TOPIC 29 RELIABILITY AND ITS RELATIONSHIP TO VALIDITY

A test is said to be **reliable** if it yields *consistent* results. It is easy to see what this means by considering an extreme example. Suppose your professor writes a midterm exam on research methods that contains only four multiple-choice items. The items are on four different important concepts that were emphasized during instruction. Thus, the exam is valid in the sense that it covers appropriate content. Even students who have thoroughly mastered the course content, however, should be concerned about taking such a test because it would be very easy to misinterpret a question or to miss a key term in it and get it wrong, yielding a score of 3 out of 4 right, or 75% correct. On the other hand, students who have moved through the semester in a fog—not understanding even basic concepts—should be pleased at the prospect of taking this exam. With only four items, the odds of getting a few right by guessing and thus passing the test, are reasonably high.

Now suppose some students complain about their scores on the midterm, so your professor writes four new multiple-choice items, and again, they are all on appropriate content. After administering the test at the next class meeting (without announcing there would be a second test, so students are not motivated to study again), should your professor expect to obtain the same scores as he did the first time? In all likelihood, no. Some students who were lucky in guessing the first time will have their luck wash out. Other students who misinterpreted a key term in a question will not do so on the new set of items. Examining the scores from the two tests provides your professor with information on the *consistency of results* or *reliability*. In this case, he or she would probably find that the scores are rather inconsistent from one test to the other.

What can your professor do to increase the reliability of his or her midterm? Obviously, your professor can increase the length of the test. This reduces the effects of the occasional ambiguous item and the effects of guessing. After realizing this principle, your professor instructs a graduate assistant to prepare a 100-item test overnight. The assistant, being pressed for time, takes the easy route and pulls a standardized test off the shelf. Although it has 100 items, they are on educational psychology, which includes some research concepts but also much material not covered in the research methods class. Administering this test should give highly reliable results. If we administer it twice, for example, those who do not know the name of the "father of educational psychology" the first time the test is administered will not know his name the second time.[1] Likewise, those who have a good command of educational psychology should do well on both tests. Also, those who do not know the material will have little chance of getting a good grade by guessing on such a large number of items. However, the professor has created a new problem: The test lacks **validity** because it covers the wrong content (see Topics 25 through 28). This example illustrates an important principle: *A test with high reliability may have low validity*.

Here is another example of this principle: An employer wants to reward the best employees with end-of-year bonuses. The employer decides that to be perfectly fair, a completely objective method for determining who should get the bonuses should be used. To do this, the employer examines the employees' time cards, and selects those who were never late for work during the previous year to receive bonuses. Notice that this method of measurement is highly reliable: Another person could independently perform the same procedure and, if careful, would identify exactly the same employees for bonuses, yielding consistent (reliable) results. But is the procedure valid? Probably only minimally so because the employer's measurement technique is limited to only one characteristic. Those who are outstanding in a number of other ways (such as identifying more effective ways to advertise products) but who were late to work even once are excluded from getting bonuses. Thus, the procedure is reliable, but it is of questionable validity.

This brings us to the next principle: When evaluating instruments, *validity is more important than reliability*. This should be clear from considering the example of the employer basing bonuses on employees' time cards. A complex measure involving subjective judgments of employees' performances that taps a variety of important types of behavior and achievement on the job would be much more valid (even if it turned out to be only modestly

[1] Edward Thorndike of Teachers College (Columbia University) is widely cited as the father of educational psychology primarily because of his influence in promoting the use of empirical methods for studying education.

reliable) than a highly reliable measure that considers only punctuality.

Finally, there is a third principle: *To be useful, an instrument must be both reasonably valid and reasonably reliable.*

To understand the complex relationship between reliability and validity, consider Figures 1 through 4 below.

In Figure 1, the gun is aimed in a valid direction (toward the target), and all the shots are consistently directed, indicating that they are reliable.

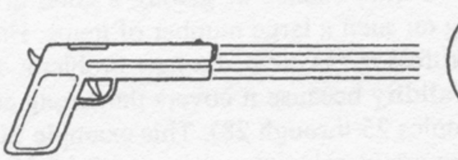

Figure 1. Reliable and valid.

In Figure 2, the gun is also aimed in the direction of the target, but the shots are widely scattered, indicating low consistency or reliability. The poor reliability makes it unlikely we will hit the target. Thus, poor reliability undermines an attempt to achieve validity.

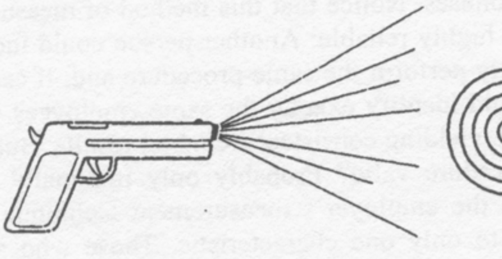

Figure 2. Unreliable, which undermines the valid aim of the gun. Not useful.

In Figure 3, the gun is not pointed at the target, making it invalid, but there is great consistency in the shots, indicating that it is reliable. (In a sense, it is very reliably invalid.)

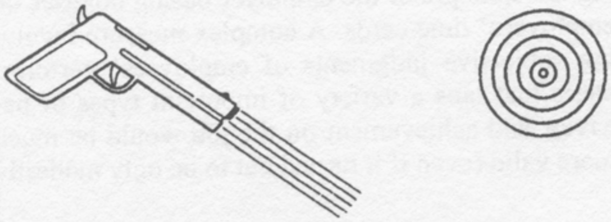

Figure 3. Reliable but invalid. Not useful.

In Figure 4, the gun is not pointed at the ta making it invalid, and the lack of consistency i direction of the shots indicates its poor reliabilit

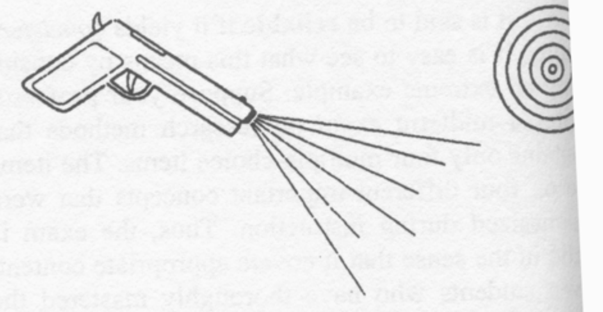

Figure 4. Unreliable and invalid. Not useful.

Of course, Figure 1 represents the ideal in mea urement. For most measures in the social and b havioral sciences, however, we should expect th direction of the gun to be off at least a sma amount, indicating less-than-perfect validity. W should also expect some scatter in the shots, indi cating less-than-perfect reliability.[2] Clearly, ou first priority should be to point the gun in the cor rect *general direction*, which promotes validity Then, we should work on increasing reliability.

[2] Examination of the technical manuals for publishe tests indicates that professional test makers tend to b more successful in achieving high reliability than i achieving high validity. This is because it is relativel easy to increase reliability by increasing the number c objective-type test items, while the tasks that need to b undertaken to increase validity vary greatly from con struct to construct and may not be obvious.