2023, Volume 55, Issue 3, 171-179, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

The Association Between Length of Training and Family Medicine Residents' Clinical Knowledge: A Report From the Length of Training Pilot Study

Patricia A. Carney, PhD, MS[a]; Steele Valenzuela, MS[a]; Annie Ericson, MA[a]; Lars Peterson, MD, PhD[b]; Dang H. Dinh, MS[a]; Colleen M. Conry, MD[c]; James C. Martin, MD[d]; Karen B. Mitchell, MD[e]; Stephanie E. Rosener, MD[f]; Miguel Marino, PhD[a]; M. Patrice Eiff, MD[a]

AUTHOR AFFILIATIONS:

[a] Oregon Health & Science University, Portland, OR

[b] American Board of Family Medicine, Lexington, KY

[c] University of Colorado, Denver, CO

[d] Long School of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX

[e] American Academy of Family Physicians, Leawood, KS

[f] United Family Medicine Residency, St Paul, MN

CORRESPONDING AUTHOR:

Patricia A. Carney, Oregon Health & Science University, Portland, OR, carneyp@ohsu.edu

HOW TO CITE: Carney PA, Valenzuela S, Ericson A, et al. The Association Between Length of Training and Family Medicine Residents' Clinical Knowledge: A Report From the Length of Training Pilot Study. *Fam Med.* 2023;55(3):171-179.
doi: 10.22454/FamMed.2023.427621

ABSTRACT

Background and Objective: The associations between training length and clinical knowledge are unknown. We compared family medicine in-training examination (ITE) scores among residents who trained in 3- versus 4-year programs and to national averages over time.

Methods: In this prospective case-control study, we compared the ITE scores of 318 consenting residents in 3-year programs to 243 who completed 4 years of training between 2013 through 2019. We obtained scores from the American Board of Family Medicine. The primary analyses involved comparing scores within each academic year according to length of training. We used multivariable linear mixed effects regression models adjusted for covariates. We performed simulation models to predict ITE scores after 4 years of training among residents who underwent only 3 years of training.

Results: At baseline postgraduate year-1 (PGY1), the estimated mean ITE scores were 408.5 for 4-year programs and 386.5 for 3-year programs, a 21.9 point difference (95% CI=10.1–33.8). At PGY2 and PGY3, 4-year programs scored 15.0 points higher and 15.6 points higher, respectively. When extrapolating an estimated mean ITE score for 3-year programs, 4-year programs would still score 29.4 points higher (95% CI=15.0–43.8). Our trend analysis revealed those in 4-year programs had a slightly lesser slope increase compared to 3-year programs in the first 2 years. Their drop-off in ITE scores is less steep in later years, though these differences were not statistically significant.

Conclusions: While we found significantly higher absolute ITE scores in 4 versus 3-year programs, these increases in PGY2, PGY3 and PGY4 may be due to initial differences in PGY1 scores. Additional research is needed to support a decision to change the length of family medicine training.

INTRODUCTION

In-training examination (ITE) scores provide formative assessments of residents' progression toward developing clinical knowledge needed to practice independently, and provide residency programs with comparative data to help determine if a program is meeting its educational objectives.[1] In family medicine, the ITE has been found to be predictive of performance on the American Board of Family Medicine (ABFM) Certification Examination,[2] also reported in other disciplines.[3,4] Studies that have examined factors that predict ITE scores have found that being married and having higher prior examinations scores (eg, United States Medical Licensing Exam Step 1 and Step 2) were predictive of higher ITE scores,[5–8] while having a high debt load, being an international medical school graduate, having trained in an osteopathic versus allopathic program, and underrepresented race/ethnicity were predictive of lower ITE scores.[9,10]

Studies of ITE in family medicine have examined how predictive scores are when taken in the first year of residency compared to the second year[11] or beyond,[2] the impact of educational interventions for at-risk residents,[12] and how educational innovations in residency training affected ITE scores.[13] O'Neill et al[2] specifically examined how resident performance on the ITE differed over time and found that exam scores tend to increase annually, though the average increase lessens in each successive year. Shokar examined an educational intervention

for at-risk residents, which increased ITE scores, though not statistically,[12] and Waller, et al examined the impact of educational innovations as part of the Preparing the Personal Physician for Practice project, which found that residency education redesign did not negatively affect ITE scores.[13]

The Length of Training Pilot Study (LoTP) in family medicine has as one of its research questions, "What associations exist between length of training and residents' clinical knowledge?"[14] To address this question, we partnered with the ABFM to examine ITE scores of residents undertaking 3 versus 4 years of training at an LoTP site to explore the hypothesis that no significant differences in clinical knowledge scores would be found among residents who underwent 3 versus 4 years of training.

METHODS

The Length of Training Pilot

The LoTP is a mixed-methods prospective case-control pilot study running from 2013-2023 designed to assess several associations between the length of residency training in family medicine and learner outcomes, such as scope of practice, preparedness for independent practice and clinical knowledge.[14] Residency programs that had already transitioned to 4 years of training or that were planning to do so applied for the pilot in 2012. Those selected included six civilian programs and four Navy programs. The 4-year (4YR) civilian programs were matched to 3-year programs (3YR) based on region, size, and continuity clinic setting. Because of the large size of one 4YR program, two 3YR programs were matched to it to ensure equivalent numbers of residents in 3YR and 4YR groups.

A total of 17 residency programs, all in good standing with the Accreditation Council for Graduate Medical Education and who agreed to participate in required evaluation activities, were selected to participate (seven 3YR civilian programs, six 4YR civilian programs, and four Navy programs). We excluded Navy programs in these analyses because their training setting and content differs from civilian programs. The 4YR programs included two university-based programs, which were located at and administered by Universities that include a medical school as well as residency programs. It also included four community-based programs, which are sponsored by their local hospitals but have an affiliation with medical schools in their region. They ranged in size from six to 22 residents per year. Four of the six 4YR programs required 4 years of training for all graduates, while two offered an optional fourth year of training where residents knew at the time of entry to the program that completing a fourth year was possible. Alterations in curriculum varied in the programs undertaking 4 years of training. 3YR programs included two that were university based, four that were community based, medical school affiliated, and one community based, nonaffiliated, and ranged in size from six to 11 residents per year.

All LoTP evaluation activities are overseen by researchers in the Department of Family Medicine at Oregon Health & Science University (OHSU). All LoTP programs obtained local Institutional Review Board (IRB) approval, and OHSUs IRB granted an educational exemption to obtain data from the study sites (IRB #9770). All participating residents were invited to consent to allow their deidentified data on their ITE scores to be shared with OHSU under a data use agreement between OHSU and the ABFM.

The Family Medicine In-Training Examination and Data Ascertainment

The ITE consists of 200 multiple-choice questions written by ABFM board-certified family physicians who are in private practice or work in an academic setting.[1] Before administration, all questions are reviewed by a committee consisting of current or former residency program directors. The ITE is administered using an online format to approximately 10,000 residents from just over 700 residency programs each year in late October, and the number of residents in each year of residency is fairly evenly distributed.[1] The possible range of scores for the ITE is 200-800.

We obtained ITE scores for all consenting residents in the LoTP programs for 2013-2019 from the ABFM via a secure, password-protected file. This included 278 consenting residents in 3YR programs and 322 in 4YR programs. Thirteen residents (4.5%) did not consent from 3YR programs and 16 (4.7%) did not consent from 4YR programs, and were excluded, leaving data on 600 (90.8% of the full sample). We included resident cohorts for those in an LoTP residency program between 2013 and 2019 and categorized them as PGY1, PGY2, PGY3, and PGY4 for each examination year. Residents' demographic information included age, gender identity, race, ethnicity, marital and parental status, attended US medical school, and debt load.

Statistical Analyses

We used descriptive statistics to characterize residents' demographic information by length of training group, including means, standard deviations, frequencies, and percentages. We analyzed continuous variables comparing the two groups using independent samples t tests and χ^2 tests for categorical variables. To test the association between ITE scores and length of training, we used two analytic approaches. The first was conducted at the program level and utilized an intent-to-treat analysis[15] where residents in 3YR control programs were compared to residents in 4YR programs at baseline (PGY1), year 2 (PGY2), and year 3 (PGY3). The second approach was conducted at the resident level, and utilized an as-treated analysis (16) where only residents enrolled in and who completed 4 years of training are included. We removed 44 (13.7%) residents in 4YR programs who graduated after 3 years of training (though they are shown in Appendix Figures 1 and 2 showing mean ITE scores over time for comprehensiveness).

Data visualizations of ITE scores were composed of: (1) residents who completed 3 years of training in 3YR control programs, (2) residents who completed 3 years of training even though they trained in 4YR programs; and (3) residents who completed 4 years of training in 4YR programs, according to

training year. Additional visualizations included ITE score by training year among individual programs to assess variance.

We compared the ITE scores between study groups during their last year of training (third year is the last year of a 3YR program, the last year of a 4YR program includes those who graduate in either their third or fourth year in the intent-to-treat scenario) and during their third year of training (third year of a 4YR program and third year of a 3YR program). Unadjusted differences in the mean and standard deviation of ITE scores were reported, along with the mean difference between 4YR and 3YR programs (along with their 95% confidence interval). Lastly, we reported whether differences of mean ITE scores between two groups were meaningful using the approach identified by Norman, et al,[16] defined as when a mean difference in ITE scores is greater than one-half of the pooled standard deviation.

Next, we utilized a mixed-effects linear regression model where ITE scores were denoted as the dependent variable. In particular, we assumed that ITE scores follow a quadratic trend as number of training year progresses. We included the interaction term between two programs (4YR and 3YR programs) and training time (by year) in the model to assess the slopes' difference between study groups after adjusting for age, race, ethnicity, marital and parental status, status as a US medical school graduate, debt load, and examination year. We accounted for repeated measures by random intercepts at the individual participant level. We reported the linear slope and quadratic slope terms only, the remaining covariates and their estimations can be seen in Appendices A and B. To assess differences in ITE scores at PGY1 through PGY3, we derived estimated marginal means and 95% confidence intervals (CI) from the aforementioned models and were estimated for each training year and program. For PGY4, the fourth-year's ITE scores from 3YR program participants were extrapolated to compare against the fourth year of 4YR program participants.

National ITE data were analyzed by the ABFM (years 2013 through 2019). Residents from the LoTP programs are included in the national data, including nonconsenting residents and those in Navy programs. The increases noted between 2013 and 2015 in national data reflect the implementation period of the fourth year of training. The *P* values reflect a trend analysis illustrating how exam scores changed from year to year.

RESULTS

Residents in 3YR programs were older than those in 4YR programs on average (29.6 years vs 28.9 years), and residents in all study groups were predominantly female, non-Hispanic White, single, not parents, US medical school graduates, and had a debt load greater than $150,000, though other than age and debt load, none of these findings were statistically different in either the program-level analysis or the resident-level analysis (Table 1).

Residents training in a 4YR program scored higher than those in 3YR programs in the first year, PGY1 (4YR of 431.6 vs 3YR of 406.4; Appendix Figure 1). 4YR program residents continued scoring higher than their counterparts from 3YR programs in PGY2 and PGY3; and their scores increased in their last year of training to an average score of 525.4 compared to 484.0 among residents in 3YR programs in their last year of training. Although residents in both programs consistently increase their ITE scores as time progressed, visually, the slope of 3YR programs begins to flatten after PGY2, whereas the slope of 4YR programs flattens after PGY3.

Appendix Figure 2 presents the mean unadjusted ITE scores in the program-level analysis (intent-to-treat) according to training year, and Appendix Figure 3 presents the mean unadjusted ITE scores in the resident-level analysis (as-treated). The large dots in both figures show individual resident mean scores and the smaller dots represent individual program mean scores. Both analytic approaches produce similar findings, wherein all training years, residents in 4YR programs scored higher than residents in 3YR programs on average and the increase in scores between training years flattened in the final year of training for both groups.

Unadjusted mean differences in ITE scores in the last training year were 37.5 points to 41.4 points higher for 4YR programs compared to 3YR programs in both the intent-to-treat and as-treated scenarios, differences that were clinically meaningful (Table 2). In the third year comparison, the 4YR programs' scores were 30.8 points higher than the 3YR programs in the intent-to-treat analysis and 31.5 points higher in the as-treated analysis.

The covariate-adjusted linear mixed-effects models for the program level analysis (intent-to-treat) show the estimated model mean ITE scores and their 95% CI (Table 3). Model output prior to postestimation procedures are shown in Appendix Tables A and B. At baseline (PGY1), the estimated mean ITE scores were 407.7 for 4YR programs and 387.7 for 3YR programs, a 20.0 point difference (95% CI=9.0–31.0). At PGY2 and PGY3, 4YR programs scored 14.5 points higher and 15.4 points higher, respectively. Lastly, an extrapolated mean ITE score for the fourth year from 3YR programs was 28.7 points lower (95% CI=14.9–42.6) compared to the fourth year of 4YR programs.

The covariate-adjusted regression models for the resident-level analysis (as-treated) show the estimated marginal mean ITE scores and their 95% CI in Table 3. We observed similar differences between groups at each timepoint as in the intent-to-treat analyses.

Table 4 shows the linear slope term and the quadratic slope term from full models. Full-model outputs with all covariates are shown in Appendix Tables A and B. In the intent-to-treat scenario, the main effect of the quadratic term (estimate=-13.6, *P*<.001) suggests that 3YR programs had a curvilinear increase in ITE scores over time (as demonstrated in Appendix Figure 2) where the increase in ITE scores were rapid in the first 2 years and then leveled off in years 3 and 4. The coefficient and P-value of the interaction between the quadratic term and 4YR indicator (estimate=3.2, *P*=.397) suggests that participants in the 4YR program saw a similar curvilinear trend as participants in the

TABLE 1. Characteristics of Residents According to Length of Training at the Program Level and at the Resident Level

	Length of Training Program-Level Analysis (Intent to Treat)			Length of Training Resident-Level Analysis (As Treated)		
Characteristic	**3YR† (n=318)**	**4YR~ (n=322)**	***P* Value***	**3YR† (n=318)**	**4YR†† (n=243)**	***P* Value***
Mean Age in Years (SD)	29.6 (4.1)	28.9 (2.8)	.011	29.6 (4.1)	28.9 (2.5)	.022
Gender Identity	**n (%)**	**n (%)**	.095	**n (%)**	**n (%)**	.465
Male	111 (34.9)	137 (42.5)		111 (34.9)	94 (38.7)	
Female	204 (64.2)	184 (57.1)		204 (64.2)	148 (60.9)	
Nonbinary	1 (0.3)	0 (0.0)		1 (0.3)	0 (0.0)	
Missing	2 (0.6)	1 (0.3)		2 (0.6)	1 (0.4)	
Race/Ethnicity	**n (%)**	**n (%)**	.557	**n (%)**	**n (%)**	.599
Non-Hispanic White	202 (63.5)	217 (67.4)		202 (63.5)	160 (65.8)	
Hispanic	20 (6.3)	17 (5.3)		20 (6.3)	9 (3.7)	
Non-Hispanic Black	14 (4.4)	9 (2.8)		14 (4.4)	8 (3.3)	
Non-Hispanic Asian/PI	53 (16.7)	52 (16.1)		53 (16.7)	44 (18.1)	
Non-Hispanic AI/AN	0 (0.0)	2 (0.6)		0 (0.0)	1 (0.4)	
Non-Hispanic other	17 (5.3)	17 (5.3)		17 (5.3)	15 (6.2)	
Multiracial	10 (3.1)	7 (2.2)		10 (3.1)	5 (2.1)	
Missing	2 (0.6)	1 (0.3)		2 (0.6)	1 (0.4)	
Marital Status	**n (%)**	**n (%)**	.448	**n (%)**	**n (%)**	.520
Single	162 (50.9)	181 (56.2)		162 (50.9)	135 (55.6)	
Married/partnered	150 (47.2)	136 (42.2)		150 (47.2)	104 (42.8)	
Separated	1 (0.3)	0 (0.0)		1 (0.3)	0 (0.0)	
Divorced	3 (0.9)	3 (0.9)		3 (0.9)	2 (0.8)	
Widowed	0 (0.0)	1 (0.3)		0 (0.0)	1 (0.4)	
Missing	2 (0.6)	1 (0.3)		2 (0.6)	1 (0.4)	
Parental Status (Have Children)	**n (%)**	**n (%)**	.361	**n (%)**	**n (%)**	.562
No	272 (85.5)	267 (82.9)		272 (85.5)	203 (83.5)	
Yes	43 (13.5)	53 (16.5)		43 (13.5)	38 (15.6)	
Missing	3 (0.9)	2 (0.6)		3 (0.9)	2 (0.8)	
US Medical School Graduate	**n (%)**	**n (%)**	.156	**n (%)**	**n (%)**	.797
Yes	272 (85.5)	288 (89.4)		272 (85.5)	211 (86.8)	
No	44 (13.8)	33 (10.2)		44 (13.8)	31 (12.8)	
Missing	2 (0.6)	1 (0.3)		2 (0.6)	1 (0.4)	
Debt Load (Dollars)	**n (%)**	**n (%)**	.055	**n (%)**	**n (%)**	.047
None	41 (12.9)	41 (12.7)		41 (12.9)	30 (12.3)	
<25k	14 (4.4)	7 (2.2)		14 (4.4)	4 (1.6)	
25k - 74k	16 (5.0)	32 (9.9)		16 (5.0)	27 (11.1)	
75K - 149K	51 (16.0)	43 (13.4)		51 (16.0)	30 (12.3)	
150k - 249k	86 (27.0)	103 (32.0)		86 (27.0)	67 (27.6)	
>=250k	108 (34.0)	94 (29.2)		108 (34.0)	83 (34.2)	
Missing	2 (0.6)	2 (0.6)		2 (0.6)	2 (0.8)	

Abbreviations: PI, Pacific Islander; AI, American Indian; AN, Alaska Native.
*P value does not include missing category
†Matched third-year residents
~Received either 3 or 4 years of training at a 4 year program
††Received 4 years of training at 4-year program

https://doi.org/10.22454/FamMed.2023.427621

TABLE 2. Unadjusted Mean Differences in In-Training Exam Score by Last Year and Third Year Comparisons

		4YR Program, Mean (SD)	3YR Program, Mean (SD)	Mean Difference (Pooled SD)	95% CI of Mean Difference	Meaningful [5]
Intent-to-Treat [1]	Last Year[3] Comparison	521.5 (68.1)	484.0 (70.0)	37.5 (69.1)	22.8 – 52.2	Yes
	Third Year[4] Comparison	514.8 (72.8)	484.0 (70.0)	30.8 (71.5)	16.5 – 45.1	No
As-Treated [2]	Last Year[3] Comparison	525.5 (63.7)	484.0 (70.0)	41.4 (67.7)	25.4 – 57.4	Yes
	Third Year[4] Comparison	515.5 (71.6)	484.0 (70.0)	31.5 (70.8)	16.3 – 46.7	No

[1]For intent-to-treat analysis, 4YR program mean includes residents who graduated in 3 years and those who graduated in 4 years.
[2]For as-treated analysis, 4YR program mean includes only residents who graduated in 4 years.
[3]For last-year comparison, we compared the last year of ITE scores in 4YR programs to 3YR programs.
[4]For third-year comparison, we compared the third year of ITE scores in 4YR programs to 3YR programs.
[5] Meaningful = mean difference in ITE scores is greater than one-half of the pooled standard deviation.

TABLE 3. Adjusted†Mean Differences Between Length of Training Programs and Exam Year Performance

Exam Year	4YR	3YR	Difference	95% CI
Length of Training Program – Intent-to-Treat Analysis				
Baseline (PGY1)	407.7	387.7	20.0	9.0 – 31.0
PGY2	462.6	448.1	14.5	3.2 – 25.8
PGY3	496.7	481.2	15.4	3.4 – 27.5
PGY4	509.4	487.1***	28.7	14.9 – 42.6
Length of Training Program – As Treated				
Baseline (PGY1)	408.5	386.5	21.9	10.1 – 33.8
PGY2	461.8	446.7	15.0	2.9 – 27.2
PGY3	495.3	479.7	15.6	2.7 – 28.5
PGY4	509.1	485.6***	29.4	15.0 – 43.8

†Adjusted for exam year, age, race, ethnicity, marital status, parental status, status as US medical school graduate and debt load.

3YR program. In other words, the change in ITE scores were not significantly different between the 4YR and 3YR programs. We observed similar findings in the as-treated sample.

Table 5 compares mean ITE exam scores among all LoTP participants to those nationally from 2013 to 2019, withP values for trend indicating differences in exam scores from year to year. National ITE scores include all residents, including those who did not consent to be in the LoTP and those in the Navy programs, which explains the dissimilarity among numbers. Residents in any LoTP study group scored higher than residents nationally during PGY 1, 2, and 3 for all study years where relevant test scores are available. Significant variability in terms of exam scores over time is evident among residents in 3YR programs and nationally, while this finding is not evident among residents in 4YR programs until they are combined with LoTP residents in 3 YR programs (Table 5).

DISCUSSION

This study explored the hypothesis that no significant differences in clinical knowledge scores would be found among residents who underwent three versus four years of training. Findings indicate that ITE scores sharply increased between PGY1 and PGY2 for both groups, with residents in 4YR programs starting with higher scores at baseline compared to residents in 3YR programs and maintain this difference in each subsequent year.

The slope analysis found that scores in 3YR programs started to flatten sooner than scores in 4YR programs in both the intent-to-treat and as-treated analyses, indicating that knowledge growth was slightly higher in 4YR programs and this increase continued into the fourth year, though scores were flatter in the final year in both 4YR and 3YR programs. It may be that a focus on finding a future job is distracting in that final year or that residents are reinforcing clinical knowledge learned in the prior years, and therefore do not continue on their previous trajectory of learning as reflected in ITE scores. It may also be that the last year of residency training is focused on factors not included in the ITE, such as practice management and leadership development. If this is the case, then it is imporatant to ensure the last year of residency training has a significant impact.

We found no statistically significant differences in knowledge scores after baseline (PGY1) between those who underwent 3 compared to 4 years of training. This suggests that differences noted between the two study groups in PGY2, PGY3, and PGY4 may be due to the initial difference in PGY1 scores.

TABLE 4. Slope Analysis of In-Training Exam Score

Effect		Estimate	95% CI	*P* Value
Intent-to-Treat Slope Estimates				
Linear term[a]	3YR program	74.0	60.4 – 87.6	<.001
Linear term x 4YR program	3YR program	-8.7	-24.8 – 7.4	.290
Quadratic term[a]	3YR program	-13.6	-20.4 – -6.8	<.001
Quadratic term x 4YR program	3YR program	3.2	-4.2 – 10.7	.397
As-Treated Slope Estimates				
Linear term[a]	3YR program	73.7	59.9 – 87.6	<.001
Linear term x 4YR program	3YR program	-10.6	-27.6 – 6.4	.221
Quadratic term[a]	3YR program	-13.6	-20.5 – -6.7	.001
Quadratic term x 4YR program	3YR program	3.7	-4.0 – 11.4	.343

[a] In-training exam score increases from PGY1 to PGY4 were assumed to follow a quadratic model, therefore, two slope terms are required.

It may be that the higher scores among residents in the 4YR programs is related to how those programs recruit or rank residents, though in a prior analysis of the match in LoTP programs, including applicant type, number, match positions filled, matched applicant type, and ranks to fill did not differ between 3YR and 4YR residencies.[17] However, those motivated to apply for 4 years of training did report a desire for more flexibility in training and to learn additional skills beyond clinical skills. It may also be that those more skilled at test taking chose to apply to 4YR programs.[17]

The relationship between clinical knowledge attainment and length of training is complex. The knowledge family physicians need for effective clinical practice is continually expanding. Educational innovations often influence training approaches,[11,18] and several factors including gender and marital status affect examination scores in residency,[19–22] all of which were accounted for in our analyses. It is likely that PGY1 scores reflect factors that predate residency, such as medical school curricular content, teaching methods, and emphasis on test preparation.

Analyses conducted by the ABFM indicate that when all LoTP trainees are included and compared to national data, mean ITE scores of both 3YR and 4YR residents are higher than mean ITE scores nationally for all years included in the study. We did observe significant variation affected ITE scores in certain years, where residents performed higher compared to other years; however, this finding is not related to the psychometric properties of the ITE[23], indicating some other reason resulted in residents scoring higher in those years. We found it interesting that trends assessed by the ABFM for residents in 4YR training programs produced more stable scores than occurred nationally or among residents in 3YR programs, though this could be related to the cell sizes in those groups or the test taking-abilities of those who chose to apply to and were selected by 4YR programs. A weakness of the national data is the lack of covariates that were available as part of the LoTP study; thus, it was not possible to determine how adjustment for key characteristics may have affected national data.

Though this study found significant differences in knowledge scores according to length of training, the increases in PGY2, PGY3 and PGY4 may be due to initial difference in PGY1 scores. In addition, this is a single pilot study and should not be used alone to make a decision regarding the length of training in family medicine, a topic that has been intensely debated for more than a decade.[24–27] Several questions remain unanswered. For example, we do not know what effects an additional year of independent clinical practice may have had on clinical knowledge. Though several papers indicate that ITE scores are predictive of board certification scores,[2–4] the exams are not equivalent for direct comparison. We also are unable to determine what specific curricular elements in 4YR programs may be most impactful in terms of knowledge gains. Those who chose to undertake 4 years of training may plan to practice full-scope family medicine, which is not always available to family physicians due to health system policies or geographic locations. Those wanting a broader scope may have performed better on the ITE because of this focus. Peterson et al, found a broader scope of practice was associated with higher board scores among practicing family physicians.[28]

The strengths of this study include data capture of more than 90% of residents participating in the LoTP as well as our ability to conduct several analytic approaches to explore the study hypothesis. We included analyses at the program level (intent-to-treat) and at the resident level (as-treated) to parse out the effects of actually receiving 4 years of training from receiving training in a program where the fourth year was optional. We also collected key variables that have been known or hypothesized to affect examination scores to include in our regression models, so they could be adjusted for in analyses.

TABLE 5. **Mean In-Training Exam Scores of LOT Residents by Length of Training Compared to All Residents Nationally**

		LOTP and National Residents' ITE Scores According to Study Year (2013-2019)							
Trainees	**Program Length**	**2013** LoTP: n=86Nat: n=10,041	**2014** LoTP: n=195Nat: n=10,276	**2015** LoTP: n=302 Nat: n=10,547	**2016** LoTP: n=358 Nat: n=10,685	**2017** LoTP: n=353 Nat: n=11,203	**2018** LoTP: n=365 Nat: n=11,791	**2019** LoTP: n=362 Nat: n=12,766	***P* Value†**
LoTP PGY-1 Residents Mean Composite Scores (95%CI)	3YR n=488	408.2 (387.2, 429.2)	413.4 (397.0, 429.8)	391.7 (372.9, 410.6)	376.5(358.4, 394.6)	432.8 (415.6, 450.0)	450.2 (433.3, 467.2)	451.3 (436.2, 466.4)	<.0001
	4YR n=256	460.5 (436.5, 484.5)	446.0 (425.4, 466.6)	412.2 (393.3, 431.0)	400.0(380.2, 419.8)	437.0 (410.9, 463.1)	424.8 (398.6, 451.1)	448.1 (429.2, 466.9)	.3263
All PGY-1 Residents Nationally Mean Composite Scores (95% CI)	N/A	395.7 (393.3, 398.1)	383.6 (381.0, 386.1)	365.9 (363.9, 368.4)	356.1 353.7, 358.5)	392.7 (390.4, 395.0)	401.8 (399.5, 404.0)	414.0 (412.0, 416.0)	<.0001
LoTP PGY-2 Residents Mean Composite Scores (95% CI)	3YR n=405	–	455.9 (433.4, 478.4)	457.4 (438.4, 476.5)	443.0 (421.6, 464.4)	476.7 (461.1, 492.3)	486.9 (470.0, 503.8)	503.5(489.4, 517.6)	<.0001
	4YR n=218	–	499.8 (476.9, 522.7)	477.9 (458.5, 497.4)	438.5 (419.0, 458.0)	485.5 (467.7, 503.3)	504.0 (479.6, 528.4)	472.5(446.1, 498.9)	.7396
All PGY-2 Residents Nationally Mean Composite Scores (95% CI)	N/A	447.6 (445.1, 450.0)	441.1 (438.4, 433.7)	427.7 (425.2, 430.2)	419.6 (417.1, 422.2)	447.8 (445.6, 450.1)	454.1 (451.8, 456.4)	463.3 (461.2, 465.4)	<.0001
PGY-3 Residents Mean Composite Scores (95% CI)	3YR n=317	–	–	464.8(444.8, 484.8)	474.4 (455.1, 493.6)	504.8 (488.2, 521.5)	515.2 (500.2, 530.3)	520.0 (503.2, 536.8)	<.0001
	4YR n=184	–	–	531.9(505.1, 558.7)	500.6 (476.3, 524.8)	508.8 (492.5, 525.0)	519.5 (497.4, 541.5)	527.3 (501.1, 553.6)	.9718
All PGY-3 Residents Nationally Mean Composite Scores (95% CI)	N/A	475.7 (473.1, 478.2)	474.7 (471.9, 477.4)	460.3 (457.6, 462.9)	456.8 (454.2, 459.4)	479.6 (4773, 481.9)	489.2 (486.8, 491.6)	488.5 (486.2, 490.7)	<.0001
PGY-4 Residents Mean Composite Scores (95% CI)	4YR n=153	–	–	518.0 (439.0, 597.0)	522.7 (500.6, 544.9)	535.0 (518.7, 551.3)	526.3 (508.9, 543.8)	548.0 (528.1, 567.9)	.1174
All LoTP PGY Residents Mean Composite	3YR n=1,210	412.1 (391.3, 432.9)	429.7 (416.1, 443.4)	434.8 (422.6, 446.9)	430.6 (418.1, 443.2)	468.9 (458.5, 479.2)	481.8 (471.7, 491.9)	492.7 (483.1, 502.2)	<.0001
Scores Combined (95% CI)	4YR n=811	460.5 (436.5, 484.5)	475.3 (458.7, 492.0)	474.3 (458.5, 490.1)	465.3 (452.0, 478.6)	493.4 (482.3, 504.6)	494.6 (481.6, 507.6)	497.7 (484.4, 511.0)	<.0001
All Residents Nationally Mean Composite Scores (95% CI)	N/A	438.3 (436.7)	431.7 (430.0, 433.4)	416.8 (415.2, 418.5)	409.8 (408.1, 411.4)	438.9 (437.4, 440.4)	446.7 (445.3, 448.2)	454.2 (452.8, 455.5)	<.0001

Abbreviation: LoTP, Length of Training Pilot.
† *P* value for trend.

Limitations of this study include that some 4YR programs had an optional fourth year, which resulted in residents undertaking 3 years versus 4 years of training by choice. This introduced selection bias, which is an issue across the board in this study because it is not possible to randomly assign residents to their training program and we could not assign which residencies would transition to 4 years of training. We addressed this by adjusting analyses for several covariates that could have affected our outcome to account for these inherent biases. Another weakness involves the small number of training programs that enrolled in the LoTP study. Converting from 3 to 4 years of training is a considerable endeavor, likely requiring time commitments for planning and implementation. The programs that chose to undertake such an effort may have greater resources or resilience compared to other training programs across the nation, though we matched 3YR programs to 4YR programs based on geographic location, size, and continuity clinic setting, and our insignificant findings between the study groups suggests our matching strategy was successful.

In conclusion, we found significantly higher absolute ITE scores in 4- versus 3-year programs, but the increases in PGY2, PGY3, and PGY4 may be due to initial difference in PGY-1 scores. Additional research about associations between length of family medicine training and other aspects of clinical practice, including practice setting, continuity of care, clinical preparedness, and scope of practice will be forthcoming and are needed to inform future decisions about the optimal training model.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions made by Samuel Jones, MD, who was a member of the Length of Training Pilot Executive Committee.

Financial Support: The Length of Training Pilot is sponsored by the Accreditation Council for Graduate Medical Education and is funded by the American Board of Family Medicine Foundation. None of the authors have a conflict of interest to declare regarding this article.

REFERENCES

1. ABFM. In-training Examination. *American Board of Family Medicine.* 2021. https://www.theabfm.org/become-certified/acgme-program/in-training-examination.
2. O'neill TR, Li Z, Peabody MR, Lybarger M, Royal K, Puffer JC. The predictive validity of the ABFM's In-training Examination. *Fam Med.* 2015;47(5):349-356.
3. Indik JH, Duhigg LM, Mcdonald FS. Performance on the cardiovascular in-training examination in relation to the ABIM Cardiovascular Disease Certification Examination. *J Am Coll Cardiol.* 2017;69(23):2862-2868.
4. Yen D, Athwal GS, Cole G. The historic predictive value of Canadian orthopedic surgery residents' Orthopedic In-training Examination scores on their success on the RCPSC certification examination. *Can J Surg.* 2014;57(4):260-262.
5. Kreitz T, Verma S, Verma AA, K. Factors predictive of Orthopaedic In-training Examination performance and research productivity among orthopaedic residents. *J Am Acad Orthop Surg.* 2019;27(6):286-292.
6. Carmichael KD, Westmoreland JB, Thomas JA, Patterson RM. Relation of residency selection factors to subsequent Orthopaedic In-training Examination performance. *South Med J.* 2005;98(5):528-532.
7. Miller BJ, Sexson S, Shevitz S, Peeples D, Sant SV, Mccall WV. US Medical Licensing Exam scores and performance on the Psychiatry Resident In-training Examination. *Acad Psychiatry.* 2014;38(5):627-631.
8. Peterson LE, Boulet JR, Clauser B. Associations between medical education assessments and American Board of Family Medicine Certification Examination score and failure to obtain certification. *Acad Med.* 2020;95(9):1396-1403.
9. Phillips JP, Peterson LE, Kovar-Gough I, et al. Family medicine residents' debt and certification examination performance. *PRiMER.* 2019;3:7-7.
10. Wang T, Neill O, Eden T, A. Racial/ethnic group trajectory differences in exam performance among US family medicine residents. *Fam Med.* 2022;54(3):184-192.
11. Sloychuk J, Szafran O, Duerksen K, Babenko O. Association between family medicine residents' mindsets and in-training exam scores. *PRiMER.* 2020;4:33-33.
12. Shokar GS. The effects of an educational intervention for "at-risk" residents to improve their scores on the in-training exam. *Fam Med.* 2003;35(6):414-417.
13. Waller E, Eiff MP, Dexter E. Impact of residency training redesign on residents' clinical knowledge. *Fam Med.* 2017;49(9):693-698.
14. *Training Pilot Study. LoTPilot.* 2021.
15. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res.* 2011;2(3):109-112.
16. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care.* 2003;41(5):582-592.
17. Eiff MP, Ericson A, Waller E. A comparison of residency applications and match performance according to 3 years versus 4 years of training in family medicine. *Fam Med.* 2019;51(8):641-648.
18. O'neill TR, Peabody MR. *ITE Score Results Handbook.* 2013. .
19. Mainous AG, Iii, Fang B, Peterson LE. Competency assessment in family medicine residency: observations, knowledge-based examinations, and advancement. *J Grad Med Educ.* 2017;9(6):730-734.
20. Klein R, Ufere NN, Rao SR. Gender Equity in Medicine workgroup. Association of gender with learner assessment in graduate medical education. *JAMA Netw Open.* 2020;3(7):2010888-2010888.
21. Dayal A, Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. *JAMA Intern Med.* 2017;177(5):651-657.
22. Error in figure labels. Correction. *JAMA Intern Med.* 2017;177(5):747-747.
23. *Personal Communication with Lars Peterson. MD, PhD, Vice President for Research.* 2021. .
24. Fields KB. More on the 4-year FM residency program. *Fam Med.* 2005;37(1):8-8.
25. Scherger JE. Residencies: heal thyself before extending. *Fam Med.* 2006;38(3):158-159.

26. Carek PJ. The length of training pilot: does anyone really know what time it takes. *Fam Med.* 2013;45(3):171-172.
27. Sairenji T, Dai M, Eden AR, Peterson LE, Mainous AG, Iii. Fellowship or further training for family medicine residents. *Fam Med.* 2017;49(8):618-621.
28. Peterson LE, Blackburn B, Peabody M, Neill O, R T. Family physicians' scope of practice and American Board of Family Medicine recertification examination performance. *J Am Board Fam Med.* 2015;28(2):265-270.