Interpreting Performance on Bar Examinations—
Which Score Types Make the Grade?

By Douglas R. Ripkey

In the bar examination process, several different types of scores can be generated to quantify candidates' performances. All of these scores have the potential to provide useful information, but some types of scores are more informative than others. An accurate understanding of exam performance requires knowledge of the proper uses and limitations of each score type. This article seeks to identify the most common types of scores and to explain issues that affect their proper interpretation.

Raw scores

The most basic (and least useful) type of score is the ***raw score***. For the multiple-choice tests that are common components of bar examinations, this score represents the total number of items a candidate answered correctly. Raw scores, taken alone, provide only a limited amount of useful information. If someone reports getting 125 correct on a test, a logical next question might be "Out of how many questions?" A raw score of 125 might indicate adequate knowledge of the subject matter if it were on a 150-item test in the local jurisdiction, but would be less impressive if it were on the 200-item Multistate Bar Examination (MBE). In the case of the MBE, a raw score of 125 indicates that the candidate got more than half of the questions correct. However, it is unclear what level of performance the score represents without other contextual information.

While a bar examiner might recognize that a raw score of 125 may be near the pass/fail point in that examiner's jurisdiction, a candidate who receives her MBE score for the first time may not realize that her performance is "borderline." In addition, the small differences that occur in overall exam difficulty across administrations must be taken into account. For example, that raw score of 125 might convert to a passing score on an extremely difficult version of the MBE, but it could also represent failure if earned on an administration that was composed of relatively easy questions. Thus, only limited interpretations can be made from raw scores unless additional information about test difficulty is available.

A related example of how raw scores provide only limited information can be found when reviewing raw scores on different content areas of the MBE. A candidate who earned a raw score of 20 on Torts and 25 on Contracts (out of a possible total of 34 in each category) might conclude that he performed better on Contracts than Torts. If this candidate needed to repeat the examination, this view of the information might lead the candidate to spend extra time reviewing Torts. However, perceptions about the original relative performances might change if it

were true that the Contracts items on that particular administration were as a group relatively easy (e.g., most other examinees earned raw scores of at least 27), and that the Torts items were relatively hard (e.g., most others earned a raw score of less than 17). With the additional information, this candidate might focus future study efforts on Contracts since it is the area where the candidate's performance was below average.

Similar misperceptions may also occur when trying to interpret raw scores across administrations. If our candidate repeated the exam after placing extra effort studying Contracts and earned a raw score of 25 on the Contracts section of the new exam, he might conclude that studying didn't help because his score didn't change. However, if the average raw score on Contracts were 22 on this new administration, then the candidate's performance after studying appears to improve since he scored above average on the new exam after scoring below average on the first one. (It is important to note that this perception of improvement would reflect an actual increase in knowledge only if the Contacts items combined were no easier and the overall candidate group was no less knowledgeable on the second administration than on the first.) These examples show that raw score interpretation is very difficult because of the need for contextual information about test difficulty and the performance of other examinees.

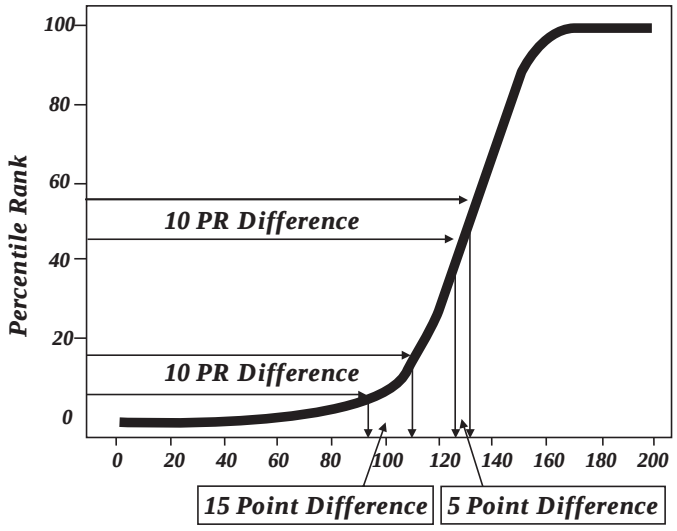
PERCENTILE RANKS

Recognition that context is important to score interpretation leads some jurisdictions to report scores that show a candidate's performance relative to others who took the same examination. The ***percentile rank*** is one type of performance indicator that reflects a candidate's position in the raw score distribution (typically on the normal or bell curve) for a defined reference group. More specifically, a percentile rank represents the percentage of test-takers whose raw scores fall below that of the candidate. For example, a report stating that a candidate has a percentile rank of 75 indicates that that person's achieved raw score is better than 75 percent of the other candidates taking the same examination and is the same as or worse than 25 percent of the same group.

Percentile ranks are known as derived or transformed scores since they are based on other scores. This conversion of a set of scores with a normal distribution, like raw scores, to one with a uniform distribution, like percentile ranks, affects score interpretations. The problem resulting from this transformation is that relative differences between scores don't remain constant. The result of this particular transformation process is that relatively large differences in percentile ranks in the middle of a distribution represent relatively small differences in raw scores, and relatively small changes in percentile ranks at the tails of the distribution are associated with relatively large differences in raw scores. Figure 1, which plots each raw score with its associated percentile based on data from one large U.S. jurisdiction on the July 2004 MBE, illustrates that the raw score differences are much greater for two people whose percentile ranks are 5 and 15 (about 15 raw score points, from 95 to 110) than those whose ranks are 45 and 55 (about 5 raw score points, from 125 to 130).

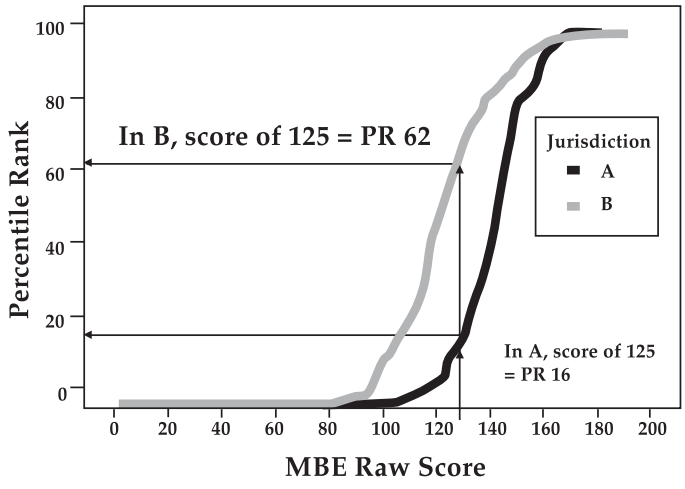
In addition, percentile ranks have many of the same drawbacks as raw scores. As with raw scores, additional information about context is needed to meaningfully understand the scores. When using percentile ranks, one needs to know the composition of the reference group upon which the score distribution is based. Again using the MBE as an

FIGURE 1. DIFFERENCES IN MBE SCORES AS A FUNCTION OF PERCENTILE RANK



example, percentile rank information for the total test and subscores is based on the score distributions in the individual jurisdictions. Using separate score distributions for each of the jurisdictions can, and often does, result in the same raw score being associated with somewhat different percentile ranks in different jurisdictions. Figure 2 shows an example by plotting the raw scores with their associated percentile ranks for two jurisdictions from the July 2004 MBE. Using a raw score of 125 as a reference point, the associated percentile rank is 16 in Jurisdiction A and 62 in Jurisdiction B. The difference remains relatively constant with the only exceptions occurring at extremely low or high raw scores.

FIGURE 2. PERCENTILE RANK DIFFERENCES ACROSS TWO JURISDICTIONS ON ONE MBE ADMINISTRATION



Within a single jurisdiction the associations can also change over time. For example, in one jurisdiction, a raw score of 125 corresponded to a percentile rank of 16 in July 2004 and a percentile rank of 23 on the previous February administration, in large part because the July group was more proficient than the February group. Because of the limitations with comparisons across groups, percentile ranks are best restricted to use for local evaluations at a single point in time.

SCALED SCORES

The most important type of score used in the bar admission process is known as a ***scaled score***. It is a derived score that provides information about a candidate's performance relative to a particular collection of candidates that is specified as the "reference group."

Scaled scores represent raw scores that have been transformed from their original scale to a new scale with a mean and standard deviation that matches that of the reference group. As a consequence of the scaling process, any particular scaled score will represent the same level of performance from administration to administration.[1]

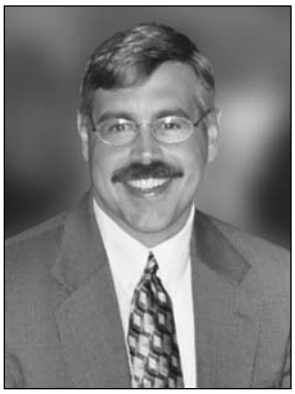
This scaling process has several practical implications that are important to performance interpretation. First, a particular scaled score may not correspond to the same raw score across multiple test administrations. For example, a scaled MBE score of 145.0 might be associated with a raw score of 141 on a relatively easy test and a raw score of 135 on a

relatively hard one, accurately reflecting that a student with a fixed level of knowledge will have a lower raw score on a hard test than on an easy one. Second, the consistent meaning provided by the scaled score implies that observed differences in scaled scores across groups or across time reflect differences in candidate ability and not in difficulty of test material. It would be reasonable to conclude that candidates who sat for the July 2004 MBE tended to be more knowledgeable than those who sat in February 2004 since the mean scaled score in July was 140.8 and the mean scaled score in February was 135.9. Finally, the distribution of scaled scores allows them to be compared and/or combined with other different measures (i.e., the scaled score on an MBE can be readily combined with essay scores that have been scaled to the MBE). Overall, the consistent properties of scaled scores make them the best ones to use in making pass/fail decisions.

Each type of score described above provides some information about candidate performance, but some are more informative than others. Raw scores tend to have the most limited use because their interpretation is dependent on contextual information like test difficulty and performance of other examinees. Percentile ranks provide somewhat more information about relative performance, but their interpretation should be limited to a single group at a fixed point in time. The scaled score is the most useful because of its consistent meaning across time. Knowledge of the properties associated with each score type should help those involved in the bar admission process make the most informed decisions about candidate performance.

ENDNOTE

1. ***See*** Deborah J. Harris, Equating the Multistate Bar Examination, 72 BAR EXAMINER 3:12 (August 2003).



DOUGLAS R. RIPKEY is the Associate Director of Testing for the National Conference of Bar Examiners.