Chapter 5

Reliability Theory for Total Test Scores

The basic motivation for classical true-score theory is to provide a workable method for estimating the precision of measurement of a test score. Generically, this is the problem of *test reliability*. The mathematics of the theory is extremely simple. The application of the theory can be problematic.

Let us begin by considering the contrasting case of estimating the precision of a physical measurement—length. Given a graduated rule, if we wish to estimate the error of measurement of the length of an object, we will replicate the measurement operation a number of times. We take the mean of the measurements as our best estimate of its length, and their *SD* is our best estimate of the error in the measuring process. This procedure rests on the reasonable assumption that the replications are independent trials, so the errors are independent of each other, and therefore uncorrelated. There are, of course, further assumptions implicit in this process. Thus it is assumed that the graduated instrument contains no source of constant error, and that the length of the object does not change over the time period during which the repeated measurements are taken. (If, for example, we were measuring lengths of railroad track with a plastic tape, we might have to allow for expansion of the object with temperature.) But with appropriate precautions, such replicated measurements give us the amount of error with which the test instrument measures the specified property of the object, in the metric defined by the instrument—say, centimeters or inches. (This is the very simplest case in physical measurement. A deeper analysis would show that there are nontrivial problems to be found in all physical measurement, but these do not concern us here.)

As soon as we imagine a problem in psychological measurement, we see difficulties over replicating the measurement. If we propose to administer a psychological test a large number of times to a single examinee, we see how unlikely it is that the replications could constitute independent trials and yield uncorrelated errors. The closer together in time the retests are given, the more similar the test scores will be, due to factors such as memory of previous responses. After no more than, say, three replications of the test procedure, the motivation of the examinee will probably decline, stereotyped responses will ensue, and the results will not constitute independent trials. Further, few of the attributes we measure will be strictly constant over time. An extreme case is mood. We only have to consult personal experience of self or others to recognize that emotional states are characterized by lability, and some people are more labile than others. Thus a short time interval between retests will make the responses spuriously alike, while a long time interval will allow change in the attribute to be measured. And there is no clear way to choose an intermediate time interval at which the attribute is unchanged and the repeated measures are statistically independent. No trait is stable from the cradle to the grave. Some states are stable enough to be treated as though they are traits.

There is a further point of contrast with the simpler forms of physical measurement. In the case of length, it does not seem possible to question whether we are measuring the property of the object we wish to measure. Indeed, we may find it hard to distinguish conceptually between the graduation mark we read off the scale and the length of the object it has been laid against. In the case of a psychological attribute we can see a clear distinction between the concept—extraversion, intelligence, attitude toward gun control, positive affect—and the total score on a specific set of items selected to measure it. There have been three apparently distinct ways to conceive the relationship between the test score and the attribute:

1. It has been treated as the precision with which the test score measures the attribute—the *reliability* of the test as a composite indicator of the attribute.
2. It has been treated as the extent to which the test measures the attribute it was designed to measure—the *validity* of the test as a composite indicator of the attribute.
3. It has been treated as the extent to which the composite test score generalizes beyond the specific items chosen to form the composite, to the domain of further indicators that might have been used—the question of *generalizability*, from items we have to items we do not have.

These distinctions are generally hard to sustain in practice.

One conclusion that might seem to follow by now is that the problem in the opening statement—how to provide "a workable method for estimating the precision of measurement of a test score"—has no solution. Another is that the problem as stated contains ambiguities requiring analysis. We can postpone the conceptual issues by treating the classical true-score model as a piece of pure mathematics, and separating the question of whether it can be applied to empirical data, and under what conditions. For this purpose it can be illustrated with artificial, simulated data. The strategy adopted in this chapter is to describe the classical true-score model as a piece of mathematics, illustrated with random numbers, in the next section. The following section then provides an account of the necessary assumptions (and attendant problems) for the two main techniques for applying the theory—test–retest and alternate-form methods.

THE TRUE-SCORE MODEL FOR TEST SCORES

We simulate the process of drawing a single examinee at random from a population of interest, administering a test consisting

of m items, and form a total number-right or number-keyed score Y. We suppose that Y consists of the sum of two parts—a random component T and a random component E, which is independent of T. We can call T a *true score* and E an *error*, but for the present these are just quantities in a mathematical equation. Thus we write simply

$$Y = T + E. \tag{5.1}$$

For the moment we do not interpret these variables. Table 5.1, row 1, presents a simulation of the process of drawing a series of random numbers T from a table and, for each T, adding independent random numbers E. The process was stopped at 10 random drawings, but is imagined to continue indefinitely.

I chose numbers T and numbers E with variances σ_T^2 and σ_E^2 known to me, of course, but not to you. From this information alone, even if you had the "population," because I have hidden the numbers making up Y, you can draw up the distribution of Y, compute its mean and variance, and examine its shape, but that is about all. By their construction, you know some general properties of the components T and E. You know the following facts:

TABLE 5.1
True-Score Model Simulation

		"Subject"										
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	· · ·
P_1	Y	48	100	62	14	44	66	40	94	34	78	· · ·
	Y'	32	100	72	14	44	54	48	96	24	78	· · ·
P_2	Y	89	91	76	73	76	89	81	81	84	90	· · ·
	Y'	100	87	86	73	76	77	89	83	74	90	· · ·

1. T and E are measured on the scale of Y and are bounded within the range of Y, having the same floor and ceiling.
2. T and E are uncorrelated, that is,

$$\rho_{TE} = 0, \tag{5.2}$$

because E is chosen independent of T.

3. The variance of Y is the sum of the variances of T and of E. That is, with σ_T^2, and σ_E^2 for the respective variances of T and of E, and σ_Y^2 for the total variance of Y.

$$\sigma_Y^2 = \sigma_T^2 + \sigma_E^2 \tag{5.3}$$

(because variances of uncorrelated variables add).

4. The variances of T and of E are both less than and at most equal to the variance of Y. That is,

$$\sigma_T^2 \leq \sigma_Y^2 \quad \text{and} \quad \sigma_E^2 \leq \sigma_Y^2.$$

5. The ratio of the variance of T to the variance of Y,

$$\rho_{r} = \sigma_T^2/\sigma_Y^2 = \sigma_T^2/(\sigma_T^2 + \sigma_E^2) \tag{5.4}$$

is bounded by zero and unity. That is,

$$0 \leq \rho_{r} \leq 1. \tag{5.5}$$

The variance ratio r is by definition the *reliability coefficient* of Y. This is a fundamental theoretical quantity in test theory.

But these properties are not very informative. Now suppose I give you a second total score Y, from each of the randomly drawn "examinees" where Y (to be read as "Y prime") is the sum of the same T as before and an independent E also with variance σ_E^2, and so we write

$$Y' = T + E'. \tag{5.6}$$

In row 2 of Table 5.1(a) I have simulated this process, drawing further numbers E from a table of random numbers, and adding them to the same numbers T that were used to construct the first row. So Y and Y have the same randomly drawn T

value, and independently drawn E and E values. By construction, therefore, E and E are uncorrelated with T and with each other. That is,

$$\rho_{TE} = \rho_{TE'} = \rho_{EE'} = 0. \tag{5.7}$$

Also by construction they have equal variances, i.e.,

$$\sigma_{E'}^2 = \sigma_E^2. \tag{5.8}$$

It follows that

$$\sigma_{Y'}^2 = \sigma_Y^2, \tag{5.9}$$

so Y also yields properties 1–5 like Y.

A further property now follows, namely, that

$$\rho_{YY'} = \rho_r, \tag{5.10}$$

where

$$\rho_r = \sigma_T^2/\sigma_Y^2, \tag{5.11}$$

and ρ_{YY} is the correlation between Y and Y. The important consequence is that ρ_r can now be computed from observations, and estimated from a finite sample. That is, the correlation coefficient between Y and Y gives the reliability coefficient of Y (or, equally, of Y). (We would not normally expect a variance ratio to equal a correlation coefficient. This will seem less strange when it is noted that

$$\rho_r = \rho_{YT}^2 = \rho_{Y'T}^2, \tag{5.12}$$

that is, the reliability coefficient is the square of the correlation between Y and T or Y and T. The latter correlation is known as the *reliability index*. Although of interest to psychometricians, it is not commonly quoted by users of reliability theory.)

*** Proof of (5.10): By Appendix A,

$$\begin{aligned}\mathrm{Cov}\{Y, Y'\} &= \mathrm{Cov}\{(T + E)(T + E')\} \\ &= \sigma_{TT} + \sigma_{TE} + \sigma_{TE'} + \sigma_{EE'} \\ &= \sigma_T^2,\end{aligned}$$

and $\mathrm{Var}\{Y\} = \mathrm{Var}\{Y\}$. $$$

Note. The subscript r on ρ_r is preferred here because it draws attention to the facts that (a) ρ_r is, by definition, the reliability coefficient, and (b) it is, by definition, a ratio of variances. Some accounts use ρ_{YY} as representing both the variance ratio and the correlation from which it is obtained in applications. This is quite correct, given the mathematical equivalence, but it loses useful conceptual distinctions.

In the simulation in Table 5.1, you are still unable to derive the three component scores, T, E, and E, from just two numbers Y and Y. Two knowns cannot give three unknowns. But you can estimate the correlation ρ_{YY}, and this is an estimate of the variance ratio ρ_r. In fact, I chose numbers, with equal probability, from the 10 numbers 0, 10, 20, …, 90, for T and two sets of numbers, with equal probability, from 0, 2, 4, 6, …, 18, for E and E, by multiplying random digits 0 through 9 by 10 and by 2 respectively. This means I know precisely that $\sigma_T^2 = 825$, $\sigma_E^2 = 33$, and so $\sigma_Y^2 = 825 + 33 = 858$, and $\rho_r = .9615$. If you estimate these quantities from the rather small sample in the first two rows of Table 5.1, you will find out how close it is to .96. You can also check the sample variances of Y and Y, using (3.6) and correcting for bias, to see how close they are to each other and to the population value 858.

The reliability coefficient is merely a means to an end. The ultimate object is to obtain an estimate of the variance of E in the metric of the test score. The expression for this,

$$\sigma_E^2 = \sigma_Y^2(1 - \rho_{YY'}), \tag{5.13a}$$

is just a rearrangement of (5.4). It gives us what generally is not known, σ_E^2, in terms of two quantities, σ_Y^2 and ρ_{YY}, that can be estimated from samples. Using the population values—as though we had taken an infinite sample—we have

$$\sigma_E^2 = 858(1 - .9615) = 33.0.$$

Given data such as those in Table 5.1, we can compute estimates $s_Y^2, r_{YY'}$ of σ_Y^2 and $\rho_{YY'}$, as already suggested, and now we can get an estimate s_E^2 of σ_E^2 from these.

As a further step, we define the *standard error of measurement* of the test score Y, as

$$SEM\{Y\} = \sqrt{\sigma_E^2}. \tag{5.13b}$$

It is estimated by a sample counterpart

$$SEM\{Y\} = \sqrt{s_E^2}. \tag{5.13c}$$

In the artificial example, with known population values, we have $SEM\{Y\} = 5.74$ ($= \sqrt{33}$). If the test score had understandable units—say number right out of a hundred—we could read the $SEM\{Y\}$ as approximately six correct answers.

In interpreting the true-score model we think of T as characteristic of the examinee, and E as characteristic of the test. Suppose now that in a second population of interest, variability of T is much smaller, but the variability of E is unaltered. In Table 5.1, we think of the first two rows as being from population P1. To begin a simulation of a population P2, with smaller σ_T^2, I took the previous values of T, divided them by 10, added 68, and added the same E and E to get rows 3 and 4 in Table 5.1. This divides σ_T^2 by 100, to give 8.25, and it leaves σ_E^2 unaltered, giving $\sigma_Y^2 = 8.25 + 33 = 41.25$ and $\rho_{YY'} = .20$. But we still obtain $\sigma_E^2 = 41.25(1 - .2) = 33.0$. As an exercise, the student should estimate $\rho_{YY'}$, σ_Y^2, $\sigma_{Y'}^2$, and σ_E^2 from rows 3 and 4.

In applications of this model, we can usually expect that the reliability coefficient will vary according to the population sampled. This is because in practice the variance of T represents the variability in each population of the attribute we are measuring. Under reasonable assumptions the variance of E will remain approximately invariant, even though the reliability coefficient varies with the population sampled.

APPLICATIONS OF THE MODEL

To apply the simple model (5.1), we wish to interpret T as the true score of an examinee, and E as an error of measurement. A possible link to applications has been described by Lord and Novick[1]:

> The correlation between truly parallel measurements taken in such a way that the person's true score does not change between them is often called the *coefficient of precision*. For this coefficient, the only source contributing to error variance is the unreliability or imprecision of the measurement procedure. This is the variance ratio that would apply if a measurement were taken twice and if no practice, fatigue, or memory factor affected repeated measurements. In most practical situations, other sources of error variation affect the reliability of measurement, and hence the coefficient of precision is not the appropriate measure of reliability.

The quoted remarks in effect permit two distinct views. One is that there is an ideal situation—"taking measurements" without practice, fatigue, memory, and other effects in which the correlation between the two measurements gives the precision of the test score. The other is that when other sources of "error" are possible, the coefficient of precision, and the resulting "error" variance, are not what is wanted. And we do not yet have a workable definition of "error."

The three main recognized methods for estimating the reliability coefficient of an objective test from real data are (a) test–retest methods, (b) parallel or alternate-form methods, and (c) internal analysis. The first two of these rest on the correlation between two total test scores, and directly apply the theory of the preceding section. The third requires theory concerning relations between the items constituting the test, and is dealt with in Chapter 6.

In retest methods, a test of m items is administered to a large sample of examinees at two points in time, yielding pairs of scores Y and Y'. It is proposed that we use these paired scores to estimate $\rho_{YY'}$ and equate this to r. We can then use (5.13) to estimate σ_E^2, and obtain the SE of measurement. This proposal might be justified in two ways. The first is to admit that we are making the strong assumption that the true values of the examinees' scores do not change between administrations. If so, the errors are independent, so the ideal situation yielding $\rho_{YY'}$ as the coefficient of precision has been closely approximated. A problem with this position is that nothing distinguishes a case where the assumption holds and a case where it does not. The second is to say that the *retest true score* is defined as the component of the observed score that does not change "between administrations." Some writers at least implicitly adopt this second option and actually define the resulting reliability coefficient as a *coefficient of stability*. (The danger of conflating this interpretation with the first must then be carefully watched. If an attribute is very unstable over time, and gives a low retest reliability, it may be a mistake, as we demonstrate later, to regard the retest coefficient as the precision with which the attribute is measured.)

The main problem with the second position is that what is observed is a small fraction of a possible larger and more informative study of the behavior of the test score over time. By retesting at a sequence of time intervals we could graph the stability coefficient, as just defined, as a function of time. (To do this, we can take a large cohort of examinees at an initial time and retest subsets of them once only at a series of time intervals. This will avoid the effects of many replications.) Generally, the longer the time interval, the lower the coefficient. We might take into account the stability or otherwise of

relevant environmental factors. We would certainly plot the scores against time for systematic changes—individual curves of growth or decay of the performance measured by the *m* items in the given test. Such a study can be very informative, especially if situational factors allow conclusions about causes of change. However, once this point is reached in the investigation we do not have a single "retest true score" and we do not have a single coefficient that we can call *the* retest reliability or coefficient of stability. Of course, retest correlations contain useful information about the stability or lability of an attribute. They tell us the extent to which it is traitlike or statelike, so to speak. This can be important information. But it is best to obtain a set of retest correlations, over a series of increasing time intervals, if we wish to study either the stability of the measurement or the course that it follows through time.

Accordingly, there are good reasons for conducting longitudinal studies involving repeated administrations of a test. But we do not yet have good reasons for relating such data to the ideal coefficient of precision. This is not to say that it is impossible, merely that it is generally difficult to motivate such a step. Lord and Novick (1968) stated that any coefficient of stability underestimates the coefficient of precision because the "error" variance includes unstable "true" variance. Certain physiological functions, certain sensory or motor tasks, may approximate the conditions for the ideal coefficient of precision. These may also be cases where we do not have a psychological attribute for which "test theory" is necessary or appropriate. Instead, it may become possible to estimate error by a large number of replications.

In *alternate form* methods, two tests—the alternate forms—contain disjoint, that is, nonoverlapping, sets of items, and these are administered, usually close together in time, to a large sample of examinees, to yield pairs of scores Y, and Y. To treat them by the true-score model we must be able to suppose that their variances σ^2_Y, $\sigma^2_{Y'}$, are the same in the population of interest. At least they should not be significantly different as tested in the sample. This is not a very stringent condition. As in the case of test-retest data, we use the scores to estimate $\rho_{YY'}$ we interpret it as r, and we obtain an estimate of σ^2_E, and the *SE* of measurement, by the equations of the preceding section.

In making the transition from retest reliability to alternate-form reliability we should note that the retest reliability of the total score from *m* specific items has no necessary relation to the precision with which we measure the psychological attribute itself. The *m* items are just one set of indicators of it, and possibly not closely related to it. A set of items may have high stability and low alternate-form reliability, or low stability and high alternate-form reliability.

We might regard it as an assumption that each of two alternate forms equally measures the examinees' true scores and that they differ by independent errors of measurement. But then nothing defines the true score or the error, so as to distinguish a case where the assumption holds and a case where it does not. Instead we might define the *alternate-form true score* to be a component in the two total test scores that is common to the forms we have constructed. Then their errors are components of the total test scores that are unique to each form. (This statement should be intuitively understandable already. It should also become clearer after discussion of the common factor model in Chap. 6.) Writers who choose this option define the resulting reliability coefficient as a *coefficient of equivalence* of the alternate forms that have been constructed.

In terms of the mathematics of the model, a given test form can have as many alternate forms as there are tests of the same variance to correlate it with, and as many coefficients of equivalence. These correlations could be thought of as measuring how much the test measures in common with each other test. In applications, restrictions on content will be imposed in the construction of equivalent forms. The hope would be that the coefficient of equivalence will become a coefficient of precision—corresponding to a decomposition into "true" and "error" parts—because the alternate forms will then measure the same attribute. Thus, Lord and Novick (1968) stated that generally a coefficient of equivalence will be less than the (ideal) coefficient of precision because the "error" component will include true-score variability due to lack of parallelism of the tests, but "when conditions of the two administrations are equivalent and the intervening time is short," the alternate forms method produces "a coefficient of equivalence which is close to the coefficient of precision." [2] In the simple classical true-score model of this chapter, *parallel forms* are tests that give parallel measurements, and two parallel measurements are just two test scores with equal variance from which we choose to compute a coefficient of equivalence. We meet other, more stringent requirements for parallel forms later.[3]

In applications, we would expect that conditions will be placed on the substantive content of the items composing each form. We would expect to correlate the scores on two algebra tests to get a coefficient of equivalence, rather than correlating algebra with geometry or with English vocabulary. (But note that algebra and geometry measure mathematical ability in common, and algebra and vocabulary measure scholastic ability in common.) We could require the items in each form to be in some sense equivalent to the items in the other. We want them to be similar yet not identical. This is not a precise requirement. Consider a word-fluency test consisting of one frequency-count item, namely, "Write down as many words as you can think of beginning with the letter E" (time, 5 minutes). Similar items can be obtained by substituting other initial letters. Similar items can also be obtained by the format "Write down as many items as you can think of whose 2nd, 3rd, …, last letter is E." Similarity is, in a sense, multidimensional; different principles of similarity will yield different coefficients of equivalence.

A distinction can be made between content-parallel test forms and content-equivalent test forms. *Content-parallel test forms* are two forms containing the same number of items, in which the items are paired to be similar in content, while distinct items within each form may be less similar. Consider, for example, a general test of intelligence in the forms:

Form L	Form M
(1) What day of the week is it?	(1) What month is it?
(2) If I buy 4 cents' worth of candy and pay 10 cents what change do I get?	(2) If I buy 12 cents' worth of candy and pay 15 cents what change do I get?
(3) Repeat in reverse order 6–5–2–8	(3) Repeat in reverse order 3–6–2–9

(These items are taken from an old version of the Stanford-Binet test, year 9.) The principle should be clear. Having written the first form to measure an attribute, the test constructor tries to write a closely similar but not identical alternate for each item. This requires judgment.

One reasonable conception of *content-equivalent test forms* would be two tests, whose items can be recognized as *content-homogeneous* when they are combined to make a test. Clearly the construction of such test forms requires judgment, and it involves an element of idealization. Consider the test in Table 5.2. Consulting nothing more than our experience of life and our understanding of the words of these statements, we can agree, I hope, that all 14 indicate or exemplify an attribute that could be called "satisfaction with life." [4] (Items 2, 8, and 12 are negative indicators, measuring in the direction of dissatisfaction.) We have met the items marked A–E already.

As will commonly happen, there are a number of ways to form subsets of these items that indicate something more specific than satisfaction with life, and more general than the meaning of any one item. Items 1, 4, 5, 7, 12, 13, and 14 seem to indicate satisfaction with one's present life. Items 2, 6, 9, and 11 seem to indicate satisfaction with one's past life. Items 3, 8, and 10 perhaps indicate emotional lability as opposed to items of a more purely evaluative character. The reader is free to debate these suggestions and to find other reasonable subsets in the same way. The fact that it can be debated does not deny that a set of items can be judged to be content-homogeneous—that judgments can be offered. Having made such a judgment, we can draw two subsets from it to make content-equivalent forms.

The important general point is this: In applications, it is reasonable to select alternate forms from a set of items that are indicators of a common attribute. A judgment of homogeneity of content, where possible, will be strong evidence that an item-set has this property. Statistical evidence requires the methods of Chapter 6.

TABLE 5.2
Satisfaction With Life Items

You should agree or disagree with each item using the 1–7 scale below. Place a number from 1 to 7 next to each item on the answer sheet to indicate your degree of agreement with that item.

7. Strongly agree

6. Agree

5. Slightly agree

4. Neither agree nor disagree

3. Slightly disagree

2. Disagree

1. Strongly disagree

A	1.	In most ways my life is close to my ideal.
	2.	I frequently think about unhappy times or events of my past.
	3.	I am a person who can feel happy very easily.
E	4.	The conditions of my life are excellent.
	5.	I am satisfied with the current state of affairs in my life.
	6.	I like the life I have led.
C	7.	I am satisfied with my life.
	8.	I frequently experience intense negative emotions that make me unhappy.
D	9.	So far I have gotten the important things I want in life.
	10.	When something makes me happy, this emotion usually lasts a long time.
B	11.	If I could live my life over, I would change almost nothing.
	12.	My life does not live up to the standards I have for a good life.
	13.	I am satisfied with my present life.
	14.	If I imagine the most desirable life for myself (the ideal), my life is very close to that point.

Suppose we have tried to create alternate forms that are content-parallel or content-equivalent. We return to the question: In what sense, and to what extent, do two total test scores—from test- and retest forms, from content-equivalent forms, or from content-parallel forms—give information about the precision of measurement of a test?

Perhaps enough has been said to establish that a test-retest correlation—a coefficient of stability—generally bears no clear relation to anything we would regard as the precision of measurement of the test. Even if a test-retest correlation may approximate the coefficient of precision—the extent to which the responses to the given items are stable across replications in unrealizable conditions (no effects of previous responses, fatigue, etc.)—the coefficient of precision is commonly not of interest. In most cases the quantity of interest is the precision with which the attribute itself is measured. This is not the sum score on a particular set of items chosen to measure it. The study of stability as such may be very important in some applications of tests. Note that getting respondents to provide retest measures is generally an expensive research procedure, and may require a good research motive.

Intuitively, it can be seen that the correlation between alternate forms—their coefficient of equivalence—is a possible measure of the precision of measurement of the attribute itself. This requires the condition that all the items in the two forms are indicators of just that attribute. And it might be sufficient evidence for this condition that the items are judged content-homogeneous. Each form contains a distinct set of indicators, and the indicators are related to each other because they are related to their common attribute. In effect, the "true score" T is the common trait measured by all the indicators, measured

in the same units as the raw scores *Y* and *Y*. One limitation of this intuitive interpretation is that there are many ways we could assign 2m homogenous items to two test forms. The coefficient of equivalence would vary over different assignments. Given item data, we would use the methods of Chapter 6. We may accept published alternate-form reliabilities if these are the best figures made available to us, and if enough has been said to let us evaluate the choice of the alternate forms.

The case of content-parallel test forms is rather curious. It actually requires a complicated measurement model. Matched item pairs measure something in common across forms that they do not measure in common with other items in their own form. What they measure in common with their own set is an attribute of a higher level of abstraction from behavior than what is measured by each pair. In the example already given, pairs measure (a) understanding the calendar, (b) problem arithmetic, (c) digit span backward, and so on. Hopefully, following Binet and Terman, we may suppose they combine in either form as indicators of "general intelligence." In consequence, the coefficient of equivalence will be spuriously high as a measure of the ratio of variance due to the trait—in this case "intelligence"—represented by either form to total variance. This is because the numerator includes a sum of shared variances of the paired items. This effect rapidly becomes negligible as *m* becomes large, until content-parallel forms behave just like content-equivalent forms.

To continue, we need theory at the level of items, or at least of subtests. That is one topic of the next chapter. Classical true-score theory as we have considered it may seem too unsatisfactory to be worth our attention. It is introduced here because the basic true-score model does need to be understood, and can be considered the foundation of the treatment by internal analysis of the item relationships. We keep the model but improve the method. It is also necessary for the student to be aware that both retest and alternate form methods have had a large role in the history of psychological tests, and their use still continues.

REVIEW GUIDE

General problems in assessing precision of psychological measurements need to be understood. These include:

1. Repeated measures are not generally independent.
2. The attribute changes over time.
3. The attribute is not the same as the score from any given set of items used to measure it.

The classical true-score model (5.1) gives the reliability coefficient—a variance ratio—defined by (5.4). The reliability coefficient is equivalent to the correlation between two measurements satisfying the assumptions of the model. Given an acceptable reliability coefficient, we can compute the error variance of the test score using (5.13), and the *SE* of measurement using (5.13b) and (5.13c). Generally, the measurement error variance will be unaltered by a choice of population, whereas the reliability coefficient varies.

Both retest methods and parallel-form methods for estimating a reliability coefficient from a correlation between measures are problematic. They can be interpreted respectively as a measure of stability of the test score over time, and as a measure of the equivalence of the test forms.

END NOTES

General: See Lord and Novick (1968) for a more comprehensive, and more technical, account of classical reliability theory. Further resources for followup reading could include Feldt and Brennan (1989).

1. Lord and Novick (1968, p. 134).
2. Lord and Novick (1968, p. 137).
3. In the classical accounts of alternate form reliability, two tests are *parallel* if they have the same variance and the same covariances with every other test they might be correlated with. This second condition is untestable, and in practice it is not subjected even to limited testing.
4. Diener et al. (1985).