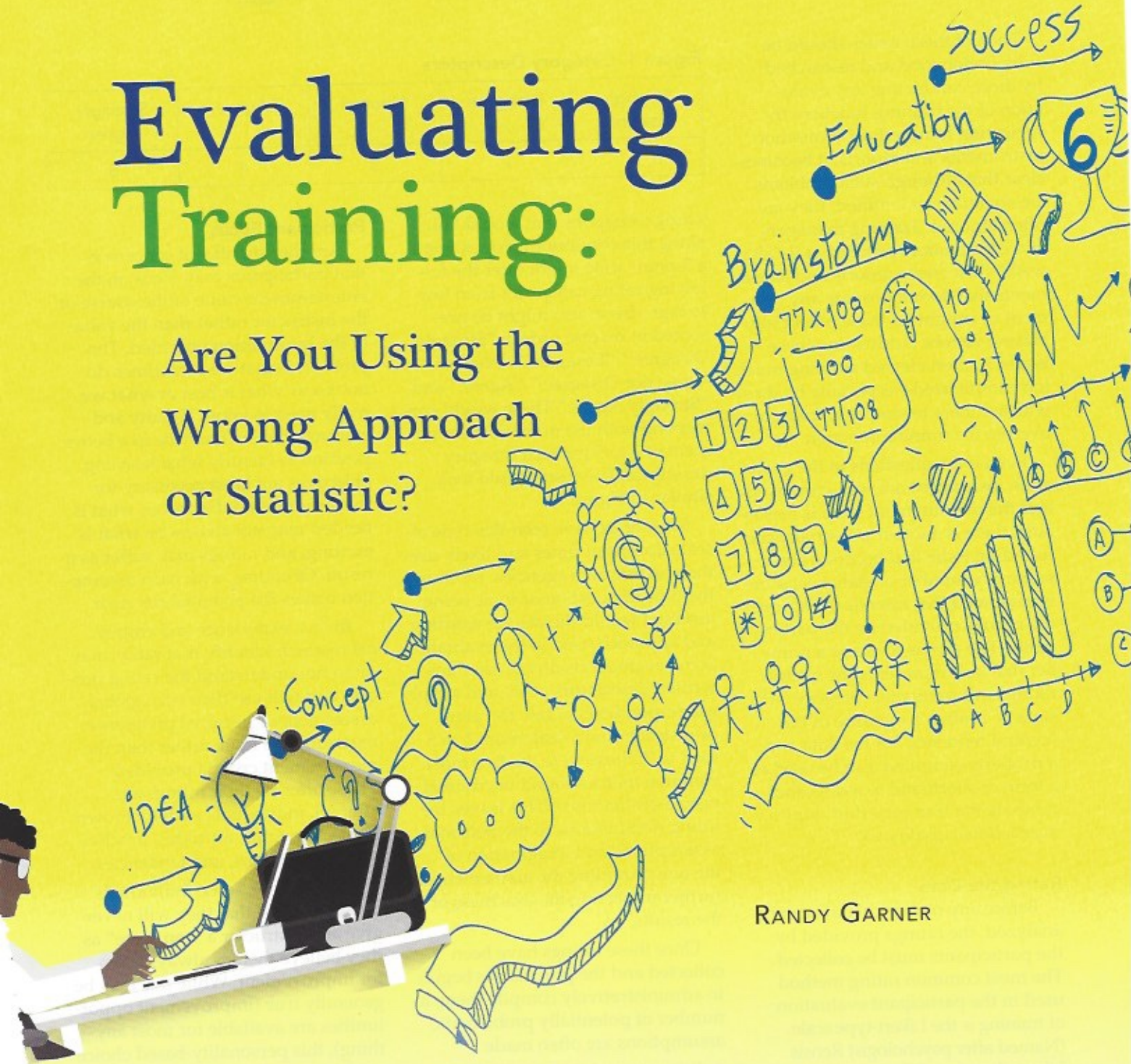
Evaluating Training:

Are You Using the Wrong Approach or Statistic?



RANDY GARNER

Effectively evaluating agency training efforts is important, but doing this well can often be elusive. There are a number of factors that can impact the success of any program evaluation. There can be significant concerns with the way the evaluation is formatted, the manner in which the questions are formed or asked, the evaluation of the so-called "feel good" or "afterglow" aspect of the training rather than something more meaningful, the scale or rating methods used, or the statistical tools employed to assess the training.

Each of these issues should be fully understood and researched by those conducting any evaluation of the training function. In reviewing hundreds of evaluation instruments and reports, it becomes clear that although such problems are sometimes identified, they are often ignored. Training managers and coordinators proceed as if "all is well," then create reports and assessments that do not provide any comment or insight on these potentially biasing factors. Addressing these issues is often viewed as being much too complicated—especially by those who are there to simply manage or oversee the training function.

Even if the questions in the evaluation instrument are well-conceived and the measurement scale is reasonable, there is another potentially confounding issue that occurs when a summary number or statistical score is generated that intends to capture the ratings provided by the participants. The goal is to create a number that best represents the overall ratings in a particular category. This statistic is often a measure of the "central tendency" of the data and, if properly compiled, can be used as a form of shorthand to assess and compare the training effort, topics, instructional quality, etc.

Rating the Data

Before any numerical data can be analyzed, the ratings provided by the participants must be collected. The most common rating method used in the participant evaluation of training is the Likert-type scale. (Named after psychologist Rensis Likert, there is a difference between a Likert scale and a Likert-type scale. Because the prescribed process for development of the scale is typically not followed, the latter term is correct.)

After a training event, participants are typically asked to rate a topic, issue, or performance based on an interval scale that offers increasing or decreasing evaluative terms such as "Strongly Agree" to "Strongly Disagree." These may be further divided into 5 to 7 distinct rating categories that would fall along this continuum. For example, a 5-point scale might offer the following rating categories from low to high (these also might be presented in reverse order): "Strongly Disagree," "Disagree," "Neither Agree nor Disagree," "Agree," and "Strongly Agree." These would correspond with the numerical assignment of 1 to 5 for each category descriptor. The result could look similar to Figure 1.

Figure 1. Category Descriptors

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

Most of us have seen this type of scale dozens of times and likely use them in our own agencies. Of course, the type of question or topic being assessed would dictate the wording and approach used in the evaluation. Some evaluations do not offer any written descriptors at all, and simply ask respondents to rate the issue or instructor on a scale from 1 to 5 (or 1 to 7; these were the two most common metrics found in criminal justice evaluations). This leaves the rating designations to the personal assumptions and imagination of the one providing the assessment, further impacting the exactness of the results.

Once these ratings have been collected and the process has begun to administratively compile them, a number of potentially problematic assumptions are often made:

- It is assumed that those offering the ratings understand the rating process and descriptors.
- It is assumed that the participants are fully present, attentive, and can adequately assess the training.
- It is typically assumed that people will be fair in the process.
- It is assumed the raters will follow the instructions and solely focus on the exact question being asked.

Participant Focus

Another significant concern is that participants may focus on the entertainment value of the course or the instructor rather than the value of the information provided. The problem is that we sometimes do not know what is best or what we really need to know. Faculty and instructional staff may be in a better position to identify what learning objectives are most necessary for the topic offered. However, what is needed may not always be what is exciting, and ratings may suffer as a result. Of course, with each assumption comes the potential for error.

In fact, experience and empirical research identify that raters may often not understand the rating process and will use their own global assessment or personal preference regarding the issue rather than the measurement criteria provided. Those offering their ratings are, after all, individuals with their own biases, thoughts, attitudes, proclivities, assumptions, and tendencies.

For example, some individuals have indicated that they will never give any instructor a "top score" as they believe there is always room for improvement. While this may be generally true (improvement opportunities are available for most anything), this personality-based choice can be in conflict with the design or intention of the measure. This could be analogous to a teacher declaring he or she will never assign any student an "A" because there is always the possibility for improvement.

Realizing that people arrive for a training event with varying backgrounds and degrees of knowledge on the subject, one may ask that certain assessments of a topic be based on what might be beneficial for a larger group or for a greater

purpose, rather than what might suit any single individual. Despite this admonition, some will continue to focus on their personal preference—not what might be more helpful generally.

As an example, consider an introductory class in understanding statistical issues in decision-making. Though this class has been well-received in general and can help individuals better understand some of the statistical information that they must work with every day, a participant who is already very familiar with the subject may focus on his or her personal perception that he or she did not find as much added value from the class. Rather than assess what might be of benefit from a more global perspective or the potential benefit for others that could be derived from that training, someone with a mathematical background may focus on his or her particular perspective and offer a marginal or unflattering assessment.

Additionally, an individual may have a hidden agenda—such as not wanting to participant in training. As a result, he or she makes that position known through an errant evaluation of a particular training event or class rather than honestly assessing the event itself. Obviously, this adversely impacts the assessment of that particular training and skews the results. Others may provide an evaluation of something completely unrelated to the question asked.

A recent review of evaluation instruments found that individuals were replying to questions on specific topics with responses that were well off the intended target. Responses such as "the room was too cold" or "the water fountain is not working" were found to questions inquiring about the quality of the instructor—obviously feedback related to the physical environment was inconsistent with the intent of the question. Unfortunately, those processing the evaluations took no steps to parcel out these unrelated responses, and those evaluations and rating scores were included in the overall assessment of the instructor.

When considering these variables, it does not take a significant leap to realize that evaluation instruments may not always assess what was intended. The problem is that training personnel often ignore these potential complications and biasing influences and simply "run the numbers." Naturally, the way in which one quantifies and analyzes the numerical ratings can have the potential to further exacerbate the situation.

Averages

Once a training event occurs and evaluations are collected, the typical procedure identified by most training coordinators is to simply average the individual scores and come up with a number (a statistic) that represents the collective rating. Of course, the issue of averaging is a bit more complex and deserves greater consideration by the training coordinator. One of the identified problems is with the use of the term "average," because there are many statistical approaches that can be identified as an average. The goal is to fairly arrive at one number that is a measure of the central tendency (the typical or descriptive value) of the data and that best represents all of the data points collected.

Measures of central tendency can include the *mode* (most frequently occurring rating), the *median* (the data point that is at the center of the numerical array) and the *mean* (adding all ratings and dividing by the number of raters). There are many other averages, including quadratic averages, harmonic averages, and so forth. All of these can be properly categorized as an average; however, they do not all measure the data in the same way and can produce potentially dramatic differences.

Unfortunately, the overwhelming evidence demonstrates that evaluations of training events almost exclusively rely on the use of the mean as the preferred (and often sole) method of assessing the data. But it is simply inappropriate to repeatedly use the mean as the exclusive statistical measure to analyze evaluation data. To ensure that one is using the proper statistical approach, one must first assess the data to derive the most appropriate statistic to employ. In fact, the mean is precisely the wrong tool to use when the data might contain outliers.

An outlier is a statistical term for data that is inconsistent with the other observations. The training data must be reviewed to determine if there are outliers present before one can reasonably decide on the appropriate analysis. If this is not done, there is the potential for creating a depiction of the training event that is both inaccurate and misleading.

To illustrate this point, consider Figure 2, a fictional array of numbers representing the evaluation of your instructional effort at a train-

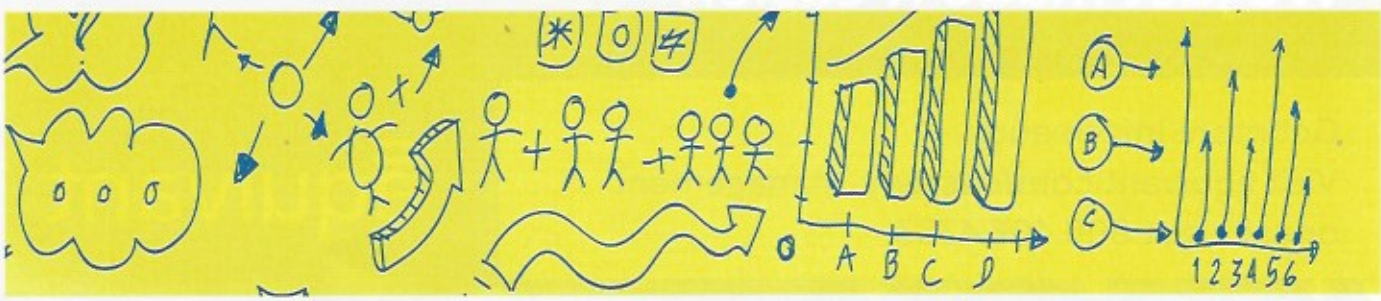


Figure. 2. Example of Instructor Ratings

7	7	6	7	5	7	6	7	1

ing event. Each number represents how an individual participant rated you on a 1 to 7 scale where 7 is "Excellent" and 1 is "Poor."

In this case, the normally used mean would indicate that your typical teaching score was a 5. Not a bad score; however, also not truly representative of the data or your efforts. By using the mean, your teaching will not be categorized in the top tier as "Excellent." Clearly, there is a single data point—and outlier—among the evaluations that is inconsistent with the others in the group. This person may have been having a bad day, may dislike the topic, may have been dealing with family issues, or may object to you personally for some reason. In any case, this single score does not fit with the others.

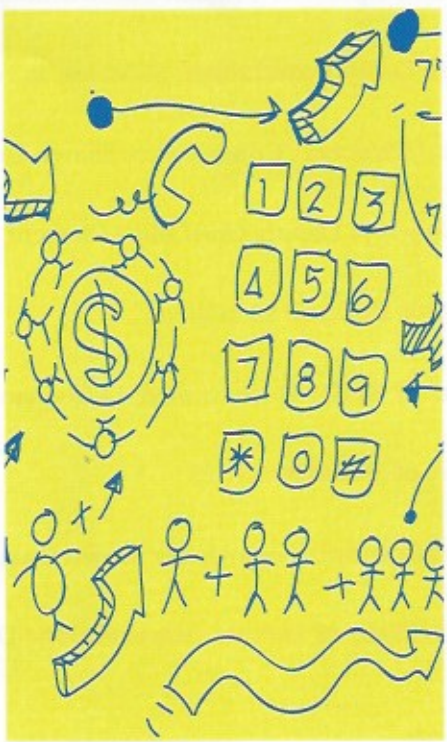
As a result, when one uses the mean to calculate the central tendency or typical score with this errant data point, it pulls the mean downward and does not necessarily provide a true picture. However, if one were to recalculate these data using a different measure of central tendency—either the mode or the median, which are not similarly impacted by this single outlier—the actual teaching score would be more properly represented as a 7 ("Excellent"). For simplicity, this was a very small data set to illustrate the point; however, the result of using the wrong analysis can easily become much more dramatic.

Another option could be the appropriate statistical response to exclude outliers from the analysis and recalculate. If this is done, it should be clearly noted so there is no suggestion that any attempt was made to improperly or unduly influence the results. If one is uneasy with eliminating any data from consideration, the best approach would be to use a statistical technique other than the mean when outliers are present.

A Tool

A statistic is simply a tool, and like any tool, it can be used well or it can be used poorly. It is not the number itself (e.g., mode, median, mean) that has meaning; it is what that number represents that is important. The reason for developing any statistic is to generate sound information from which to base decisions. These decisions might be related to how a course is conducted, what topics are offered, or who is selected to teach in the future—all important issues.

Obviously, the desire is to have the best data possible, so that the best possible decisions can be made. Simply doing what has always been done by calculating a quick average and reporting that as the holy grail of analysis is at best flawed and at worst disingenuous. Decision-makers should have the most reliable and valid data possible. Increased attention to issues such as those presented is equally owed to the instructors who are professionally conducting the training classes.



It is recommended that those responsible for the evaluation of training and other such programs reconsider their assessment process and protocols. Providing additional and well-targeted analysis can often give the best picture of the collected data. Instead of simply reporting the traditional mean, report the mode and median as well. Though a topic that is more in-depth than current space allows, it is also recommended that one calculate, review, and report the standard deviation. This statistic measures the amount of variance or deviation occurring in the data. In fact, assessing the standard deviation can tell one that something is skewed with the data and provide information that it may be necessary to dig a bit deeper to get the true statistical picture.

None of these techniques requires extraordinary effort or sophisticated, expensive software. In fact, one can do an online search for "Mode, Median, Mean, Standard Deviation Calculator" and identify a number of sites that allow one to simply enter the data and hit "calculate." The results will be available instantaneously.

Ultimately, success depends on making sound decisions and properly assessing your training efforts. By considering some of the biasing influences outlined here, better measures, improved analytic techniques, and enhanced evaluative protocols can be developed to ensure that important training decisions are based on the best possible representation of the data. ■

Randy Garner, PhD, is a Professor of Behavioral Sciences in the College of Criminal Justice at Sam Houston State University in Texas and is a primary instructor in the National Jail Leadership Command Academy. He can be contacted at *rgarner@shsu.edu*.