AS SEEN IN GRADUATE MANAGEMENT NEWS

Demystifying the GMAT: Where Do Scale Scores Comes From?

By Lawrence M. Rudner

GMAT scaled scores convey the same level of ability over time, and GMAT percentiles convey the competitiveness of scores relative to today's GMAT test takers. In an earlier column, I discussed the role of the GMAT scaled scores and percentiles. Here, I get more technical and discuss how GMAT scaled scores were developed. Special attention is given to the scale for the new Integrated Reasoning section, which launches June 5.

Raw Scores

Raw scores, such as the number or percentage of correct answers, are sometimes used to report test results, but their interpretation is limited. They tell you how well an individual answered a specific set of questions, and they also give you an idea how one test taker did relative to others who answered the same set of questions. Yet raw scores rarely convey regular intervals of ability. In other words, the difference in ability between a person who got 95 percent correct and one who got 90 percent correct on a test is not the same as the difference between a person who got 85 percent correct and one who got 80 percent correct.

Scaled Scores

To overcome key issues with raw scores, facilitate the interpretation of test results, and permit comparisons across test administrations, tests such as the GMAT exam use scaled scores. First, a scale is defined to convey the range of ability measured and the precision of the scale. Results from subsequent test administrations can then be mapped back, through a process called equating, to the original scale. The GMAT Quantitative and Verbal sections are computer adaptive, and the results are computed based on the entire pattern of responses and the difficulties of the questions using Item Response Theory (perhaps a subject for another column). The Integrated Reasoning section will use different test forms designed to measure the same skills, and the results will be based on the number of correctly answered questions.

The GMAT Total score scale was originally defined so that scaled scores would be normally distributed, with an initial mean of 500 and an initial standard deviation of 100. The GMAT scores could therefore be interpreted using facts about the normal distribution: A GMAT score of 600 was one standard deviation above the mean for the reference group and had a percentile rank of 84 in the reference group.

Percentiles

As the overall ability levels of those taking the test as well as the test itself evolved over the years, the mean and standard deviation have changed slightly, so the normative interpretation cannot be followed exactly today. Whereas a scaled score from years ago should mean a similar level of ability as the same scaled score today, how that ability level compares with others taking the test today may differ. Therefore, the GMAT exam also reports the score percentile, or the percentage of tests ranking below a given score in the past three years.

Integrated Reasoning Score Scale

Launching the new IR section presented a few challenges. Unlike the computer adaptive Quantitative and Verbal Sections, Integrated Reasoning will have different fixed test forms. And unlike the Quant and Verbal, whose current scales were developed after a long history of paper testing, IR needed to have a score scale defined before launch. Percentiles of motivated test takers could not be computed in advance, a problem common to all new tests.

Unlike Quant and Verbal, which have 37 and 41 questions, IR has just 12 questions measuring the ability to integrate data to solve complex problems. Because integration is a key, many of the questions require multiple responses, and test takers must get all responses correct to receive credit for a question. With just 12 questions, a scale of 1 to 8 was chosen because it reflects the available level of precision, does not look like AWA or any previous GMAT scale, and because it provides a slightly higher degree of reliability.

GMAT Quant, Verbal, Total, and AWA percentiles are based on three-year rolling averages. For IR, percentiles will be based on cumulative distributions of tests taken starting on June 5. Percentile data will be updated monthly for the first six months and then annually at the same time as the other percentiles are updated. We do not anticipate much fluctuation after the first three months.

Pilot testing of IR questions defined the relative difficulty of questions in the initial question bank. This, in turn, allows us to develop numerous test forms that cover the same content and are of near-equal difficulty. The equating process will assure that IR scaled scores will be like the other GMAT scaled scores and convey the same level of ability over time.

Lawrence M. Rudner, PhD, MBA, is vice president of research and development and chief psychometrician for the Graduate Management Admission Council. He can be reached at lrudner@gmac.com.