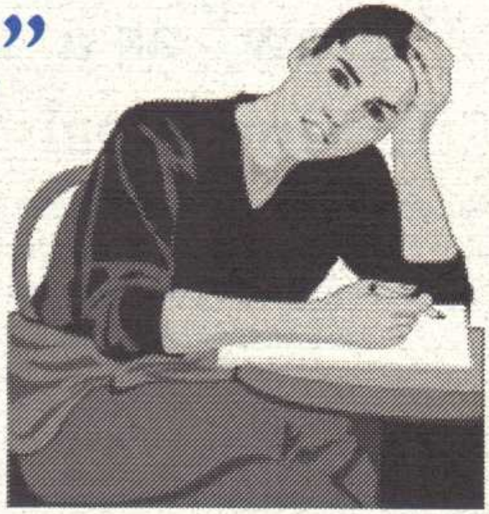
What is the "Right" Test Length?



The "right" test length is more folklore and accident than intention. Anastasi assures us that "other things being equal, the longer a test, the more reliable it will be." Unfortunately "other things" are never equal. Nunnally mandates that for "settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated." Unfortunately he does not explain how to determine the test length that gets a .90. That's because reliability is an awkward amalgam of the length and targeting of the test, and the spread of the examinees who happen to take this test.

What's wrong with a one-item test?

1) Content Validity

To be useful a test must implement the one intended dimension. We assert our singular intention through the formulation of test items. But each item, in all its reality, inevitably invokes many dimensions. No matter how carefully constructed, the single item will be answered correctly (or incorrectly) for numerous reasons. The unidimensional intention of a test only emerges when this intention is successfully replicated by essentially identical, yet specifically unique test items. Whether an item requiring Jack and Jill to climb a hill contributes to test score as a reading, physics, or social studies item depends on the other items in the test.

2) Construct Validity

The various items in a useful test replicate our singular intention sufficiently to evoke singular manifestations we can count on to bring out the one dimension we seek to measure. Arithmetic addition is usually intended to be easier than multiplication. We could write hard multiple-digit additions that would be more difficult to answer than simple single-digit multiplications. But such a test would not realize our intention to measure increasing arithmetic skill in an orderly and easy-to-use way. Once we have successfully implemented our construct, the qualifying items define our variable, and their calibrations provide its metric benchmarks.

3) Fit

A useful test gives examinees repeated opportunity to demonstrate proficiency. An examinee may guess, make a careless error, or have unusual knowledge. One, two, or even three items provide too little evidence. We need enough replications along our one dimension to resolve any doubts about examinee performances. As doubts are resolved, the relevance of each response to our understanding of each examinee's performance becomes clear. We can focus attention on the responses that contribute to examinee measurement, reserving irrelevant responses (guesses, scanning errors, etc.) for qualitative investigation.

4) Precision

A useful test must measure precisely enough to meet its purpose. The logit precision (standard error) of an examinee's measure falls in a narrow range for a test of L items: $2/L < \text{SEM} < 3/L$. Doubling precision (halving the standard error) requires four times the items. The placement of examinee measures and confidence intervals (±SEM) on the calibrated variable shows us immediately whether the test has provided enough precision for the decisions we need to make.

When there is a criterion point, it is inevitable that some measures will be close enough (less than 2 SEM) to leave doubt whether the examinee has passed or failed. In these cases, an honest, but statistically arbitrary, pass-fail decision may have to be made. There is no statistical solution. Increasing the number of items increases test precision, but we always reach a point at which we no longer believe the added precision. If your bathroom scale reports your weight to the nearest pound, you could weigh yourself 1000 times and get an estimate of your weight to within an ounce. But you would not believe it. Your weight varies more than an ounce and, indeed, more than a pound over the course of a day.

So what is the "right" test length?

1) Enough items to clarify the test's intention and replicate out a unidimensional variable.

2) Enough person responses to each item to confirm item validity and provide a calibrated definition of the variable.

3) Enough item responses by each examinee to validate the relevance of this examinee's performance.

4) Enough responses by each examinee to enable precise-enough inferences for the decisions for which the test was constructed and administered.

Ben Wright