Performance on the National Board of Medical Examiners Part I Examination by Men and Women of Different Race and Ethnicity

Beth Dawson, PhD; Carrolyn K. Iwamoto, MEd; Linette Postell Ross, MA; Ronald J. Nungester, PhD; David B. Swanson, PhD; Robert L. Volle, PhD

Objective.—To investigate the performance of men and women from various racial and ethnic backgrounds on the National Board of Medical Examiners Part I examination, controlling for any differences in measures of educational background and academic performance before entering medical school.

Design.—A retrospective analysis of existing records from the National Board of Medical Examiners and the Association of American Medical Colleges.

Setting.—National Board of Medical Examiners.

Participants.—All students taking the June administration of Part I for the first time in 1986, 1987, or 1988 and who were 2 years from graduation from an accredited medical school.

Methods.—Multiple regression methods were used to estimate Part I examination group differences in performance that would be expected if all students entered medical school with similar Medical College Admission Test scores, undergraduate grade point averages, and other prematriculation measures.

Main Outcome Measure.—Performance on the Part I examination.

Results.—There were substantial differences in performance, with white students scoring highest, followed by Asian/Pacific Islanders, Hispanics, and blacks; within all racial and ethnic categories, women scored lower than men. Controlling for dissimilarities in academic background greatly reduced Part I differences among most racial and ethnic groups, except Asian/Pacific Islander men; unexplained differences remained between men and women. Results were consistent for the 3 years examined.

Conclusions.—The results of this study do not imply that physician performance varies among racial and ethnic groups or between men and women; no written examination can measure all the abilities that may be desirable to assess. Validity research investigating reasons for the reported gender and racial and ethnic differences in performance on the National Board examinations should be continued.

(*JAMA*. 1994;272:674-679)

THE LAST two decades have seen an emphasis on increasing the numbers of women and underrepresented ethnic minorities in medicine.[1] Thus, recruitment and retention of qualified women and minority students have been goals at many medical schools. During the same period, differences in the performance on standardized examinations by different examinee groups, often referred to as test bias or fairness,[2,3] have been the focus of attention as well.

Scores on the National Board of Medical Examiners (NBME) Part I examination have often been used by medical schools to make promotion and retention decisions and by residency program directors for selection.[4,5] The Standards for Educational and Psychological Testing state that examinations should be fair for all examinee groups.[3] Although differences in scores do not necessarily mean a test is biased or unfair, reasons for these differences should be investigated.

The NBME has undertaken a research program to review differential performance on its examinations. This article investigates the performance on Part I by men and women medical students from different racial and ethnic backgrounds. The purposes of this study were to explore (1) whether differences exist in the mean performance on Part I of men and women from different racial and ethnic backgrounds and (2) whether differences may be explained by measures of educational background and academic performance before entry into medical school.

METHODS

The Part I Examination

Part I was the first in the three-examination sequence for NBME certification at the time of the study. Approximately 16 000 examinees took the 2-day examination each year, of whom most were students concluding their second year of medical school. The NBME Part I examination consisted of 900 to 1000 multiple-choice questions, with similar numbers of questions in each of seven basic science subjects (anatomy, physiology, biochemistry, pathology, microbiology, pharmacology, and behavioral science). The examination scores were highly reproducible; the internal consistency reliability (Cronbach's α) ranged from .96 to .97.

For editorial comment see p 713.

Study Population

Subjects were all students who took the June administration of Part I in 1986, 1987, or 1988 and met the NBME's reference group criteria: first takers who were candidates for NBME certification and were 2 years from expected graduation from a medical school accredited by the Liaison Committee on Medical Education. Fewer students took the September administrations of Part I as first

From the Department of Internal Medicine, Southern Illinois University School of Medicine, Springfield (Dr Dawson), and Divisions of Psychometrics (Dr Nungester and Mss Iwamoto and Ross) and Evaluation Programs (Dr Swanson) and the Office of the President (Dr Volle), National Board of Medical Examiners, Philadelphia, Pa. Dr Volle is now retired.

Reprint requests to Department of Internal Medicine, Southern Illinois University School of Medicine, PO Box 19230, Springfield, IL 62794-9230 (Dr Dawson).

takers, and performance levels of these students were substantially lower; therefore, analysis was restricted to the more homogeneous student population meeting the reference group criteria.

Data Sources

Data were obtained from a joint Association of American Medical Colleges and NBME data file containing Medical College Admission Test (MCAT) scores and other pertinent undergraduate information as well as Part I scores. The data file was anonymous, however, with no names or student identification numbers and no socioeconomic information. Part I scores were standardized to have a mean of 500 and SD of 100 based on the performance of the four previous June reference groups. A score of 380 or higher was needed to pass.

The MCAT data included each student's first scores on biology, chemistry, physics, reading, and quantitative skills, and the number of times each student took the MCAT. The scale for MCAT scores ranged from 1 to 15. Undergraduate variables included science and nonscience grade point averages (GPAs), each on a 4-point scale; the number of credit hours in science and in nonscience subjects; undergraduate major (biology, another science, or nonscience); and an index of the selectivity of the student's undergraduate college. This index was determined by the Higher Education Research Institute, University of California at Los Angeles, and was equal to the sum of the mean Verbal and Quantitative Scholastic Aptitude Test (SAT) scores of students at the college. This index was used in the current study to provide a measure of control for differences in levels of academic achievement among undergraduate schools.

Demographic variables included the student's age, gender, and racial and ethnic background as reported on the application to medical school. Racial and ethnic categories included white, black, Asian/Pacific Islander (referred to hereinafter as Asian), five subgroups of Hispanic, and American Indian/Alaskan Native. Because of small numbers, Hispanic subgroups were combined, and American Indian/Alaskan Native students were not included in the study. Information was missing on one or more variables for approximately 10% of the students. Approximately 60% of the examinees had missing undergraduate information, and 30% had no MCAT scores or information on age or race and ethnicity. Scores on Part I, however, were similar for students with and without missing information; therefore, those with missing data were omitted from the analysis.

Statistical Analysis

The statistical analyses were replicated on all 3 years. Means and SDs were calculated for each subgroup of examinees. Because the variables were measured on different scales, descriptive comparisons among groups were made in terms of effect size.[6] Multiple regression methods were used to predict Part I scores; dummy (indicator) variables were formed for ethnicity-by-gender combinations (with white men as the comparison or reference group) and for undergraduate major. The unique contribution of each independent variable to the regression equation, independent from all other independent variables, was determined (the difference between R^2, the squared multiple correlation coefficient, for the full model regression with all variables included, and R^2 for the restricted regression model with the variable of interest eliminated from the equation). This unique contribution was the basis for concluding statistical significance: a variable was significant if the inclusion of the variable in the regression model (after all other variables were in the model) increased R^2 with probability P less than .001 because the large sample size could easily lead to statistical significance at less conservative levels. The interaction between the selectivity index and each GPA was examined for any systematic relation that would be important to include in the prediction equation.

The regression equations were used to estimate the mean difference between each gender-ethnic group and white men on Part I that would be expected if all students entered medical school with similar academic credentials (analysis of covariance method). Interactions between each gender-ethnic group combination and the remaining variables were investigated before using the covariance method. Because performance at the pass-fail point on Part I is of particular interest to medical educators, a discriminant function analysis predicting pass-fail for each year was also performed, using the same variables and significance level (P<.001) as in the regression equations.

RESULTS

Complete data were available for 90% of reference group examinees: 10 809 in 1986, 10 873 in 1987, and 10 403 in 1988. The findings were consistent for all 3 years, and details are reported for 1988 only. (Data for all 3 years can be obtained from B.D.) Approximately one third of the examinees were women (Table 1). When the students were categorized by race and ethnicity, approximately 85% were white, 9% were Asian, 3.5% were Hispanic, and 5.5% were black. The proportions of students from different racial and ethnic backgrounds were different for men and women (χ^2=77.6; 3 *df*; P<.001); in particular, the proportion of black women (8%) was larger than the proportion of black men (4%).

Table 1.—1988 Reference Group* Scores on the National Board Part I Examination

	No.	Mean Score†	SD	Pass, %
Women	3526	455	95	79.4
Asian/Pacific Islander	351	458	93	78.9
Hispanic	129	386	97	55.8
Black	277	369	87	44.0
White	2769	467	90	84.1
Men	6877	492	100	87.2
Asian/Pacific Islander	619	485	99	86.6
Hispanic	239	447	109	71.6
Black	271	392	97	53.9
White	5748	499	97	89.5
Total	**10 403**	480	100	84.5

*First takers at accredited medical schools taking Part I for certification 2 years from expected graduation.

†Mean=500 and SD=100 in the previous four June Part I reference groups.

Descriptive Results

In 1988, mean scores on Part I were 480, and the pass rate was 84.5%; men's mean scores were 37 points higher than women's. Differences can be converted to effect sizes by dividing by the pooled SD; for Part I comparisons, the pooled SD was approximately 100. Differences also occurred at the pass-fail point, with approximately 8% more women than men failing the examination. The score distributions for 1988 for men and women are given in Fig 1 and illustrate the lower trend in scores for women compared with men.

Asian students had mean scores 15 to 20 points lower than white students during the 3 years analyzed in the study, Hispanics had mean scores 60 points lower, and blacks had mean scores 100 to 120 points lower. In 1988, pass rates were 88% for whites, 84% for Asians, 66% for Hispanics, and 49% for blacks. Score distributions for each race and ethnicity for men and women separately are given in Figs 2 and 3, respectively. The score distributions for black and Hispanic men were notably shifted to the left and were more peaked for black men and approximately normal for Hispanic men. Among women, the score distributions for both black and Hispanic women were peaked and positively skewed.

The relation among MCAT scores and undergraduate variables were similar in 1986, 1987, and 1988; details are again given for 1988 only (Table 2). Mean MCAT scores were approximately 10 points on biology, chemistry, and physics and approximately 9 points on reading and quantitative skills. Correlations among the three science sections on the MCAT ranged between .50 and .60; cor-

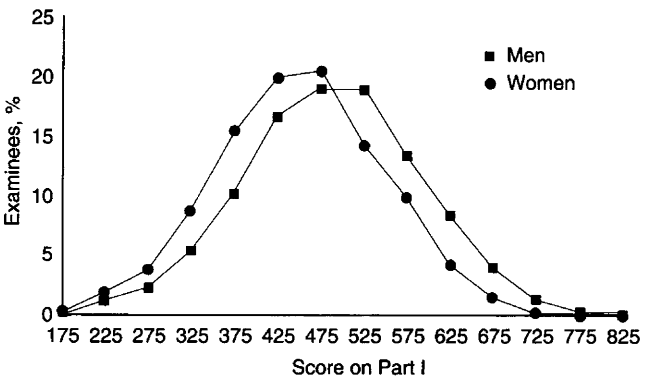

Fig 1.—Frequency distributions of the 1988 National Board of Medical Examiners Part I scores for men and women.

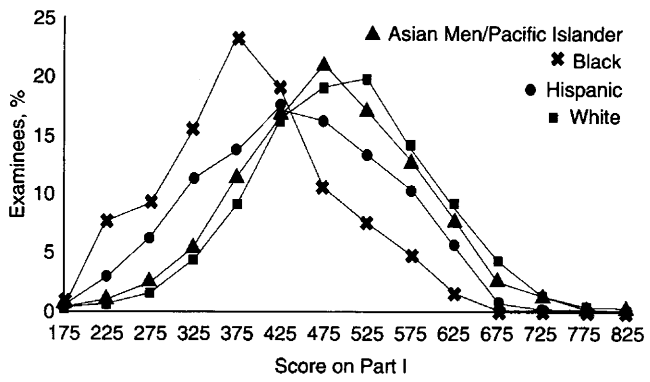

Fig 2.—Frequency distributions of the 1988 National Board of Medical Examiners Part I scores for Asian/Pacific Islander, Hispanic, black, and white men.

relations with reading and quantitative skills sections were slightly lower.

Differences on the MCAT between men and women and among racial and ethnic groups indicated a need to compare groups in terms of their adjusted mean scores rather than the observed scores. Men had higher mean scores than women on biology, chemistry, physics, and quantitative skills; differences ranged from 0.2 to 0.5 SD. Women had slightly higher reading scores. Asian students outperformed all other groups on all MCAT sections (except reading), especially on chemistry and physics. Differences between Hispanic and white students ranged from 0.4 SD on biology to 0.7 SD on quantitative skills. Between blacks and whites, the smallest differences were noted in biology and reading (1.0 to 1.2 SD), and the largest differences were noted in quantitative skills and physics (1.3 to 1.5 SD). Students took the MCAT an average of 1.5 times before matriculation. Women and black students tended to take the MCAT a greater number of times than did men and nonblack students.

The GPAs in science subjects ranged from 3.44 in 1986 to 3.40 in 1988 (4.0 scale); the mean nonscience GPA was consistently 3.54 or 3.55. Students had taken, on average, 60 undergraduate credit hours in various science subjects and 44 hours in nonscience subjects. Approximately 60% reported a biology major and 20% a major in another science or mathematics. Across gender, men had higher GPAs in science subjects and more often majored in a science other than biology, while women had higher nonscience GPAs and were more often biology majors. Compared with white students, Asians had slightly higher GPAs, while those of Hispanic and black students were substantially lower. Asians took fewer nonscience hours than blacks, and higher proportions of Hispanic students majored in biology. Asian students graduated from undergraduate schools at which mean SAT scores were generally higher; mean SAT scores at black students' undergraduate schools were generally lower.

Men averaged 23.7 years of age at entry to medical school, while women tended to be about 6 months older. Asian students, on average, were 1 year younger than others when they entered medical school.

Prediction Models

The variables analyzed in this study accounted for 37% of the variation in students' Part I scores in 1988 (R^2 in bottom row of Table 3); they accounted for 31% in 1986 and 36% in 1987. (The sum of unique R^2 values for individual variables does not equal the total value for R^2, because the sum does not include the variation in Part I scores accounted for by more than one of the variables.) All correlations between MCAT sections and Part I scores were positive. Biology, chemistry, physics, and reading MCAT scores were significant predictors of performance on Part I. The number of times the MCAT was taken, with a negative relation to scores on Part I, was also significant. Overall, the MCAT variables accounted for the largest single block of variation in performance on the NBME Part I examination. Comparatively, the undergraduate measures of performance played a less significant role; only science GPA and biology major contributed to prediction of Part I performance. Although not shown in Table 3, interactions between the selectivity index and GPAs were not significant, indicating no systematic tendency for students with high GPAs at more selective undergraduate schools to perform better than students with high GPAs from less selective undergraduate schools.

A key finding was that Part I score differences between white men and the other gender-ethnic groups were not explained solely by differences in their performance on the MCAT examination or in undergraduate school. After controlling for undergraduate performance differences, the mean scores of Asian women, Hispanic women, white women, and Asian men continued to differ significantly from those of white men in all 3 years of the examination analyzed in this study.

With only minor exceptions, the same variables significant in predicting Part I scores were also significant in the discriminant analysis prediction of pass-fail. However, the predictions were only marginal improvements over the baseline percentage of passing Part I (84.5%), given no knowledge about past performance. Pass-fail predictions were 97% accurate for students who passed but only 25% accurate for those who failed.

Adjusted mean differences provide an estimate of how a given ethnic or gender group's mean performance would compare with that of white men if they had similar MCAT scores and levels of undergraduate performance. To the extent that predictor variables account for observed group differences in Part I performance, adjusted mean differences should be reduced to zero. Among men, the adjusted mean difference for Hispanic students was almost 40 points lower than observed differences, reduced from 52 to 13 points (Fig 4). An even larger reduction in mean difference of 103 points was observed for black men. Thus, both Hispanic and black men performed, on average, about as well as white men who had the same MCAT scores and undergraduate performance characteristics. Among Asian men, the adjusted mean difference was greater than the observed difference, 26 compared with 14, indicating that, as a group, they scored slightly lower on Part I than would be predicted based on their MCAT scores and undergraduate performance.

For Asian and white women, adjust-

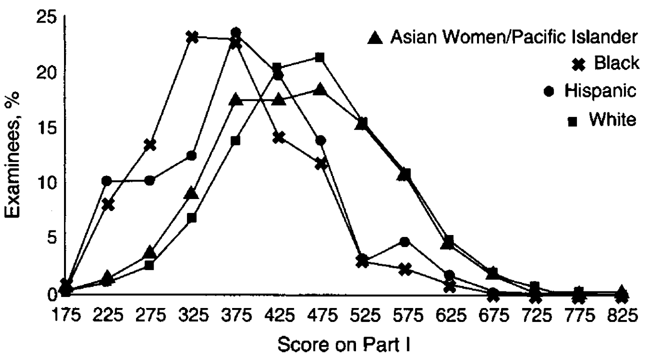

Fig 3.—Frequency distributions of the 1988 National Board of Medical Examiners Part I scores for Asian/Pacific Islander, Hispanic, black, and white women.

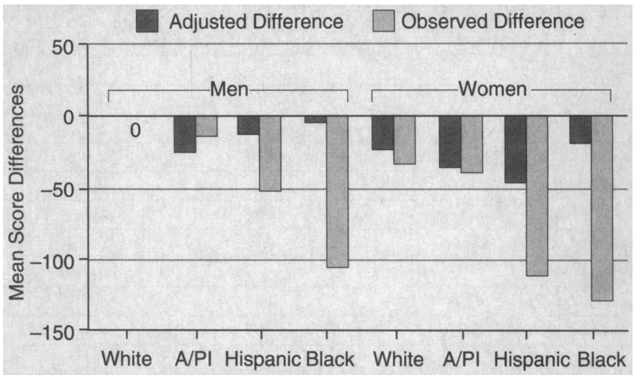

Fig 4.—Observed and adjusted mean scores on the 1988 National Board of Medical Examiners Part I examination for Asian/Pacific Islander (A/PI), Hispanic, and black men, and A/PI, Hispanic, black, and white women, all compared with white men.

Table 2.—Means and Percentages for Reference Group Examinees* Taking the June 1988 Part I Examination

	Men					Women					All				
	Asian/ Pacific Islander	Hispanic	Black	White	All	Asian/ Pacific Islander	Hispanic	Black	White	All	Asian/ Pacific Islander	Hispanic	Black	White	All
No. of examinees	619	239	271	5748	6877	351	129	277	2769	3526	970	368	548	8517	10 403
MCAT mean score†															
Biology	10.5	9.6	8.1	10.2	10.1	10.2	9.0	7.8	9.8	9.7	10.4	9.4	8.0	10.1	10.0
Chemistry	10.9	9.4	7.8	10.2	10.1	10.3	8.1	7.3	9.5	9.4	10.7	8.9	7.6	10.0	9.9
Physics	11.2	9.3	7.7	10.4	10.3	10.1	8.0	6.9	9.3	9.1	10.8	8.8	7.3	10.0	9.9
Reading	8.8	8.2	7.0	9.1	9.0	8.9	8.2	7.3	9.5	9.2	8.9	8.2	7.1	9.2	9.1
Quantitative	9.5	8.2	6.4	9.4	9.3	8.9	7.1	6.0	8.9	8.6	9.3	7.8	6.2	9.2	9.0
No. of MCATs taken	1.6	1.7	1.8	1.5	1.5	1.7	1.8	1.9	1.6	1.6	1.6	1.7	1.8	1.5	1.5
Undergraduate means and percentages‡															
Science (BCPM) GPA	3.46	3.17	2.86	3.44	3.41	3.45	3.10	2.83	3.45	3.39	3.46	3.25	2.85	3.45	3.40
Nonscience (AO) GPA	3.54	3.42	3.21	3.53	3.52	3.60	3.44	3.27	3.60	3.57	3.56	3.43	3.24	3.56	3.54
Science (BCPM) GPA	61	60	61	61	61	61	64	63	61	61	61	61	62	61	61
Nonscience (AO) hours	39	43	46	44	43	41	44	47	45	45	39	43	46	44	44
Biology major, %§	66	69	72	60	61	70	73	69	65	66	67	70	70	61	63
Other science major, %§	25	20	17	23	23	17	11	14	14	14	22	17	16	20	20
Selectivity index¶	1116	1069	1013	1072	1073	1119	1066	1042	1080	1080	1117	1068	1027	1074	1076
Demographic															
Age at matriculation, y	23.0	23.7	24.0	23.8	23.7	22.9	23.6	24.0	24.5	24.3	23.0	23.7	24.0	24.0	23.9

*First takers at accredited medical schools taking Part I for certification 2 years from expected graduation.

†Standard deviations of the Medical College Admission Test (MCAT) scores are approximately 2 points.

‡BCPM indicates biology, chemistry, physics, and mathematics; AO, all other subjects; SDs are approximately 0.35 for grade point averages (GPAs) (based on a 4-point scale) and 18 for hours.

§Yes coded 1; no coded 0.

¶Sum of mean Scholastic Aptitude Test Verbal and Quantitative scores at student's undergraduate institution; SDs are approximately 150.

ing Part I scores for differences between them and white men on the MCAT or in undergraduate performance had little effect (Fig 4); in 1988, unexplained differences of 38 points remained for Asian women and unexplained differences of 23 points remained for white women, compared with observed differences of 41 and 32, respectively. For Hispanic women, adjusting for differences on the MCAT and in undergraduate performance reduced the observed mean difference of 113 points, but a relatively large adjusted difference of 47 points remained. Similar reductions were obtained for black women, from the observed difference of 130 points to an adjusted difference of 19 points.

COMMENT

No published report of differential performance by students from different racial and ethnic backgrounds was found in the literature. In 1973, Weinberg and Rooney[7] reported lower scores for women than men on both the MCAT and the NBME Part I examination, but not on NBME Part II. Weinberg and Rooney hypothesized that differences might disappear as larger numbers of women enter medical school. It appears they have not. Our investigations indicate that (1) there are both ethnic and gender differences in Part I performance, and (2) prior academic performance is sufficient to explain a large part of the observed differences among underrepresented racial and ethnic groups but not differences between men and women.

Table 3.—Regression Results for Predicting 1988 Part I Reference Group* Test Score

Variable†	Correlation (r)	Regression Coefficient	Unique R^2	P
MCATs			.169	
Biology	.45	11.0	.021	<.0001
Chemistry	.49	9.6	.015	<.0001
Physics	.45	2.9	.001	<.0001
Reading	.35	9.2	.016	<.0001
Quantitative	.38	0.5	<.001	.30
No. of MCATs taken	−.27	−9.4	.005	<.0001
Undergraduate			.022	
Science (BCPM) GPA	.34	41.6	.014	<.0001
Nonscience (AO) GPA	.19	4.1	<.001	.20
Science (BCPM) hours	−.01	<−0.1	<.001	.47
Nonscience (AO) hours	−.08	0.1	<.001	.003
Biology major‡	−.01	8.6	.001	.0002
Other science major‡	.04	5.6	<.001	.04
Selectivity index§	.13	<0.1	<.001	.07
Demographic			.014	
Age at matriculation, y	−.07	0.4	<.001	.14
Women				
Asian/Pacific Islander‡	−.04	−38.5	.005	<.0001
Hispanic‡	−.10	−46.5	.003	<.0001
Black‡	−.18	−18.7	.001	.0004
White‡	−.08	−22.7	.008	<.0001
Men				
Asian/Pacific Islander‡	.01	−25.6	.003	<.0001
Hispanic‡	−.05	−13.0	<.001	.02
Black‡	−.14	−4.3	<.001	.41
Constant		−5.8		.75
R^2			.369	

*First takers at accredited medical schools taking Part I for certification 2 years from expected graduation.

†MCAT indicates Medical College Admission Test; BCPM, biology, chemistry, physics, and mathematics; AO, all other subjects; and GPA, grade point average (based on a 4-point scale).

‡Yes coded 1; no coded 0.

§Sum of mean Scholastic Aptitude Test Verbal and Quantitative scores at student's undergraduate institution.

Results of this research indicate that observed differences in Part I performance are particularly large for black and Hispanic students, two minority groups underrepresented in medical schools. These differences in Part I performance are not surprising, since differences in MCAT scores and undergraduate performance were also present before entry to medical school. In part, the observed racial and ethnic differences reflect the lower mean MCAT scores and GPAs of underrepresented minority students.[8]

The magnitudes of the differences reported in this study reinforce the need to continue developing programs for students who apply to medical school with relatively poor undergraduate preparation. Programs aimed at enhancing students' academic preparation before medical school and improving their performance while in medical school are in place in a number of medical schools. Prematriculation programs are increasingly available to help students gain the skills to overcome, at least partially, gaps in their preparation.[9] During medical training, programs teaching students how to organize their time and develop good study skills can help them avoid falling behind in their studies. Some medical schools provide special tutoring for students who need extra help; others have instituted alternative curricular approaches believed to benefit students with weaker academic backgrounds.[10] Despite the adverse financial ramifications for students, some programs permit students to decompress the first 2 years by taking an extra year to complete the course of study. Increasingly, medical schools have programs to help students prepare for the Part I examination.[11] All of these approaches may have merit, but research investigating their effectiveness is needed, particularly in light of major efforts to recruit underrepresented minority students.[12]

Statistically controlling for differences at the time of matriculation to medical school through regression analysis substantially reduced Part I performance differences between white students and underrepresented minority students. However, this was not the case for Asian students. Based on prematriculation measures, Asian students should have performed better on Part I than white students did. Instead, Asians students performed slightly worse, and controlling for prematriculation differences increased the magnitude of the Part I performance differences for Asian men. It is unclear why this occurred, although cultural differences may offer a possible explanation.

By comparison, statistically controlling for differences present at the time of matriculation to medical school reduced the observed differences in Part I performance between men and women by only 50%, a smaller percentage than for underrepresented minority students. This study provides few insights into reasons for these differences. A common explanation is that men and women enter medical school with different levels of academic skills, but this phenomenon was not observed in this study: differences in the range of 0.2 to 0.4 SDs on the Part I scale remained after adjustments. Of course, it is also possible that the prematriculation measures used in this study were not sufficiently sensitive to differences in academic preparation directly tied to medical school course work, such as undergraduate credit hours in specific science courses. Although it seems unlikely that the medical education environment should induce such differences after 2 years, attention should be given to identifying any forces that depress performance in the basic sciences of women and Asian men. Differences among students in their interest in science, leisure-time activities, and other possible moderating characteristics might be investigated,[13-15] along with the proportion of variance they might explain performance on the National Board Examinations. For example, the Educational Testing Service[16] and other researchers[17] have suggested possible factors contributing to differences in performance by girls and boys on standardized tests; some of these factors, such as differences in what teachers do to encourage students to excel, may be relevant in medical education and worth investigation.

Regardless of why performances differed on Part I, these differences will affect examinees unequally[15]; some schools require passing Part I for promotion,[18] and highly selective residency training programs rely on the examination results.[19,20] Like the Educational Testing Service[16] and Cole and Moss,[21] the NBME believes in a comprehensive approach to evaluating information relevant to testing decisions and emphasizes the need for tests to be fair to all examinee groups.

The need for investigations of the validity of test scores takes on added importance because of changes in the medical licensure system. At the time of this study, the National Board examinations and the Federation Licensing Examinations (FLEX) constituted two routes to licensure. During the early 1990s, these two routes were merged into the US Medical Licensing Examination.[22] Thus, examinees beginning the licensure process in 1992 had only a single route to licensure. Because phase-in of the US Medical Licensing Examination will take several years, validity studies will also require several years for completion.

When a new, high-stakes examination is introduced, in-depth studies of the validity of test scores and pass-fail decisions take on added importance.[3] Two measures of validity of an examination are (1) whether the examination content represents important knowledge or skills and (2) how well performance is associated with other measures of performance, both current and future.[23]

Content validity is established by demonstrating the relation between the content on an examination and the knowledge and skills needed by those who take the examination. For a medical licensing examination, content validity studies should include careful review of test content to ensure its relevance for the current and future practice of medicine. Some basic science concepts are much more important to the practice of medicine than others,[24] and this is the knowledge most important to test in the basic science component of the licensure examination system.

Judgments about content alone do not provide a sufficient basis for validity of inferences and decisions based on test scores.[23] Few studies of physician performance in practice settings have been reported,[25-27] and comparisons of examination results with performance in medical school, residency training, and practice are clearly needed. Studies such as the current one, comparing the performance of gender-ethnic groups, can reveal important information for judging the validity of test scores. If examination facets extraneous to the assessment of relevant performance result in lower scores for women and Asian men, they can be investigated. Such work takes many years to perform (and is never really completed). The NBME has begun a broad program of validity research into several areas, and initial reports are available.[28-31]

The current study was limited to the reference group of medical students who chose the NBME certification route at a time when the FLEX provided an additional mechanism for achieving licensure. Some of the conclusions might change if all students were included and if ultimate pass rates were considered. In addition, the variables in this study were those available in existing databases, thus precluding the use of some measures that might be desirable, such as the student's socioeconomic status and participation in special courses to prepare for the examination.

In conclusion, it is important to recognize that no examination provides a gold standard for all the abilities it may be desirable to assess. In particular, tests of knowledge base cannot assess many important clinical skills. It is unclear whether the differences among student groups observed in this study reflect the actual magnitude of academic differences or result from other factors. Despite these limitations, the findings from this study point to three areas for continued investigation: (1) the enhancement of educational programs for students at academic risk, (2) the investigation of reasons for differences in performance on the National Board examinations that remain unexplained after differences in academic background are statistically controlled, and (3) the relation between test performance and subsequent success in the practice of medicine. Research on these areas is under way in many medical schools as well as at the NBME.

We are indebted to Bernadette J. Brennan, Paul R. Kelley, PhD, Donald E. Melnick, MD, and Robin D. Powell, MD, from the National Board of Medical Examiners, Philadelphia, Pa; Susan C. Day, MD, Steven M. Downing, PhD, and John J. Norcini, PhD, from the American Board of Internal Medicine, Philadelphia; Judy A. Shea, PhD, from the University of Pennsylvania, Philadelphia; Charles Killian, MA, from the Association of American Medical Colleges, Washington, DC; Candice S. Rettie, PhD, from the Robert Wood Johnson Medical School at Camden, NJ; and Gerry R. Schermerhorn, PhD, from Southern Illinois University School of Medicine, Springfield.

References

1. Petersdorf RG, Turner KS, Nickens HW, Ready T. Minorities in medicine: past, present, and future. *Acad Med.* 1990;65:663-679.
2. Green DR. What does it mean to say a test is biased? *Educ Urban Soc.* 1975;8:33-52.
3. American Psychological Association. *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association; 1985.
4. Wagoner NE, Suriano JR, Stoner JA. Factors used by program directors to select residents. *J Med Educ.* 1986;61:10-21.
5. Volle RL. Using National Board of Medical Examiners scores in selection of residents. *JAMA.* 1988;259:266.
6. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* Revised ed. New York, NY: Academic Press; 1987:43-44.
7. Weinberg E, Rooney JF. The academic performance of women students in medical school. *J Med Educ.* 1973;48:240-247.
8. Jolly P. Academic achievement and acceptance rates of under-represented minority applicants to medical school. *Acad Med.* 1992;67:765-769.
9. Watts VG, Harris CT, Pearson W. Course selections and career plans of black participants in a summer intervention program for minority students. *Acad Med.* 1989;64:166-167.
10. Bridgham RG, Scarborough S. Effects of supplemental instruction in selected medical school science courses. *Acad Med.* 1992;67(suppl):S69-S71.
11. Borek D. Unchanging dilemmas in American medical education. *Acad Med.* 1989;64:240-244.
12. Association of American Medical Colleges. *Project 3000 by 2000 Technical Assistance Manual: Guidelines for Action.* Washington, DC: Association of American Medical Colleges; 1992.
13. Calkins EV, Arnold LM, Willoughby TL. Gender differences in predictors of performance in medical training. *J Med Educ.* 1987;62:682-685.
14. Jacobi M. Mentoring and undergraduate academic success: a literature review. *Rev Educ Res.* 1991;61:505-532.
15. Alper J. The pipeline is leaking women all along the way. *Science.* 1993;260:409-411.
16. Educational Testing Service Board of Trustees. *Toward a Better Understanding of Gender and Testing.* Princeton, NJ: Educational Testing Service; 1989.
17. Sadker M, Sadker D. *Failing at Fairness: How America's Schools Cheat Girls.* New York, NY: Charles Scribner's Sons; 1994.
18. Swanson D, Case S, Melnick D, Volle R. Impact of USMLE Step I on teaching and learning in the basic biomedical sciences. *Acad Med.* 1992;67:553-556.
19. Nungester RJ, Dawson-Saunders B, Kelley PK, Volle RL. Score reporting on NBME examinations. *Acad Med.* 1990;65:723-729.
20. Case S, Swanson D. Validity of NBME Part I and Part II scores for selection of residents in orthopaedic surgery, dermatology, and preventive medicine. *Acad Med.* 1993;68(suppl):S51-S56.
21. Cole NS, Moss PA. Bias in test use. In: Linn RL, ed. *Educational Measurement.* 3rd ed. New York, NY: Macmillan Publishing Co Inc; 1989:201-219.
22. Swanson D, Case S, Kelley P, et al. Phase-in of the NBME Comprehensive Part I examination. *Acad Med.* 1991;66:443-444.
23. Messick S. Validity. In: Linn RL, ed. *Educational Measurement.* 3rd ed. New York, NY: MacMillan Publishing Co Inc; 1989:13-103.
24. Dawson-Saunders B, Feltovich PJ, Coulson RL, Steward DE. A survey of medical school teachers to identify basic biomedical concepts medical students should understand. *Acad Med.* 1990;65:448-454.
25. Norcini JJ, Fletcher SW, Quimby BB, Shea JA. Performance of women candidates on the American Board of Internal Medicine Certifying Examination, 1973-1982. *Ann Intern Med.* 1985;102:115-118.
26. Stillman PL, Regan MB, Swanson DB, Haley HA. Gender differences in clinical skills as measured by an examination using standardized patients. In: Hart IR, ed. *More Developments in Assessing Clinical Competence.* Montreal, Quebec: Can-Heal Publications; 1992.
27. Day SC, Norcini JJ, Shea JA, Benson JA. Gender differences in the clinical competence of residents in internal medicine. *J Gen Intern Med.* 1989;4:309-312.
28. Becker D, Swanson D, Case S, Nungester R. Results of the initial administrations of the NBME Comprehensive Part I and Part II examinations. *Acad Med.* 1992;67(suppl):S16-S18.
29. Case S, Becker D, Swanson D. Relationship between scores on NBME basic science subject tests and the first administration of the newly designed NBME Part I examination. *Acad Med.* 1992;67(suppl):S13-S15.
30. Swanson D, Case S, Waechter D, et al. A preliminary study of the validity of pass/fail standards for USMLE Steps 1 and 2. *Acad Med.* 1993;68(suppl):S19-S21.
31. Case S, Swanson D, Becker D. Performance of men and women on NBME Part I and Part II: the more things change. . . . *Acad Med.* 1993;68(suppl):S25-S27.