The Impact of Postgraduate Training on USMLE® Step 3® and its Computer-Based Case Simulation Component

Richard A. Feinberg, MA^{1,2}, Kimberly A. Swygert, PhD¹, Steven A. Haist, MD MS¹, Gerard F. Dillon, PhD¹, and Constance T. Murray, RN MSN¹

¹National Board of Medical Examiners, Philadelphia, PA, USA; ²University of Delaware, School of Education, Newark, DE, USA.

BACKGROUND: The United States Medical Licensing Examination® (USMLE®) Step 3® examination is a computer-based examination composed of multiple choice questions (MCQ) and computer-based case simulations (CCS). The CCS portion of Step 3 is unique in that examinees are exposed to interactive patient-care simulations.

OBJECTIVE: The purpose of the following study is to investigate whether the type and length of examinees' postgraduate training impacts performance on the CCS component of Step 3, consistent with previous research on overall Step 3 performance.

DESIGN: Retrospective cohort study

PARTICIPANTS: Medical school graduates from U.S. and Canadian institutions completing Step 3 for the first time between March 2007 and December 2009 ($n=40,588$).

METHODS: Post-graduate training was classified as either broadly focused for general areas of medicine (e.g. pediatrics) or narrowly focused for specific areas of medicine (e.g. radiology). A three-way between-subjects MANOVA was utilized to test for main and interaction effects on Step 3 and CCS scores between the demographic characteristics of the sample and type of residency. Additionally, to examine the impact of post-graduate training, CCS scores were regressed on Step 1 and Step 2 Clinical Knowledge (CK) scores. Residuals from the resulting regressions were plotted.

RESULTS: There was a significant difference in CCS scores between broadly focused ($\mu=216$, $\sigma=17$) and narrowly focused ($\mu=211$, $\sigma=16$) residencies ($p<0.001$). Examinees in broadly focused residencies performed better overall and as length of training increased, compared to examinees in narrowly focused residencies. Predictors of Step 1 and Step 2 CK explained 55% of overall Step 3 variability and 9% of CCS score variability.

CONCLUSIONS: Factors influencing performance on the CCS component may be similar to those affecting Step 3 overall. Findings are supportive of the validity of the Step 3 program and may be useful to program directors and residents in considering readiness to take this examination.

KEY WORDS: USMLE; Step 3; computer-based case simulation (CCS); postgraduate training; graduate medical education.

J Gen Intern Med 27(1):65–70

DOI: 10.1007/s11606-011-1835-1

© Society of General Internal Medicine 2011

INTRODUCTION

The United States Medical Licensing Examination® (USMLE®) Step 3® examination is a computer-based examination composed of multiple choice questions (MCQ) and computer-based case simulations (CCS). Step 3 is the last in a series of examinations representing the pathway to medical licensure in the United States for individuals who have obtained an M.D. or equivalent degree, and “assesses if a physician possesses and can apply the medical knowledge and understanding of clinical science considered essential for the unsupervised practice of medicine, with emphasis on patient management in ambulatory care settings.”¹ On the CCS portion of Step 3, examinees are exposed to interactive patient-care simulations where they receive patient status feedback and manage simulated time in which the case unfolds.

Previous research indicates that Step 3 scores are related to the type and length of postgraduate training; specifically, examinees in residency programs who encounter a variety of patients under a variety of conditions (broadly focused programs) obtain higher scores than those in programs with more limited exposure to a variety of patients and conditions (narrowly focused programs), and that those in broadly focused residencies show score increases when Step 3 is taken later in residency training.^{2–4}

OBJECTIVE

The purpose of this study is to update previous investigations and to extend the research by including separate analyses for CCS. As type and length of postgraduate training has already been shown to be related to performance on Step 3 as a whole, it is hypothesized that this association will generalize to CCS. Therefore, the current study will investigate the effect of examinees' residency specialty and length of postgraduate training on the performance on the CCS component of Step 3, in addition to replicating previous research.

Received September 13, 2010

Revised June 17, 2011

Accepted August 2, 2011

Published online August 31, 2011

Understanding the relationship between CCS and postgraduate training would be useful to both residents and program directors interested in factors that impact Step 3 performance. The CCS component provides an assessment format which more closely mirrors clinical encounters and patient management than does the multiple choice format. With the recent growing concern for safety and quality of patient care, performance on this patient simulation format seems worthy of further exploration.⁵⁻⁷ Furthermore, a demonstration of a similar relationship between CCS and length of postgraduate training, provides validity for the CCS component and can support additional uses beyond a contribution to the total test score.

Additionally, the effect of postgraduate training on CCS and Step 3 overall may provide useful information for the current national debate on how to redesign internal medicine residency training and the optimal length of training.⁸⁻¹⁰ For instance, there has been an ongoing debate about the necessary length of internal medicine residency; with some support to decrease the length of training to allow earlier differentiation of residents into future practice (cardiology fellowship, hospitalist, ambulatory specialist).^{11,12} Alternatively, there has been support for increasing the length of internal medicine residency because of perceived lack of competency to begin practice in certain settings.¹³

METHOD

Data

A group of 40,588 medical school graduates from U.S. and Canadian institutions completing Step 3 for the first time between March 2007 and December 2009 were chosen. All examinees had passed all previous Step 1 and Step 2 examinations, had graduated between 2000 and 2009, and completed fewer than 85 months of postgraduate training when taking Step 3. The median testing year for Step 1 first attempt was 2005, and the median testing year for Step 2 Clinical Knowledge (CK) first attempt was 2006. Demographic information was self-reported with 50% of the examinees indicating male as their gender, and 91% reported English as their native language.

Each category of residency training program identified in the study had at least 470 examinees, and was classified as having either a broad or narrow focus. Broadly focused residencies are more aligned with the Step 3 content blueprint and concentrate on general areas of medicine, while narrowly focused residencies concentrate more on specific areas of medicine. The broadly focused residencies were combined medicine-pediatrics, emergency medicine, family medicine, internal medicine, and pediatrics. The narrowly focused residencies were anesthesiology, obstetrics-gynecology, pathology, psychiatry, radiology, surgery, and surgery subspecialties (composed of neurological surgery, orthopedic surgery, otolaryngology, plastic surgery, and urology). For the purposes of this study, surgery subspecialties were combined into a composite specialty in order to increase sample sizes.

Since the USMLE program does not capture verified data on type and amount of residency training completed, self-reported

data must be used. For purposes of this study, length of training was calculated by subtracting the self-reported residency start date from the Step 3 test date. Although this is a less than ideal measure of length of training because it cannot account for whether there were significant interruptions in training or whether the training was completed well in advance of Step 3 testing, it is felt that this variable is adequate for purposes of the study. Most examinees will take Step 3 during residency to prevent a delay in continuing their career growth in pursuit of licensure. The vast majority of the U.S. and Canadian medical school graduates used in the study reported having started residency training within three years of their Step 3 test date, making it unlikely that disruptions in training or early completion will have a meaningful impact on the general trends under investigation.

Study participants were told, upon registration for USMLE, that their de-identified data may be used for research. Analytic data sets and derivative analyses were stored securely under conditions that preserve privacy and prevent release to any third party. The American Institutes for Research (AIR) Institutional Review Board (IRB) conducted a retrospective review of this manuscript, and "...based on the submitted materials and correspondence with AAMC, the IRB has determined that ..."Effect of Postgraduate Training on the USMLE" would have been determined to be approved had the IRB been asked to review it prospectively."

Statistical Analyses

All analyses were on the reported score scale for the overall Step 3 (MCQ and CCS combined) and CCS scores, although numeric values are not reported, CCS score profiles were reported to examinees. A 2 (residency focus) x 2 (gender) x 2 (native language) multiple analysis of variance (MANOVA) was calculated on Step 3 overall and CCS scores as dependent variables to investigate whether there were meaningful differences in CCS scores between the broadly and narrowly focused residencies, and to explore the demographic context of the sample. Next, two ordinary least squares (OLS) multivariate regression analyses were conducted, one with performance on Step 3 overall as the dependent variable, and the other with performance on the CCS component of Step 3 as the dependent variable. The analysis with total Step 3 score was intended to replicate previous findings as well as allow a comparison with the CCS results from this study.

The predictor variables for both regression analyses were first-attempt scores on Step 1 and Step 2 CK, where those scores are covariates to control for differences in examinee performance at the onset of postgraduate training. The regression analyses were run on a baseline group of examinees defined as family medicine residents who completed between 6 and 18 months of training ($n=2,532$). This group was chosen because family medicine curricula are most likely to mirror the broad-based content outline of Step 3. The resulting regression equations were then used to compute predicted Step 3 and CCS scores for all examinees in the data set. To determine the influence of the focus of residency specialty (broad or narrow) and length of training, unstandardized residuals were calculated as the difference between the actual and predicted scores. Step 3 overall and CCS residuals were compared for

each residency specialty to illustrate whether both are similarly impacted by the type and length of postgraduate training. Finally, pass/fail outcomes were analyzed to assess if the differences in performance associated with residency type and length of training had an impact at the decision point.

RESULTS

Descriptive statistics for reported Step 1, Step 2 CK, Step 3 overall, and CCS scores by residency focus and specialty are shown in Table 1. Table 2 contains the Step 3 overall and CCS descriptive statistics by examinee demographic characteristics and residency type. The tables show that narrowly focused residents had similar Step 2CK scores and slightly higher Step 1 scores by 0.09 standard deviations as compared to the broadly focused residents.

The number of examinees testing by length of post-graduate training for both broadly and narrowly focused residencies is plotted in Figure 1. For the overall sample, 55% of examinees tested with 6 to 18 months of training, 93 % tested within 36 months of training, and 99% tested within 60 months of training. The patterns of when examinees are testing for broadly and narrowly focused residencies were similar, generally reflecting the relative differences in overall counts, until around 36 months. After this point the counts are relatively small and slightly higher for narrowly focused residencies, which is not surprising as most broadly focused residencies are 3 years in duration while narrowly focused residencies are longer, closer to 4 and 5 years.

The three-way between-subjects MANOVA showed a significant difference in CCS scores for broadly focused ($\mu=216$, $\sigma=17$) and narrowly focused ($\mu=211$, $\sigma=16$) residencies; ($F=257.404$, $df [1, 40,580]$, $p<0.001$, $\eta_p^2=.006$). These results suggest that examinees training in broadly focused residencies perform better on CCS. Similarly, and consistent with previous research, there was also a significant difference in Step 3 overall scores for broadly focused ($\mu=218$, $\sigma=16$) and narrowly focused ($\mu=209$, $\sigma=16$) residencies; ($F=776.233$, $df [1,$

40,580], $p<0.001$, $\eta_p^2=.019$). Focus of residency accounted for approximately 1% of the variance in CCS scores and approximately 2% of the variance in Step 3 overall scores, representing small effects based on effect size standards in the literature.¹⁴

Examinees who self-reported English as their native language ($\mu=215$, $\sigma=17$) outperformed those who did not ($\mu=208$, $\sigma=16$) for Step 3 overall scores ($F=724.466$, $df [1, 40,580]$, $p<0.001$, $\eta_p^2=.018$). Native language explained approximately 2% of the variance in Step 3 overall scores indicating a small effect.

Significant differences were also found between examinees who self-reported English as their native language ($\mu=214$, $\sigma=16$) and those who did not ($\mu=213$, $\sigma=16$) for CCS scores ($p<0.001$) and between examinees who self-reported their gender as female ($\mu=216$, $\sigma=17$) compared to those who indicated male ($\mu=214$, $\sigma=17$) for Step 3 overall scores ($p<0.001$). However, both these findings had relatively no effect ($\eta_p^2=.001$), accounting for only 0.1% of the variance. Therefore, significance is likely due to the large sample size rather than a practical difference in performance. All remaining main effect and interactions were non-significant.

Step 1 and Step 2 CK scores explained 55% of the variability in Step 3 overall and 9% of the variability in CCS scores. After coefficients from the baseline models were applied to all examinees in the data set, predicted scores and unstandardized residuals were calculated. Plots of the unstandardized Step 3 and CCS residuals by months of postgraduate training for each residency specialty are shown in Figure 2. Positive residuals indicate better-than-predicted performance, while negative residuals represent lower-than-predicted performance, as compared to the baseline family medicine group.

The figure shows that for both Step 3 and CCS, the broadly focused residency scores improved as length of postgraduate training increased, whereas narrowly focused residency scores remained relatively flat and indifferent to length of postgraduate training. Anesthesiology and psychiatry residencies displayed a negative trend as postgraduate training increased in the CCS plot, suggesting that these examinees are performing further below what is expected. This may be the result of these examinees becoming further separated from medical school curriculum while concurrently receiving training that is not as congruent

Table 1. Descriptive Statistics for Reported Scores by Residency Specialty

Focus	Residency Specialty	N	Step 1		Step 2CK		Step 3*		CCS†	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
Broad	Combined medicine-pediatrics	839	220	21	227	23	226	16	219	17
	Emergency medicine	3,034	219	19	226	21	220	15	218	17
	Family medicine	3,810	207	22	215	24	216	16	214	16
	Internal medicine	12,421	222	22	227	23	219	16	216	17
	Pediatrics	5,191	215	21	224	23	215	16	213	16
	Total	25,295	218	22	224	23	218	16	216	17
Narrow	Anesthesiology	1,556	217	19	222	21	209	16	211	16
	Obstetrics-gynecology	2,536	212	21	222	22	210	16	210	15
	Pathology	927	221	22	224	23	207	17	210	16
	Psychiatry	1,952	208	22	213	24	206	17	211	17
	Radiology	498	231	18	230	21	215	16	213	17
	Surgery	4,238	220	21	223	22	208	16	210	15
	Surgery subspecialties	3,586	233	17	232	20	212	15	211	15
	Total	15,293	220	22	224	23	209	16	211	15

* Step 3 reported scores include both MCQ and CCS components

† CCS scores are reported in a profile format; separate numerical scores are not reported

Table 2. Step 3 Overall Scores, Pass Rates, and CCS Scores by Residency Focus, Gender, and Native Language

Group	Level	N	Step 3 [*]			CCS [†]	
			Mean	SD	% Pass	Mean	SD
Residency Focus	Broad	25,295	218	16	97	216	17
	Narrow	15,293	209	16	93	211	15
Gender	Male	20,447	214	17	95	213	16
	Female	20,141	216	17	96	214	16
Native Language	English	36,927	215	17	96	214	16
	Non-English	3,661	208	16	92	213	16

* Step 3 reported scores include both MCQ and CCS components

† CCS scores are reported in a profile format; separate numerical scores are not reported

with the broad-based focus of Step 3. There is a difference in scale between the Step 3 and CCS residuals, with a greater spread for Step 3. However, the broadly and narrowly focused residencies in both plots demonstrate similar trends reflecting a consistent impact of the type and length of postgraduate training regardless of whether using Step 3 overall or CCS scores.

Step 3 pass rates by length of postgraduate training and residency type are plotted in Figure 3. During the initial months of postgraduate training, little difference is observed between pass rates for the broadly and narrowly focused residencies. However, as length of training begins to increase, narrowly focused residencies begin to show a decrease in pass rate that is not matched by the broadly focused residencies until after 36 months of postgraduate training had been completed prior to testing.

DISCUSSION

The purpose of this study was to extend previous research on the impact of postgraduate training on Step 3, specifically to the CCS component of the examination. The results show a similar relationship between residency specialties for both Step 3 overall and CCS, with higher scores

associated with examinees who receive training from broadly focused residencies. As length of postgraduate training increases, broadly focused residency examinees performed increasingly better, which is likely due to the residencies' more diverse and exam-relevant curriculum compared to their narrowly focused residency counterparts. For example, family medicine residents will have substantial clinical experiences in adult medicine, emergency medicine, pediatrics, obstetrics and gynecology and surgery as well as a minimum of 1650 patient visits in their longitudinal ambulatory clinic with exposure to patients with a variety of mental disorders.¹⁵ Thus, family medicine residents' curriculum includes diverse clinical experiences in all of the disciplines covered on Step 3 including an emphasis on ambulatory medicine.

Performing better, exemplified by larger positive residuals, indicates that examinees are being underpredicted by the regression equations; they are scoring better than expected compared to the baseline group. All broadly focused residencies begin with negative residuals early in training (overpredicting) and then become positive, at different crossing points later in training (underpredicting). This result illustrates how training in the broadly focused residencies can improve scores for both Step 3 overall and CCS.

MANOVA results suggested statistically significant differences between type of postgraduate training for both Step 3 overall and CCS and that those differences represent small effects. The effect size indicates magnitude of the results, and can help identify statistically significant findings mainly due to large samples and not meaningful differences such as the findings between native language and CCS scores and between gender and Step 3 overall scores. The other statistically significant result, with a small effect, was between native language and Step 3 overall scores. Examinees who spoke English as their native language scored higher than those who did not. This is a somewhat intuitive result, as the non-native English language examinees may have additional challenges in understanding and managing the content of Step 3.

Additionally, an analysis of Step 3 pass rates by length of postgraduate training and residency type demonstrated a practical significance. As length of postgraduate training increased, Step 3 pass rates for narrowly focused residencies decreased. This result is similar to the performance decrease

Figure 1. Number of Step 3 examinees by length of postgraduate training and residency type.

Figure 2. Plots of Step 3 and CCS residuals by months of training for each residency specialty.

Figure 3. Impact on Step 3 pass rates by length of postgraduate training and residency type.

observed through residuals from the regression analysis. This result reinforces the importance of recognizing the factors which may influence performance on the Step 3 program.

Two limitations of these findings must be noted. First, there was a relatively low amount of CCS variance explained in the regression, which may be a function of lower reliability for CCS compared to Step 3 overall. Examinees complete relatively few CCSs compared to the number of MCQs on Step 3, which contributes to the lower reliability. The CCS residuals were smaller than for Step 3 overall, but it is encouraging that the trends in residency specialty over months of postgraduate training were similar between the two. A second limitation is the accuracy of the residency specialties and start date information used to calculate months of training. These values were self-reported by examinees, and examinees could switch residencies or complete a preliminary residency before entering the specialty they were in at the time of completing the Step 3 application. Furthermore, anticipated future residencies would influence the fourth-year medical students electives, potentially adversely influencing their ability to perform at a high level on Step 3.

Future research should assess if these findings generalize to international medical graduates, a more heterogeneous group representing approximately 40% of the Step 3 examinees.¹⁶ Additionally, future research could investigate the impact of Step 1 and Step 2CK on residency selection. Descriptive statistics in Table 1 shows greater variation in residency means for Step 1 compared to Step 2CK, suggesting that Step 1 may play a more critical role in discriminating students by preferred residency, or illustrate how Step 1 could be used as a screening tool for particularly competitive residencies.

In summary, the results from this study contribute to the understanding of the relationship between CCS performance and postgraduate training. The findings provide evidence for the validity of the Step 3 program overall and for its CCS component because of the clear correspondence between higher performance on the broad-based content of Step 3 and the breadth of experience received in postgraduate training. The findings are useful to both program directors and future examinees interested in how their postgraduate experience could impact their scores. Additionally, this study may influence the advice residency programs directors give to their residents. For instance, a program director for a medicine-pediatrics residency may suggest to an incoming intern with borderline Step 1 and Step 2 CK scores to wait to take Step 3 during their first half of their fourth year (the residuals between months 36–42 are +10 to +12).

These findings can also inform the debate of optimal residency length. The regression results indicate that broadly focused residents perform better than expected as training increases, which could support maintaining or increasing length of training. On the other hand, with 75% of broad-based residents in this sample testing with less than 24 months of training and with the high Step 3 pass rate, one may question the necessity of additional training in relation to licensure requirements. Further research on this debate is warranted, particularly weighing the knowledge

gained and residency experiences beyond 24 months versus the cost to examinees, institutions and society of keeping medical graduates in training.

Acknowledgment: The authors would like to thank Polina Harik of the National Board of Medical Examiners for her support and input on this project. Part of this work was presented at the 2011 Annual Meeting of the American Educational Research Association.

Conflict of Interest None disclosed.

Corresponding Author: Richard A. Feinberg, MA; National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, USA (e-mail: rfeinberg@nbme.org).

REFERENCES

1. Step 3 Content Description and General Information. Available at: http://www.usmle.org/examinations/step3/step3_content.html. Accessed August 2, 2011.
2. Andriole DA, Jeffe DB, Hageman HL, Whelan AJ. What predicts USMLE Step 3 performance? *Acad Med*. 2005;80(10):S21–4.
3. Dillon GF, Henzel TR, Walsh WP. The impact of postgraduate training on an examination for medical licensure. In: Scherpbier AJJA, Van der Vleuten CPM, Rethans JJ, Van der Steeg AFW, eds. *Advances in Medical Education*. Dordrecht, the Netherlands. Kluwer Academic Publishers: Dordrecht, the Netherlands; 1997:146–8.
4. Sawhill AJ, Dillon GF, Ripkey DR, Hawkins RE, Swanson DB. The impact of postgraduate training and timing on USMLE Step 3 performance. *Acad Med*. 2003;78(10):S10–2.
5. Wenghofer E, Klass D, Abrahamowicz M, et al. Doctor scores on national qualifying examinations predict quality of care in future practice. *Med Educ*. 2009;43:1166–73.
6. Pingleton SK, Horak BJ, Davis DA. Is there a relationship between high-quality performance in major teaching hospitals and residents' knowledge of quality and patient safety? *Acad Med*. 2009;84(11):1510–5.
7. Harik P, Cuddy MM, O'Donovan S, Murray CT, Swanson DB, Clauser BE. Assessing potentially dangerous medical actions with the computer-based case simulation portion of the USMLE Step 3 examination. *Acad Med*. 2009;84(10):S79–82.
8. Gustin W, Batra R, Amin A, Rucker L. Education first: Reforming the first-year curriculum of the internal medicine residency. *Acad Med*. 2009;84(3):368–73.
9. Meyers FJ, Weinberger SE, Fitzgibbons JP, Glassroth J, Duffy FD, Clayton CP. Redesigning residency training in internal medicine: The consensus report of the Alliance for Academic Internal Medicine education redesign task force. *Acad Med*. 2007;82(12):1211–9.
10. Thomas KG, West CP, Popkave C, Weinberger SE, Kolars JC. Internal medicine resident perceptions of optimal training duration. *Acad Med*. 2007;82(10):996–9.
11. Fuster V, Hirshfeld JW, Brown AS, et al. Working group 8: Defining the different types of cardiovascular specialists and developing a new model for training general clinical cardiologists. *J Am Coll Cardiol*. 2004;44:267–71.
12. Huddle TS, Heudebert GR. Internal medicine training in the 21st century. *Acad Med*. 2008;83(10):910–5.
13. The Hospitalist: Should Internal-Medicine Residency Training be Extended? Available at: http://www.the-hospitalist.org/details/article/1044663/Should_Internal-Medicine_Residency_Training_Be_Extended.html. Accessed August 2, 2011.
14. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.; 1988.
15. ACGME Program Requirements for Graduate Med Educ in Family Medicine. Available at http://www.acgme.org/acWebsite/downloads/RRC_progReq/120pr07012007.pdf. Accessed August 2, 2011.
16. NBME 2010 Annual Report. Available at: <http://www.nbme.org/Publications/index.html#annrep/2010%20Annual%20Report.pdf>. Accessed August 2, 2011.