THE FAILURE RATE OF CANDIDATES FOR BOARD CERTIFICATION: AN EDUCATIONAL OUTCOME MEASURE

BY **Denis M. O'Day MD,*** AND Chun Li PhD

ABSTRACT

Background: Because most residents eventually become board certified, the overall certification rate for the American Board of Ophthalmology (ABO) is not a discriminating educational outcome measure. We have evaluated two related measures: (1) first-time failure (FTF) in the written examination, or FTF in the oral examination after passing the written examination the first time, and (2) failure to certify within 2 years of graduation (FC2).

Methods: We used the tracking system at the ABO to access and analyze information from 1998-2005 on resident performance from program match to certification.

Results: Ninety-seven percent of graduates entered the certification process. The FTF rate was 28%. The program FTF rate ranged from 0% to 89% (median, 28%). Programs with fewer than 16 graduates per 5 years were significantly more likely to have higher FTF rates than larger programs. The FC2 rate was 21%. Thirty-two programs accounted for 50% of the FTFs and 27 for 50% of the FC2s. Residents who voluntarily transferred programs performed significantly worse than nontransferring residents by both measures.

Conclusion: The FTF and FC2 rates are potentially useful outcome measures. However, the small size of many programs contributes to some imprecision. The rates should be used only in conjunction with other factors when assessing programs. These data provide an insight into the state of ophthalmic education in the United States. Although the eventual certification rate was high, graduates from a substantial minority of programs appeared inadequately prepared to sit the Board's examinations.

Trans Am Ophthalmol Soc 2006;104:129-142

INTRODUCTION

Almost alone among developed countries, the United States has separated the responsibility for training of specialists from the evaluation of their qualifications. Training for residents in ophthalmology in the United States is under the supervision of the Accreditation Council for Graduate Medical Education (ACGME), whereas a totally independent entity, the American Board of Ophthalmology (ABO), evaluates and certifies.

Board certification is a voluntary process. For residents, board certification is also an expected outcome of training, as it is becoming increasingly difficult to practice medicine without it. Training programs play an important role in the certification process, both in preparing residents for certification and in attesting to the acquisition of clinical skills that cannot be evaluated in the Board's examinations.

Candidates for Board certification must pass two examinations. The written qualifying examination (WQE) evaluates cognitive skills, whereas clinical management skills are tested in the oral examination. In this study, we examined the performance of residents from training programs in the United States from the perspective of their ability to pass both examinations at the first attempt as a measure of the effectiveness of their training. We also evaluated candidates' ability to certify within 2 years of graduation.

METHODS

DATA COLLECTION

The ABO maintains a tracking system for residents in training using data provided by the Ophthalmology Matching Program (OMP) and individual training programs .The tracking system enables the Board to ensure that residents fulfill the training requirements for certification consistent with Board regulations. It was established as a response to the problem of candidates whose entry into the certification process was being disrupted by irregularities or deficiencies in training discovered at the time of application for Board examinations.

Shortly after the resident match is completed in January each year, the OMP provides to the Board a complete list of applicants and their matching programs. The data include the individuals' names, the program to which they matched, and medical school information. Thereafter, the Board maintains the database in the tracking system using the list of accredited programs provided by the ACGME, which contains the names of each program identified by a unique number agreed to by the ABO and ACGME. Beginning the year the residents match, the Board queries each program about the PGY-1 training of each matched individual. Programs are also asked to provide information on current residents from internship to graduation, including promotion from year to year, transfers in and out, and resignation or removal from the program. At the time of application to enter the certification process, the Board requests

From the Department of Ophthalmology and Visual Sciences (Dr O'Day) and Department of Biostatistics (Dr Li), Vanderbilt University School of Medicine, Nashville, Tennessee. Supported in part by a grant from Research to Prevent
Blindness, Inc, New York, New York.
*Presenter.
Bold type indicates 𝔄𝔒𝔖 member.

a formal attestation of satisfactory completion of training from the program. While in the certification process, the Board tracks examination performance in a separate database by type of examination (written or oral), date of the examination, score, outcome of examination, and date of initial certification.

For this study, we abstracted data for residents who completed residency training in 118 accredited programs during the period 1999 to 2003.

The overwhelming majority of residents graduate in June each year. The application period for the WQE runs from March 1 through August 1, and the examination is held the following spring. Candidates who pass the WQE are scheduled for an oral examination the fall of the same year or the spring of the next (Figure 1.). Thus candidates who register for examination immediately upon graduation should expect to be certified within 2 years. For this study, analysis of the performance of residents completing training from 1999 to 2003 was determined using data collected through the completion of the oral examination in the fall of 2005. To ensure confidentiality, the identity of the residents was masked.

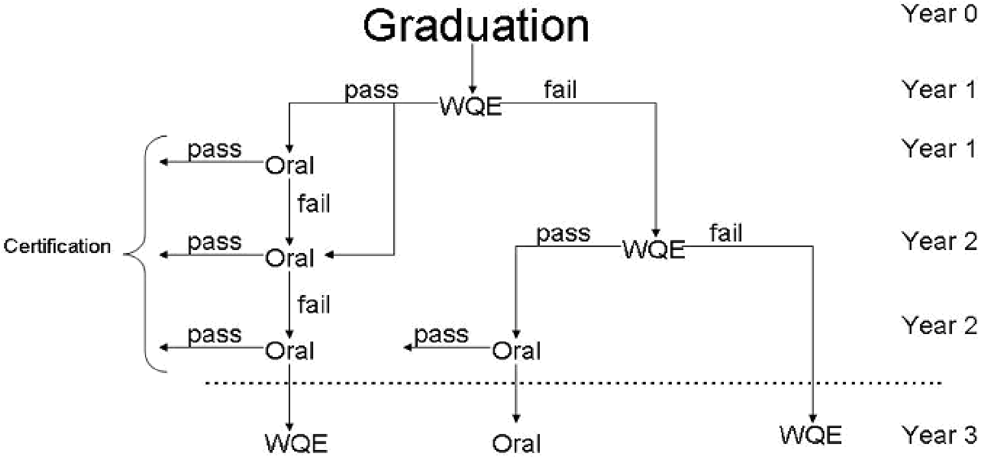

FIGURE 1

Algorithm of examinations for candidates for certification by the American Board of Ophthalmology. The horizontal dotted line indicates the end of the 2-year window for this study. Candidates who fail the oral examination three times are required to retake the written qualifying examination (WQE).

OUTCOME MEASURES

We used two statistics for this analysis. One, the first-time failure (FTF), is defined as a candidate who fails the WQE at the first attempt or, having passed the WQE at the first attempt, fails the oral examination at the first attempt. We calculated the FTF rate for the entire cohort of candidates, for transferring residents, and for individual programs. For the evaluation of program performance, we counted residents graduating from the program, including transfers into the program. When residents transferred from a program, they were not included in that program's tally.

As a second measure of program performance, we used the same approach in calculating the number of candidates who failed to certify in the 2 years following graduation (FC2) for each of the five graduating years (1999-2003). Because there might be a variety of reasons for graduates not sitting the examinations within 2 years, we excluded them from both analyses and called those who took any examination simply graduates.

STATISTICAL METHODS

Empirical Bayes Method

Suppose program i (i = 1, …, 118) has a true but unknown failure rate θ_i. Let n_i be the number of graduates from program i. Then the number of failures x_i for program i can be viewed as a random number from a binomial distribution with n_i draws and failure probability θ_i. One estimate of θ_i is the raw or naive rate $\hat{\theta}_i = x_i / n_i$. In a Bayesian approach, the 118 θ_i's are assumed to come from a common distribution, which can be modeled as a beta distribution beta (α, β). In the empirical Bayes method, the two parameters α and β may be estimated using the data.[1,2] This beta distribution reflects the distribution of possible failure rates for a program, and it serves as a prior distribution as if it were known prior to analyzing data. This prior information, together with a program's data, is used to derive a new distribution of failure rates for the program (posterior distribution) reflecting updated knowledge about the distribution of the unknown failure rate for the program. This new distribution is called the posterior distribution, and each program will have its own posterior distribution. Because additional information has been used, the posterior distributions will now show smaller variation than the prior distribution, and the variation of a large program will be smaller than that of a small program. In this model, the posterior distribution is also a beta distribution.[3-5]

For our data, the parameters α and β were estimated to be $\hat{\alpha} = 2.84$, $\hat{\beta} = 6.82$.

A program's failure rate was estimated by using the mean of its posterior distribution (posterior mean). The value lies between the program's naive rate and the overall failure rate due to a phenomenon called shrinkage.[2,4] The magnitude of shrinkage reflects the level of uncertainty in current knowledge about a program's failure rate. For large programs, the naive rate and the posterior mean tend to be similar to each other, whereas for small programs, they may differ to a greater extent.

Programs were ranked by calculating the expected ranks. An expected rank for a program is an average of all its potential ranks weighted by the associated probabilities.[4] The expected rank can also show shrinkage.[4]

All analyses were carried out in the statistical software R (www.r-project.org).

RESULTS

For the 5-year period spanning 1999 to 2003, a total of 2,163 residents graduated from the 118 programs. Eleven residents did not complete training in the specified time. Ninety-nine percent of those who applied to the Board sat the WQE the year following graduation. Of the 2,106 who applied to sit the examinations in this 5-year period, 1,491 (70.8%) achieved certification without a failure in either examination (Table 1). Twenty-five passed the WQE but had yet to sit the oral examination.

TABLE 1. AMERICAN BOARD OF OPHTHALMOLOGY EXAMINATION PERFORMANCE STATISTICS FOR 2,163 GRADUATES FROM 118 PROGRAMS OVER THE 5-YEAR PERIOD 1999-2003

VARIABLE	NO. (%)
Applied	2,106 (97.36%)
Total certified	1,832 (86.99%)
Certified by 2 years	1,660 (78.82%)
Certified without an FTF	1,491 (70.8%)
Total FTFs	590 (28.01%)
FTFs in written qualifying examination	414 (19.66%)
FTFs in oral examination	176 (8.36%)

FTF, first-time failure.

Of those who sat one or both examinations before the end of 2005, 414 failed the WQE at the first attempt. One hundred and seventy-six failed the oral examination at the first attempt, having passed the WQE at the first attempt. We considered these two types of FTF together. Thus, there were a total of 590 FTFs. The overall FTF rate was 590/2,106 = 28%.

The programs ranged in size from seven to 42 graduates (Figure 2A). Fifty-nine programs had 15 or fewer graduates, and 59 had 16 or more. Eighty-nine programs (75%) had 21 or fewer graduates. The numbers of FTFs per programs also varied considerably across programs from 0 to 17 (Table 2). When programs are ranked according to the number of FTFs, about half of the FTFs came from the bottom 32 programs. The naive FTF rates ranged from 0% to 89%, with median 27% (Figure 2B). The first and third quartiles were 17% and 41%.

Large programs tended to have a low failure rate (Table 3). A Pearson's chi-squared test of independence for the data in Table 2 gave a P value of .017. We also divided the programs into quarters based on their sizes and calculated the failure rate for each quarter (Table 4). The same trend can be seen with a trend test P value of 2.4×10^{-6}.

Because the size of programs varied, estimates of their naive failure rates had different precisions. In general, variation among small programs was higher than among large programs (Figure 3A). Smaller programs tended to be near to the two ends, whereas larger programs did not exhibit this tendency (Figure 3B).

When the posterior mean is plotted as an estimate of the true failure rate vs program size using the empirical Bayesian method, the effect of shrinkage is apparent (Figure 3C). Figure 3D is the plot of the expected rank vs program size. The tendency of small programs to be near to the ends is no longer present. Compared with Figure 3B, the expected ranks of the small programs tended to be near the center, whereas those of the large programs tended to be similar to the ranks based on naive rate.

The four programs with zero FTFs were tied in first place with zero naive FTF rates. However, they had different numbers of graduates (Table 5). All four programs appeared to be equally the best based on their naive rates. The empirical Bayes method could differentiate these programs by taking into account the different uncertainties resulting from different program sizes.

The largest program, which had 42 graduates and only one FTF, had a naive rate of 0.024 and would be ranked fifth based on the naive rate. However, because it was the largest program, we had the highest certainty about it. Accordingly, its failure rate as estimated by the posterior mean was 0.074, and its rank was 2 based on posterior mean. Its expected rank was 5.7, the highest expected rank. The program was ranked fifth based on the naive rate because of its single failure. However, it would be hard for this program to maintain zero failure for 42 graduates compared with the four programs that had zero failures but a smaller group of graduates. The empirical Bayes method did a fairer comparison between this program and the other programs, making it stand out as the best.

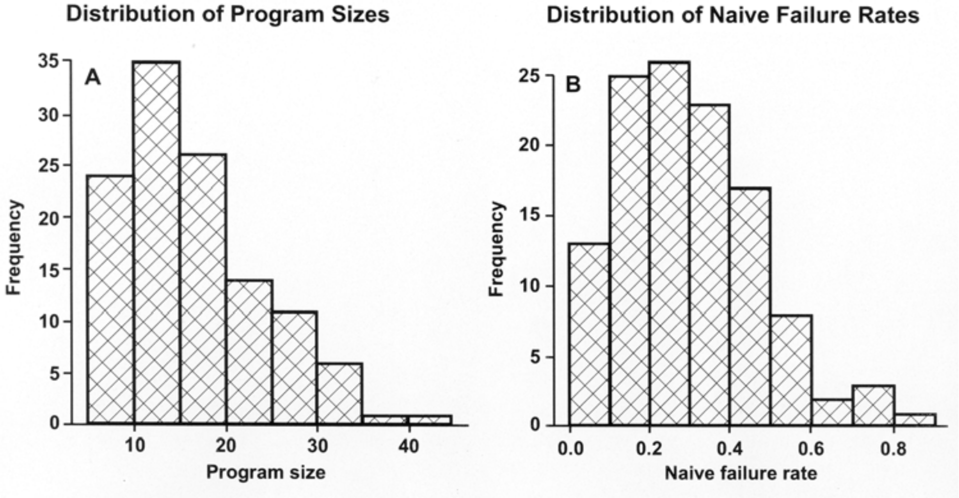

FIGURE 2

Distributions of program sizes and first-time failure rates (naive failure rates) in the American Board of Ophthalmology examinations for graduates from 1999 to 2003.

TABLE 2. DISTRIBUTION OF THE NUMBER OF FIRST-TIME FAILURES (FTFS) BY NUMBER OF PROGRAMS FOR THE AMERICAN BOARD OF OPHTHALMOLOGY EXAMINATIONS FOR GRADUATES FROM 1999-2003

No. of FTFs	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	17
No. of programs	4	8	15	14	15	22	13	5	5	3	7	2	2	1	1	1

TABLE 3. AMERICAN BOARD OF OPHTHALMOLOGY EXAMINATION FIRST-TIME FAILURE (FTF) RATES GROUPED BY NUMBER OF GRADUATES FROM THE PROGRAMS (1999-2003)

NO. OF GRADUATES	NAIVE FTF RATE ≥27%	NAIVE FTF RATE ≤27%
≤15	36	23
≥16	23	36

TABLE 4. RATES OF FIRST-TIME FAILURES (FTFs) IN THE AMERICAN BOARD OF OPHTHALMOLOGY EXAMINATIONS FOR GRADUATES FROM 1999-2003 GROUPED IN QUARTERS BY NUMBER OF GRADUATES FROM THE PROGRAMS

QUARTER	NO. OF PROGRAMS	RANGE OF NO. OF GRADUATES	TOTAL NO. OF GRADUATES	TOTAL NO. OF FTFS	FTF RATE
1	30	7 – 12	296	106	0.358
2	29	13 – 15	413	134	0.324
3	30	16 – 22	567	157	0.277
4	29	23 – 42	830	193	0.233

Figure 4 shows the ranks based on naive rates and the expected ranks, and their associated 90% confidence intervals. The small programs tended to have wider confidence intervals than larger programs because of the higher levels of uncertainty we had in them. When compared based on the expected ranks, the small programs tended to be nearer to the center than when they were compared based on the naive rates.

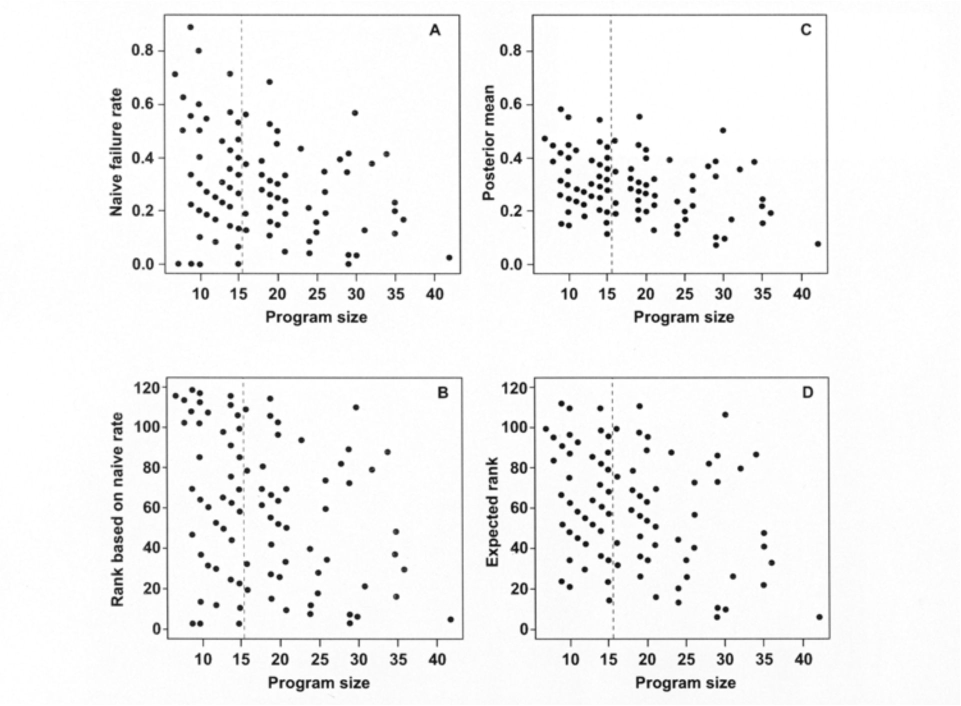

FIGURE 3

American Board of Ophthalmology examination first-time failure rates (naive rates) for candidates graduating from 1999 to 2003. A, Naive rate vs program size. B, Rank based on naive rate vs program size. C, Posterior mean vs program size. D, Expected rank vs program size. The vertical lines indicate the separation of the programs into 59 small and 59 large ones.

Figure 5A shows the comparisons between the ranks based on naive rates and expected ranks. They agree quite well in general, with a correlation coefficient of 0.99. Using the expected ranks, small programs were shrunk toward the middle.

TABLE 5. THE EFFECT OF USING THE EMPIRICAL BAYES METHOD TO RANK TRAINING PROGRAMS BASED ON FIRST-TIME FAILURE (FTF) RATES IN THE AMERICAN BOARD OF OPHTHALMOLOGY EXAMINATIONS FOR GRADUATES FROM 1999-2003

PROGRAM	NO. OF GRADUATES IN 5 YEARS	FTFS	FAILURE RATE ESTIMATED BY POSTERIOR MEAN	RANK	EXPECTED RANK	RELATIVE POSITION BASED ON EXPECTED RANK
P	29	0	0.073	1	5.9	2
Q	15	0	0.115	6	14.0	6
R	10	0	0.144	9	20.9	9
S	9	0	0.152	10	22.9	10

TRANSFERRING RESIDENTS

In the period 1999 to 2003, three programs were closed and one program was merged into another program. Residents in those programs had to transfer to other programs (involuntary transfer). Some other residents also transferred for various reasons (voluntary transfer). Among the 2,163 graduates from the 118 programs, 46 had changed programs. They transferred from 29 programs (including the four closed or merged programs) and into 34 programs. No one transferred to a closed or merged program. Table 6 shows the distribution of FTFs.

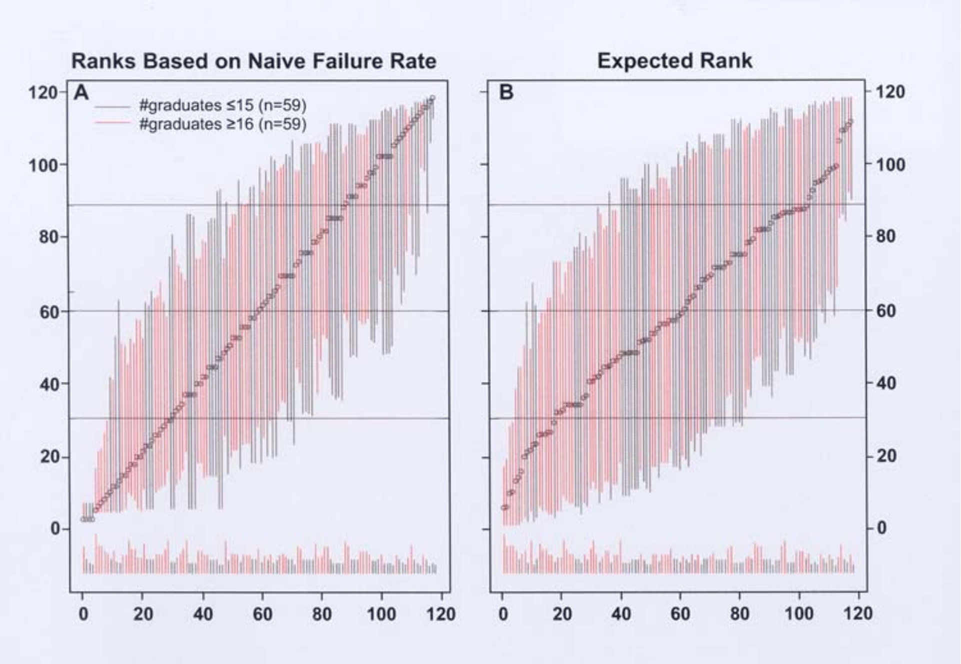

FIGURE 4

Ranks of programs and associated 90% confidence intervals based on first-time failure rates of candidates graduating from 1999 to 2003 for certification by the American Board of Ophthalmology. The 59 smaller programs are in black and the 59 larger programs are in red. The circles indicate the ranks, and the vertical bars around the circles indicate the 90% confidence intervals of the ranks. The vertical bars on the bottom denote the small (black) and large (red) programs and their sizes. The horizontal bars indicate the three quartiles of the ranks.

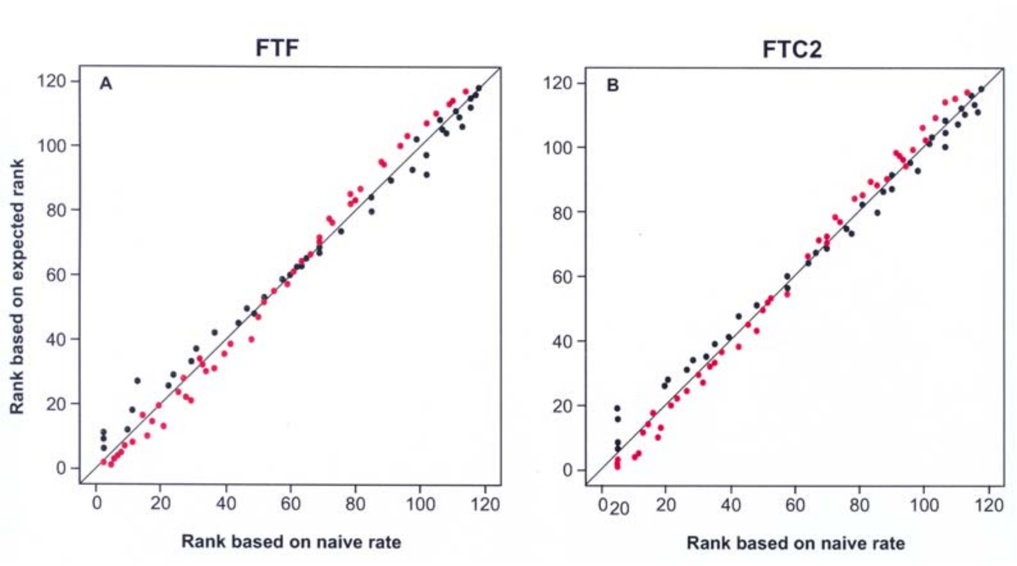

FIGURE 5

Comparisons between ranks for first time failure (FTF) and failure to certify in 2 years (FTC2) for American Board of Ophthalmology certification examinations based on naive rates and those based on expected ranks for candidates graduating from 1999 to 2003. The diagonal lines are equal lines. Red dots are the 59 larger programs and black dots are the 59 smaller programs.

Only 41 of these transferred graduates took the WQE or both examinations, with 17 FTFs. The FTF rate was 17/41 = 41.5%, higher than the overall failure rate of 28%. The voluntary transferring residents mainly contributed to this higher failure rate, as suggested in Table 4. A test of independence between FTF and voluntary status among the transferred graduates gave a marginally significant *P* value of .096, presumably due to the small counts. When we compared FTF counts between graduates who did not transfer and those who transferred voluntarily, in fact, there is a strong correlation between FTF and transfer from a nonclosed/nonmerged program (*P* value .006). Among the 2,163 graduates, only 57 had not sat any examination (2.6%), whereas among the 32 residents whose transfer was not due to program closure or merger, five had not sat any examination (15.6%). This discrepancy also was significant, with *P* value 1.1×10^{-4}.

TABLE 6. PARTICIPATION IN AMERICAN BOARD OF OPHTHALMOLOGY EXAMINATIONS AND FIRST-TIME FAILURE (FTF) RATES OF GRADUATES (1999-2003) WHO TRANSFERRED FROM ONE PROGRAM TO ANOTHER DURING TRAINING

CATEGORY	TRANSFERRING RESIDENTS (INVOLUNTARY)	TRANSFERRING RESIDENTS (VOLUNTARY)
FTF	3	14
No FTF	11	13
No examination	0	5

We investigated the effect of these transferring residents on a program's ranking. Among the 34 programs that accepted the transferring residents, two were not affected because the transferring resident had not sat any examination, 17 programs benefited because all transferring residents passed, 15 programs did not benefit (transferring residents all failed for 12 programs, one pass and one fail for two programs, and three pass and three fail for one program).

GRADUATES WHO FAILED TO CERTIFY IN TWO YEARS

We also collected data on the number of graduates who failed to be certified within 2 years of graduation. The number of FC2s and the number of FTFs are different. A graduate could have failed the WQE or oral examination the first time, but eventually was certified by the end of the 2-year period. Such a graduate would contribute to the FTF count but not the FC2 count. It also is possible that a graduate passed the WQE examination but had not taken the oral by the time the 2 years had elapsed. Such a graduate would contribute to the FC2 count but not the FTF count. Therefore, these two measures reflect different aspects of a program's performance.

We repeated the same analyses on the numbers of FC2s. Among the 2,106 graduates who took the written examination or both examinations before the end of 2005, 446 failed to be certified at the end of 2005, and the overall FC2 rate was 446/2,106 = 21%. The numbers of FC2s also varied a lot, from 0 to 15 failures. The distribution of the number of FC2s is in Table 7.

TABLE 7. DISTRIBUTION BY NUMBER OF PROGRAMS OF THE NUMBER OF CANDIDATES WHO FAILED TO ACHIEVE AMERICAN BOARD OF OPHTHALMOLOGY CERTIFICATION WITHIN 2 YEARS OF GRADUATION (FC2)*

FC2	0	1	2	3	4	5	6	7	8	9	10	11	12	15
No. of programs	10	13	26	18	13	13	8	4	3	3	3	1	1	2

*All candidates graduated between 1999 and 2003.

If the programs are ranked according to the number of FC2s, about half came from the bottom 27 programs. Ten programs had no graduates who failed to be certified. Sixty-seven programs had three or less FC2s, and 51 programs had four or more FC2s. As expected, there was a strong relationship between a program's number of FTFs and number of FC2s (Figure 6A). Seventy-six programs had smaller FC2s than FTFs, 34 programs had equal numbers of FC2s and FTFs, and eight programs had higher FC2s than FTFs.

We calculated the naive FC2 rates for the 118 programs by dividing the number of FC2s by the number of graduates. The FC2 naive rates ranged from 0% to 80%, with median 20%. The first and third quartiles were 10% and 31%. Figure 6B plots the FC2 naive rates vs the FTF naive rates.

Large programs still had a strong tendency to have low FC2 rates. However, this could not be seen if we just dichotomized the program size and FC2 naive rate as shown in Table 8, because this process of categorization lost too much information.

When we divided the programs into quarters based on their sizes and calculated the FC2 naive rate for each quarter (Table 9), the trend was apparent and a trend test on data in Table 8 gives a *P* value of 5.9×10^{-6}.

We compared the expected ranks calculated based on FTFs and those based on FC2s (Figure 6C). Overall, they agreed with each other quite well, with correlation coefficient of 0.89. Owing to the differences in numbers of FTFs and FC2s, most programs had different values in their expected ranks. Sixty-five programs had differences of no more than 10, 40 programs had differences larger than 10 but no more than 20, and 13 programs had differences of more than 20. The biggest positive difference was 36, and the corresponding program had 15 graduates, eight FTFs, and only three FC2s. The biggest negative difference was 27, and the corresponding program had 28 graduates, 11 FTFs, but 15 FC2s. Figure 6D shows the comparisons when we ranked the programs according to their expected ranks. Figure 6B shows the comparisons between the ranks based on naive rates and expected ranks. As in the case of FTFs, in general they agree quite well, with a correlation coefficient of 0.99. Again, using the expected ranks, shrinkage toward the middle was noted for small programs.

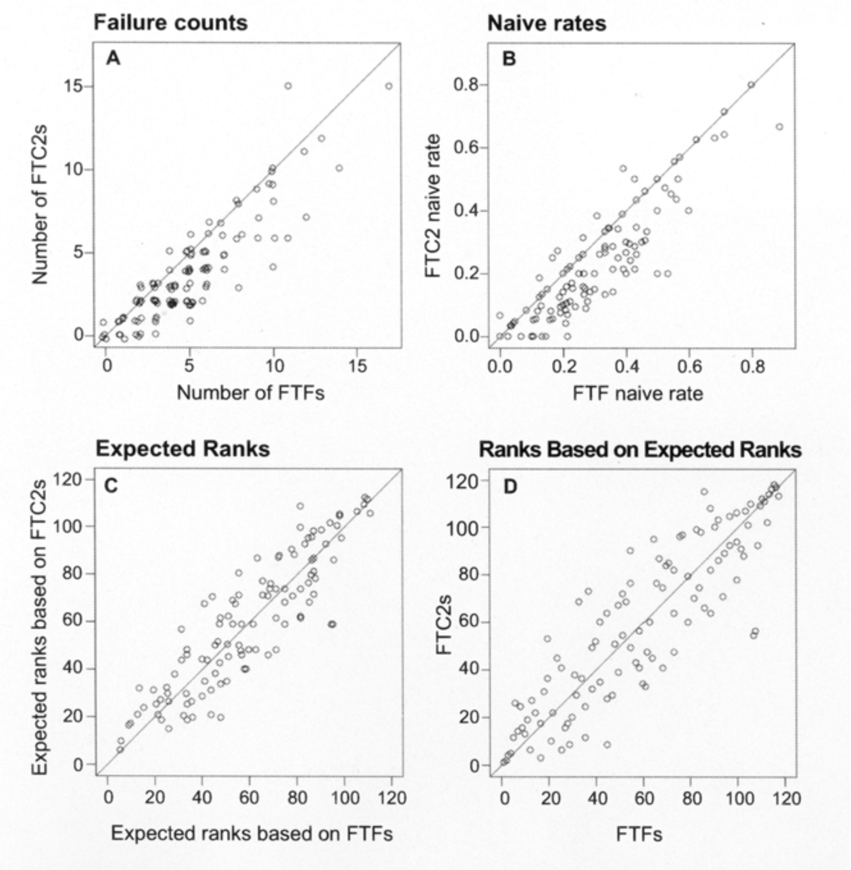

FIGURE 6

Comparisons of first-time failures (FTFs) and failures to certify (FTC2) in 2 years for American Board of Ophthalmology certification examinations. Candidates graduated from 1999 to 2003. A, Comparison by actual numbers. B, Comparison by rates. C, Comparison by expected ranks. D, Ranks by expected ranks. The diagonal lines are equal lines. Jittering was done for A so that programs with the same FTF values and the same FTC2 values are plotted as different dots.

TABLE 8. RATES OF FAILURE TO ACHIEVE AMERICAN BOARD OF OPHTHALMOLOGY CERTIFICATION WITHIN 2 YEARS OF GRADUATION (FC2) BASED ON NUMBER OF GRADUATES FROM PROGRAMS*

FC2 naive rate	≥ 20%	≤ 20%
No. of graduates ≤15	30	29
No. of graduates ≥16	26	33

*All candidates graduated between 1999 and 2003.

TABLE 9. RATES OF FAILURE TO ACHIEVE AMERICAN BOARD OF OPHTHALMOLOGY CERTIFICATION WITHIN 2 YEARS OF GRADUATION (FC2) FROM 1999 TO 2003 GROUPED IN QUARTERS BY NUMBER OF GRADUATES FROM THE PROGRAMS

QUARTER	NO. OF PROGRAMS	RANGE OF NO. OF GRADUATES	TOTAL NO. OF GRADUATES	TOTAL NO. OF FC2S	FC2 RATE
1	30	7 – 12	296	90	0.304
2	29	13 – 15	413	93	0.225
3	30	16 – 22	567	118	0.208
4	29	23 – 42	830	145	0.175

DISCUSSION

The educational performance of individual training programs for Ophthalmology remains largely unexamined. It is generally conceded that because most residents eventually achieve certification, measuring overall rates of certification does not distinguish between programs and is thus not a discriminating educational outcome measure. The ACGME through its accreditation mechanisms monitors the educational milieu and applies corrective action for demonstrated deficiencies through the issuance of citations, more frequent site visits and the threat of probation, and ultimately the loss of accreditation. Recently, with the advent of the Competencies and the development of psychometrically robust tools for their assessment, the ACGME has begun to probe more deeply the effectiveness of the educational process for the individual resident.[6-8] However, specific outcome measures for programs are still lacking.

Prior to the establishment of the Resident Tracking Program by the ABO in 1998, it was generally believed that almost all residents apply for Board certification and eventually become certified. This opinion was an educated guess, based partly on the numbers elevated to fellowship status in the American Academy of Ophthalmology, for which Board certification is a requirement. Because the numbers in training could not be tracked, the Board did not become aware of graduates from training programs unless they presented for the examinations. Thus the true number of graduating residents was unknown. In fact, our data for this cohort show that the actual number applying for certification is in accordance with previous estimates. Over 97% of graduates applied to sit the examinations, and only 1% delayed their application a year or more. Among those not applying for the certifying examinations are some international medical graduates who have returned to their native countries upon graduation. These numbers confirm the importance of Board certification from the perspective of residents in Ophthalmology. The public also has an interest in the ability of physicians to demonstrate achievement of a credible standard of knowledge and skills, according to a recent Gallup poll.[9] Thus, the degree to which residents are prepared for the Board's examination is an important educational outcome measure for training programs.

The WQE and the oral examination are designed to assess different skills (cognitive for the WQE and clinical management for the oral examination). A resident properly prepared should expect to pass both in the first attempt. To be fair to the program, in calculating FTF rates for the WQE and oral examination, subsequent failures or passes were not attributed to the program because remediation was beyond its control. For a program to achieve a perfect outcome score, all the residents would have to pass each examination at the first attempt.

In interpreting the data in this study, several issues need to be borne in mind. A central question is whether Board certification achieved by these examinations is a defensible standard. In a recent review, Brennan and associates[9] concluded that the available evidence supports the validity of examinations administered by member boards of the American Board of Medical Specialties (ABMS). For many years the WQE has been a psychometrically robust examination with high reliability and discrimination values (unreported data, American Board of Ophthalmology). In common with other ABMS Boards, the ABO constructs and administers the examination to ensure the test is relevant to the prior educational experience of candidates. It derives content that is related to clinical practice and assesses performance according to current standards of competence. Thus, the WQE is constructed using a content matrix. Test items are periodically reviewed by board certified ophthalmologists in practice as a measure of clinical relevance. To set the passing score, the Board uses procedures developed by Angoff[10] and Hofstee,[11] which are commonly used in certification programs and are accepted by psychometricians as contributing to a sound testing program. Finally, the examination is equated to ensure stability in the passing score from year to year.[10]

The oral examination by its nature is harder to assess psychometrically, although it is designed to adhere to basic psychometric principles for testing. In so doing, the Board has taken a series of steps to enhance confidence in the validity, reliability, and fairness of the examination, including the following: it is constructed according to a content matrix; content is evaluated for relevance; each candidate is examined by six independent experts; a candidate's score is based on the cumulative performance in the six tests; and the passing score is set by the same Angoff and Hofstee processes. However, unlike the WQE, the oral examination cannot be equated, so fluctuations in the passing standard may occur from year to year.

Another important consideration is the possible influence of confounding factors on the outcome of the examinations. In addition to the quality of the training program, many variables may affect a candidate's performance.[12-14] Among these, innate ability and education prior to residency are obviously important. Indeed, it might be argued that the better programs attracted students who would do well regardless of the teaching environment. A study of the performance in cognitive examinations similar to the WQE and administered by the American Board of Internal Medicine has examined this hypothesis.[13] The authors were able to separate out the effect of the residency training from other factors. They showed that although achievements prior to residency are important predictors of the outcome of the examinations, candidates perform better or worse than expected depending on the quality of the residency training program.

Validity of ranking is also an issue. Performance comparison is in wide use. The simplest forms include league tables in sports and medicine and the magazine rankings of universities. Its more sophisticated forms include value-added assessment of hospitals, public schools, and accountability programs. These rankings tend to have multiple effects, including positive or adverse publicity for the program and a "reference" for future trainees or customers. Causal interpretation often follows, even though these are observational studies, and administrators and legislators are tempted to assign praise or blame on the basis of "hard data" and "true measures." Comparisons based on a simple measure of performance such as the naive rate can be unfair because the simple measure often cannot take some important factors into account. For a fairer comparison, many statistical challenges must be overcome, and even then, sophisticated methods cannot guarantee a fair comparison.

These statistical challenges have attracted the attention of many statisticians.[15-17] Performance comparison is so important that recently the *Journal of Educational and Behavioral Statistics* devoted a whole issue on value-added assessment (2004 Spring; 29[1]).

We chose to calculate the expected rank as being the optimal means to use. Our decision was based on the work of Laird and Louis,[3] who showed the superiority of the expected rank over other types of ranking schemes. Experience with two programs in the dataset illustrates this concept. Program A's size was 15 with one FTF, and program B's size was 24 with two FTFs. Their corresponding naive rates were 1/15 = 0.067 and 2/24 = 0.083, and based on the naive rates, they were ranked 10th and 11th, respectively. However, we had higher uncertainty about program A's true failure rate than program B's. Program A's true rank could vary from 1 to 118, with various probabilities; the 5% and 95% percentiles were 3 and 61 (12th vertical bar in Figure 3B). Program B's true rank also could vary from 1 to 118, but with a more concentrated probability distribution; the 5% and 95% percentiles were 3 and 50 (8th vertical bar in Figure 3B). The expected rank for a program was calculated as an average of the ranks weighted by the associated probabilities. Program A's expected rank was 23.1, whereas program B's expected rank was 19.6, and they were the 12th and 8th best expected ranks.

Program A's posterior mean was 0.156, and program B's posterior mean was 0.144, and they were the 12th and 8th smallest posterior means. Both the expected rank and the posterior mean gave a reverse order of the programs compared with the rank based on the naive rate. These two programs also demonstrate the effect of shrinkage as a result of using expected ranking, with the movement toward the middle by smaller programs. In fact, no program could have expected to rank near No. 1 unless its superiority to other programs were strongly established, which was almost impossible unless the program sizes were much bigger.

With these cautions in mind, our analysis provides a glimpse of the state of ophthalmic education not afforded by simply examining the overall certification rates. We have been able to look at the performance of both graduates and programs. Although the high percentage of graduates who eventually achieved the standard meriting board certification is reassuring, 29% were insufficiently prepared at the end of training by reason of deficiencies in both cognitive and clinical management skills as evaluated in the examinations. Thus they required additional study to pass the examination beyond the ophthalmic education provided by the program. Twenty-seven percent of the training programs contributed to 50% of these failures.

For this study, we also observed the performance of candidates over the 2 years following graduation with the expectation that a resident applying to sit the Board's examinations immediately upon the conclusion of training would have sufficient time to achieve certification, even allowing for one or more failed attempts (Figure 1). This 2-year time frame is biased to a degree against candidates who fail the written as opposed to the oral examination, because the WQE is administered only once a year. It might also be expected to provide a reduced opportunity to certify for those few candidates who graduate in the fall after registration closes. In fact, none were adversely affected (O'Day, unreported data). Of the two measures used in this study, however, FTC has less precision because the possibility of repeated attempts to certify, even given the restricted time period, tends to maximize false-positives and minimize false-negatives. With these caveats, we believe that the 2-year certification rate period is another useful outcome measure. Our analysis showed that more than a fifth (21%) had not certified within 2 years of graduation, and again about 50% came from the 27 bottom-ranked programs. Although the correlation with expected ranking based on FTFs was quite high, the shift in rankings for most programs based on FC2s indicates the two statistics are measuring different aspects of performance.

The group of residents transferring to another program during training is of special interest. The reasons for the transfer varied. For those whose programs closed, the transfer was necessary to continue training. For the rest, we can infer a degree of dissatisfaction with the program as the motivation. Although the numbers are small, the performance of the group who voluntarily changed programs was significantly poorer than the entire cohort, suggesting that changing programs is not an effective response to a perceived educational problem.

The wide range in naive rates among the programs is striking (0% to 89% FTF and 0% to 80% FC2 rates). This disparity in program performance remains large even with expected ranking. As shown in Figure 4, for FTFs (data for failures to certify in 2 years not shown), the magnitude of the confidence intervals does provide hints about the difference among programs, especially the top and bottom performers. However, because these programs were not identified prior to the analysis, but through the data, any *P* value calculated on the same data could potentially inflate the difference. In this study, program size emerges as an important influence, not just as a limit on the precision of the ranking but because of its significant association with program performance. A study of graduates from Internal Medicine programs also found that candidates from larger programs performed better than those from smaller ones. However, when other program characteristics were factored in, the impact of size lessened, though it was still significant.[13]

The Board system in the United States was founded on the belief that resident physicians selected and trained appropriately should be able to demonstrate their competence by passing the certifying examinations. Implicit is the notion of a high pass rate so that few uncertified physicians are in practice and the public can be assured that each diplomate of the Board has the requisite knowledge and skills to provide quality care. Although the study confirmed that the great majority of residents do eventually become certified, the existence of a pool of physicians whose certification is delayed in addition to those who never certify has two consequences: the public lacks a measure of the physician's knowledge and skills, and the physician's entry into Maintenance of Certification is delayed or prevented by reason of a lack of initial certification. For the physician, failure can be costly, not just because of the additional fees. There is also the impact on credentialing and health plan participation to consider.

It is important to keep the data from this study in perspective. Because of the small size of many programs, we used aggregated data over a 5-year period. Norcini and Day,[18] in a study of Internal Medicine programs, concluded that this was a defensible strategy to increase precision. However, even with the aggregated data, the total number of candidates from many programs was still small. In such programs as we note, a single high- or low-performing candidate can have a greater effect on its rank position than a similarly performing candidate from a larger program. Although ranking does provide a range in performance achievable by a majority of disparate programs, bounded on one side by programs that appear to excel and on the other by those that fall far short, the confidence intervals are wide. Because of this and other uncertainties, we believe assessment of the quality of individual programs based on FTF and FC2 rates should be approached with caution. We do not believe that the public ranking of training programs has utility. The measures and ranks generated from statistical analyses can provide estimates and insights about the performance differences among programs that can aid improvement, but they should not be relied upon to the exclusion of other sources of information. A better way is to use these statistics as part of a global evaluation of a program so that the context and, in particular, the effect of program size are better understood.

Regardless of the precise ranking of programs, the numbers of candidates who were unable to pass the examinations at the first attempt and those who were still not certified after 2 years seem unduly high. Both statistics suggest a degree of unevenness in the quality of training among a substantial minority of these ACGME accredited programs that goes well beyond what might be expected of competent educational institutions. This finding deserves our attention as we consider the state of ophthalmic education in the United States.

A discussion of the possible causes is beyond the scope of this paper. However, the clustering of these failures in certain programs but not others provides an opportunity to determine the characteristics that are associated with better outcomes. The goal of performance comparison and ranking is not to reward or punish some programs, but to identify the factors that make a program succeed or fail and then to use those factors to engineer improvement.

REFERENCES

1. Efron B, Morris CN. Stein's estimation rule and its competitors—an empirical Bayes approach. *J Am Stat Assoc* 1973;68:117-130.
2. Morris CN. Parametric empirical Bayes inference: theory and applications (with discussions). *J Am Stat Assoc* 1983;78:47-65.
3. Laird NM, Louis TA. Empirical Bayes confidence intervals based on bootstrap samples (with discussions). *J Am Stat Assoc* 1987;82:739-757.
4. Laird NM, Louis TA. Empirical Bayes ranking methods. *J Educ Stat* 1989;14:29-46.
5. Shen W, Louis TA. Triple-goal estimates in two-stage hierarchical models. *J R Stat Soc B* 1998;60:455-471.
6. Leach DC. The ACGME competencies: substance or form? *J Am Coll Surg* 2001 Mar;192:396-398.
7. Cremers SL, Ciolino JB, Ferrufino-Ponce ZK, Henderson BA. Objective Assessment of Skills in Intraocular Surgery (OASIS). *Ophthalmology* 2005;112:1649-1654.
8. Golnik KC, Goldenhar L. The ophthalmic clinical evaluation exercise: reliability determination. *Ophthalmology* 2005;112:1649-1654.
9. Brennan TA, Horwitz RA, Duffy F, et al. The role of physician specialty board certification status in the quality movement. *JAMA* 2004;292:1038-1043.
10. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. *Educational Measurment.* 2nd ed. Washington, DC: American Council on Education; 1971.
11. Hofstee WKB. The case for compromise in educational selection and grading. In: Anderson SB, Helmick JS, eds. *On Educational Testing.* San Francisco: Josey-Bass; 1983.
12. Norcini JJ, Shea JA, Webster GD, Benson JA Jr. Predictors of the performance of foreign medical graduates on the 1982 certifiying examination in internal medicine. *JAMA* 1986;256:3367-3370.
13. Norcini JJ, Grosso LJ, Shea JA, Webster GD. The relationship between features of residency training and ABIM certifying examination performance. *J Gen Intern Med* 1987;2:330-336.
14. Shea JA, Norcini JJ, Day SC, et al. Performance of Caribbean medical school graduates on the American Board of Internal Medicine Certifying Examinations, 1984-1986. *Proc Annu Conf Res Med Educ* 1987;26:83-88.
15. Aitkin M, Longford N. Statistical modeling issues in school effectiveness studies (with discussions). *J R Stat Soc A*ssoc 1986;149:1-43.
16. Raudenbush SW, Willms JD. The estimation of school effects. *J Educ Behav Stat* 1995;20:307-335.

17. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance (with discussions). *J R Stat Soc A*ssoc 1996;159:385-443.
18. Norcini JJ, Day SC. Guidelines for interpretation of some common indicators of residency program performance. *Am J Med* 1995;98:285-290.

PEER DISCUSSION

DR SUSAN H. DAY: A paradigm shift is occurring in both educational and medical realms in which outcomes are increasingly emphasized. In part this is occurring because of a heightened demand for accountability. Simple *trust* that the system is doing what it should be doing is no longer acceptable.

The American Council for Graduate Medical Education (ACGME) oversees graduate medical education (GME) in this country, in essence through making sure that the educational centers which train house staff are providing a sound and complete education. Individuals seeking to become board certified through the conventional American Board of Medical Specialties (ABMS) boards must successfully complete a program accredited by the ACGME. Both the ACGME and the ABMS- who respectively assess programs and individuals- are modifying their criteria in response to the "outcomes" paradigm shift.

From the perspective of residency programs, it will no longer be possible to rely upon proof that a given number of surgeries, of faculty, of lectures, and so forth will satisfy program requirements. A careful analysis of *outcomes* will evolve: how successful was resident surgery? Did patients with diabetic retinopathy receive proper treatment? Did a resident introduce him/herself and wash hands? Measuring "outcomes" will become critical, and these outcomes will be identified as important components of each of the 6 competencies (patient care, practice-based learning and improvement, medical knowledge, interpersonal and communication skills, professionalism, and systems based practice). Some predict that a specified portfolio of outcomes for any given individual will become more important to fulfill than a specific amount of time for a residency.

In response to this paradigm shifts, the ACGME has clarified the use of board examination results in its periodic review of programs.

It is a common ACGME program requirement (PR) for *all* accredited GME programs that they monitor graduates' abilities to pass board certification examinations. Section VII.C.2 of the ACGME PR states that "the program should use resident performance and outcome assessment in its evaluation of the educational effectiveness of the residency program. Performance of program graduates on the certification examination should be used as one measure of evaluating program effectiveness. The program should maintain a process for using assessment results together with other program evaluation results to improve the residency program."[1]

Additionally, some specialties have included additional language regarding Board scores in its specialty-specific requirements. As examples, orthopedic surgery requires a 75% pass rate on the first examination; anesthesiology and internal medicine require that at least 70% of graduates become board certified; pediatrics specifies that programs will be deficient if the pass rate on first attempt is less than 60% over a five year period and/or if less than 80% of graduates take the certifying examination; neurosurgery even requires that the "primary" board examination be passed prior to completion of the residency.[1]

Ophthalmology's response to this shift has gradually evolved within its Residency Review Committee (RRC). As with all RRCs, the executive director of the American Board of Ophthalmology (ABO) is an ex-officio member of this important committee. It is the RRC which decides how to use board scores in the accreditation process.

This paper addresses the authors' attempts to define outcome measures which will assist in this accreditation process, using assessment tools already in place within other RRCs. The first time failure (FTF) criterion is intuitively a reasonable way of assessing a program's ability to produce fine ophthalmologists. We all know the limits of examinations, yet we live in a society where aptitude tests and entrance exams are widely used. I still personally struggle with the concept introduced by O'Day and Li that programs are inherently responsible only for a graduate's board performance at the first time of taking, yet understanding the logic that a finer measurement is needed than that of eventual board certification. In my opinion, the environment for a residency sets the stage for the entire career. What role models were there? How are patients treated? What is collegial behavior? How does one continue to learn? Conversely, the intangible of preparation for board examination takes so many factors into account: Did studying occur? Was fellowship training a help or a hindrance to taking tests of a comprehensive training? Were the 2 year old twins crying at night?

I would suggest that the authors might consider the psychological difference of using first time *failure* rate versus using the first time *pass* rate. As a program director, I would rather know that 85% of my graduates passed the board examination than that 15% failed it.

It would be interesting to know whether programs with high FTF rates fell in this category because of performance on the written or oral examination. Analyzing this data might provide a hint at outcomes for *different* competencies if indeed the failures for given institutions clustered in either the written or the oral examination.

The failure to certify in 2 years (FC2) criterion is more troublesome intuitively, particularly given the time sequence of individuals' abilities to take the board examination. Such a criterion sets a "clock" of 2 years which does not give equal opportunity to each person. In the simplest of examples, an individual who fails the written exam the first time will be unable to become certified 2 years after graduation. Conversely, an individual who passes the written may- by sheer virtue of geography- have more than one opportunity to pass the oral exam. In essence, the FC2 criterion gives greater valence to the WQE and allows a variable number of opportunities to subsequently pass the oral exam within a given time frame.

The introduction of the Bayesian method into this paper is a fascinating concept in which impact of program size on the FTF and FC2 is incorporated into rankings. I can only trust that the principle for this statistical method is sound and that the calculations are accurate. However, I am concerned that the conclusion implies that a person who is in a small program is more likely to fail the board exam on the first time even though three of the four programs who had zero FTF's were "small" by the authors' definition. Program size requires approval by the RRC; in granting such approval, adequacy of volume of clinical exposure is *the* critical component. Thus, size in part reflects the RRC's ability to define the critical masses necessary for quality education. The inference that big equals good, and small equals bad, fails to acknowledge that differences in educational models are valid on an individual basis. Inevitably, as outcomes become increasingly more visible, published data in which size is used as an outcome predictor might be as influential as program ranking.

Finally, both the ACGME and the ABMS groups must come to terms with what it is telling people that is important (the 6 competencies for a physician) and what the ABMS measures. The board certification process- as important as it is- has heavily emphasized cognitive knowledge, just one of the competencies. Fortunately for ophthalmology, the oral examination more completely addresses patient care and other competencies. Yet other competencies, including professionalism, communication skills, and systems based practice are evaluated very minimally (if at all) in the certification process. Each umbrella organization must do a better job in looking at and addressing all 6 competencies if the public will continue to provide them with the ability to accredit programs and certify individuals.

REFERENCE:

1. The Accreditation Council for Graduate Medical Education. Common Program Requirements, VII, C.2. Available at: http://www.acgme.org/acWebsite/dutyhours/dh_dutyhoursCommonPR.pdf. Accessed April 5, 2006.

DR ROBERT C. DREWS: Do you assume that the programs take in residents of equal competence when you are judging outcomes as a basis of program efficacy? Should there be a sub-analysis of language competence, for example for those candidates whose primary language is not English?

DR EDWARD L. RAAB: I have a question about the difference between failing on the written and the first time failing on the oral exam. The exams are different, but the written exam acts as sort of a gatekeeper for getting to the second exam. Does the Board have information that would allow it to be confident that a failure on the written exam would be predictive of a failure on the oral exam? If this correlation is highly predictable, you do not really need the breakout as Dr. Day suggested.

DR IRENE H. LUDWIG: Have you considered comparing entry-level medical board test scores versus the performance on the ophthalmology boards? Perhaps the smaller programs are admitting people who did not perform as well, and they're just poor test takers to begin with. Have you considered any way of measuring clinical competence as compared with any of these tests? Do these tests measure any useful future clinical abilities in ophthalmology?

DR TRAVIS A. MEREDITH: A word of caution about the use of this data. I do not think they tell us very much about the current status of the different programs. I say this as someone who has taken over program that has changed dramatically over these past five years. This is a method and measure that tells us what happened in the program about five years ago. For example if you look at the residents who have matched with a program in this last year, they will be taking the Board exams in about 2010. I think this is a rearview mirror test, and should be used very cautiously by the ACGME or in providing this information to the ACGME as a measure of the current status of the education offered by the program.

DR IVAN R. SCHWAB: This data is very powerful and potentially dangerous. How do we provide the data anonymously and yet try to assist the programs that need assistance with these data? Maybe the programs that are not doing well as a result of the fact that they are not as popular or well known as other programs. How do we factor in the candidate who is less enthusiastic about learning? It's easy to open the door as a teacher, but I cannot make the students walk through.

DR DAVID J.WILSON: I know it is your philosophy to use this in a non-punitive fashion, as a way that residencies can improve the quality of their educational programs; and the data seems powerful in that regard. The WQE at least is a general knowledge examination. People have commented on other general knowledge examinations and the oral examination may be a general knowledge examination. There are of course limitations to those types of tests, which have been infamously portrayed in the book, The Bell Curve. Once we open that type of Pandora's box, there are issues that could potentially come into play, as have been outlined in The Bell Curve.

DR MYLAN R. VAN NEWKIRK: Did you look at the relationship of basic science education within these training programs?

DR GEORGE L. SPAETH: I am concerned by your suggestion that there is an apparent causal relationship between size and performance. There was an association, certainly, but you did not demonstrate in any way that it was causal. If it were causal, you would expect smaller programs to be more completely separated from the larger ones. But as you point out, some of the smaller programs did beautifully. So it does not look as though the size of the program actually has a causal relationship to the performance of the program. I was wondering how you would tend to address that issue of causation, not just association.

DR PAUL E. TORNAMBE: I have been the Chief of a medical staff on two occasions. How do we measure someone who is not Board certified? It also might be valuable if we looked at those people that failed the Boards the first time. Does it really mean anything? Does it tell us anything about how they practice? Perhaps you can look at malpractice cases to determine if doctors who are Board certified get sued less. What percentage of doctors who fail the boards go on to fellowships, and is that important? At the moment, unfortunately, we are not measuring or accrediting fellowships and we are not certifying fellows. What happens to these after they finish their residency or fellowship?

DR RICHARD P. MILLS: The Bayesian statistical approach usually assumes you know something about an issue at hand before you start. Then you throw in a variable and find out something about the issue afterwards. Your approach is sort of a reverse use of the statistic, in the sense that you're trying to get an adjusted ranking after taking out the variable of size of the program. But in so doing, you are making the basic assumption (which may incorrect) that size of the program is not an independent predictor of first-time failure, if I understand your statistical approach.

DR DAN B. JONES: What is the relative proportionality in terms of failure on the written versus pass, and then failure on the oral exam. Since there has been a history of cooperation between the Academy and the Board, relative to preparing for the WQE with the OKAP, is there any intent to look at program performance by some of the parameters you mention on the OKAP versus the WQE?

DR ALLAN J. FLACH: Bayesian statistics has been around for a long time but yet I have heard it being used frequently recently. What has made Bayesian statistics in vogue all of a sudden?

DR JOHN R. HECKENLIVELY: I was surprised at the high level of first-time failure rate. Most of the residents get their training for the written exam from the Basic and Clinical Science Course (BCSC) of the academy and most programs rely it on very heavily. 20 years ago residents were expected to read a little summary in the BCSC and then read the actual articles and the actual books whereas now there are large summary in the BCSC books, which is supposed to explain everything for them. Most of them assume that reading the BCSC is enough. Is there some way of breaking out the data to determine if we should be emphasizing more basic source material in order to improve the education of the residents?

DR DENIS M. O'DAY: Central to all the issues raised is the validity of the examinations as a measure of competence. This is the first study in ophthalmology of program and candidate performance; however it has been studied in other disciplines. A study for the American Board of Internal Medicine showed that the abilities assessed by the ABIM cognitive examination did correlate with other independent measures of clinical competence.

To address Dr Tornambe's question there are also a number of studies showing that physicians certified by ABMS boards do a better job in terms of outcomes across quite a wide spectrum of different conditions. These include, the provision of preventative services in the management of diabetic disease, a lower peptic ulcer mortality; a lower mortality amongst cardiac patients managed by cardiologists; colon resection surgery; punitive actions of disciplinary bodies to physicians, and also the rate of malpractice suits. All of these are independent, perhaps indirect measures, but they do support the value of ABMS board examinations as measures of clinical competence.

Our finding of an association between candidate performance and size of a program is troubling perhaps because of how it can be interpreted. Even though the effect of size is highly significant, there are some large programs in the poorer performing area, as well as much smaller programs among those with a better record. There are obviously many factors to consider besides size in the educational process. I would suggest that program size is a surrogate for the multiplicity of these other factors. The effect of size has also been studied in internal medicine training programs. The same significant association was found.

The relevance of the quality of the training program to the educational process is fundamental as several of the questioners intimated. We have not yet studied this aspect. It is interesting to note however, that when the Board system was being formed back in 1915, the American Ophthalmological Society in collaboration with the American Medical Association and the American Academy of Ophthalmology and Otolaryngology stated: "The Board shall be authorized to prepare lists of medical schools, hospitals and private instructors, recognized as competent to give their applied instruction to ophthalmology." Clearly, what our forefathers believed then is that high educational standards are essential to improving care. We know also that the residents, when they enter a training program, go there with the expectation that they will become Board certified. Finally, the overwhelming majority of residents sit the written examination within a year of graduation. The proximal relationship of training to the examination is unavoidable.

Are all candidates the same? Clearly there is a wide spectrum of ability. Given a high IQ and innate ability one might argue that anybody can pass the Board's examinations. It has been shown however, that when you account for these variables of ability and past performance, the quality of the educational program contributes significantly to the success rates of the candidates in the examination.

Dr. Drews raised the issue of language competency and its influence on the outcome of the examination. While we do not have any supportive data, it is likely that the number of candidates for whom English is not their first language is increasing. All of them however, will have been successful in previous similar examinations. A parallel unexplored question is familiarity with multiple choice as opposed to the traditional examination essay format.

With regard to Dr Jones' comment about performance in the OKAP examination, we have to be cautious because it is designed to help residents discover the gaps in their knowledge and so improve. It is an in-training formative examination, not a performance-critical examination and is not well suited to looking at program performance.

I want to thank Dr Susan Day for her insightful discussion and I appreciate the many thoughtful questions. We do hope to explore in more detail the issues raised here today.