4

Validity and reliability studies

One must go seek more facts, paying less attention to techniques of handling the data and far more to the development and perfection of the methods of obtaining them. (Hill 1953)

INTRODUCTION

The serious adverse effects of the use of invalid exposure measurements have been described in Chapter 3. Selecting or developing an accurate measurement instrument is obviously a critical step in designing an epidemiological study. First, the available literature on the validity and reliability of instruments which measure the exposure of interest should be reviewed. Then, if a new instrument is to be developed which differs substantially from other methods, its reliability or, preferably, its validity should be assessed.

The term *reliability* is generally used to refer to the reproducibility of a measure, that is, how consistently a measurement can be repeated on the same subjects. Reliability can be assessed in a number of ways, of which only two are covered in this chapter. *Intramethod reliability* is a measure of the reproducibility of an instrument, either applied in the same manner to the same subjects at two or more points in time (test–retest reliability) or applied by two or more data collectors to the same subjects (inter-rater reliability). For example, a comparison could be made of exposure information from two data abstractors who extracted information from the medical records of the same group of subjects. *Intermethod reliability* is a measure of the ability of two different instruments which measure the same underlying exposure to yield similar results on the same subjects. Generally an intermethod reliability study compares a measurement method to be used in an epidemiological study with a more accurate but more burdensome method. For example, a questionnaire might be compared to an exposure diary for a group of subjects. Intermethod reliability studies of this type are sometimes called *validity studies*. Technically, however, an error-free comparison method of measurement is needed to directly measure validity, so the term intermethod reliability has been preferred. In most fields of study, the term reliability refers to *intramethod reliability*, and less work has been done on the design and interpretation of intermethod reliability studies. Intermethod reliability studies are dealt with in detail in this chapter because of their potential importance in epidemiology.

The first topic to be covered in this chapter is the relationship of measures

of reliability to measures of validity. Measures of reliability are primarily important for what they reveal about the validity of a measurement, for, as shown in Chapter 3, the bias in an epidemiological study is a function of the validity of the exposure measure. The second section covers additional issues in the design of reliability and validity studies, and the third covers the statistical analysis of reliability and validity studies.

Many of the examples in this chapter are based on real data. Most are focused on dietary measurements, in particular the reliability of a food frequency estimate of fat intake. Part of this focus is a reflection of the increased interest in reliability studies, which is due to interest in assessing diet and the difficulties it presents. Additionally, by focusing on a single exposure, the reader can observe how measures of reliability are a function of the design of the reliability study, as well as of the accuracy of the instrument itself.

THE INTERPRETATION OF MEASURES OF RELIABILITY

This section covers the interpretation of measures of reliability in terms of measures of validity, and is meant to provide some general concepts for the interpretation of reliability studies. It is limited to continuous exposure measures. The results presented assume measures are obtained from an infinite population; that is, issues of sampling error are ignored.

A model of reliability and measures of reliability

Suppose each person in a population of interest is measured twice, either with one instrument or two instruments that purport to measure the same exposure. If two instruments are used, X_1 will denote the measure of interest, that is, the one to be used in the epidemiological study, and X_2 the comparison measure. For a given subject i, two (continuous) exposure measurements, X_{i1} and X_{i2}, are obtained. A simple model that could apply to intermethod or intramethod reliability studies is

$$X_{i1} = T_i + b_1 + E_{i1}$$

$$X_{i2} = T_i + b_2 + E_{i2}$$

where $\mu_{E_1} = \mu_{E_2} = 0$. The model can also be written

$$X_{ij} = T_i + b_j + E_{ij},$$

where X_{ij} is the observation on subject i of measure X_j.

This model states that subject i's first measure, X_{i1}, is equal to the true value of exposure for subject i, T_i, plus the constant bias of the first

instrument in the population, b_1, plus the error for subject i on measure 1, E_{i1}. The second measure, X_{i2}, is equal to the same true value, T_i, plus the bias of the second instrument, b_2, plus a second error, E_{i2}.

In the population, X_1, X_2, T, E_1, and E_2 are random variables with distributions. The population mean of X_1 is denoted by μ_{X_1}, the variance by $\sigma^2_{X_1}$, etc. Because the bias of X_1 in the population is expressed as a constant b_1 and the bias of X_2 as b_2, it follows that the population means of the subject error terms, E_1 and E_2, are 0.

In a reliability study, information is available on X_1 and X_2 for each subject, but not on T. A reliability study can yield estimates of μ_{X_1}, μ_{X_2}, and the correlation between the two measures, $\rho_{X_1X_2}$, termed the *reliability coefficient*.

In Chapter 3, two measures of the validity of a continuous exposure measure were shown to be important in assessing the impact of measurement error: the bias and the validity coefficient. The primary question is, if X_1 is the measure of interest, what can the estimates of μ_{X_1}, μ_{X_2} and $\rho_{X_1X_2}$ from a reliability study tell us about the bias in X_1, b_1, and its validity coefficient, ρ_{TX_1}?

The measurement and interpretation of the bias in a measure

Reliability studies often cannot provide information on the bias in X_1 or X_2. In a reliability study based on the above model, only the *difference* between the biases of X_1 and X_2 can be observed:

$$(b_1 - b_2) = \mu_{X_1} - \mu_{X_2}. \qquad [4.1]$$

This equation states that the difference between the population means of the two measures is a measure of the difference between their biases. This difference is often not very informative. If a similar degree of bias is present in both measures—for example, if the same miscalibrated scale is used to weigh each subject twice—the difference between the means of the two measures can be close to 0 even when there is considerable bias in both measures. However, if X_2 is an unbiased measure of T ($b_2 = 0$), then

$$b_1 = \mu_{X_1} - \mu_{X_2}. \qquad [4.2]$$

Thus, only when the comparison measure X_2 is a perfect measure or when X_2 can be assumed to be unbiased (e.g. a well-calibrated scale), can a reliability study yield information about the bias in X_1.

As discussed in Chapter 3, differential bias in the exposure measure between cases and controls can have undesirable effects in an epidemiological study. (Differential precision may also be a concern, but is not discussed in this chapter.) To assess differential bias, a reliability study would need to measure X_1 and X_2 in a population of cases and a population

of controls to yield estimates of the means of X_1 and X_2 among those with disease ($\mu_{X_{1D}}$, $\mu_{X_{2D}}$) and among the non-diseased group ($\mu_{X_{1N}}$, $\mu_{X_{2N}}$). The difference between the bias in X_1 between cases and controls, $b_{1D} - b_{1N}$, can be measured *only* if there is non-differential bias in the comparison measure X_2 ($b_{2D} = b_{2N}$). Then, if the simple additive model given above holds for both cases and controls,

$$(b_{1D} - b_{1N}) = (\mu_{X_{1D}} - \mu_{X_{2D}}) - (\mu_{X_{1N}} - \mu_{X_{2N}}). \qquad [4.3]$$

Example. To assess differential bias between colon cancer cases and controls in a retrospective food frequency estimate of fat intake (X_1), a reliability study could be conducted within an existing cohort study of colon cancer. X_1 could be compared to prospective information on fat intake (X_2) with reasonable certainty that any bias in X_2 is equal for cases and controls. If cases reported 40 per cent energy from fat on average prospectively and 42 per cent retrospectively, and controls reported 39 per cent prospectively and 38 per cent retrospectively, then the differential bias could be estimated from Equation 4.3 as

$$\widehat{(b_{1D} - b_{1N})} = (42 - 40) - (38 - 39)$$
$$= 3\% \text{ energy}.$$

The inability of many reliability study designs to yield information on bias or differential bias is a major limitation. It should be recalled, however, that under non-differential measurement error (and certain other assumptions), the attenuation equations depend only on the validity coefficient and not on the bias. Thus, measures of reliability may be used to estimate at least some of the effects of measurement error in the absence of a measure of bias. When non-differential measurement error can be assumed, reliability can be assessed in a single population representative of the population in which the epidemiological study is to be conducted.

Relationship of reliability to validity under the parallel test model

When certain assumptions are met, reliability studies can yield information about the validity coefficient. One such set of assumptions is the model of parallel tests (Lord and Novick 1968; Nunnally 1978; Allen and Yen 1979; Carmines and Zeller 1979; Bohrnstedt 1983). The model is the same as the general model above, but with some additional assumptions:

$$\rho_{E_1} = \rho_{E_2} = 0$$
$$\sigma^2_{E_1} = \sigma^2_{E_2} = \sigma^2_E$$
$$\rho_{E_1E_2} = 0.$$

The first assumption of the parallel test model is that the error variables, E_1 and E_2, are not correlated with the true value T. It is further assumed that E_1 and E_2 have equal variance, σ_E^2. This also implies that X_1 and X_2 have equal variance and that X_1 and X_2 are equally precise ($\rho_{TX_1} = \rho_{TX_2}$) (see Equation 3.2). This is usually a reasonable assumption in intramethod studies, since X_1 and X_2 are measurements from the same instrument. Finally, it is assumed that E_1 is not correlated with E_2. This important (and restrictive) assumption implies, for example, that an individual who has a positive error, E_1, on the first measurement is equally likely to have a positive or a negative error, E_2, on the second measurement. These assumptions are often summarized by saying that two measures are parallel measures of T if their errors are equal and uncorrelated. The parallel test model generally includes the assumption that $b_1 = b_2 = 0$, but this assumption is not needed for the results in this chapter.

Under the assumptions of parallel tests it can be shown that (Allen and Yen 1979):

$$\left.\begin{aligned} \rho_{X_1X_2} = \frac{\sigma_T^2}{\sigma_{X_1}^2} = 1 - \frac{\sigma_E^2}{\sigma_{X_1}^2} = \rho_{TX_1}^2 = \rho_{TX_2}^2 \\ \text{or equivalently} \\ \rho_{TX_1} = \rho_{TX_2} = \sqrt{\rho_{X_1X_2}}\,. \end{aligned}\right\} \quad [4.4]$$

These equations state that the reliability coefficient, $\rho_{X_1X_2}$, is equal to the square of the validity coefficient for X_1 or X_2. This result is important, because it shows that if the assumptions are correct, the reliability coefficient, which is a measure of the correlation between two imperfect measures, can be used to estimate the correlation between T and X_1, *without* having a perfect measure of T. The correlation of X_1 with X_2 is less than the correlation of X_1 with T, due to the error in X_2.

Example. In a test–retest reliability study serum cholesterol concentration was measured twice, one year apart (Shekelle *et al.* 1981). Suppose the 'true measure' of interest was each subject's average serum cholesterol over the year separating the two measurements. The two methods were identical, so that the assumptions of equal variances would be appropriate. The other assumptions are also considered to be met, including the assumptions of no correlation between the errors and the true measure or between the errors on the two measures. The correlation of X_1 with X_2 in the reliability study was 0.65. Then, by Equation 4.4 the correlation of X_1 with T (or the correlation of X_2 with T) can be estimated as 0.8.

The definition of the reliability coefficient of X_1 as the correlation between X_1 and X_2, two parallel measures of T, is one definition of reliability.

Based on Equation 4.4, the results in the last chapter which were expressed in terms of ρ_{TX}^2 could have been (and often are) expressed in terms of $\rho_{X_1X_2}$. These expressions apply only when the meaning of the reliability coefficient is restricted to the correlation between parallel measures of T. However, we use the term reliability coefficient to refer to the correlation between measures of the same exposure, $\rho_{X_1X_2}$, even when the assumptions of parallel tests do not hold. This means, for a given instrument, X_1, applied to a given population, that the reliability coefficient will vary with the choice of X_2, depending on the extent to which the assumptions of parallel tests do or do not apply, for a given X_2.

In real reliability studies, the assumptions of parallel tests are often incorrect. Two common violations will be discussed: unequal variances of E_1 and E_2, and correlated errors. Even when these assumptions are violated, the correlation between X_1 and X_2 can still provide some information about the validity coefficient of X_1.

Relationship of reliability to validity under unequal variances of E_1 and E_2

In the model of parallel tests, the variances of E_1 and E_2 are assumed to be equal, which implies that X_1 and X_2 are equally precise ($\rho_{TX_1} = \rho_{TX_2}$). This assumption is incorrect for certain reliability studies, particularly for many intermethod reliability studies. First consider a true validity study where X_1, the exposure measure of interest, is compared to a perfect measure of exposure, termed X_2 ($X_2 = T$). Then, by definition,

$$\rho_{X_1X_2} = \rho_{TX_1}. \quad [4.5]$$

However, a perfect measure is often not available, so the exposure measure of interest, X_1, is often compared with an imperfect but more precise measure, X_2. This implies that $\rho_{TX_2} > \rho_{TX_1}$. If the other assumptions of the parallel tests model hold, including the assumption of uncorrelated errors, then

$$\rho_{X_1X_2} < \rho_{TX_1} < \sqrt{\rho_{X_1X_2}}. \quad [4.6]$$

This equation states that when X_2 is more precise than X_1, and the errors in X_1 and X_2 are not correlated, the reliability coefficient $\rho_{X_1X_2}$ can be used to yield an upper and lower bound for the validity coefficient of X_1. The lower bound for the validity coefficient of X_1 is the interpretation as if X_2 were a perfect measure (Equation 4.5), and the upper bound is the interpretation as if X_2 had equal error variance (Equation 4.4). The more accurate X_2 is, the closer the *lower* bound is to ρ_{TX_1}.

Example. Willett *et al.* (1985) conducted an intermethod reliability study to evaluate a food frequency questionnaire estimate of average daily

fat intake over the preceding year (X_1). The comparison measure was an estimate of average daily fat intake from four 1-week diet diaries spread over the year (X_2). The observed correlation between the food frequency estimate and the diary estimate (energy adjusted) among 173 subjects was $\rho_{X_1X_2} = 0.5$.

One might argue that the errors in the estimate of fat from a diet diary are not correlated with those on the food frequency questionnaire. The primary source of error on a food frequency questionnaire may be poor recall, while on the diet diaries it may be whether 4 weeks are fully representative of yearly intake. It was assumed that the four diaries yielded a more accurate measure of fat and, in fact, the variance from the diary estimate appeared to be smaller than the variance from the food frequency estimate. Then Equation 4.6 might apply:

$$0.5 < \hat{\rho}_{TX_1} < 0.7,$$

which suggests that the validity coefficient for X_1 would be between 0.5 and 0.7. Willett argued that the use of four 1-week diaries is a near-perfect criterion (e.g. there was little increase in $\rho_{X_1X_2}$ when X_2 was based on four diaries rather than two). This suggests that ρ_{TX_1} is near the lower limit 0.5.

In an effort to find a comparison measure, X_2, with error uncorrelated with the error in X_1, the comparison measure may be less accurate than X_1 ($\rho_{TX_2} < \rho_{TX_1}$). Then if the other assumptions of the parallel tests model hold:

$$\rho_{TX_1} > \sqrt{\rho_{X_1X_2}}. \qquad [4.7]$$

If it is not known whether X_1 or X_2 is more accurate, it can still be assumed that (Allen and Yen 1979)

$$\rho_{TX_1} \geq \rho_{X_1X_2}. \qquad [4.8]$$

In other words, the correlation of X_1 with even a poor measure X_2 with uncorrelated errors gives a lower limit for the correlation of X_1 with the true measure. For example, if a blood measure of fat intake is available and the correlation of a food frequency measure of fat intake (X_1) with the blood measure (X_2) is 0.2, then Equation 4.8 shows that if there were no sources of correlated errors between X_1 and X_2 a lower limit for P_{TX_1} would be 0.2.

A model of reliability allowing for correlated errors

One assumption of the model of parallel tests that is often violated is the assumption of uncorrelated errors. Often $\rho_{E_1E_2} > 0$. In other words, the

error in one measure is positively correlated with the error in the other. Correlated errors occur when the sources of error in the first measurement on a subject tend to repeat themselves in the second. For example, weight may be consistently under-reported by some subjects on re-administration of a questionnaire.

A model for reliability that makes explicit the correlated errors is

$$X_{ij} = T_i + b_j + E_{ij},$$

where $E_{ij} = B_i + F_{ij}$. The error terms E_{i1} and E_{i2} for a given subject are the sum of two parts: a part that repeats itself on each measure of subject i, B_i, termed the *within-subject bias*; and a part that varies between measures (around a mean of 0 for subject i), F_{ij}, termed the *random error* (see Figure 4.1). E_1 and E_2 are correlated because they both measure the within-subject bias.

To simplify the reliability coefficient under this model, let S_i be that part of X_1 and X_2 that is consistently measured for subject i on both instruments. S_i would be the sum of T_i plus B_i (plus the average bias across measures). Then the model of reliability can be rewritten as

$$X_{ij} = S_i + m_j + F_{ij},$$

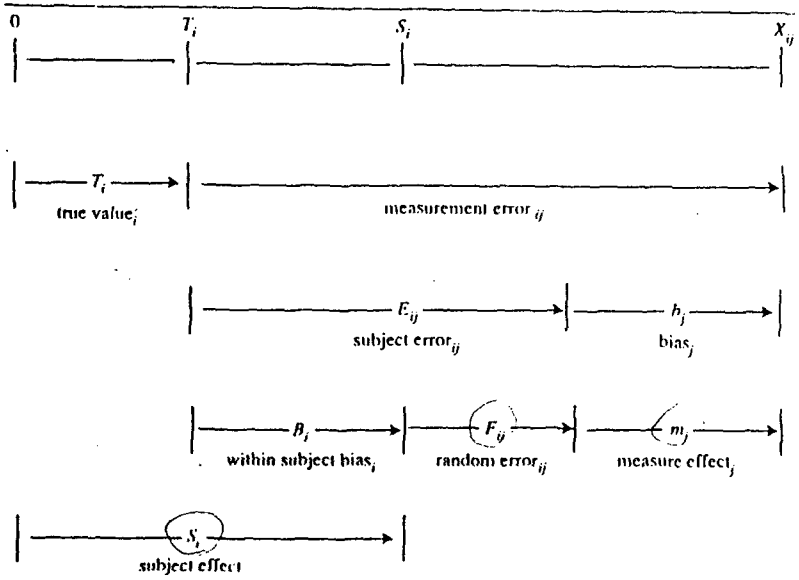

Figure 4.1 Measurement error in X_{ij}, the jth measure on subject i.

where S_i represents the effect of subject i and m_j the effect of measure j on X_{ij} (the m_j are the b_j minus the average bias across measures, such that $\Sigma m_j = 0$). If X_1 and X_2 are equally precise (and $\rho_{TF_1} = \rho_{TF_2} = 0$), then X_1 and X_2 meet the assumptions of parallel measures of S (not T). The reliability coefficient under this model is the proportion of the variance of X_1 explained by S, the part that is common to the repeated measures. That is, (Allen and Yen 1979)

$$\rho_{X_1X_2} = \frac{\sigma_S^2}{\sigma_{X_1}^2} = \frac{\sigma_S^2}{\sigma_{X_2}^2} = 1 - \frac{\sigma_F^2}{\sigma_{X_1}^2}. \qquad [4.9]$$

The right-hand side of this equation shows that when there are correlated errors, the reliability coefficient reflects only the random component of error in X_1.

Equation 4.9 is essentially identical to the classical definition of the *reliability* of a measure X (Allen and Yen 1979; Dunn 1989):

$$\rho_X = \frac{\sigma_S^2}{\sigma_X^2}. \qquad [4.10]$$

ρ_X is conceptualized as the correlation between two repeated applications of X and is consistent with Equation 4.9 when X_1 and X_2 are replications of the same measurement. S_i is conceptualized as subject i's expected value of X (mean over infinite repeated measures). (S_i is usually denoted by T_i and termed the subject's *true score*, but S is used here to distinguish 'the subject's expected score on the instrument' from T, 'the subject's true value of the exposure of interest').

Relationship of reliability to validity under correlated errors

When the errors E_1 and E_2 of the measures in a reliability study are positively correlated, then the reliability study can only yield an *upper limit* for the validity coefficient. Specifically, when X_1 and X_2 are equally precise (or X_2 is more precise than X_1) and the assumptions of the above model hold, then the validity coefficient is less than the square root of the reliability coefficient (Walker and Blettner 1985):

$$\rho_{TX_1} < \sqrt{\rho_{X_1X_2}}. \qquad [4.11]$$

Thus a measure can be reliable (repeatable) even if it has poor validity. While a low reliability coefficient implies poor validity, a high reliability does not necessarily imply a high validity coefficient. The high reliability may be due instead to repeated within subject errors. The reliability coefficient is only diminished by the random component of error, whereas the validity coefficient is a measure of both the random error and the within subject bias.

To interpret a reliability study, one should evaluate whether there are potential sources of correlated errors between the two or more measures. As outlined in Table 3.1, there is a wide range of sources of measurement error, and most of these could be sources of correlated errors.

Example. A test–retest reliability study was also conducted in the study by Willett *et al.* (1985) presented in the previous example. The observed correlation between the estimates of average daily fat intake from two administrations of the food frequency questionnaire, 1 year apart, was 0.6.

Sources of correlated errors between two administrations of a food frequency questionnaire include the following:

(a) Some subjects may consistently tend to report their 'best' diet rather than their usual diet.

(b) Certain high-fat foods eaten frequently by a few subjects may have been omitted from the questionnaire. Those subjects would have their fat intake consistently underestimated.

(c) The nutrient database used to convert foods to grams of fat may be incorrect for certain subjects. For example, those subjects who reduce the fat content of standard recipes such as stews, lasagna, etc., would have the fat content of their diet consistently overestimated.

(d) The time period assessed by the instrument (diet in last year) may differ from the true time period of interest (e.g. diet over the last 5 years). Then, those who have lowered the fat in their diet in recent years will have their fat intake underestimated on both administrations of the instrument compared with their true 5-year average fat intake.

Under this strong likelihood of correlated errors, Equation 4.11 would apply in interpreting this reliability study,

$$\rho_{TX_1} < 0.8,$$

so only an upper limit on the validity coefficient of the measurement can be estimated. This outcome clearly provides less information about the validity of the food frequency measure of fat than did the outcome of the intermethod reliability study described in the last example.

Correlated errors commonly occur in intramethod studies, but they could occur in intermethod studies as well. In the example of an intermethod study of a food frequency estimate of fat intake compared with a diet diary

estimate, it was argued that the errors on the two instruments were unlikely to be significantly correlated. However, of the four sources of correlated errors noted in the above example, at least two could lead to correlated errors between a food frequency measure and a diet diary measure (the tendency of some subjects to report their best diet, and the issue relating to the time period of measurement). Reliability studies in which the errors of the measures are correlated cannot provide estimates of a lower bound for ρ_{TX_1}. Estimates of a lower bound depend on uncorrelated errors, so lower bounds should be interpreted cautiously.

Interpretation of the value of the reliability coefficient

Some authors have provided guidance on whether to consider the reliability of a measure poor, fair, or good from the value of the reliability coefficient. It may be more appropriate first to consider what information the reliability study yields about the validity of the measure, as discussed above. Then consideration could be given to the effect of the estimated measurement error in the exposure on the epidemiological study which will use the measure, based on the tables and equations in the last chapter.

For example, suppose a reliability study which complied with the parallel test model yielded $\rho_{X_1X_2} = 0.64$, leading to an estimate of the validity coefficient, $\hat{\rho}_{TX} = 0.8$. If the true odds ratio were 2.0, and if the assumptions of non-differential measurement error and the other assumptions of Equation 3.6 were reasonable, then the estimated observable odds ratio would be

$$\hat{OR}_O = OR^{\hat{\rho}_{TX}} = 2^{0.8} = 1.7.$$

This might be considered to be acceptable attenuation. On the other hand, if a reliability study of a different instrument produced the same reliability coefficient (0.64) but, due to correlated errors between the measures, only an upper bound for ρ_{TX} of 0.8 could be estimated, then applying Equation 3.6 would yield

$$\hat{OR}_O < 1.7.$$

This estimate of the attenuated odds ratio includes only the attenuation due to the random error in X, and the actual attenuation would be greater. This instrument might not be acceptable.

Other violations of the assumptions of the model

In some situations the assumption of the basic model, which states that both X_1 and X_2 measure T with additive errors E_1 and E_2, is incorrect. One alternative to this model is that X_1 and X_2 or both may be a linear function of T:

$$X_{ij} = c_j T_i + b_j + E_{ij}.$$

(In intramethod studies it is assumed that c_1 and c_2 are equal, i.e. X_1 and X_2 have the same scale.) Reliability studies such as these can still yield information about the validity coefficient, but not about bias. The results presented so far in this chapter for interpreting the reliability coefficient $\rho_{X_1X_2}$ in terms of the validity coefficient ρ_{TX_1} would also apply to the above model, because $\rho_{X_1X_2}$ is not affected by linear transformations of X_1 or X_2. For example, a food frequency questionnaire measure of β-carotene (X_1) could be compared with serum β-carotene (X_2) and interpreted by eqn. 4.7 or 4.8 (if there were no sources of correlated errors) even though serum β-carotene is not measured in the same units as β-carotene intake.

Further violations of the assumption that T and E are uncorrelated are beyond the scope of this book, but several practical points should be noted. First, a transformation of X, such as the logarithmic transformation, may reduce the dependence of E on T (Altman and Bland 1983). Second, when T and E are negatively correlated, the measure X will have a variance less than $\sigma_T^2 + \sigma_E^2$, possibly even a smaller variance than T: therefore the variances of two measures should not be compared to determine which is more precise.

Finally, the equations given so far have assumed a population of infinite size. In practice, there will be sampling error in the estimate of $\rho_{X_1X_2}$. The confidence interval around estimates of $\rho_{X_1X_2}$ should be taken into consideration when using the equations in this section to interpret reliability in terms of validity.

Interpretation of reliability studies of categorical variables

Most of the concepts presented for continuous variables apply in a qualitative way to interpretation of reliability studies of categorical variables. For example, when two imperfect categorical measures of exposure are being compared, part of the agreement between them could be due to repeated error. Mathematical relationships between measures of reliability and measures of validity for categorical variables are not straightforward. However, some reliability study designs can yield estimates of the sensitivity and specificity of each measure or can be used to estimate the bias in the odds ratio that would result from the misclassification (Hui and Walter 1980; Walter 1984; Clayton 1985; Kaldor and Clayton 1985; Walter and Irwig 1988; Dunn 1989).

ISSUES IN THE DESIGN OF VALIDITY AND RELIABILITY STUDIES

There are several issues that need to be considered in the design of reliability studies (Fleiss and Shrout 1977; Carmines and Zeller 1979; Dunn 1989;

Willett 1990). Most of these issues are also important in interpreting reliability studies carried out by others.

Purpose and timing of the reliability study

When a new instrument is to be developed for an epidemiological study, or when an existing one is to be applied in a substantially different population, a validity or reliability study of the instrument should be carried out first. Estimates of the validity or reliability coefficient and, when possible, the bias of the instrument can then be used to decide whether it is necessary to develop a more accurate instrument. If the measure is shown to be reasonably reliable, and by inference reasonably valid, this will increase confidence in the outcome of the epidemiological study.

Reliability studies conducted before the main epidemiological study, or early in its course, can be used not only to evaluate but also to improve the instrument. For example, an inter-rater reliability study could identify interviewers, abstractors, or laboratory personnel who need more training or should be dropped from the study. Such a study could be done by comparing three or more raters who collect data on the same subset of subjects. Computation of reliability coefficients for each pair of raters may reveal an individual rater who compares poorly with the others. In addition, the researcher should investigate the situations in which discrepancies between the repeated measurements have occurred. This can often lead to improvement of the instrument or the protocol for its use. For example, if disagreements on the variable marital status of subjects usually involved divorced subjects being erroneously classified as 'single', the category 'single' might be clarified by the label 'never married'.

An additional use of reliability studies is to estimate the impact of exposure measurement error on the results of a study after the parent epidemiological study has been completed. Information from a reliability study conducted on a subset of subjects concurrently with the epidemiological study can yield information about the validity of the exposure measure. This information can be used to adjust the observed odds ratio for the effects of measurement error. Adjustment procedures are discussed in Chapter 5.

Choice of comparison measures

Many types of comparison measures have been used in reliability studies. An instrument can be compared with a re-administration of the same instrument at a different time, by a different rater, or with variation of some other condition of interest, for example proxy respondent versus index subject. For intermethod studies, questionnaire data have been compared with medical records (Harlow and Linet 1989), physical or biochemical measures of exposure (Jarvholm and Sanden 1987; Siconolfi *et al.* 1985; Willett *et al.* 1983), interviews by experts such as nutritionists, industrial hygienists, or physicians (Eskenazi and Pearson 1988), exposure diaries (Willett *et al* 1985; Williams *et al.* 1989), and with direct observation (Klesges *et al.* 1985; Decker *et al.* 1986). Information from medical records has been compared with physician interviews and direct observation (Gerbert *et al.* 1988).

How is an appropriate comparison method selected? The issues discussed in the section on interpretation of measures of reliability (page 5~~000~~) should provide some guidance in selecting a comparison measure. Ideally, measurements from the instrument whose accuracy is to be determined are compared with those provided by a perfect, or near perfect, measure of exposure. This type a study, a *validity study* allows one to estimate both dimensions of measurement error, the bias and the validity coefficient (Equations 4.2 and 4.5).

If a validity study is not possible, then one should consider comparing the instrument of interest to a measure of exposure with uncorrelated errors. A good choice is a comparison measure, X_2, that is more precise than the measure of interest, X_1, and has error unlikely to be correlated to X_1. Intermethod studies comparing questionnaire measures to records (e.g. comparing a questionnaire on oral contraceptive use to complete medical or pharmacy records) or to multi-week diaries (e.g. comparing a questionnaire on leisure physical activty over the last year to six 1-week diaries), often meet these criteria. Equations 4.1 and 4.6 can aid in interpreting such studies. If a comparison measure can be selected with equal error as well as error uncorrelated to the instrument of interest, this can yield good information on the validity coefficient of the instrument (Equation 4.4). Test-retest studies of biochemical measures can often be assumed to have equal and uncorrelated errors if the replicates are sampled over the entire time period to which the exposure measure is intended to relate. Often questionnaire measures of behaviours can be compared with relevant physical or biochemical measures under the assumption of uncorrelated errors (e.g. a questionnaire on physical activity can be compared to a treadmill test), but such comparisons are often limited because the physical or biochemical measure may be a poor measure of the behaviour. Equations 4.7 or 4.8 can be useful in interpreting such studies. However, if both the questionnaire and the biochemical test reflect recent exposure, and the instrument is intended to represent exposure over a longer period of time, the errors could be correlated.

When a comparison measure with uncorrelated error is not available, then only part of the measurement error can be assessed in the reliability study. The part of the error that is repeated cannot be measured. The researcher can attempt to select the comparison measure, X_2, so that the main sources of error in the measure of interest, X_1, are not repeated in X_2. For example, in assessing a questionnaire covering diet 10 years in the past, long-term recall might be the greatest concern. One reliability study of this issue

selected subjects who had answered a diet questionnaire years earlier and compared their measurements from the questionnaire of interest with those on the earlier version (Wu *et al.* 1988). This design would permit assessment of the main sources of error, that due to poor recall and that due to random variation, even though some sources of error (e.g. omission of certain foods on both questionnaires) could be repeated on both questionnaires. By careful selection of X_2, correlated errors between X_1 and X_2 can be minimized and the reliability study may then yield more information about the validity of X_1.

Finally, a simple test–retest reliability study is often quick and inexpensive to undertake. If there are sources of correlated error, the study will yield only an upper limit for validity (Equation 4.11). Nonetheless, if the reliability coefficient proves to be *low*, the instrument should probably be abandoned or extensively revised.

In reviewing reliability studies by others, the same issues should be considered. The key questions are:

- Was the comparison method used close to perfect?
- Was there an imperfect comparison measure but with uncorrelated errors?
- If two or more measures with correlated errors were used, were the errors likely to be strongly or weakly correlated?

The answers to these questions will guide the interpretation of the reliability study.

Separate studies on diseased and non-diseased groups

The researcher must decide whether or not to attempt to measure differential error. To assess differential measurement error, the reliability study needs to be conducted on a sample of cases and a sample of controls, and the comparison measure needs to be carefully selected.

Differential *bias* is a particular concern (Chapter 3), therefore a comparison of the bias in the measure of interest, X_1, between cases and controls would be of major interest. For reliability studies to assess differential bias in X_1, a comparison measure X_2 needs to be selected that is unbiased or that can be assumed to have non-differential bias. Then the differential bias in X_1 can be estimated by Equation 4.3. For example, comparison of recall of a specific medication (X_1) with an abstract of medical records (X_2) among cases versus a similar comparison among controls may be a good way to assess differential bias: any bias in records is unlikely to be related to the disease under study, provided that a sufficient period prior to diagnosis is excluded. On the other hand, a test–retest reliability of recall of medications (assessed separately on cases and controls) would not reveal any differential bias between cases and controls, because the bias would occur in both measures.

Selection of subjects for reliability studies

Ideally, subjects in a reliability study should be a random sample of those in the population in which the epidemiological study will be carried out. This is because there are problems in generalizing reliability studies conducted on one population to another. These problems include the following:

- Reliability or validity studies which use self-selected volunteers might find the instrument to be more valid than it would be in the intended population, due to the higher level of motivation among the volunteers.
- Differences between populations in education, age, sex, and other factors could influence the validity and reliability of the instrument.
- Differences in the *distribution* of the true exposure between populations can influence the validity and reliability coefficients. Recall that ρ^2_{TX} (also $\rho_{X_1X_2}$ under parallel tests) is equal to

$$\frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E} = \frac{1}{1 + \sigma^2_E/\sigma^2_T}.$$

Thus even if the same instrument had the same error variance, σ^2_E, in two populations, the validity coefficient (or reliability coefficient) would be smaller in the population with least variation in true exposure, σ^2_T. For this reason, the use of the validity coefficient (and the reliability coefficient) to express measurement error has been criticized (Altman and Bland 1983). However, these are appropriate statistics in that they can be used to estimate the effect of measurement error on the bias in the observed odds ratio in the population of interest, which depends on the ratio σ^2_T/σ^2_X. Nonetheless, a validity or reliability coefficient assessed in one population may not apply to another with a different distribution of exposure.

Timing and order of measures

Correlated errors between measures in a reliability study may occur when subjects recall at the second or later testing the responses they gave on earlier tests. This recall can be minimized by separating the measures over time, usually by a month at least. However, when the two periods of testing are well separated, the two measures of exposure may refer to different time periods. Thus some lack of correlation between them may be due to true change in exposure over time. For example, a test–retest reliability study of a food frequency questionnaire estiamte of fat intake over the last year, with the two administrations separated by a year, would not yield a perfect cor-

relation even if the instrument was perfect, because of the one-year shift in the time period covered. However, this issue may not be a problem; depending on the true exposure of interest, it may be appropriate to include the variation in a measure over time as a source of measurement error.

In intermethod reliability studies, the instrument to be evaluated is generally given first because the comparison measure is usually less prone to error and may, therefore, be less affected by recall of the prior measurement. Knowledge that a measure is to be validated can influence subjects' responses (e.g. self-report of weight may be influenced by knowledge that they will be weighed), so the invitation to subjects to participate in the second measure should be given after the first measurement has been completed. In intramethod reliability studies, the order of measures (e.g. raters) should be randomized (although X_1 would always be the measure by rater 1, X_2 by rater 2, etc.).

Review of studies using the instrument

There is an addtional approach to assessing the validity of an instrument, which is similar to the concept of *predictive validity* from psychology (Nunnally 1978; Carmines and Zeller 1979). If an exposure measure has been shown to be associated with a disease or other outcome in several epidemiological studies, this provides some evidence that it is a valid measure. Specifically, it can be seen from Equation 3.8 that $\rho_{TX} > \rho_{XY}$. That is, the correlation between a continuous exposure variable X and a continuous outcome Y is a lower limit for the validity coefficient of X. The assumption is that the errors in the outcome Y and exposure X are uncorrelated. This approach of reviewing other epidemiological studies can also be applied in a qualitative way to studies with dichotomous disease outcomes. For example, if an estimate of vitamin A from a dietary questionnaire was significantly associated with disease in several studies, in agreement with prior hypotheses, this provides some evidence for the validity of the instrument. However, caution should be exercised in this approach; the disease-exposure association may be due to confounding between the exposure measure and some other risk factor for disease, or to other sources of bias in the study.

ANALYSIS OF VALIDITY AND RELIABILITY STUDIES

Selecting the appropriate measures of validity or reliability

This section covers some common approaches to the statistical analysis of validity and reliability studies.

The selection of the appropriate analysis depends on certain aspects of the

design of the study (Fleiss 1981, 1986; Kelsey *et al.* 1986; Maclure and Willett 1987; Dunn 1989). First, is the exposure measure a continuous variable, a nominal (including dichotomous) categorical variable, or an ordered categorical variable? All of the statistical techniques to be discussed are for reliability and validity studies in which the two or more measures of exposure are on the same type of scale.

Second, is the study an intermethod reliability/validity study or an intramethod reliability study? In an intermethod reliability study or a validity study, the instruments usually differ, the variances of the measures may not be equal, and even the units of measure may not even be the same, for example a measure of β-carotene intake may be compared with serum β-carotene concentration. Further, our discussion of intermethod reliability is limited to the comparison of only two measures at one time. In the intramethod type of study the instruments used are essentially the same, possibly with some difference in administration; for example, two interviewers. Intramethod reliability studies can be performed using more than two measures per subject. One assumption of the analytical methods for intramethod studies is that the variances of the measures, $X_1, X_2, \ldots, X_k$, are equal for continuous measures, or that the measures are 'equally precise' (Fleiss 1986) for categorical exposures.

The issue of correlated errors does not influence the choice of analytical method for reliability studies, only the interpretation of the results.

Table 4.1 gives an overview of methods for the analysis of validity and reliability studies. The upper half gives methods for intermethod studies, and the lower half approaches for intramethod studies. These techniques will be described in this section, with emphasis on intramethod reliability. It is assumed that the reader is already familiar with the assumptions and computations of the Pearson product-moment correlation coefficient, the one- and two-sample t-test, and the analysis of variance (ANOVA) (Armitage and Berry 1987).

The emphasis in the analysis of reliability and validity studies is on parameter estimation; for example, estimation of $\rho_{X_1X_2}$. Confidence intervals also add useful information. Statistical tests are less important, for it should almost be a 'given' that X_1 and X_2 are not related by chance.

Except for the evaluation of differential bias between cases and controls, the statistical techniques given in this section are for a single population. Many could be extended to the comparison of cases and controls, but these extensions are beyond the scope of this book.

Validity and intermethod reliability studies of continuous measures

Intermethod reliability studies and validity studies can be analysed using common statistical techniques.

In the analysis of continuous exposure variables under the model of

Table 4.1 Analysis of validity and reliability studies

Type of exposure measures	Statistical measure of reliability or validity	Condition under which statistic equals 1
Intermethod reliability and validity studies		
Continuous	Pearson correlation coefficient and $\bar{X}_1 - \bar{X}_2$	$X_{i1} = cX_{i2} + d$
Nominal or binomial	misclassification matrix	(not a summary measure)
Ordered categorical	misclassification matrix	(not a summary measure)
	or	
	Pearson correlation coefficient and $\bar{X}_1 - \bar{X}_2$	$X_{i1} = cX_{i2} + d$
	or	
	Spearman correlation coefficient	ranking by X_1 same as ranking by X_2
Intramethod reliability studies		
Continuous: $X_1, \ldots, X_k$ are essentially the same measure	intraclass correlation coefficient: one-way (R_1)	$X_{i1} = X_{i2} = \ldots = X_{ik}$
Continuous: $X_1, \ldots, X_k$ are the k measures (e.g. raters) to be used in epidemiological study	intraclass correlation coefficient: two-way fixed effects (R_2)	$X_{i1} = X_{i2} = \ldots = X_{ik}$
Continuous: $X_1, \ldots, X_k$ represent k of many measures (e.g. raters) that will be used in a study	intraclass correlation coefficient: two-way random effects (R_3)	$X_{i1} = X_{i2} = \ldots = X_{ik}$
Continuous: same as two-way fixed effects except difference between means of measures not a source of error in parent study	intraclass correlation coefficient with difference in means excluded in variance of X (R_4)	$X_{i1} = X_{i2} + d_2 = \ldots X_{ik} + d_k$
Nominal or binomial	Cohen's κ	$X_{i1} = X_{i2}$
Ordered categorical	weighted κ	$X_{i1} = X_{i2}$

additive independent errors, the difference between the biases of the two measures can be estimated as the difference between the sample means of X_1 and X_2 (from Equation 4.1):

$$\widehat{(b_1 - b_2)} = \bar{X}_1 - \bar{X}_2.$$

For a validity or reliability study in which X_2 is unbiased, $\bar{X}_1 - \bar{X}_2$ is an estimate of b_1, the bias in X_1. The value of t computed through a one-sample t-test on the variable $(X_{i1} - X_{i2})$ computed for each subject can be used to compute a confidence interval.

For reliability studies in which cases are compared with controls, if X_2 has non-differential bias then the difference in bias can be estimated as (from Equation 4.3):

$$\widehat{(b_{1D} - b_{1N})} = (\bar{X}_{1D} - \bar{X}_{2D}) - (\bar{X}_{1N} - \bar{X}_{2N}).$$

The value of t from a two-sample t-test on the variable $(X_{i1} - X_{i2})$ can be used to compute a confidence interval for the difference in b_1 between the two groups.

The Pearson product-moment correlation and its confidence interval can be used to estimate $\rho_{X_1 X_2}$ for intermethod or validity studies. If X_1 and X_2 are, at least, positively associated, then the Pearson correlation coefficient would range from 0 to 1. The correlation is equal to 1 when X_1 is a perfect linear transformation of X_2 for all subjects:

$$X_{i1} = cX_{i2} + d.$$

For example, if subjects in an epidemiological study were to be weighed on a portable scale which was validated against a highly accurate scale, a Pearson correlation coefficient close to 1 would suggest that the portable scale was highly precise. If there were a consistent difference between the two measures, (e.g. if the portable scale was miscalibrated 2 kg too heavy, or even if the portable scale weighed in pounds and the comparison scale was a kilogram scale), this would not reduce the correlation $\rho_{X_1 X_2}$.

Example. In an intermethod reliability study on 110 women, an estimate of percentage of energy intake from fat (X_1) from a food frequency questionnaire was compared with a four-day diet record (X_2). The results of the study were:

	mean (% energy)	standard deviation
X_1	39.1	6.6
X_2	37.5	6.0

Pearson correlation coefficient = 0.45

An estimate of the difference in the biases of X_1 and X_2 is:

$$\widehat{(b_1 - b_2)} = 39.1 - 37.5 = 1.6\% \text{ dietary energy.}$$

That is, the food frequency questionnaire overestimates percentage dietary energy from fat by 1.6 per cent compared with a food record. The estimated reliability coefficient is based on the Pearson correlation coefficient, $\hat{\rho}_{X_1X_2} = 0.45$.

In assessing the relationship between X_1 and X_2, one might also consider adjustment for potentially confounding factors that may explain the association of X_1 and X_2 other than by way of their relationship with T.

Analysis of validity and intermethod studies of categorical measures

Several methods can be used to analyse validity or intermethod reliability studies of categorical exposure variables. For a nominal categorical variable the validity or intermethod reliability can be described by the misclassification matrix (or, for a dichotomous variable, by the sensitivity and specificity) as described in Chapter 3.

The misclassification matrix is also appropriate for ordered categorical variables. Depending on the distribution of the ordered categorical variable, the difference in means and the Pearson product-moment correlation coefficient between the two categorical measures or the Spearman rank correlation coefficient might be used. Recall from Chapter 3 that the effect of measurement error in an ordered categorical variable can also be described (under certain assumptions) in terms of the validity coefficient of the underlying continuous variable from which the categorical variable was created. This means that the difference in means and the correlation between the two underlying continuous variables could be appropriate.

Because the misclassification matrix is not a summary measure, κ (described below) is often used to analyse intermethod reliability studies of categorical variables. However, these methods were developed under assumptions more appropriate to intramethod reliability studies.

Analysis of intramethod reliability studies: the concept of interchangeable measures

The primary distinction we have made between intermethod and intramethod reliability studies is that intermethod studies use two instruments and intramethod studies involve repeated applications of one instrument. However, the most important distinction in terms of selecting an analytic method is that in intermethod studies only one measure, X_1, is to be used in the full epidemiological study: we are interested in the reliability of X_1. In intramethod studies, the two or more measures compared are to be used interchangeably as a single exposure measurement in the epidemiological study. For example, in an intramethod reliability study, X_1 may refer to a measure by one interviewer and X_2 to one by a second interviewer, but in the parent epidemiological study each subject will be questioned by one or other of the two interviewers. In intramethod studies, we are interested not in the reliability of X_1 or X_2 but in the reliability of the interchangeable measure 'X', the measure to be used in the epidemiological study.

There is a key difference between intermethod and intramethod reliability. In intermethod reliability, any systematic difference between X_1 and X_2 reflects a consistent bias which affects all subjects in the parent study, and thus does not affect the precision of X_1. In intramethod reliability, on the other hand, a systematic difference between measures contributes to a lack of precision in X because it affects some subjects but not others. For example, if one interviewer weighs subjects on a correctly calibrated scale and a second rater's scale is miscalibrated 2 kg too heavy, this source of error will affect only the subjects measured by the second rater. Thus any consistent difference between study interviewers would increase the variance of the exposure measure (σ_X^2) in the full study and decrease the reliability compared with the use of only one interviewer. The Pearson product-moment correlation is not appropriate for intramethod studies, because systematic differences (in bias or scale) between X_1 and the comparison measure X_2 are not reflected in the Pearson correlation.

Special analytical methods have been developed for intramethod reliability studies, particularly in the context of inter-rater reliability studies. For continuous variables, the reliability of X, ρ_X, is estimated by a version of the intraclass correlation coefficient (R). The *intraclass correlation coefficient* is an estimate of ρ_X as defined in Equation 4.10. That is,

$$\hat{\rho}_X = R = \frac{\hat{\sigma}_\xi^2}{\hat{\sigma}_X^2}.$$

The variance of X in the full epidemiological study, σ_X^2, is estimated under the assumption that in the full study each subject will be *randomly* assigned one measure (e.g. one interviewer). The intraclass correlation reflects is diminished by the error in X due to systematic differences between measures $X_1, \ldots, X_k$ as well as that due to random error. Thus the intraclass correlation (except for R_a discussed below) is equal to 1 only when there is *exact* agreement between measures, that is, when $X_{i1} = X_{i2} = \ldots X_{ik}$ for each subject.

The intraclass correlation coefficient can be interpreted by Equation 4.4 or Equation 4.11, depending on whether there are correlated errors between X_1 and X_2 or not. The bias in X cannot be estimated in an intramethod study. The mean difference between measures, $(\bar{X}_1 - \bar{X}_2)$, can be used to reflect the systematic difference between measures, but any consistent difference between X_1 and X_2 beyond chance also contributes to a lower estimate of ρ_X by the intraclass correlation coefficient.

Four versions of the intraclass correlation coefficient are discussed here.

This discussion is followed by a presentation of an analogous statistic, κ for intramethod reliability studies of categorical variables. Books by Fleiss (1981, 1986) and by Dunn (1989) contain excellent discussions of the intraclass correlation coefficient and κ.

Intraclass correlation for a simple replication study

Intramethod reliability studies of continuous exposure measures are analysed by analysis of variance (ANOVA) techniques. The selection of the appropriate version of the intraclass correlation coefficient depends on the reliability study design, within the context of ANOVA models.

First, consider a simple replication reliability study in which there is no characteristic that distinguishes the first and second measure across all subjects. Examples of this type of design include a study in which blood from each subject is analysed three times in the laboratory, or a study in which medical records are abstracted twice for each subject by two randomly selected abstractors from a pool of three or more abstractors. In studies of this type, the order of the measures can be considered arbitrary. This study design is analysed by a one-way random effects model ANOVA.

Each of n subjects is measured k times, with X_{ij} being the jth measure on subject i, $\bar{X}_i$ the mean for subject i, and $\bar{X}$ the overall mean. The computations for a one-way ANOVA appear in Table 4.2. The model is $X_{ij} = S_i + F_{ij}$, where S is the subject effect and F the random error as described on page 000. The reliability coefficient of X can be estimated by the intraclass correlation coefficient from the one-way random effect model termed here R_1:

$$\hat{\rho}_X = R_1 = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_X^2} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (k-1)\text{WMS}},$$

where BMS is the between-subjects mean square and WMS the within-subjects mean square.

Example. Table 4.3 presents a summary of the data and an analysis of variance for a test–retest study of a food frequency questionnaire measure of percentage energy from fat. The two measures were derived from two administrations of the questionnaire to 110 subjects 6 months apart. If this were considered to be a simple replication study, the reliability coefficient would be estimated as:

$$\hat{\rho}_X = R_1 = \frac{64.37 - 15.81}{64.37 + 15.81} = 0.61.$$

A lower 100 $(1 - \alpha)$ per cent confidence interval for R_1 can be estimated, under the assumption of normality of X and of the random error, from

Table 4.2 One-way analysis of variance and two-way analysis of variance for the computation of intraclass correlation coefficients

One-way ANOVA

Source of variance	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS = SS/df)	Expected mean square
Between subjects	$n - 1$	$k\sum_i(\bar{X}_i - \bar{X})^2$	BMS	$\sigma_F^2 + k\sigma_S^2$
Within subjects (random error)	$n(k-1)$	$\sum_i\sum_j(X_{ij} - \bar{X}_i)^2$	WMS	σ_F^2
Total	$nk - 1$	$\sum_i\sum_j(X_{ij} - \bar{X})^2$		

Two-way ANOVA

Source of variance	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS = SS/df)	Expected mean square
Between subjects	$n - 1$	$k\sum_i(\bar{X}_i - \bar{X})^2$	SMS	$\sigma_F^2 + k\sigma_S^2$
Between measures	$k - 1$	$n\sum_j(\bar{X}_j - \bar{X})^2$	MMS	$\left\{\begin{array}{l}\sigma_F^2 + \dfrac{n}{k-1}\sum m_j^{2\,a} \\ \sigma_F^2 + n\sigma_M^{2\,b}\end{array}\right.$
Random error	$(n-1)(k-1)$	by subtraction	EMS	σ_F^2
Total	$nk - 1$	$\sum_i\sum_j(X_{ij} - \bar{X})^2$		

[a] Fixed effects model
[b] Random effects model

$$R_1 \geq \frac{\frac{\text{BMS}}{\text{WMS}} - f}{\frac{\text{BMS}}{\text{WMS}} + (k-1)f}$$

where f denotes the $(1-\alpha)$ centile from tables of the F distribution with $n-1$ and $n(k-1)$ degrees of freedom.

Fleiss (1986) gives a method for the analysis of reliability studies in which the number of measures, k, can vary across subjects.

Intraclass correlation for subjects by measures (two-way) design

The remaining designs to be considered are intramethod reliability studies in which the two or more measures may have different characteristics; that is, the order of measures is not arbitrary. An example would be an inter-rater reliability study in which all subjects are interviewed by the same two interviewers, with X_1 being the measure by rater 1, X_2 by rater 2. When the k raters in the reliability study are the same as the k raters who will participate in the epidemiological study, the intraclass correlation coefficient from the two-way fixed effects ANOVA model (R_2) is appropriate. When the k raters in the reliability study are a *sample* from a population of raters to be used in the epidemiological study, the intraclass correlation coefficient from the two-way random effects ANOVA model (R_3) applies. An example of this would be when two interviewers participate in the reliability study as representative of the several interviewers who will participate in the full study.

The ANOVA table for this two-way (subjects by measures) design under the assumption of no interaction between subjects and measures, is given in the lower half of Table 4.2. X_{ij} is the jth measure on subject i, $\bar{X}_i$ is the mean for subject i, and $\bar{X}_j$ the mean for measure j. The computations of the mean squares given in the table are the same for the fixed and random effects models, but the estimate of σ_X^2 and therefore of R differs. SMS, MMS, and EMS are the mean squares for subjects, measures, and error respectively.

The two-way fixed effects model, where m_j is the fixed effect of measure j, is

$$X_{ij} = S_i + m_j + F_{ij}.$$

Under this model, the intraclass correlation coefficient (R_2) is estimated by

$$R_2 = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_X^2} = \frac{n(\text{SMS} - \text{EMS})}{n\text{SMS} + (k-1)\text{MMS} + (n-1)(k-1)\text{EMS}}.$$

No simple method is available for a confidence interval.

Under the two-way random effects model $X_{ij} = S_i + M_j + F_{ij}$, where M_j is the random effect of measure j, the intraclass correlation coefficient (R_3) is (Bartko 1966):

$$R_3 = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_X^2} = \frac{n(\text{SMS} - \text{EMS})}{n\text{SMS} + k\text{MMS} + (nk - n - k)\text{EMS}}.$$

A lower $100(1-\alpha)$ per cent confidence limit for R_3 has been derived by Fleiss and Shrout 1978 (see Fleiss 1986).

The estimates of the intraclass correlation for both fixed and random effects models include the variation between measures (e.g. between raters) as a source of variance in X. R_3 is generally less than R_2 when applied to the same data. This is because the error in X is estimated to be larger when the measures (e.g. raters) in the reliability study are only a sample of the measures to be used in the full study.

Example. Suppose that two interviewers administered a food frequency questionnaire to each subject in a reliability study, and the same two interviewers will be employed in the full study. Then the fixed effects model applies, and R_2 is an appropriate estimate of ρ_X. The computation is illustrated on the same data as in the previous example (Table 4.3). From the two-way ANOVA:

$$R_2 = \frac{110(64.37 - 13.81)}{110(64.37) + 233.81 + 109(13.81)} = 0.63.$$

Table 4.3 Example of an analysis of variance for a test–retest study of percentage energy from fat estimated from a food frequency questionnaire

Variable	N	Mean (% kcal)	Standard deviation
% energy at baseline (X_1)	110	37.5	5.96
% energy at 6 months (X_2)	110	35.5	6.53

One-way ANOVA

Source of variance	Sum of squares	Degrees of freedom	Mean square
Between subjects	7015.89	109	64.37
Within subjects (random error)	1739.52	110	15.81
Total	8755.41	219	39.98

Two-way ANOVA

Source of variance	Sum of squares	Degrees of freedom	Mean square
Between subjects	7015.89	109	64.37
Between measures	233.81	1	233.81
Random error	1505.71	109	13.81
Total	8755.41	219	39.98

Intraclass correlation which excludes the mean differences between measures

For some reliability study designs, the systematic differences between the measures, the differences between the $\bar{X}_j$, do not need to be included as a source of variance in X. This is the case when it is intended to adjust for the systematic difference between measures in the epidemiological study. For example, if interviewers produced different mean estimates of exposure, one might adjust in the epidemiological study for interviewer effects. Under these circumstances the intraclass correlation coefficient, R_4, is used. R_4 may also be appropriate for reliability studies in which the difference between X_1 and X_2 can be explained by a 'learning effect' or other factors that will not add error (variance) to the measure X in the full study.

When the measurement effects are to be excluded as a source of variance in X, the intraclass correlation coefficient for the fixed effects model is

$$R_4 = \frac{\text{SMS} - \text{EMS}}{\text{SMS} + (k-1)\text{EMS}}.$$

It has a lower $100(1-\alpha)$ confidence limit:

$$R_4 > \frac{\frac{\text{SMS}}{\text{EMS}} - f}{\frac{\text{SMS}}{\text{EMS}} + (k-1)f}$$

where f denotes the $(1-\alpha)$ centile of the F distribution with $(n-1)$ and $(n-1)(k-1)$ degrees of freedom.

R_4 will equal 1 when the measures are identical for each subject except for a constant difference between measures:

$$X_{i1} = X_2 + d_2 = \ldots X_k + d_k.$$

(R_4 differs from the Pearson correlation coefficient, in that R_4 includes any differences in scale, ~~that is a difference in units~~, as a source of variance in X.)

When the k measures, $X_1, \ldots, X_k$, have identical distributions, all four versions of R are essentially the same and are equal to the Pearson product-moment correlation coefficient.

Cohen's κ for binary or nominal variables

The intramethod reliability of nominal categorical variables, including dichotomous variables, is measured by Cohen's κ (Cohen 1960). This can be computed from a reliability study in which n subjects have each been measured twice where each measure is a nominal variable with k categories. (Note that k here refers to number of categories, not number of measures

Table 4.4 Layout of data for computation of Cohen's κ and weighted κ

		Measure 2					
		1	2	.	.	k	Total
Measure 1	1	p_{11}	p_{12}	.	.	p_{1k}	r_1
	2	p_{21}	p_{22}	.	.	p_{2k}	r_2
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	k	p_{k1}	p_{k2}	.	.	p_{kk}	r_k
	Total	s_1	s_2	.	.	s_k	1

per subject.) It is assumed that the two measures are equally accurate. To compute κ the data are laid out as a $k \times k$ table as in Table 4.4. The p_{ij} are the proportions of subjects who fall into the ith category in measure 1 and the jth category in measure 2. Note that the p_{ij} proportions in the table sum to 1 over the entire table. The r_i and s_j are the marginal proportions for the first and second measure respectively.

An obvious measure of agreement between two measures is the proportion of subjects for whom there was agreement. The observed proportion of agreement, P_o, is the sum of the proportions on the diagonal:

$$P_o = \sum_{i=1}^{k} p_{ii}.$$

However, this simple measure does not take into consideration the agreement that would be expected by chance. For example, suppose one interviewer classifies 5 per cent of subjects as exposed and 95 per cent as unexposed, but a second interviewer (who perhaps skipped the question) classified all subjects as unexposed. Then the percentage agreement would be 95 per cent, which does not reflect the poor reliability of the measure.

κ is a measure of agreement that corrects for the agreement that would be expected by chance. The expected agreement (on the diagonal), P_e, is:

$$P_e = \sum_{i=1}^{k} r_i s_i.$$

κ is estimated as the observed agreement beyond chance divided by the maximum possible agreement beyond chance:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e}.$$

$\hat{\kappa}$ is equal to 1 when there is exact agreement between the two measures for all subjects. It is greater than 0 when agreement is greater than chance, but

can be less than 0 if agreement is less than expected by chance.

An approximate lower $100(1-\alpha)$ per cent confidence bound for $\hat{\kappa}$ (if $\kappa \neq 0$) is (Fleiss *et al.* 1969; Fleiss 1981):

$$\hat{\kappa} - Z_\alpha \times \text{s.e.}(\hat{\kappa})$$

where Z_α is the value of the $(1-\alpha)$ centile of the standard normal variable and s.e.$(\hat{\kappa})$ is the estimated standard error of $\hat{\kappa}$:

$$\text{s.e.}(\hat{\kappa}) = \sqrt{\frac{a+b-c}{(1-P_e)^2 n}}$$

where

$$a = \sum_{i=1}^{k} p_{ii}[1-(r_i+s_i)(1-\hat{\kappa})]^2,$$

$$b = (1-\hat{\kappa})^2 \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} p_{ij}(r_i+s_j)^2,$$

and

$$c = [\hat{\kappa} - P_e(1-\hat{\kappa})]^2.$$

Example. Consider the reliability study described above in the intraclass correlation examples, and suppose subjects were to be divided into only two categories of fat intake: those above the median in percentage energy from fat and those below. Cross-classifying the 110 subjects by the two measures yields the following table, with the proportions in brackets:

		2nd measure: Upper half	2nd measure: Lower half	
1st measure	Upper half	40 (0.364)	15 (0.136)	55 (0.5)
	Lower half	15 (0.136)	40 (0.364)	55 (0.5)
		55 (0.5)	55 (0.5)	110 (1.0)

Then

$$P_o = 0.364 + 0.364 = 0.727,$$

$$P_e = (0.5 \times 0.5) + (0.5 \times 0.5) = 0.50,$$

$$\hat{\kappa} = \frac{0.727 - 0.5}{0.5} = 0.45.$$

When the number of categories is greater than two, the source of unreliability may become clearer by computing a κ for each category compared with all other categories combined. When the number of measures per subject is greater than two, another version of κ has been derived, under the assumption that there is no order to the measures (Landis and Koch 1977; Fleiss 1986). The assumptions and resulting κ are similar to the one-way ANOVA intraclass correlation coefficient.

Weighted κ for ordered categorical variables

The κ presented above for nominal categories is a measure of exact agreement, with all disagreements considered to be equally serious. For example, if rater 1 categorizes a subject as falling into category 1, and rater 2 disagrees, κ will be the same whether the second rating is category 2, 3, 4, etc. When the measure of interest in an intramethod reliability study is an ordered categorical variable, the use of κ is not appropriate. Instead κ_w, weighted κ (Cohen 1968) is used instead, as this measure yields a higher reliability when disagreements between raters are small compared with when they are large. In other words, weighted κ gives 'partial credit' for close but not exact agreement.

Weighted κ is estimated by

$$\hat{\kappa}_w = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the weighted observed proportion of agreement (across the entire table):

$$P_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij},$$

P_e is the weighted expected proportion of agreement (across the entire table):

$$P_e = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} r_i s_j,$$

and p_{ij}, r_i and s_j are the proportions shown in Table 4.4.

The usual weight applied for one measure yielding category i and the other category j is:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}.$$

This gives a weight of 1 for exact agreement and a weight of 0 when one measure yields the lowest category and the other the highest (kth) category.

A confidence interval for $\hat{\kappa}_w$ can be computed based on the large sample estimate of the standard error of $\hat{\kappa}_w$ (for $\kappa_w \neq 0$):

$$\text{s.e.}(\hat{\kappa}_w) = \sqrt{\frac{a - b}{(1 - P_e)^2 n}}$$

where

$$a = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} [w_{ij} - (\bar{w}_i + \bar{w}_j)(1 - \hat{\kappa}_w)]^2,$$

$$b = [\hat{\kappa}_w - P_e(1 - \hat{\kappa}_w)]^2,$$

$$\bar{w}_i = \sum_{j=1}^{k} s_j w_{ij},$$

and

$$\bar{w}_j = \sum_{i=1}^{k} r_i w_{ij}.$$

For the reliability study used in the previous examples, if percentage calories from fat were divided into four equal ordered categories, κ_w would be 0.57.

Ordered categorical variables are often created by categorizing a continuous variables. In these situations, the intraclass correlation coefficient could also be computed on the underlying variable.

Interpretation and limitations of κ and weighted κ

Certain similarities allow κ and κ_w to be interpreted as reliability coefficients. κ for binary variables and κ_w are equal to the intraclass correlation coefficient based on the two-way random effects model (R_3), except for a term that goes to 0 as n increases (Fleiss and Cohen 1973; Fleiss 1975; Dunn 1989), when the categories are numerically coded 1 for category 1, 2 for category 2, etc. κ_w is also equal to the Pearson product-moment correlation coefficient if the marginal distributions of the two measures are identical (Cohen 1968).

There are several limitations to the interpretation of κ and κ_w (Maclure and Willett 1987). The value of κ_w varies with the number of exposure categories. In the reliability study of percentage energy from fat used in the previous examples, κ was 0.45 when the measure was divided into two categories and κ_w was 0.57 when four categories were used. (The intraclass correlation was 0.61 when percentage energy from fat was treated as a continuous variable.)

In addition, the value of κ or κ_w depends on the distribution of exposure in the population. Thus κ cannot be used to compare the reliability of two instruments measuring the same underlying exposure if the two reliability studies were conducted in populations which may have different distributions of the true exposure. This is similar to the problem of comparing reliability coefficients across populations which differ in the variance of exposure. However, this dependence of κ on the prevalence of exposure may be a desirable property for, as noted in Chapter 3, the attenuation of the odds ratio depends on the exposure prevalence as well as the sensitivity and specificity of the measurement. When κ is derived from a study in which the two dichotomous measures compared have equal sensitivity, equal specificity and independent error probabilities (similar to the parallel test model), κ has been shown to be crudely related to the attenuation of the odds ratio under non-differential misclassification (for a limited range of parameters) (Thompson and Walter 1988):

$$\text{OR}_0 \doteq (\text{OR}_T - 1)\kappa + 1.$$

In addition, since κ_w could be interpreted as an intraclass correlation coefficient, the interpretations given for $\rho_{X_1X_2}$ in terms of ρ_{TX_1} in this chapter (and subsequently in terms of the attenuation of the odds ratio given in the last chapter) might crudely apply to κ_w, depending on the degree of violation of the assumptions of the error model.

Neither κ nor weighted κ_w is sufficient to detect differential misclassification between cases and controls. κ is a single summary measure of the misclassification in the measure, while the assessment of differential misclassification requires estimates of sensitivity and specificity for cases and controls (or the misclassification matrices for $k > 2$). κ could be similar for cases and controls even when there is differential misclassification, that is when the underlying sensitivity and specificity of the exposure measurement differs substantially between the two groups. This is analogous to the problem that intramethod reliability studies of continuous variables can provide information only on precision but not on bias, and therefore cannot assess differential bias between cases and controls.

Other types of analysis of reliability studies

Some authors (Liu *et al.* 1978) present the reliability of a continuous measure X in terms other than the reliability coefficient ρ_X. The *ratio of the within-subject variance, σ^2_W, to the between-subject variance, σ^2_S,* is sometimes used, where S in the subject effect and σ^2_W is defined as $\sigma^2_X - \sigma^2_S$. This ratio is a simple transformation of ρ_X:

$$\frac{\sigma_W^2}{\sigma_S^2} = \frac{1 - \rho_X}{\rho_X},$$

where ρ_X is based on the appropriate measure of the reliability coefficient. Because the ratio of within- to between-subject variance provides the same information as the reliability coefficient, the equations presented in Chapters 3, 4 and 5 as functions of ρ_X (or of ρ_{TX}^2 when ρ_X can be considered as an estimate of ρ_{TX}^2) could be presented in terms of σ_W^2/σ_S^2 by substituting

$$\rho_X = 1/[(\sigma_W^2/\sigma_S^2) + 1].$$

One additional analytical technique for reliability studies of continuous measures deserves mention: the *coefficient of variation* (Garber and Carey 1984). For laboratory measures, reliability is often assessed by repeated analysis of a single reference material with known true measurement t. For example a fluid with a known concentration of retinol might be repeatedly analysed to yield measures of X, the measured retinol concentration. (This type of study only assesses the laboratory error, of course, and excludes errors due to storage and handling of specimens, and error due to the variation in the measure over time within individuals.) In such studies, the mean and variance of X can be used to assess the reliability of X. The bias of the measure can be estimated as

$$b = \bar{X} - t.$$

Because t is a constant, the variance of X in the reliability study is equal to the variance of the random error F:

$$\hat{\sigma}_F^2 = \hat{\sigma}_X^2.$$

A reliability coefficient cannot be estimated, because the comparison measure is constant (t) for each measurement of X. Instead a coefficient of variation, CV, defined as the estimated standard deviation divided by the mean of $X \times 100$, is often used:

$$\text{CV\%} = \frac{\hat{\sigma}_X}{\bar{X}} \times 100.$$

A small CV is considered to indicate a reliable measure. However, it may be more informative to relate the variance of the random error, $\hat{\sigma}_F^2$, to the expected variance of X in the population of interest (see right-hand side of Equation 4.9), to yield information closer to the reliability coefficient of X in the population of interest.

Reliability study designs may be more complex than those covered in this chapter, in order to yield more information about the measurement error. For example, an intramethod study could have two interviewers question each subject with each interview coded by two coders, or an intermethod study could be conducted of three or more instruments which measure the same exposure. Dunn (1989) provides other approaches to the analysis of simple intermethod and intramethod studies, and also gives methods for more complex designs.

In addition, reliability studies might include evaluation of several exposures simultaneously for, as noted in Chapter 3, the bias in the odds ratio for one exposure depends on the measurement error in the covariates as well as the primary exposure. Procedures have been developed that incorporate information from reliability studies of multiple exposures into the analysis of the parent epidemiological study; these are briefly described in Chapter 5. These procedures yield an estimate of the effect of the unreliability of the exposure measure(s) on the odds ratio in the particular research context of interest; this could be more informative than a reliability coefficient or κ.

Finally, reliability studies of some instruments do not require a comparison measure. In the social sciences, the reliability of a total score on a test is often assessed by the reliability of parts of the test, e.g. by the correlation between test items or the correlation between two halves of the test (Carmines and Zeller 1979; Dunn 1989). This is termed *internal consistency reliability*, and the Spearman–Brown formula given in Chapter 5 (Equation 5.3) or related equations are used to assess the reliability of the total score. This approach may be useful for some applications in epidemiology. (One note of caution: in some packaged programs, SPSS in particular, the term reliability is often used to refer to the reliability of the sum of the measures. If you are interested in the reliability of the individual measures X_1, X_2, etc., the reliability coefficient that is reported by the program may not be the statistic you are interested in.)

Sample size for reliability studies

The computation of the required sample size depends on the design and aim of the reliability study. For an intermethod reliability study conducted to assess differential bias between cases and controls, the required sample size could be based on a two-sample comparison of means (Kelsey *et al.* 1986), where the variable of interest is $(X_{i1} - X_{i2})$. For a validity or intermethod reliability study conducted to estimate $\rho_{X_1X_2}$, the sample size should be that needed to evaluate the correlation between two variables (Willett 1990). Note that the null hypothesis to be tested is not $\rho_{X_1X_2} = 0$ (for it should be assumed that X_1 and X_2 are at least positively correlated); rather the study should have sufficient power to detect whether $\rho_{X_1X_2}$ is greater than some minimum value. For intramethod reliability studies, Donner and Eliasziw (1987) have presented methods for selecting the appropriate number of subjects and number of measurements per subject for studies which will be analysed using the intraclass correlation coefficient, and Jannarone *et al.* (1987) have given an approach for studies which estimate κ.

SUMMARY

Reliability studies can be designed to provide information about the validity of a measure if the comparison measure is carefully selected. If a comparison measure without differential bias between cases and controls is chosen, then a reliability study can yield estimates of differential bias in the measure of interest. Useful information about the validity coefficient can be obtained from a comparison of the measure of interest with an equally accurate or more accurate measure when the errors between the two measures are uncorrelated. When the measures in a reliability study have correlated errors, the reliability coefficient can provide only an upper limit to the validity coefficient.

The choice of an analytical technique for a validity or reliability study depends on whether the exposure is measured as a continuous variable, as a nominal categorical (or dichotomous variable), or as an ordered categorical variable. The choice also depends on whether the two or more measures in the reliability study will be used interchangeably in the full epidemiological study or only one will be used, and on other design issues. For example, for a reliability study in which a continuous exposure measure from proxy respondents was compared with the same measure from the subjects themselves, the Pearson correlation coefficient might be appropriate for the analysis if all interviews in the full study were to be from proxy respondents; a version of the intraclass correlation coefficient (R_2) might be used if both proxies and index cases were to be included in the full study; and another version (R_4) might be used if both were included but a factor indicating whether the interview was by index or proxy respondent was to be adjusted for in the full study. Versions of Cohen's κ are most commonly used to summarize the information obtained in reliability studies involving nominal or ordered categorical variables.

REFERENCES

Allen, M.J. and Yen, W.M. (1979). *Introduction to Measurement Theory*, pp. 1–117. Brooks/Cole, Monterey.

Altman, D.G. and Bland, J.M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, 32, 307–17.

Armitage, P. and Berry, G. (1987). *Statistical methods in medical research*, (2nd edn). Blackwell Scientific Publications, Oxford.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3–11.

Bohrnstedt, G.W. (1983). Measurement. In *Handbook of survey research*, (ed. P. Rossi, J. Wright, and A. Anderson), pp. 70–121. Academic Press, Orlando, Florida.

Carmines, E.G. and Zeller, R.A. (1979). *Reliability and validity assessment*. Sage, Beverly Hills, California.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–20.

Clayton, D. (1985). Using test-retest reliability data to improve estimates of relative risk: an application of latent class analysis. *Statistics in Medicine*, 4, 445–55.

Decker, M.D., Booth, A.L., Dewey, M.J., Fricker, R.S., Hutcheson, R.H., and Schaffner, W. (1986). Validity of food consumption histories in a foodborne outbreak investigation. *American Journal of Epidemiology*, 124, 859–63.

Donner, A. and Eliasziw, M. (1987). Sample size requirements for reliability studies. *Statistics in Medicine*, 6, 441–448.

Dunn, G. (1989). *Design and analysis of reliability studies*. Edward Arnold, London, and Oxford University Press, New York.

Eskenazi, B. and Pearson, K. (1988). Validation of a self-administered questionnaire for assessing occupational and environmental exposures of pregnant women. *American Journal of Epidemiology*, 128, 1117–29.

Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31, 651–9.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*, (2nd edn), pp. 188–236. John Wiley and Sons, New York.

Fleiss, J.L. (1986). *The design and analysis of clinical experiments*, pp. 1–32. John Wiley and Sons, New York.

Fleiss, J.L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613–9.

Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323–7.

Fleiss, J.L. and Shrout, P.E. (1977). The effects of measurement errors on some multivariate procedures. *American Journal of Public Health*, 67, 1188–91.

Fleiss, J.L. and Shrout, P.E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*, 43, 259–62.

Garber, C.C. and Carey, R.N. (1984). Laboratory statistics. In *Clinical chemistry: theory, analysis, and correlation*, (ed. L. Kaplan and A. Pesce), pp. 290–2. C.V. Mosby, St. Louis, Missouri.

Gerbert, B., Stone, G., Stulbarg, M., Gullion, D.S., and Greenfield, S. (1988). Agreement among physician assessment methods. *Medical Care*, 26, 519–35.

Harlow, S.D. and Linet, M.S. (1989). Agreement between questionnaire data and medical records. The evidence for accuracy of recall. *American Journal of Epidemiology*, 129, 233–48.

Hill, A.B. (1953). Observation and experiment. *New England Journal of Medicine*, 248, 995–1001.

Hui, S.L. and Walter, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167–71.

Jannarone, R.J., Macera, C.A., and Garrison, C.Z. (1987). Evaluating inter-rater agreement through 'case-control' sampling. *Biometrics*, 43, 433–7.

Jarvholm, B. and Sanden, A. (1987). Estimating asbestos exposure: a comparison of methods. *Journal of Occupational Medicine*, 29, 361–3.

Kaldor, J. and Clayton, D. (1985). Latent class analysis in chronic disease epidemiology. *Statistics in Medicine*, 4, 327–35.

Kelsey, J.L., Thompson, W.D., and Evans, A.S. (1986). *Methods in observational epidemiology*, pp. 277, 285–308. Oxford University Press, New York.

Klesges, R.C., Klesges, L.M., Swenson, A.M., and Pheley, A.M. (1985). A valida-

tion of two motion sensors in the prediction of child and adult physical activity levels. *American Journal of Epidemiology*, **122**, 400–10.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–74.

Liu, K., Stamler, J., Dyer, A., McKeever, J., and McKeever, P. (1978). Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. *Journal of Chronic Diseases*, **31**, 399–418.

Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores*, pp. 13–278. Addison-Wesley, Reading, Massachusetts.

Maclure, M. and Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, **126**, 161–9.

Nunnally, J.C. (1978). *Psychometric theory*, pp. 190–255. McGraw-Hill, New York.

Shekelle, R.B., Shryock, A.M., Paul, O., Lepper, M., Stamler, J., Lui, S., and Raynor, W.J. (1981). Diet, serum cholesterol, and death from coronary heart disease: The Western Electric Study. *New England Journal of Medicine*, **304**, 65–70.

Siconolfi, S.F., Lasater, T.M., Snow, R.C.K., and Carleton, R.A. (1985). Self-reported physical activity compared with maximal oxygen uptake. *American Journal of Epidemiology*, **122**, 101–5.

Thompson, W.D. and Walter, S.D. (1988). Variance and dissent: A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, **41**, 949–58.

Walker, A.M. and Blettner, M. (1985). Comparing imperfect measures of exposure. *American Journal of Epidemiology*, **121**, 783–90.

Walter, S.D. (1984). Commentary on 'Use of dual responses to increase validity of case-control studies.' *Journal of Chronic Diseases*, **37**, 137–9.

Walter, S.D. and Irwig, L.M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, **41**, 923–37.

Willett, W. (1990). *Nutritional epidemiology*, pp. 34–126. Oxford University Press, New York.

Willett, W.C., Stampfer, M.J., Underwood, B.A., Speizer, F.E., Rosner, B., and Hennekens, C.H. (1983). Validation of a dietary questionnaire with plasma carotenoid and alpha-tocopherol levels. *American Journal of Clinical Nutrition*, **38**, 631–9.

Willett, W.C., Sampson, L., Stampfer, M.J., Rosner, B., Bain, C., Witschi, J., Hennekens, C.H., and Speizer, F.E. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology*, **122**, 51–65.

Williams, E., Klesges, R.C., Hanson, C.L., and Eck, L.H. (1989). A prospective study of the reliability and convergent validity of three physical activity measures in a field research trial. *Journal of Clinical Epidemiology*, **42**, 1161–70.

Wu, M.L., Wittemore, A.S., and Jung, D.L. (1988). Errors in reported dietary intakes. II. Long-term recall. *American Journal of Epidemiology*, **128**, 1137–45.

5

Reducing measurement error and its effects

In most statistical approaches to observer variability, . . . no efforts have been made to detect and remove sources of inconsistency. After noting the disagreements and quantifying them with kappa scores or other indices of concordance, investigators write the paper and depart from the analytic scene. (Feinstein 1983).

INTRODUCTION

In Chapters 3 and 4 we have described the effects of exposure measurement error and how the extent of measurement error can be assessed. The primary focus of exposure measurement, however, should not be on assessment but on the reduction of measurement error and its effects.

Several approaches to the reduction of measurement error and its effects are discussed in this chapter. The first is the use of multiple measures of exposure, an important method of reducing measurement error. Next, adjustment procedures are briefly covered; these are methods of 'correcting', study results for the effect of measurement error by using information from validity or reliability studies. Finally, minimization of error by means of quality control procedures is discussed. These procedures include a wide range of methods of reducing measurement error during each phase of a study, from instrument development through data collection and creation of the data set for analysis.

USE OF MULTIPLE MEASURES OF EXPOSURE

The use of the average (or sum) of two or more measures of the exposure for each subject in an epidemiological study can be an effective method of decreasing the measurement error, in comparison with the use of a single measurement. The measures can be repeated administrations of the same instrument, or measurements from two different instruments. For example serum cholesterol could be measured by use of an average of three measurements from samples collected over a year. Or, dietary fat could be assessed as an average of a food frequency measurement and a measurement from a 7-day diet diary.

The use of *multiple measures* refers to repeated measurement of all subjects in an epidemiological study to *reduce* measurement error; this differs from a reliability study, in which a sample of subjects would be repeatedly measured to *assess* measurement error. Nonetheless, many of the concepts introduced in the last chapter are important in understanding the benefits of multiple measures.

The use of multiple measures to increase validity under parallel tests

The improvement in the validity of an exposure variable resulting from the combination of multiple measures is easily demonstrated when the errors in the two or more measures to be averaged are equal and uncorrelated (the parallel test model) (Carmines and Zeller 1979; Bohrnstedt 1983; Fleiss 1986; Dunn 1989).

Suppose each individual in a population is measured k times, by use of parallel measures of the underlying true exposure T, yielding observations of continuous variables $X_1, \ldots, X_k$. Recall from Chapter 4 that, under the model of parallel tests, the errors of the measures ($E_1, \ldots, E_k$) are uncorrelated with each other and with T, and the variances of the errors are equal (σ_E^2). This implies that the correlation of each X_j with T is identical, ρ_{TX}. The average measure for individual i, A_i, is computed as

$$A_i = \frac{X_{i1} + \ldots + X_{ik}}{k},$$

where X_{ij} represents the observation on subject i of variable x_j. Then the variable A has a validity coefficient:

$$\rho_{TA} = \sqrt{\sigma_T^2/(\sigma_T^2 + \sigma_E^2/k)}. \qquad [5.1]$$

It can be seen from Equation 5.1, that, as k increases, the term σ_E^2/k goes to 0. This shows that the validity coefficient of A is greater than that of the individual measures, $X_1, \ldots, X_k$ (i.e. Equation 5.1 is greater for $k \geq 2$ than for $k = 1$), and that the validity coefficient of A approaches 1 as k increases.

Equation 5.1 can be rewritten as a function of the validity coefficient of the parallel measures, ρ_{TX}:

$$\rho_{TA} = \sqrt{\frac{k\rho_{TX}^2}{1 + (k-1)\rho_{TX}^2}}. \qquad [5.2]$$

If ρ_{TX} is known from a validity or reliability study, then the validity coefficient for A can be calculated from Equation 5.2. Table 5.1 gives examples of the improvement in validity one can achieve by using multiple parallel measures of T. For example, averaging two measures each with a validity

Table 5.1 Improvement in the validity of a measure by averaging k parallel measures[a]

Number of measures k	$\rho_{TX} = 0.5$ ρ_{TA}	$\rho_{TX} = 0.7$ ρ_{TA}
1	0.50	0.70
2	0.63	0.81
3	0.71	0.86
5	0.79	0.91
10	0.88	0.95

[a] ρ_{TX} is the validity coefficient of each parallel measure, X_j, of T; ρ_{TA} is the validity coefficient of A, the average of the parallel measures

coefficient of 0.7 can yield a new exposure measure with a validity coefficient of 0.8. Similarly, averaging five measures each with a validity coefficient of 0.5 can also yield a new measure with validity coefficient of 0.8.

The same concept expressed in terms of the reliability of A, ρ_A, as a function of the reliability of X is the Spearman-Brown formula (Spearman 1910; Brown 1910):

$$\rho_A = \frac{k\rho_{x_1x_2}}{1 + (k-1)\rho_{x_1x_2}}. \qquad [5.3]$$

$\rho_{x_1x_2}$ represents the common correlation between any two of the k measures, and may be estimated from the average Pearson correlation coefficient of the pairs of measures (Carmines and Zeller 1979; Bohrnstedt 1983).

The advantage of using multiple, parallel exposure measures for each subject in an epidemiological study (assuming non-differential misclassification) is that it would result in less attenuation of the observed odds ratio (or other measure of association) due to measurement error. This also implies that a smaller sample size is required for a given power. For example, in a case-control study with equal numbers of cases and controls, the sample size required with k measures per subject, n_k, compared to the sample size needed when only one measure per subject is used, n_1, is (Fleiss 1986):

$$n_k = \frac{1 + (k-1)\rho_{x_1x_2}}{k} n_1. \qquad [5.4]$$

Example. Suppose a case-control study is to be conducted on the relationship between serum cholesterol and colon cancer in a large health maintenance organization, where records of prediagnostic serum cholesterol levels are available. A test-retest reliability study of serum cholesterol levels over the time period of interest yields an