

THE TESTING COLUMN

MEN AND WOMEN: DIFFERENCES IN PERFORMANCE ON THE MBE

by Susan M. Case, Ph.D.

Recently we began to assemble a national database of bar applicants. Following the July 2005 MBE, the database included some biographic information such as gender as well as undergraduate GPA, LSAT score, MPRE score, and MBE score for each of 22,000 applicants.

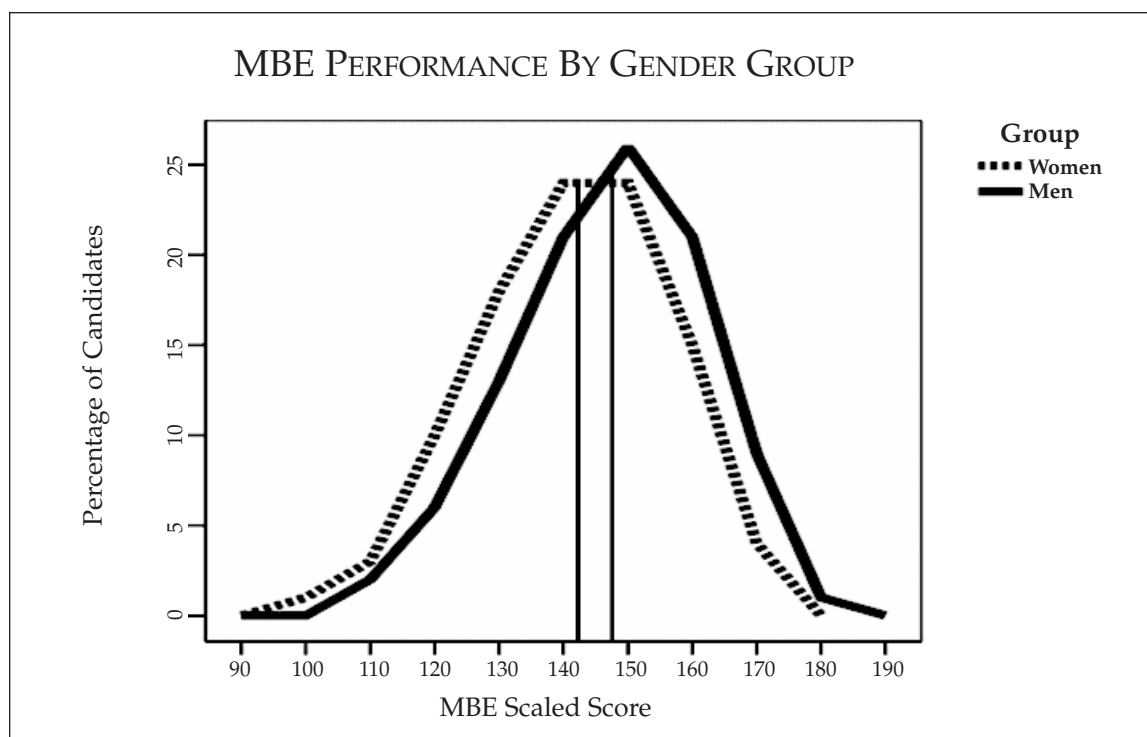


Our preliminary analysis showed that men outperform women on the MBE by about 5 points, which is about 1/3 of a standard deviation (SD). (My column in the May 2005 issue of *THE BAR EXAMINER* provides an explanation of SDs.) The graph on the next page shows the spread of scores for women and men. The curve for men is slightly to the right of the curve for women. With such a large sample size, a difference of this size is statistically significant; it is also large enough for many people to believe that it is practically significant. This result was not a complete surprise to me because I found similar results in some areas of medical licensing. As readers of this column are aware, I frequently compare our results with those found in medical licensure. There are several reasons for making these comparisons. One is that medicine is probably the most heavily researched profession, with a substantial body of published research on virtually every topic related to education, assessment, and licensing.

The second is that the profession of medicine is probably most similar to law in terms of the quality of the applicants and the rigor of the educational and licensing experiences.

Once a result is found that shows a difference in overall performance between groups, it is appropriate to undertake a series of research studies directed at various hypotheses that might explain the observed result: in this case, that men outperform women by about 1/3 of an SD. One possible explanation is that men are, in fact, more proficient than women in the knowledge and skills that the MBE is designed to assess. If men are more proficient than women in these areas, then it is appropriate that the scores reflect this difference in proficiency.

However, there are several studies that should be done before a conclusion about the validity of the result would be reached. These studies are designed to investigate whether or not other factors could account for the differences. First, the predictors of performance on the test should be investigated. This would include a look at factors such as undergraduate GPAs, LSAT scores, and MPRE scores. In this case, we found that the women had on average slightly higher undergraduate GPAs (+0.25 SD) than



the men, but slightly lower LSAT scores (-0.14 SD), and almost identical MPRE scores (-0.07 SD). These results are similar to those we found in medicine, where women outperformed men on their undergraduate GPAs, but men outperformed women on the MCAT (the medical analog to the LSAT).

The next set of analyses might look at subscores to see if the differences were consistent across the MBE content areas (Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, and Torts). In medicine, we found larger differences between men and women in some subjects such as biochemistry and anatomy, and smaller differences in other subjects such as pathology; this led us to investigate a hypothesis that men performed relatively better on subjects that were more “pure science” and less closely linked to patient care. In the case of law, I could not predict a difference between the genders on the six MBE subjects, and would be happy to hear from any of you who have hypotheses about our results. We found that men outperformed women in all six content areas of the MBE, but that the differences were lowest in

Evidence and Real Property and highest in Constitutional Law and Torts.


A third set of analyses would look at performance on individual items. This process divides examinees into two or more subgroups and looks at each subgroup’s performance on individual items along the continuum of the subgroup’s performance on the examination overall. This analysis will inevitably identify a set of questions on which the men outperformed the women to a larger extent than on the exam overall (along with, perhaps, a set of questions on which the women outperformed the men). Following traditional protocol, these items would then be reviewed by a group of content experts who would try to determine why the differences exist, in particular attempting to determine whether there is something about any flagged item that is unfair to women, something unrelated to the knowledge that the item was intended to assess. This technique often will surface problematic items on lower-level exams (such as those administered to high-school students), but is less likely to surface such items on exams administered for licensure in the professions.

On medical licensing examinations, for example, we found differences in performance, but either these differences were linked appropriately to the content or no explanation could be found. For example, on average, women performed better on gynecology questions, but we believed that this result was appropriate and accurately reflected true differences in proficiency. As noted above, men outperformed women on some of the basic sciences, and again we believed that this accurately reflected real differences in proficiency. We also invariably found items for which no explanation could be found. This technique would usually result in some number of false positives due to chance alone (perhaps 5 percent), and on the medical licensing exams, after discussion failed to reveal any problems, we concluded that the remaining items were likely to have been flagged in error. This is a common result. We performed the individual item analyses for the MBE, and found differences on less than 5 percent of the items. Review by content experts failed to reveal any explanation for the differences.

There are other possible explanations that are unrelated to content, which, if found to be true, might pose threats to the validity of the exam. Researchers have advanced several hypotheses for men outperforming women. One is that women do not perform as well as men on standardized tests, that because of format alone, women will do less well than their knowledge would predict. While this is a widely held belief, our research in medical licensing did not support this hypothesis. While women did not perform as well as men on some of the multiple-choice exam subjects, women outperformed men on some other subjects, such as pediatrics, obstetrics, and gynecology. At least three explanations arose out of the finding that men outperformed

women on the basic sciences typically taught in the first year, but that women outperformed men on at least one basic science subject typically taught in the second year and on some clinical sciences typically taught in the third year. One theory is that, on average, women have a weaker background in science at the time of matriculation, but “catch up” as a result of medical school training. The second explanation that other researchers suggested is that men and women adapt differently to stresses in medical school (including sexism) in ways that might disadvantage women in the early years, but be to their advantage later on; this hypothesis would explain men outperforming women on first-year subjects and women closing the gap in exams on clinical subjects. To my knowledge, the kind of careful longitudinal research, involving multiple schools, that would be needed to evaluate this explanation has not been undertaken in any field. Finally, some have noted that women tend to perform better in areas related to specialties attracting large numbers of women; while this phenomenon has been observed, the causal connection is unclear.

We have begun follow-up analyses of the MBE results, and would be happy to hear from any of you regarding your proposed explanations for the results as well as your hypotheses regarding subscore differences. When reviewing these results, one should keep in mind that they are preliminary, based on a first sample of cases. It is possible, though unlikely, that subsequent samples might show different results.

Results regarding racial and ethnic group performance on the MBE will be reported in a subsequent column. 

SUSAN M. CASE, PH.D., is the Director of Testing for the National Conference of Bar Examiners.