and what the appropriate time interval between tests is. Usually, for very short time intervals (hours or days), most attributes are quite stable, and a large test–retest r (r = .90 or higher) should be expected. Over longer time intervals, it is usual to expect much lower r's, depending on the attribute being measured. For example, over six months or a year, individuals' knowledge of programming languages might change. If so, there will be lower test–retest reliabilities (e.g., r = .50).

Intrarater Agreement. To calculate intrarater agreement, scores that the rater assigns to the same people in two different time periods are compared. The calculation could involve computing the correlation between the two sets of scores, or it could involve using the same formula as for interrater agreement (see Exhibit 7.8).

Interpretation of intrarater agreement is made difficult by the time factor. For short time intervals between measures, a fairly high relationship is expected (e.g., r = .80, or percentage agreement = 90%). For longer time intervals, the level of reliability may reasonably be expected to be lower.

Implications of Reliability

The degree of reliability of a measure has two implications. The first of these pertains to interpreting individuals' scores on the measure and the standard error of measurement. The second implication pertains to the effect that reliability has on the measure's validity.

Standard Error of Measurement. Measures yield scores, which in turn are used as critical inputs for decision making in staffing activities. For example, in Exhibit 7.1 a test of knowledge of programming languages was developed and administered to job applicants. The applicants' scores were used as a basis for making hiring decisions.

The discussion of reliability suggests that measures and scores will usually have some amount of error in them. Hence, scores on the test of knowledge of programming languages most likely reflect both true knowledge and error. Since only a single score is obtained from each applicant, the critical issue is how accurately that particular score indicates the applicant's true level of knowledge of programming languages alone.

The standard error of measurement (SEM) addresses this issue. It provides a way to state, within limits, a person's likely score on a measure. The formula for the SEM is

$$\text{SEM} = \text{SD}_x\sqrt{1-\text{r}_{xx}}$$

where SD_x is the standard deviation of scores on the measure and r_{xx} is an estimate of the measure's reliability. For example, if $\text{SD}_x = 10$ and $\text{r}_{xx} = .75$ (based on coefficient alpha), SEM = 5.