Section 3: USMLE Step 2 CS Score Uses and Interpretations

Section Overview

This section includes summaries of select articles published between 2006 and 2016 that focus on USMLE Step 2 CS performance. These studies examine various validity issues as they relate to Step 2 CS total and subcomponent score uses and interpretations. It is important to note that Step 2 CS has changed since some of these studies were conducted. Thus, the findings presented are based on older data and may not be entirely reflective of current patterns.

Research Summaries and Abstract Links

Cuddy MM, Winward ML, Johnston MM, Lipner RS, Clauser BE. Evaluating validity evidence for USMLE Step 2 Clinical Skills data gathering and data interpretation scores: Does performance predict history-taking and physical examination ratings for first-year internal medicine residents? Academic Medicine. 2016;91(1): 133–139.
https://www.ncbi.nlm.nih.gov/pubmed/26397703

What was the primary purpose of the study or the research question(s) that the study sought to answer?
This study examined the relationships between USMLE Step 2 CS scores and performance in internal medicine residency as indicated by American Board of Internal Medicine (ABIM) residency program director ratings. Specifically, it focused on the data gathering and data interpretation subcomponents of Step 2 CS and performance on similar tasks using similar skills in internal medicine residency.

How was the study conducted (i.e., what data and methodologies were used)?
Step 2 CS data gathering and data interpretation scores and ABIM program director ratings of residents' history-taking and physical examination skills for 6,306 internal medicine residents who took Step 2 CS in 2005 were examined using hierarchical linear models. Residency ratings were from the end of the first year of residency. Covariates at the individual level included Step 1 score, Step 2 CK score, and demographic variables, and at the program level they included selectivity and size.

What were the primary results of the study?
Two-thirds of the variation in both history-taking and physical examination ratings occurred within programs (between residents), with the remaining third occurring between programs. Final models showed that Step 2 CS data interpretation scores were modestly positively related to both history-taking and physical examination ratings, after controlling for covariates. Step 2 CS data gathering scores were unrelated to residency ratings.

What are the implications of the findings?
Findings provide some validity evidence for the current use and interpretation of Step 2 CS data interpretation scores. Data gathering scores did not add predictive value above and beyond the information provided by the other factors in the model and thus less validity evidence was found for the use and interpretation of Step 2 CS data gathering scores. This lack of relationship may reflect test-taking strategies, or may be indicative of how different residency programs approach and teach the clinical encounter. It is important to note that the assessment of history-taking is now captured via the Step 2 CS data interpretation score, rather than the data gathering score. The results of this study provide support for this enhancement.

What are the limitations of the study?
Scores were provided at different times during an examinee's medical education and training, and different types and rates of learning may have happened between the first Step 2 CS attempt and later performance in residency. Additionally, program director ratings do not have published reliability estimates, and may suffer from low reliability, low inter-rater reliability, or ceiling effects. These issues may have contributed to the relationships found. Also, while this study examined program director ratings that were provided to the ABIM, future research may want to consider ACGME milestone ratings as more current measures of residency performance.

Harik P, Clauser BE, Grabovsky I, Margolis MJ, Dillon GF, Boulet JR. Relationship among subcomponents of the USMLE Step 2 Clinical Skills Examination, the Step 1, and the Step 2 Clinical Knowledge examinations. Academic Medicine. 2006;81(10 Suppl):S21-S24. https://www.ncbi.nlm.nih.gov/pubmed/17001128

What was the primary purpose of the study or the research question(s) that the study sought to answer?

This study examined the relationships among USMLE Step 1, Step 2 CK, and Step 2 CS scores. It also examined whether or not individuals who fail Step 2 CS also fail Step 2 CK.

How was the study conducted (i.e., what data and methodologies were used)?

Data included two samples of first-time takers: students from US medical schools (n=15,800) and students from international medical schools (n=12,300). These students took the Step 2 CS examination between 2004 and 2005. Step 2 CS total and subcomponent scores were analyzed. Analysis included correlations (corrected for unreliability) and cross-tabulations.

What were the primary results of the study?

For both groups of students, Step 2 CS patient note scores were highly correlated with Step 2 CS data gathering scores. Step 2 CS communication and interpersonal skill scores and Step 2 CS data gathering scores were moderately correlated. For graduates of international medical schools, Step 2 CS spoken English proficiency scores were more highly correlated with other Step 2 CS subcomponent scores, likely due to a ceiling effect for scores from this subcomponent for students from US medical schools. For both student groups, Step 2 CS subcomponent scores were weakly related to Step 1 and Step 2 CK scores. Cross-tabulations of fail rates for Step 2 CS and Step 2 CK showed little overlap.

What are the implications of the findings?

The low correlations between Step CS subcomponent scores and Step 1 and Step 2 CK scores, along with the small overlap in fail rates, suggests that Step 2 CS measures constructs different from those assessed by Step 1 and Step 2 CK.

What are the limitations of the study?

This study focused on descriptive statistics, but other types of analyses that account for possible confounding factors may yield different results.

Swygert KA, Cuddy MM, Van Zanten M, Haist SA, Jobe AC. Gender differences in examinee performance on the Step 2 Clinical Skills data gathering (DG) and patient note (PN) components. Advances in Health Sciences Education: Theory and Practice. 2012;17(4):557-571. https://www.ncbi.nlm.nih.gov/pubmed/22041870

What was the primary purpose of the study or the research question(s) that the study sought to answer?

This study assessed whether gender differences exist in scores from the data gathering and patient note subcomponents of USMLE Step 2 CS, and whether the gender of the standardized patients encountered by examinees during the examination had an interaction effect with an examinee's gender.

How was the study conducted (i.e., what data and methodologies were used)?

The sample included 27,901 examinees who took Step 2 CS for the first time in 2009. Descriptive statistics were computed and hierarchical linear models were estimated.

What were the primary results of the study?

Unadjusted analyses showed that female examinees outperformed male examinees on the data gathering and patient note subcomponents of Step 2 CS. Differences were smaller, yet still present, once covariates were added to the models. No interaction effects were found between examinee and standardized patient gender with respect to scores.

What are the implications of the findings?

Findings mirror those from the literature that suggest that female physicians outperform males with respect to communication skills. The examinee gender effect is small, suggesting that other factors may be more influential for understanding examinee performance. Also, findings suggest that Step 2 CS data gathering and patient note scores were not impacted by the combination of an examinee's gender and the gender of the standardized patients encountered during the examination, thus providing some validity evidence for the current use and interpretation of Step 2 CS scores.

What are the limitations of the study?

The study does not examine gender within the context of other variables, such as previous academic performance, which may have implications for scores on the Step 2 CS subcomponents examined.

Winward ML, Lipner RS, Johnston MM, Cuddy MM, Clauser BE. The relationship between communication scores from the USMLE Step 2 Clinical Skills examination and communication ratings for first-year internal medicine residents. Academic Medicine. 2013;88(5):693-698. https://www.ncbi.nlm.nih.gov/pubmed/23524927

What was the primary purpose of the study or the research question(s) that the study sought to answer?

This study examined the relationship between Step 2 CS communication and interpersonal skills performance and subsequent evaluation of communication skills in internal medicine residency, as indicated by American Board of Internal Medicine (ABIM) residency program director ratings.

How was the study conducted (i.e., what data and methodologies were used)?

This study used the same sample and similar methods as the Cuddy et al. (2016) study described above. Step 2 CS communication and interpersonal skills scores were used to predict ABIM program director ratings of communication skills, controlling for individual- and program-level covariates. Like in the Cuddy et al (2016) study, a total of 6,306 residents from 238 internal medicine programs who took Step 2 CS for the first time in 2005 were analyzed using hierarchical linear models.

What were the primary results of the study?

Two-thirds of the variation in communication ratings occurred within programs (between residents); the remaining third of the variation in ratings occurred between programs. Final models show that Step 2 CS communication and interpersonal skills scores were modestly positively related to communication ratings as provided by residency program directors, after controlling for individual- and program-level covariates.

What are the implications of the findings?

Finding provide some validity evidence for the current use and interpretation of Step 2 communication and interpersonal skills scores, as they demonstrate a relationship between scores and later measures of performance in practice.

What are the limitations of the study?

Step 2 CS communication and skills scores may show a ceiling effect, as failing candidates may not have continued in the USMLE sequence and therefore not matched into a residency program. Again, program director ratings may have low reliability estimates, limiting the relationships that can be demonstrated with them. Like the Cuddy et al. (2016) paper, this study examined program director ratings that were provided to the ABIM, but future research may want to consider ACGME milestone ratings as more current measures of residency performance.