathic medicine met the study criteria and volunteered to provide data. There may also be selection bias in both the schools who participated as well as the type of student who completes both COMLEX and USMLE Step 1. Finally, there may be other factors moderating the relationship between examination scores, such as when during the curriculum each examination was taken. The linear relationship between these examinations also may be affected by each examination's scaling procedures which may help to explain the differences seen at the score extremes.