CHAPTER 4

Planning the Measurements: Precision, Accuracy, and Validity

Stephen B. Hulley, Thomas B. Newman, and Steven R. Cummings

Measurements describe phenomena in terms that can be analyzed statistically, and the validity of a study depends on how well the variables designed for the study represent the phenomena of interest (Figure 4.1). How well does a handheld glucometer measure blood glucose, for example, or an insomnia questionnaire detect amount and quality of sleep?

This chapter begins by considering how the choice of **measurement scale** influences the information content of the measurement. We then turn to the central goal of minimizing measurement error: how to design measurements that are relatively **precise** (free of random error) and **accurate** (free of systematic error), thereby enhancing the appropriateness of drawing inferences from these measurements to the phenomena of interest. We address the concept of **validity**, a qualitative relative of accuracy, before concluding with some considerations for clinical and translational research, noting especially the advantages of storing specimens for later measurements.

MEASUREMENT SCALES

Table 4.1 presents a simplified classification of measurement scales and the information that results. The classification is important because some types of variables are **more informative** than others, adding power or reducing sample size requirements, and revealing more detailed distribution patterns.

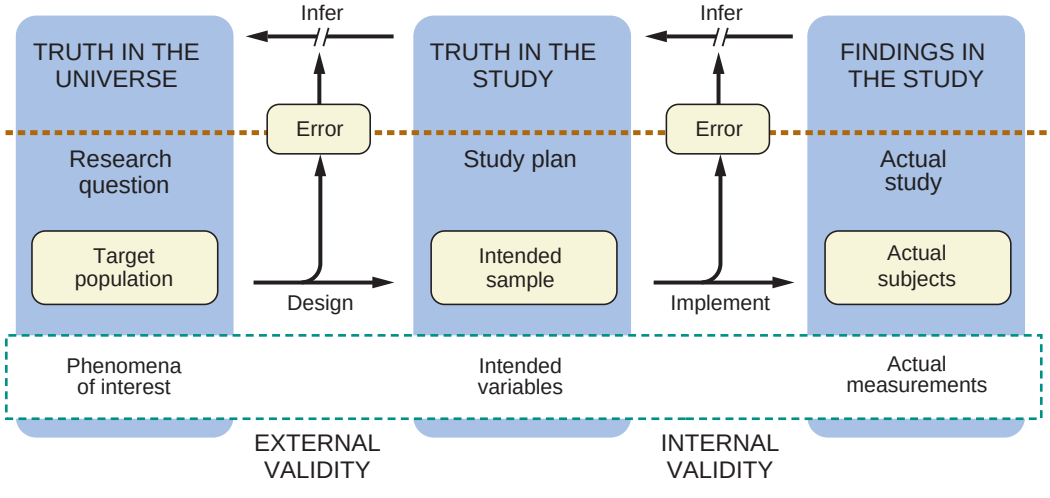

FIGURE 4.1 Designing measurements that represent the phenomena of interest.

TABLE 4.1 MEASUREMENT SCALES

TYPE OF MEASUREMENT	CHARACTERISTICS OF VARIABLE	EXAMPLE	DESCRIPTIVE STATISTICS	STATISTICAL POWER
		Categorical		
Dichotomous	Two categories	Vital status (alive or dead)	Counts, proportions	Low
Nominal	Unordered categories	Race; blood type	Same as above	Low
Ordinal	Ordered categories with intervals that are not quantifiable	Degree of pain; social class	In addition to the above: medians	Intermediate
		Numeric		
Continuous or discrete[†]	Ranked spectrum with quantifiable intervals	Weight; number of cigarettes/day	In addition to the above: means, standard deviations	High

[†]Continuous variables have an infinite number of values (e.g., weight), whereas discrete numeric variables are more limited (e.g., number of cigarettes/day). Discrete variables that have a large number of possible values resemble continuous variables for practical purposes of power and analysis.

Numeric Variables: Continuous and Discrete

Numeric variables can be quantified with a number that expresses how much or how many. **Continuous variables** quantify how much on an infinite scale; the number of possible values of body weight, for example, is limited only by the sensitivity of the machine that is used to measure it. Continuous variables are rich in information. **Discrete numeric variables** quantify how many on a scale with fixed units, usually integers, such as the number of times a woman has been pregnant. Discrete variables that have a considerable number of possible values can resemble continuous variables in statistical analyses and be equivalent for the purpose of designing measurements.

Categorical Variables: Dichotomous, Nominal, and Ordinal

Phenomena that are not suitable for quantification are measured by classifying them in categories. **Categorical variables** with two possible values (e.g., dead or alive) are termed **dichotomous**. Categorical variables with more than two categories (polychotomous) can be further characterized according to the type of information they contain. Among these, **nominal variables** have categories that are not ordered; type O blood, for example, is neither more nor less than type B blood; nominal variables tend to have an absolute qualitative character that makes them straightforward to measure. The categories of **ordinal variables** do have an order, such as severe, moderate, and mild pain. The additional information is an advantage over nominal variables, but because ordinal variables do not specify a numerical or uniform difference between one category and the next, the information content is less than that of discrete or continuous numeric variables.

Choosing a Measurement Scale

A good general rule is to **prefer continuous over categorical** variables when there is a choice, because the additional information they contain improves statistical efficiency. In a study comparing the antihypertensive effects of several treatments, for example, measuring blood pressure in millimeters of mercury allows the investigator to observe the magnitude of the change in every subject, whereas measuring it as hypertensive versus normotensive limits the

assessment. The continuous variable contains more information, and the result is a study with more power and/or a smaller sample size (Chapter 6).

Continuous variables also allow for more flexibility than categorical variables in fitting the data to the nature of the variable or the shape of the association, especially when the relationship might have a complex pattern. For example, a study of the relationship of vitamin D to various cancers would need to measure vitamin D as a continuous variable to be able to detect a possible **U-shaped pattern**, the higher mortality that has been observed in subjects with low or high levels of vitamin D than in those with intermediate levels (1). And a study of predictors of low birth weight babies should record actual birth weight rather than above or below the conventional 2,500 g **threshold**; this leaves the analytic options open, to change the cutoff that defines low birth weight, or to develop an ordinal scale with several categories of birth weight (e.g., >2,500 g, 2,000–2,499 g, 1,500–1,999 g, and <1,500 g).

Similarly, when there is the option of designing the number of response categories in an ordinal scale, as in a question about food preferences, it is often useful to provide a half-dozen categories that range from "strongly dislike" to "extremely fond of." The results can later be collapsed into a dichotomy (dislike and like), but not vice versa.

Many characteristics, particularly symptoms like pain or aspects of lifestyle, are difficult to describe with categories or numbers. But these phenomena often have important roles in diagnostic and treatment decisions, and the attempt to measure them is an essential part of the scientific approach to description and analysis. This is illustrated by the Short Form (SF)-36, a standardized questionnaire for assessing **quality of life** that produces discrete numerical ratings (2). The process of classification and measurement, if done well, can increase the objectivity of our knowledge, reduce bias, and provide a means of communication.

PRECISION

The **precision** of a variable is the degree to which it is reproducible, with nearly the same value each time it is measured. A beam scale can measure body weight with great precision, whereas an interview to measure quality of life is more likely to produce values that vary from one observer or occasion to another. Precision has a very important influence on the power of a study. The more precise a measurement, the greater the statistical power at a given sample size to estimate mean values and to test hypotheses (Chapter 6).

Precision (also called **reproducibility**, **reliability**, and **consistency**) is a function of **random error** (chance variability); the greater the error, the less precise the measurement. There are three main sources of random error in making measurements.

- **Observer variability** is due to the observer, and includes such things as choice of words in an interview and skill in using a mechanical instrument.
- **Instrument variability** is due to the instrument, and includes changing environmental factors (e.g., temperature), aging mechanical components, different reagent lots, and so on.
- **Subject variability** is due to intrinsic biologic variability in the study subjects unrelated to variables under study, such as variablity due to time of day of measurements or time since last food or medication.

Assessing Precision

Precision is assessed as the **reproducibility** of repeated measurements, either comparing measurements made by the same person (within-observer reproducibility) or different people (between-observer reproducibility). Similarly, it can be assessed within or between instruments. The reproducibility of continuous variables is often expressed as either the within-subject standard deviation or the **coefficient of variation** (within-subject standard deviation divided by

the mean).[1] For categorical variables, percent agreement, the interclass correlation coefficient, and the **kappa** statistic are often used (3–5).

Strategies for Enhancing Precision

There are five approaches to minimizing random error and increasing the precision of measurements (Table 4.2):

1. **Standardizing the measurement methods.** All study protocols should include specific instructions for making the measurements (**operational definitions**). This may include written directions on how to prepare the environment and the subject, how to carry out and record the interview, how to calibrate the instrument, and so forth (Appendix 4). This set of materials, part of the **operations manual**, is essential for large and complex studies and recommended for smaller ones. Even when there is only a single observer, specific written guidelines for making each measurement will help her performance to be uniform over the duration of the study and serve as the basis for describing the methods when the results are published.

TABLE 4.2 STRATEGIES FOR REDUCING RANDOM ERROR IN ORDER TO INCREASE PRECISION, WITH ILLUSTRATIONS FROM A STUDY OF ANTIHYPERTENSIVE TREATMENT

STRATEGY TO REDUCE RANDOM ERROR	SOURCE OF RANDOM ERROR	EXAMPLE OF RANDOM ERROR	EXAMPLE OF STRATEGY TO PREVENT THE ERROR
1. Standardizing the measurement methods in an operations manual	Observer	Variation in blood pressure (BP) measurement due to variable rate of cuff deflation (often too fast)	Specify that the cuff be deflated at 2 mm Hg/second
	Subject	Variation in BP due to variable length of quiet sitting before measurement	Specify that subject sit in a quiet room for 5 minutes before BP measurement
2. Training and certifying the observer	Observer	Variation in BP due to variable observer technique	Train observer in standard techniques
3. Refining the instrument	Instrument and observer	Variation in BP due to malfunctioning manometer	Purchase new high quality manometer
4. Automating the instrument	Observer	Variation in BP due to variable observer technique	Use automatic BP measuring device
	Subject	Variation in BP due to subject's emotional reaction to observer	Use automatic BP measuring device
5. Repeating the measurement	Observer, subject, and instrument	All measurements and all sources of variation	Use mean of two or more BP measurements

[1] When there are two measurements of a continuous variable per subject, it may be tempting to express their agreement using a **correlation coefficient**. However, because the correlation coefficient is extremely sensitive to outliers (3,4), a better approach is a "Bland-Altman" plot in which the difference between the two measurements is plotted as a function of their mean. If the absolute value of the difference between the measurements tends to increase linearly with the mean, the coefficient of variation is a better way to summarize variability than the within-subject standard deviation.

2. **Training and certifying the observers.** Training will improve the consistency of measurement techniques, especially when several observers are involved. It is often desirable to design a formal test of the mastery of the techniques specified in the operations manual and to certify that observers have achieved the prescribed level of performance (Chapter 17).
3. **Refining the instruments.** Mechanical and electronic instruments can be engineered to reduce variability. Similarly, questionnaires and interviews can be written to increase clarity and avoid potential ambiguities (Chapter 15).
4. **Automating the instruments.** Variations in the way human observers make measurements can be eliminated with automatic mechanical devices and self-response questionnaires.
5. **Repetition.** The influence of random error from any source is reduced by repeating the measurement, and using the mean of the two or more readings. Precision will be substantially increased by this strategy, the primary limitations being the added cost and practical difficulties of repeating the measurements.

For each measurement in the study, the investigator must decide how vigorously to pursue each of these strategies. This decision can be based on the importance of the variable, the magnitude of the potential problem with precision, and the feasibility and cost of the strategy. In general, the first two strategies (standardizing and training) should always be used, and the fifth (repetition) is an option that is guaranteed to improve precision when it is feasible and affordable.

TABLE 4.3 THE PRECISION AND ACCURACY OF MEASUREMENTS

	PRECISION	ACCURACY
Definition	The degree to which a variable has nearly the same value when measured several times	The degree to which a variable approximates the true value
Best way to assess	Comparison among repeated measures	Comparison with a "gold standard"
Value to study	Increase power to detect effects	Increase validity of conclusions
Threatened by	Random error (chance) contributed by	Systematic error (bias) contributed by
	The observer	The observer
	The subject	The subject
	The instrument	The instrument

ACCURACY

The **accuracy** of a variable is the degree to which it represents the true value.

Accuracy is different from precision in the ways shown in Table 4.3, and the two are not necessarily linked. If serum cholesterol were measured repeatedly using standards that had inadvertently been diluted twofold, for example, the result would be inaccurate but could still

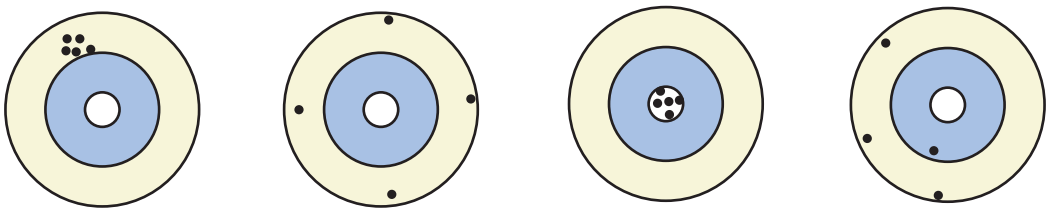

FIGURE 4.2 The difference between precision and accuracy.

be precise (consistently off by a factor of two). This concept is further illustrated in Figure 4.2. Accuracy and precision do often go hand in hand, however, in the sense that many of the strategies for increasing precision will also improve accuracy.

Accuracy is a function of **systematic error** (bias); the greater the error, the less accurate the variable. The three main classes of measurement error noted in the earlier section on precision each have counterparts here.

- **Observer bias** is a distortion, conscious or unconscious, in the perception or reporting of the measurement by the observer. It may represent systematic errors in the way an instrument is operated, such as a tendency to round down blood pressure measurements or to use leading questions in interviewing a subject.
- **Instrument bias** can result from faulty function of a mechanical instrument. A scale that has not been calibrated recently may have drifted downward, producing consistently low body weight readings.
- **Subject bias** is a distortion of the measurement by the study subject, for example, in reporting an event (respondent or recall bias). Patients with breast cancer who believe that alcohol is a cause of their cancer, for example, may exaggerate the alcohol intake they report.

The accuracy of a measurement is best assessed by comparing it, when possible, to a "**gold standard**"—a reference measurement carried out by a technique that is believed to best represent the true value of the characteristic. The decision as to what measurement approach to designate as the gold standard can be a difficult judgment that the investigator needs to make, drawing on previous work in the field.

The degree of accuracy can be expressed, for measurements on a continuous scale, as the mean difference between the measurement under investigation and the gold standard across study subjects. For measurements on a dichotomous scale, accuracy in comparison to a gold standard can be described in terms of sensitivity and specificity (Chapter 12). For measurements on categorical scales with more than two response options, the percent correct on each can be calculated.

Strategies for Enhancing Accuracy

The major approaches to increasing accuracy include the first four strategies listed earlier for precision, and three additional ones (Table 4.4):

1. **Standardizing the measurement methods.**
2. **Training and certifying the observers.**
3. **Refining the instruments.**
4. **Automating the instruments.**
5. **Making unobtrusive measurements.** It is sometimes possible to design measurements that the subjects are not aware of, thereby eliminating the possibility that they will consciously bias the variable. For example, an evaluation of the effect of placing a hand sanitizer and a hand hygiene poster in a hospital cafeteria utilized observers who blended in with cafeteria customers (6).
6. **Calibrating the instrument.** The accuracy of many instruments, especially those that are mechanical or electrical, can be increased by periodic calibration with a gold standard.
7. **Blinding.** This classic strategy does not ensure the overall accuracy of the measurements, but it can eliminate **differential bias** that affects one study group more than another. In a double-blind clinical trial the subjects and observers do not know whether active medicine or placebo has been assigned, and any inaccuracy in measuring the outcome will be the same in the two groups.

The decision on how vigorously to pursue each of these seven strategies for each measurement rests, as noted earlier for precision, on the judgment of the investigator. The considerations are the potential impact that the anticipated degree of inaccuracy will have on the conclusions of the study, and the feasibility and cost of the strategy. The first two strategies

TABLE 4.4 STRATEGIES FOR REDUCING SYSTEMATIC ERROR IN ORDER TO INCREASE ACCURACY, WITH ILLUSTRATIONS FROM A STUDY OF ANTIHYPERTENSIVE TREATMENT

STRATEGY TO REDUCE SYSTEMATIC ERROR	SOURCE OF SYSTEMATIC ERROR	EXAMPLE OF SYSTEMATIC ERROR	EXAMPLE OF STRATEGY TO PREVENT THE ERROR
1. Standardizing the measurement methods in an operations manual	Observer	Consistently high diastolic blood pressure (BP) readings due to using the point at which sounds become muffled	Specify the operational definition of diastolic BP as the point at which sounds cease to be heard
	Subject	Consistently high readings due to measuring BP right after walking upstairs to clinic	Specify that subject sit in quiet room for 5 minutes before measurement
2. Training and certifying the observer	Observer	Consistently high BP readings due to failure to follow procedures specified in operations manual	Trainer checks accuracy of observer's reading with a double-headed stethoscope
3. Refining the instrument	Instrument	Consistently high BP readings with standard cuff in subjects with very large arms	Use extra-wide BP cuff in obese patients
4. Automating the instrument	Observer	Conscious or unconscious tendency for observer to read BP lower in group randomized to active drug	Use automatic BP measuring device
	Subject	BP increase due to proximity of attractive technician	Use automatic BP measuring device
5. Making unobtrusive measurements	Subject	Tendency of subject to overestimate compliance with study drug	Measure study drug level in urine
6. Calibrating the instrument	Instrument	Consistently high BP readings due to manometer being out of adjustment	Calibrate each month
7. Blinding	Observer	Conscious or unconscious tendency for observer to read BP lower in active treatment group	Use double-blind placebo to conceal study group assignment
	Subject	Tendency of subject who knew she was on active drug to overreport side effects	Use double-blind placebo to conceal study group assignment

(standardizing and training) should always be used, calibration is needed for any instrument that has the potential to change over time, and blinding is essential whenever feasible.

VALIDITY

Validity resembles **accuracy**, but we like to think of it as adding a qualitative dimension to considering how well a measurement represents the phenomena of interest. For example, measurements of creatinine and cystatin C in the blood, two chemicals excreted by the kidneys,

might be equally *accurate* (e.g., within 1% of the true level), but cystatin C may be more *valid* as a measure of kidney function because creatinine levels are also influenced by the amount of muscle (7). In Figure 4.2, we can think of validity as describing whether the bull's-eye is in the right target.

Validity is often not amenable to assessment with a gold standard, particularly for measurements aimed at subjective and abstract phenomena such as pain or quality of life. Social scientists have created qualitative and quantitative constructs for addressing the validity of these measurement approaches.

- **Content validity** examines how well the measurement represents all aspects of the phenomena under study; for example, including questions on social, physical, emotional, and intellectual functioning to assess quality of life.
- **Face validity** describes whether the measurement seems inherently reasonable, such as measuring pain on a 10-point scale or social class by household income.
- **Construct validity** is the degree to which a specific measuring device agrees with a theoretical construct; for example, an IQ test should distinguish between people that theory or other measures suggest have different levels of intelligence.
- **Predictive validity** is the ability of the measurement to predict an outcome; for example, how well a questionnaire designed to assess depression predicts job loss or suicide.
- **Criterion-related validity** is the degree to which a new measurement correlates with well accepted existing measures.

The general approach to measuring subjective and abstract phenomena is to begin by searching the literature and consulting with experts in an effort to find a suitable **instrument** (typically a questionnaire) that has already been validated. Using such an instrument has the advantage of making the results of the new study comparable to earlier work in the area, and may simplify and strengthen the process of applying for grants and publishing the results. Its disadvantages, however, are that the validation process may have been suboptimal, and that an instrument taken off the shelf may be outmoded or not optimal for the research question.

If existing instruments are not suitable for the needs of the study, then the investigator may decide to develop a new measurement approach and validate it herself. This can be an interesting challenge and even lead to a worthwhile contribution to the literature, but it generally requires a lot of time and effort (Chapter 15). It is fair to say that the process is often less conclusive than the word "validation" connotes.

OTHER FEATURES OF MEASUREMENT APPROACHES

Measurements should be **sensitive** enough to detect differences in a characteristic that are important to the investigator. Just how much sensitivity is needed depends on the research question. For example, a study of whether a new medication helps people to quit smoking could use an outcome measure that is not very sensitive to the *number* of cigarettes smoked each day. On the other hand, if the question is the effect of reducing the nicotine content of cigarettes on the number of cigarettes smoked, the method should be sensitive to differences in daily habits of just a few cigarettes.

An ideal measurement is **specific**, representing only the characteristic of interest. The carbon monoxide level in expired air is a measure of smoking habits that is only moderately specific because it can also be affected by other exposures such as automobile exhaust. The specificity of assessing smoking habits can be increased by adding measurements (such as self-report and serum cotinine level) that are not affected by air pollution.

Measurements should be **appropriate** to the objectives of the study. A study of stress as an antecedent to myocardial infarction, for example, would need to consider which kind of stress (psychological or physical, acute or chronic) was of interest before setting out the operational definitions for measuring it.

Measurements should provide an adequate **distribution of responses** in the study sample. A measure of functional status is most useful if it produces values that range from high in some subjects to low in others. A major reason for pretesting is to ensure that the actual responses do not all cluster around one end of the possible range of responses (Chapter 17).

Whenever possible, measurements should be designed in a way that minimizes subjective judgments. **Objectivity** is achieved by reducing the involvement of the observer and by using automated instruments. One danger in these strategies, however, is the consequent tunnel vision that limits the scope of the observations and the ability to discover unanticipated phenomena. This can be addressed by including some open-ended questions, and an opportunity for acquiring subjective and qualitative data, in addition to the main objective and quantitative measurements.

In designing a study there is a tendency to keep adding items that are not central to the research question but *could* be of interest. It is true that additional measurements increase the likelihood of interesting findings, including some that were not anticipated at the outset. However, it is important to keep in mind the value of **efficiency** and **parsimony**. The full set of measurements should be designed to collect useful data at an affordable cost in time and money. Collecting too much information is a common error that can tire subjects, overwhelm the team making the measurements, and clutter data management and analysis. The result may be a more expensive study that paradoxically is less successful in answering the main research questions.

MEASUREMENTS ON STORED MATERIALS

Clinical research involves measurements on people that range across many domains. Some of these measurements can only be made during contact with the study subject, but many can be carried out later on biological **specimens** banked for chemical or genetic analysis, or on **images** from radiographic and other procedures filed electronically (Table 4.5).

One advantage of such storage is the opportunity to reduce the cost of the study by making measurements only on individuals who turn out during follow-up to have an outcome of interest. A terrific approach to doing this is the nested case–control design (Chapter 8), especially if paired blinded measurements can be made in a single analytic batch eliminating the batch-to-batch component of random error. This approach also has the advantage that scientific advances years after the study is begun may lead to new ideas and measurement techniques that can then be employed, funded by newly submitted grants.

TABLE 4.5 COMMON TYPES OF MEASUREMENTS THAT CAN BE MADE ON STORED MATERIALS

TYPE OF MEASUREMENT	EXAMPLES	BANK FOR LATER MEASUREMENT
Medical history	Diagnoses, medications, operations, symptoms, physical findings	Paper or electronic medical records
Psychosocial factors	Depression, family history	Voice recordings, videotapes
Anthropometric	Height, weight, body composition	Photographs
Biochemical measures	Serum cholesterol, plasma fibrinogen	Serum, plasma, urine, pathology specimens
Genetic/molecular tests	Single nucleotide polymorphisms	DNA
Imaging	Bone density, coronary calcium	X-rays, CT scans, MRIs
Electromechanical	Arrhythmia, congenital heart disease	Electrocardiogram, echocardiogram

The growing interest in **translational research** (Chapter 2) takes advantage of new measurements that have greatly expanded clinical research, for example, in the areas of **genetic and molecular epidemiology** (8, 9) and **imaging**. Measurements on specimens that contain DNA (e.g., saliva and blood) can provide information on genotypes that contribute to the occurrence of disease or modify a patient's response to treatment. Measurements on serum can be used to study molecular causes or consequences of disease; for example, inflammatory markers provide useful information in the pathophysiology of many diseases. It is important to consult with experts regarding the proper collection tubes and storage conditions in order to preserve the quality of the specimens and make them available for the widest spectrum of subsequent use. It is also important to obtain informed consent from participants that covers the scope of potential uses of the specimens.

SUMMARY

1. Variables are either **numerical** or **categorical**. Numerical variables are **continuous** (quantified on an infinite scale) or **discrete** (quantified on a finite scale such as integers); categorical variables are **nominal** (unordered) or **ordinal** (ordered), and those that have only two categories are termed **dichotomous**.
2. Variables that contain more **information** provide greater power and/or allow smaller sample sizes, according to the following **hierarchy**: continuous variables > discrete numeric variables > ordinal variables > nominal and dichotomous variables.
3. The **precision** of a measurement (i.e., the **reproducibility** of replicate measures) is another major determinant of power and sample size. Precision is reduced by **random error (chance)** from three **sources of variability**: the observer, the subject, and the instrument.
4. Strategies for **increasing precision** that should be part of every study are to **operationally define** and **standardize methods** in an **operations manual**. Other strategies that are often useful are **training and certifying observers**, **refining** and **automating** the **instruments**, and **repetition**—using the mean of repeated measurements.
5. The **accuracy** of a measurement is the degree to which it approximates a gold standard. Accuracy is reduced by **systematic error** (**bias**) from the same three sources: the observer, subject, and instrument.
6. The **strategies for increasing accuracy** include all those listed for precision with the exception of repetition. In addition, accuracy is enhanced by **unobtrusive measures**, by **calibration**, and (in comparisons between groups) by **blinding**.
7. **Validity** is the degree to which a measurement represents the phenomena it is intended to measure; it is commonly used for more abstract and subjective variables, and is assessed by **content validity**, **face validity**, **construct validity**, **predictive validity**, and **criterion-related validity**.
8. Individual measurements should be **sensitive**, **specific**, **appropriate**, and **objective**, and they should produce a **range of values**. In the aggregate, they should be **broad** but **parsimonious**, serving the research question at moderate cost in time and money.
9. Investigators should consider **storing images** and other **materials** for later measurements that can take advantage of **new technologies** as they are developed and the efficiency of **nested case–control** designs.

APPENDIX 4

OPERATIONAL DEFINITION OF A MEASUREMENT OF GRIP STRENGTH

The **operations manual** describes the method for conducting and recording the results of all the measurements made in the study. This example, from the operations manual of the Study of Osteoporotic Fractures, describes the use of a dynamometer to measure grip strength. To standardize instructions from examiner to examiner and from subject to subject, the protocol includes a script of instructions to be read to the participant verbatim.

PROTOCOL FOR MEASURING GRIP STRENGTH WITH THE DYNAMOMETER

Grip strength will be measured in both hands. The handle should be adjusted so that the participant holds the dynamometer comfortably. Place the dynamometer in the right hand with the dial facing the palm. The participant's arm should be flexed 90° at the elbow with the forearm parallel to the floor.

1. Demonstrate the test to the subject. While demonstrating, use the following description: "This device measures your arm and upper body strength. We will measure your grip strength in both arms. I will demonstrate how it is done. Bend your elbow at a 90° angle, with your forearm parallel to the floor. Don't let your arm touch the side of your body. Lower the device and squeeze as hard as you can while I count to three. Once your arm is fully extended, you can loosen your grip."
2. Allow one practice trial for each arm, starting with the right if she is right handed. On the second trial, record the kilograms of force from the dial to the nearest 0.5 kg.
3. Reset the dial. Repeat the procedure for the other arm.

The arm should not contact the body. The gripping action should be a slow, sustained squeeze rather than an explosive jerk.

REFERENCES

1. Michaelsson K, Baron JA, Snellman G, et al. Plasma vitamin D and mortality in older men: a community-based prospective cohort study. *Am J Clin Nutr* 2010;92:841–848.
2. Ware JE, Gandek B Jr. Overview of the SF-36 health survey and the International Quality of Life Assessment Project. *J Clin Epidemiol* 1998;51:903–912.
3. Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ* 1996;313:41–42; also, Measurement error proportional to the mean. *BMJ* 1996;313:106.
4. Newman TB, Kohn M. Evidence-based diagnosis. New York: Cambridge University Press, 2009.
5. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37– 46.
6. Filion K, Kukanich KS, Chapman B, et al. Observation-based evaluation of hand hygiene practices and the effects of an intervention at a public hospital cafeteria. *Am J Infect Control* 2011;39:464–470.
7. Peralta CA, Shlipak MG, Judd S, et al. Detection of chronic kidney disease with creatinine, cystatin C, and urine albumin-to-creatinine ratio and association with progression to end-stage renal disease and mortality. *JAMA* 2011;305:1545–1552.
8. Guttmacher AE, Collins FS. Genomic medicine: a primer. *NEJM* 2002;347:1512–1520.
9. Healy DG. Case–control studies in the genomic era: a clinician's guide. *The Lancet Neurology* 2006;5:701–707.