BRIEFING

<1032> Design and Development of Biological Assays. This General Chapter *Design and Development of Bioassays* <1032> is one of an integrated group of new General Chapters that provide guidance across several complementary bioassay topics. The others include *Biological Assay Validation* <1033>; *Analysis of Biological Assay* <1034>; and an as-yet unnumbered "roadmap" chapter that will include a glossary applicable to General Chapters <1032>–<1034>. The group arose from one source, General Chapter *Design and Analysis of Biological Assays* <111>, that is now official in the *US Pharmacopeia* (*USP*). The partitioning of chapters to different aspects of biological assays allows both greater focus and clarity and the opportunity to expand on issues. As the group of General Chapters evolves, General Chapter <111> will remain in modified form. Thus the entire group when completed will consist of five *USP* General Chapters.

The Bioassay Development ad hoc Advisory Panel of the Statistics Expert Committee encourages input from all interested parties regarding the proposed <1032>. USP's intent is to reflect the best contemporary practices regarding the design and development of bioassays. This will be achieved when members of the bioassay community take advantage of this opportunity to engage in <1032>'s development by responding to this In-Process Revision. Comments regarding <1032> should be sent to Tina S. Morris, PhD (tsm@usp.org).

<1032> DESIGN AND DEVELOPMENT OF BIOLOGICAL ASSAYS

1. INTRODUCTION

1.1 Purpose

General Chapter *Design and Development of Biological Assays* <1032> presents methodology for the development of bioassay procedures that have sound experimental design, that provide data that can be analyzed using well-founded statistical principles, and that are fit for their specific use.

General Chapter <1032> is one of a group of five planned General Chapters that focus on relative potency assays, in which the activity of a Test material is quantified by comparison to the activity of a Standard material. However, many of the principles can be applied to other assay systems.

This General Chapter is intended to guide the design and development of a bioassay for a drug product intended for commercial distribution. Although adoption of this chapter's recommended methods may be resource intensive during initial drug development, early implementation may yield benefits.

1.2 Audience

This chapter is intended for both the practicing bioassay analyst and for the statistician who is engaged in bioassay efforts. The former will find guidance for implementing bioassay structure and methodology to achieve analytical goals while reliably demonstrating the biological activity of interest, and the latter will gain insights regarding the constraints of biology that can prove challenging to balance with the most straightforward practice of statistics.

Focus on relative potency. Common applications of biological assays include characterization of a biological or biotechnology product, stability testing, and lot release. Potency assays help provide assurance of the quality and consistency of the product. Because of the inherent variability in test systems (from animals, cells, instruments, and reagents, and day-to-day and between-lab variation), an absolute measure of potency (e.g., a dose of erythropoietin that increases hemoglobin content by 1 g) may not be available. This has led to the adoption of the relative potency methodology. Assuming that the Standard and Test materials contain samples with biologically similar (if not identical) activity, the Test sample can be expected to behave like a concentration or dilution of the Standard, and *parallelism* (the quality of Test and Standard concentration–response segments being parallel) should be present. For samples that are biologically similar, the horizontal displacement between the curves is interpreted as log relative potency. Relative potency is determined by comparison of Test to Standard assay response, which means that the assay does not need to achieve a specific observable response. The potency of the Standard is usually assigned a value of 1 (or 100%). The Standard could be a material established as such by a national (e.g., USP) or international (e.g., WHO) organization, or it could be an internal Standard.

2. BIOASSAY FITNESS FOR USE

To evaluate whether an assay is fit for use, analysts must clearly specify the purposes for performing the bioassay. Common uses for a bioassay include lot release of drug substance (active pharmaceutical ingredient) and drug product; stability; qualification of Standard sample and other critical reagents; characterization of process intermediates and formulations, contaminants, and degradation products; and support of changes in the product production process. The accuracy, specificity, precision, and robustness requirements may be different for each of these potential uses. It is a good strategy to develop and validate a bioassay to support multiple intended uses. Decisions about fitness for use are based on scientific and statistical considerations, as well as practical considerations such as cost, turnaround time, and throughput requirements for the assay.

When assays are used for lot release, a linear-model bioassay may allow sufficient assessment of similarity. For bioassays used to support stability or comparability, to evaluate changes in production processes, or to qualify reference materials, critical reagents, or changes in the assay process, it may be useful to assess similarity using the asymptote of maximum response.

2.1 Process Development

Bioassays are often required in the development and optimization of product manufacturing, including formulation and scale-up processes. Bioassays can be used to evaluate purification strategies, optimize product yield, and measure product stability. Because samples taken throughout the process are often analyzed and compared, sample matrix effects should be carefully studied to determine an assay's fitness for use. For relative potency measures, the Standard material may require dilution into suitable matrices for quantitation. The bioassay's precision should be suitable for measuring process performance or for assessing and comparing the stability of candidate formulations.

2.2 Process Characterization

In order to assess an effect on drug potency specific to different stages of drug manufacture or following changes in the manufacturing process (i.e., to demonstrate product equivalence before and after process changes are made), analysts may perform bioassays distinct from routine product lot-release assays. These assays can provide information about product characteristics in the presence of process changes. Bioassays used in this type of application may be qualitative or quantitative. The assay's fitness for use depends on its sensitivity to unexpected changes in the product profile.

2.3 Product Release

Bioassays are used to evaluate the potency of the drug before product release and may predict clinical efficacy. To this end, the assay should reflect or mimic the product's known or intended mechanism of action. If the bioassay does not include the functional biology directly associated with the mechanism of action, it may be necessary to demonstrate a relationship between this bioassay's estimated potency results and those of some other assay that better or otherwise reflects biomimetic activity.

For product-release testing, a release specification is established to define the range of potency values that are acceptable for product stored under labeled conditions until the expiration date. The relative accuracy and the precision of the reportable value from the bioassay must support the number of significant digits listed in the specification (see *Biological Assay Validation* <1033>), as well as the specification range. These specifications should include the loss of potency (instability) that will be tolerated. In order to meet these specifications, manufacturing quality control will have sufficiently narrow release assay specifications in order to accommodate any loss of activity due to instability and uncertainly in the release assay. The expiration period may be calculated as the point at which the 95% confidence bound of the predicted value crosses the product specification limit. The more variable the assay and the steeper the slope of decay, the nearer in time the assigned expiration date will be.

2.4 Process Intermediates

Bioassay assessment of process intermediates can provide information regarding interference and specificity. Downstream process decisions pertaining to formulation and fill strategies may rely on bioassays in order to ensure that product will meet its established specifications. For example, unformulated bulk materials may be held and evaluated for potency. Bulks may be pooled with other bulk lots, diluted, or reworked based on the potency results. For these types of applications, the bioassay must be capable of measuring product activity in different matrices. In some cases, a separate Standard material is made and is used to calculate relative potency for the process intermediate. Thus fitness for use must demonstrate that decisions based on the evaluation of process intermediates do in fact yield the desired outcomes for downstream product, including final container.

2.5 Stability

Potency is a stability-indicating parameter of biotechnology and vaccine products. Stability studies are performed during development to establish shelf life as well as to identify and estimate the degradation products and rates of the product. Post licensure stability studies are used to monitor and maintain product stability. Stability studies are designed by selecting the number of observations and the testing intervals, as well as the potency assay format that is predicted to achieve a suitable level of uncertainty in the stability parameter of interest. Knowledge of both short-term and long-term variability of the bioassay is important for a product's stability evaluation. In addition, a trend in bias across levels in the bioassay can influence the measurement of potency over time. A limit on the trend in bias can help ensure accurate assessment of product stability.

2.6 Qualification of Reagents

The quantitative characterization of a new Standard material requires a highly precise measurement of the new material's biological activity. This value is used either to establish that the new lot is equivalent to the previous lot or to assign it a label potency to which Test samples can be compared. Replication beyond routine testing may be required to achieve a greater confidence in measuring the potency of the Standard material. In some cases, such as maintenance of the cell seed train, the bioassay may be used to qualify a reagent such as fetal bovine serum. The fitness for use in such cases is tied to the ability of the assay to screen reagent lots and to ensure that lots that may degrade the relative potency measurements are not accepted.

2.7 Product Integrity

Biotechnology, biological, and vaccine products may contain a population of heterogeneous material, including the intended predominant product material. Additionally, the product is subject to variation in manufacture, as well as degradation because of stress and storage conditions. Some process impurities and degradation products may be active, partially active, or inactive in the bioassay. Biological assays

can be used to differentiate intended product from product variants or derivatives in which changes in structure or relative composition may be manifested in subtle yet characteristic changes in the bioassay response (e.g., change in slope or asymptote). Studies that identify characteristic changes associated with variants of the intended product help ensure consistent product performance. Whenever possible, the bioassay should be accompanied by orthogonal methods that are capable of separating and detecting product variants, process impurities, and degradation products.

3. BIOASSAY FUNDAMENTALS

3.1 In Vivo Bioassays

In vivo potency assays are bioassays in which a set of dilutions of each of the Standard and Test materials is administered to animals and the dose–response relationships are used to estimate potency. For some animal assays, the end point is simple (e.g., rat body weight gain assay for human growth hormone or rat ovarian weight assay for follicle stimulating hormone), but others require further processing of samples collected from treated animals (e.g., reticulocyte count for erythropoeitin, steroidogenesis for gonadotropins, neutrophil count for granulocyte colony stimulating factor, or antibody titer after administration of vaccines). With the advent of cell lines specific for the putative physiological mechanism of action (MOA), the use of animals for the measurement of potency has greatly diminished. Cost, low throughput, ethics, and other practical issues argue against the use of animal bioassays. Regulatory agencies have encouraged the responsible limitation of animal use whenever possible (see The Interagency Coordinating Committee on the Validation of Alternative Methods, Mission, Vision, and Strategic Priorities; February 2004). When in vitro activity is not strongly associated with in vivo activity (e.g., EPO), the combination of in vitro cell-based assay and a suitable physicochemical method (e.g., IEF, glycan analysis) may substitute for in vivo assays. However, a need for in vivo assays may remain when in vitro assays cannot detect differences that may affect in vivo activity.

Animals' physiological responses to biological drugs (including vaccines) may predict patients' responses. Selection of animal test subjects by species, strain, gender, and maturity or weight range seeks to provide a representative model with which to assess the activity of Test materials. In order to minimize assay variability, bioassays involve comparison of a Test to a Standard material of defined potency, together with the application of statistical tools, specific lab techniques, and rigorous adherence to protocol.

Some methods lend themselves to the use of colony versus naïve animals. For example, pyrogen and insulin testing benefit from using experienced colony rabbits that provide a reliable response capacity. If animals recently introduced to the colony fail to respond normally to the Test compound after several doses, they should be culled from the colony so they do not cause repeated future invalid or indeterminate assay results. In the case of assaying highly antigenic Test compounds for pyrogens, however, naïve animals should be used to avoid generating inaccurate or confounded results from

given samples. Other colony advantages include common controlled environmental conditions (macro/room, and micro/rack), consistent feeding schedule, provision of water, and husbandry routine.

When a historical database of colony records and assay data is available, valuable method optimization can be obtained. The influence of biasing factors can be reduced by applying randomization principles such as distribution of weight ranges across dose groups, group assignments from shipping containers to different cages, or use of computer-generated or deck patterns for injection/dosing. A test animal must be healthy to be suitable for use in a bioassay and must have time to stabilize in its environment. Factors that combine to influence an animal's state of health include proper nutrition, hydration, freedom from physical and psychological stressors, adequate sanitization of housing components, controlled light cycle (diurnal/nocturnal), experienced handling, skillful injections and bleedings, and absence of noise or vibration. Daily observation of test animals is essential, and veterinary care must be available to minimize stress to the animals and to evaluate potential issues that might challenge the validity of bioassay results.

3.2 Ex Vivo Bioassays

Cells or tissues from human or animal donors can be cultured in the laboratory and used to assess the activity of a Test article. In the case of cytokines, the majority of assays use cells from the hematopoietic system and subsets of hematopoietic cells from peripheral blood such as peripheral blood mononuclear cells or peripheral blood lymphocytes. For proteins that act on solid tissues, such as growth factors and hormones, specific tissue on which they act can be removed from animals, dissociated, and cultured for a limited period either as adherent or semi-adherent cells. Although an ex vivo assay system has the advantage of similarity to the natural milieu, it may also suffer from substantial donor-to-donor variability, as well as challenging availability of appropriate cells.

Bioassays that involve live tissues or cells from an animal (e.g., rat hepatocyte glucagon method) require process management similar to that of in vivo assays to minimize assay variability and bias. The level of effort to manage bias (e.g., randomization) should be appropriate for the purpose of the assay. Additional factors that may affect assay results include time of day, weight or maturity of animal, anesthetic used, buffer components/reagents, incubation bath temperature and position, and cell viability.

3.3 In Vitro (Cell-based) Bioassays

Cell-based bioassays using clonal cell lines that respond to specific ligands or infectious agents can be used for lot-release assays of therapeutic proteins and vaccines. These cell lines can be derived from tumors, immortalized as factor-dependent cell lines, or engineered cell lines transfected with appropriate receptors. Such cells can be banked and used with an expectation of uniform response through some number of continuous passages. The selection of appropriate cell lines with desirable magnitude of response

and stability may be challenging. Advances in recombinant DNA technology and the understanding of cellular signaling mechanisms have allowed the generation of engineered cell lines with improved response and longer stability. The cellular responses to the protein of interest depend on the drug's MOA and the duration of exposure. Such responses include cell proliferation, cell killing, antiviral activity, differentiation, cytokine/mediator secretion, and enzyme activation. Assays involving these responses may require incubation of the cells over several days, during which time contamination, uneven evaporation, or edge effects may arise. Comparatively rapid responses based on an intracellular signaling mechanism—such as second messengers, protein kinase activation, or reporter gene expression—have proven acceptable to regulatory authorities. Lastly, most cell lines used for bioassays express receptors for multiple cytokines and growth factors. This lack of specificity may not be detrimental if other assays demonstrate the Test material's specificity.

Cell-based bioassay design should reflect knowledge of the factors that influence response of cells to the Test agent. Response variability is often reflected in parameters such as slope, EC_{50} of the dose–response curve, or the ratio of maximum to minimum response. Even though the adoption of relative potency methodology alleviates the effect of these parameters on potency estimates, such response variability could result in significant effects on system suitability, precision, and accuracy. Therefore sources of variability relevant to a given bioassay method should be identified, and control strategies to minimize their effects should be implemented.

The development of a cell-based bioassay begins with the selection or generation of a cell substrate. The cell substrate is usually the most critical reagent in a cell-based assay. To ensure an adequate and consistent supply of cells for product testing, the cell line(s) used in the assay should be banked. Several considerations arise in developing the assay cell banks.

The source of the cell line, whether generated by the product manufacturer or acquired from a collaborator, academic institution, or culture collection, must have documentation that details the cell line's history from origin to banking and supports its application for commercial use. The origin, generation, and propagation of the cell line through to the point at which the cell line is selected for use must be documented and described in detail sufficient to permit recreation of a similar cell line if necessary. Information pertaining to the cell line should be documented during assay development and before banking. This information may include but is not limited to identity (e.g., isoenzyme, phenotypic markers, genetic analysis); morphology (e.g., archived photographic images); purity (e.g., mycoplasma, bacteria, fungus and virus testing); cryopreservation; thaw and culture conditions (e.g., media components, thaw temperature and method, methods of propagation, seeding densities, harvest conditions); thaw viability (immediately after being frozen and after time in storage); growth characteristics (e.g., cell doubling times); and functional stability (e.g., ploidy).

Extensive characterization is necessary to ensure the quality and longevity of cell banks for use in the QC environment. The general health and metabolic state of the cells at

the time of bioassay can greatly influence the test results. After a cell line has been characterized and is ready for banking, analysts typically prepare a two-tiered bank (Master and Working). A Master Cell Bank is created as the source layer for the Working Cell Bank. The Working Cell Bank is derived by expansion of one or more vials of the Master Cell Bank. The size of the banks depends on the growth characteristics of the cells, the number of cells required for each assay, and how often the assay will be performed. Some cells may be sensitive to cryopreservation, thawing, and culture conditions, and the banks must be carefully prepared and characterized before being used for validation studies and for regular use in the QC laboratory.

The following factors that may affect bioassay performance are common to many cell-based bioassays. This list is not exhaustive, and analysts with comprehensive understanding and experience with the cell line should be involved during assay development. These experienced individuals should identify factors that might influence assay outcomes and establish strategies for an appropriate level of control whenever possible. Relevant factors include cell type (adherent or nonadherent); cell thawing; plating density (at thaw and during seed train maintenance) and confluence (adherent cells); culture vessels; growth, staging, and assay media; serum requirements (source, heat inactivation, gamma irradiation); incubation conditions (temperature, CO_2, humidity, culture times from thaw); cell harvesting reagents and techniques (for adherent cells, method of dissociation); cell sorting; cell counting; determination of cell health (growth rate, viability, yield); cell passage number and passaging schedule; cell line stability (genetic, receptor, marker, gene expression level); and starvation or stimulation steps.

3.4 Standard

The Standard is a critical reagent in bioassays because of its quality as a reliable material to which a Test preparation can be quantitatively compared in a relative potency assay. The Standard may be assigned a unitage or specific activity that represents fully (100%) potent material. Where possible, a Standard should be prepared using the same manufacturing process as the drug substance. If this is not possible (e.g., lyophilization is the preferred Standard storage), verify the capacity of the assay for providing accurate relative potency assessment of samples. Testing performed to qualify a Standard may be more rigorous than the routine testing used for lot release.

A Standard must be stored under conditions that preserve its full potency during the intended course of its use. To this end, the Standard may be stored under conditions that are different from the normal storage of the drug substance or drug product. These could include a different temperature (e.g., –70 °C or –20 °C instead of 2–8 °C), a different container (e.g., plastic vials instead of syringes), a different formulation (e.g., lyophilizable formulation or addition of carrier proteins such as human serum albumin, stabilizers, etc.). The Standard material should be tested for stability at appropriate intervals. System suitability criteria of the bioassay such as maximum or background response, EC_{50}, slope, or potency of assay control may be used to detect change in the

activity of the Standard. Accelerated stability studies can be performed to predict degradation rates and establish meaningful markers of Standard instability.

At later stages in clinical development, the Standard may be prepared using the manufacturing process employed in pivotal clinical trials. If the Standard formulation is different that that used in the drug product process, demonstration is required of comparable (similar) assay performance. An initial Standard may be referred to as the *primary Standard*. Subsequent Standards can be prepared using current manufacturing processes and can be designated *working Standards*. Separate SOPs are usually required for establishing these standards for each product. Bias in potency measurements sometimes can arise when the activity of the Standard gradually changes. Nonequivalence in assay response may be observed if the Standard and the Test sample are not biologically similar (e.g., if there are differences in glycosylation). In such cases, careful attention should be paid to changes in critical reagents and/or media components that differentially affect the activities of the Standard and Test sample. It is prudent to archive aliquots of each Standard lot for assessment of comparability with later Standards and for the investigation of assay drift. In this context, trend charts may be useful in identifying the cause of assay drift.

4. STATISTICAL ASPECTS OF BIOASSAY FUNDAMENTALS

4.1 Data

The statistical elements of bioassay development include consideration of data type and the bioassay model, along with statistical criteria for assessing and ensuring the quality of bioassay results. These form the framework for the bioassay system that will be used to estimate the potency of a Test article.

There are fundamentally two bioassay data types: quantitative and quantal (categorical). Quantitative data can be either continuous (e.g., collected from an instrument), counts (e.g., plaque-forming units), or discrete (e.g., endpoint dilution titers). Quantal data are typically dichotomous, e.g., survival in an animal model that uses challenge with a pathogen to measure the protection afforded by a Test article or positivity in a plate-based infectivity assay that results in destruction of a cell monolayer following administration of an infectious agent. Quantitative data can be transformed to quantal data by selecting a threshold that statistically distinguishes a positive response from a negative response. Such a threshold can be calculated from data acquired from a negative control, e.g., adding (or subtracting) a measure of uncertainty, such as two times the standard deviation of negative control responses, to the negative control average. Analysts should be cautious about transforming quantitative data to quantal data because this results in a loss of information that may affect bioassay measurement.

4.2 Assumptions

The key assumption for the analysis of most bioassays is that the Standard and Test samples contain the same effective analyte or population of analytes and thus may be expected to behave similarly in the bioassay. This is termed *similarity*. As will be shown in more detail in Chapter <1034> for specific statistical models, biological similarity implies that statistical similarity is present (the Standard and Test curves are parallel, for parallel-line and parallel-curve models; the Standard and Test curves have a common intercept for slope-ratio models). The reverse is not true. Statistical similarity (parallel lines, parallel curves, or common intercept, as appropriate) implies, but does not ensure, biological similarity (same effective analyte). However, failure to satisfy statistical similarity may be taken as evidence against biological similarity. The existence of a Standard–Test sample pair that passes the assessment of statistical similarity is thus a necessary but not sufficient test that the key assumption of biological similarity is satisfied. Biological similarity thus remains, unavoidably, an assumption. Departures from statistical similarity that are consistent in value across many assays could be indicative of matrix effects or of real differences between Test and Standard materials. This is true even if the departure from statistical similarity is sufficiently small to support determination of a relative potency.

In many assays multiple compounds will yield similar concentration–response curves. It may be reasonable to use a biological assay system to describe or even compare response curves from different compounds. But it is not appropriate to report relative potency unless the Standard and Test samples contain only the same active analyte or population of analytes. Biological products typically exhibit lot-to-lot variation in the distribution of analytes (i.e., most biological products contain an intended product and, at suitably low levels, some process contaminants that may be active in the bioassay). Assessment of similarity is then, at least partially, an assessment of whether the distribution of analytes in the Test sample is close enough to that of the distribution in the Standard sample for relative potency to be meaningful; i.e., the assay is a comparison of like to like. When there is evidence (from methods unrelated to the bioassay) that the Standard and Test samples do not contain the same compound(s), the biological assumption of similarity is not satisfied, and it is not appropriate to report relative potency.

Other common statistical assumptions in bioassay practice are constant variance of the responses around the fitted model (see Section 4.3 for further discussion), normally distributed residuals (see Section 4.4 for further discussion), and independence of the observations. Constant variance, normality, and independence are interrelated in the practice of bioassay. For bioassays with a quantitative response, a well-chosen transformation may be used to achieve near constant variance and a nearly normal distribution of residuals. For such assays, once transformation of data has been imposed, the remaining assumption, independence, then remains to be addressed with proper modeling.

Independence of measurements in bioassay is important for effective design and appropriate modeling of data. Two events are independent if they have no common factor that could cause them to vary together (co-vary). Independence can be assessed

with statistical tools, including variance component estimates and estimates of correlation among observations. A proper understanding requires knowledge of assay design and the analytical model so that experimental units are accurately identified. The statistical model should capture the design structure imposed by the presence of blocks or grouping of responses (e.g., due to serial dilution or the use of a multichannel pipette)

4.3 Variance Heterogeneity, Weighting, and Transformation

Analysis of bioassay data requires that the data be approximately normally distributed with constant variance across the range of the data. For linear and nonlinear regression models, the variance referred to here is the residual variance from the fit of the model. Constant variance is often not observed, e.g., an increase in variability with increase in response (variance heterogeneity). If the variances are not equal but the data are analyzed as if they were, then the estimate of relative potency may be reasonable. However, failure to address nonconstant variance around the fitted dose–response model results in an unreliable estimate of within-assay variance. Further, the assessment of statistical similarity may not be accurate, and standard errors and confidence intervals for all parameters (including a Fieller's Theorem–based interval for the relative potency) should not be used. Confidence intervals for relative potency that combine multiple assays will still be correct.

Constant variance can be assessed via residual plots, Box-Cox (or power law) analysis, or Bartlett's or Levene's test. With either of the tests, improvement of the value of the statistic obtained is useful as a basis for judging whether homogeneity is improved or worsened; do not rely on the *P* value. Assess variance on a large body of assay data. Using only the replication variances from the current assay is not appropriate because there are too few data to properly determine truly representative variances specific to each concentration. This approach is necessary during development, and it is prudent to re-assess variance during validation and to monitor it periodically during ongoing use of the assay.

Two methods used to mitigate variance heterogeneity are weighting and transformation. Lack of constant variance can be addressed with a suitable transformation chosen for the assay system (not on an assay-by-assay basis). If no suitable transformation is identified, weighting can be explored. Weighting is a statistical method of emphasizing precise data instead of imprecise data. Transformation is used to address nonconstant variation in the response and normality of residuals and to improve the fit of a statistical model to the data.

Variance may be proportional to concentration or may be some function of concentration. This possibility can be examined by plotting the sample variance at each concentration (preferably pooled across multiple assays) against concentration and then against a function of concentration (e.g., concentration squared). Variance will be proportional to the function of concentration where the plot follows approximately a straight line. If the data fall along a horizontal line, no weighting is needed. If such a function can be found, then the weights are taken as the reciprocal of that function.

There may be no such function, or at least not one that is easy to describe, particularly if the variation is higher at both extremes of the concentration range studied. An alternative is to develop an estimate of the variability at each concentration using historical data for the assay as currently conducted. Using only the replication variances from the current assay is not appropriate because there are too few data to properly determine truly representative variances specific to each concentration. Whether a model or historical data are used, the goal is to capture the relative variability at each concentration. Even with use of historical data, one does not necessarily have to assume that the absolute level of variability of the current assay is identical to that of the historical data but only that the ratios of variances among concentrations are consistent.

Alternatively, an approach using a power of the mean (POM) response at each concentration has been advocated for many bioassay systems.

Additionally, transformation of the bioassay data may be used to help satisfy the requirements of normality and variance homogeneity. Bioassay data are commonly displayed with concentration on a log scale (in parallel-line or parallel-curve analyses). Slope-ratio assays are displayed with concentration on the original scale. Transformation may be performed to the response data as well as to the concentration data. Common choices for a transformation of the response include log, square root (for counts), reciprocal, and, for count data with known asymptotes, logit of the percent of maximum response. Log transformations are commonly used, partly because of the ease of transforming back to the original scale for interpretation.

A log–log fit may be performed on data exhibiting nonlinear behavior. The log transformation is used to linearize a portion of the concentration–response relationship, but such a transformation often also normalizes the data, generating responses that are more uniform in their variance. Other transformations may be performed in accord with the POM principle cited for weighting; i.e., data may be transformed by the inverse of the POM function. A POM coefficient $k = 2$ corresponds to a log transformation of the data. Note that transformation of the data requires re-evaluation of the model used to fit the data. If nontransformed data are expected to follow a four-parameter logistic model, another nonlinear model might be considered to fit the transformed measurements.

In many bioassays a suitable transformation (often the logarithm) yields data that have nearly constant variance, near normal distribution of the residuals, and a good fit to a four-parameter logistic curve. From a statistical perspective there is nothing special about the original scale of measurement. Any transformation that improves accordance with assumptions is acceptable. Analysts should recognize, however, that transformations, choice of statistical model, and choice of weighting scheme are interrelated. If a transformation is used, that may affect the choice of model. That is, transforming the response, e.g., by a log or square root, may change the shape of the response curve, as may the range of concentrations for which the responses are nearly straight and nearly parallel.

Appropriate training and experience in statistical methods are essential in determining an appropriate variance-modeling strategy. Sources of variability may be misidentified if the wrong variance model is used. For example, data can have constant variation throughout a four-parameter logistic concentration–response curve but can also have appreciable variation in the EC_{50} parameter from block to block within the assay, or from assay to assay. This assay can appear to have large variation in the response for concentrations near the long-term average value of ED_{50}, if the between-block or between-assay variability is not recognized. A weighted model with low weights for concentrations near the EC_{50} would misrepresent a major feature of such an assay system.

4.4 Normality

Many statistical methods for the analysis of quantitative responses assume normality of the residuals. If the normality assumption is not met, the estimate of relative potency and its standard error may be reasonable, but the commonly calculated confidence interval for the relative potency estimate will not be valid. Most methods used in this chapter are reasonably robust to departures from normality, so the goal is to detect substantial nonnormality. During assay development, in order to discover substantial departure from normality, one can use graphical tools such as a normal probability plot or a histogram (or something similar like stem-and-leaf or box plots) of the residuals from the model fit. The histogram should appear unimodal and symmetric. The normal probability plot should approximately follow a straight line. A pattern about a straight line is one indication of nonnormality. Nonnormal behavior may be due to measurements that are log normal and show greater variability at higher levels of response. This is characterized by a concave pattern in the residuals in a normal plot.

Statistical tests of normality are not generally useful. If such tests are used, as for homogeneity of variance, improvement of the value of the statistic obtained is useful for judging whether normality is improved or worsened. Again, reliance should not be placed on the *P*-value. Evaluate normality on as large a body of assay data as possible during development, re-assess during validation, and monitor periodically during ongoing use of the assay. Important departures from normality can often be mitigated with a suitable transformation. Failure to assess and mitigate important departure from normality carries the risks of disabling appropriate outlier detection and losing capacity to obtain reliable estimates of variation. This assessment typically is performed during assay development and is not routinely conducted with each assay.

4.5 Linearity of Concentration–Response Data

Some bioassay analyses assume that the shape of the concentration–response curve is a straight line or approximates a straight line over a limited range of concentrations. In those cases, a linear-response model may be assessed to determine if it is justified for the data in hand. Difference testing methods for assessing linearity face the same problems as do standard methods for difference testing of parallelism. More data and better precision make it more likely to detect nonlinearity. Because situations when

nonlinearity does not affect potency generally are rare and occur only by chance, analysts should routinely assess departure from linearity if they wish to use a linear-response model to estimate potency.

If an examination of a data plot reveals gross departure from linearity, this is sufficient to support a conclusion that linearity is not present. High data variability, however, may mask departures from linearity. Thus a general approach for linearity can conform to that for similarity, developed in detail in Section 4.7, *Implementing equivalence testing for similarity (parallelism)*:

(1) Specify a measure of departure from nonlinearity. Possibilities are the nonlinearity sum of squares, assuming similar slopes, or the quadratic coefficients from a fitted quadratic model or from contrast estimates from a means model.
(2) Use one of the four approaches of Section (2) of *Implementing equivalence testing for similarity* (*parallelism*) to determine a range of acceptable values (acceptance interval) for the measure of nonlinearity. A common alternative is to determine the acceptable range by visual examination. Because departures from nonlinearity are typically evident visually, the acceptable range consists of those values of the measure of nonlinearity for which the nonlinearity is not evident.
(3) Determine a 90% two-sided confidence interval, following the Two One-Sided Test (TOST) procedure, and compare the result to the acceptance interval determined in Step 2.

Selecting a linear range. Often a subset of the concentrations measured in the assay will be selected in order to derive a linear concentration–response curve. The subset can often be identified graphically. If, in the final assay, the intent is to use only concentrations in the linear range, choose a range that will result in parallel straight lines for the range of relative potencies expected during routine use of the assay Otherwise, the assay will fail parallelism tests when the problem is actually that the potency produces assay response values outside the linear range, and repeat assays will be required. The repeat assays together with the valid assays may generate a biased estimate of potency because of the selective process of repeating assays when the response is in the extremes of the concentration–response curve.

The problem is more complex in assays where there is even modest variation in the shape or location of the concentration–response curve from run to run or from block to block within an assay. In such assays, which are quite common, it is appropriate to choose subsets for each sample in each assay or even in each block within an assay. Note that a fixed-effects model will mask any need for different subsets in different blocks, but a mixed-effects model may reveal and accommodate different subsets in different blocks (see Section 4.9).

Additional specific guidance about selection of data subset(s) for linear model estimation of relative potency includes the following: Use at least three and preferably

four adjacent concentrations; require that the slope of the linear segment is sufficiently steep; and require that the lines fit to Standard and Test samples are straight and that the lines are parallel. One way to derive a steepness criterion is to compute a *t*-statistic on the slope. If the slope is not significant the bioassay will have poor performance. Another aspect that supports requiring adequate steepness of slope is the use of subset selection algorithms. Without a suitable slope steepness criterion, a subset selection algorithm that seeks to identify subsets of three contiguous data points that are straight and parallel might select doses on an asymptote. These are obviously poor subsets to use to estimate potency. Requiring a slope that has a small *P*-value is a simple and appropriate way to steer the subset selection algorithm to more appropriate doses. How steep or how significant the steepness of the slope should be depends on the assay. This criterion should be set during assay development and possibly refined during assay validation.

4.6 Common Bioassay Models

Most bioassays consist of a series of concentrations or dilutions of both a Test article and a Standard material. A mathematical model is fit to the dose–response data, and a relative potency is then calculated from the parameters of the model. Choice of model may depend on whether quantitative or qualitative data are being analyzed.

For quantitative data, parallel-profile models are preferred because of their superior statistical properties. If a parallel-profile model is used, concentrations or dilutions are usually scaled geometrically, usually in 2-fold, log, or half-log increments. If a slope-ratio model is used, concentrations or dilutions should be scaled arithmetically. Several functions may be used for fitting a parallel profile model to quantitative data, including a linear function, a higher-order polynomial function, a four-parameter logistic (symmetric sigmoid) function, or a five-parameter function for asymmetric sigmoids. Such functions require a sufficient number of concentrations or dilutions to fit the model. It is preferable to have at least four more concentrations or dilutions than the number of parameters that will be estimated in the model. Thus, a linear model with two parameters might require six concentrations, a quadratic model with three parameters might require seven concentrations, a four-parameter model may utilize eight concentrations, and a five-parameter model may use nine concentrations. The four- and five-parameter sigmoid models also require concentrations that help to establish the asymptotes of the response profile. Two concentrations are recommended to support each asymptote.

A linear model is sometimes selected because of efficiency and ease of processing. Because bioassay response profiles are usually nonlinear, the laboratory might perform an experiment with a wide range of concentrations in order to identify the approximately linear region of the concentration–response profile. For data that follow a four-parameter logistic model, these are the concentrations near the center of the response region, usually between 20% to 80% response when the data are rescaled to the asymptotes. Caution is appropriate in using a linear model because for a variety of reasons the apparently linear region may shift. A stable linear region may be identified after sufficient experience with the assay and with the variety of samples tested in the

assay. Data following the four-parameter logistic function may also be linearized by transformation. The lower region of the function is approximately linear when the data are log transformed (log–log fit).

Quantal data are typically fit using more complex mathematical models. A probit or logit model may be used to estimate a percentile of the response curve (usually the 50th percentile) or, more directly, the relative potency of the Test to the Standard. Spearman-Kärber analysis is a nonmodeling method for determining the 50th percentile of a quantal dose–response curve. Note that the Spearman-Kärber analysis requires a full range (0% through 100% response) in response to yield an accurate assessment of potency.

4.7 Suitability Testing

System suitability and sample suitability assessment should be performed to ensure the quality of bioassay results. System suitability in bioassay, as in other analytical methods, consists of prespecified criteria by which the validity of an assay (or perhaps a run containing several assays) is assessed. Analysts can assess system suitability by determining that some of the parameters of the Standard response are in their usual ranges and that some properties (e.g., residual variation) of all the data are in their usual range. To achieve high assay acceptance rates it is advisable to accept large fractions of these usual ranges (99% or more) and to assess system suitability with a few relatively uncorrelated measures. System suitability parameters and their ranges may also be selected on the basis of empirical or simulation studies that measure the impact of changes in a parameter on potency estimation.

Sample suitability in bioassay is evaluated using prespecified criteria for the validity of the potency estimate of an individual Test article. System and sample validity criteria should be established after the bioassay has been developed and before bioassay validation. Where there is limited experience with the bioassay these criteria may be considered tentative.

System suitability. System suitability parameters may be selected based on the design and the model of the procedure. Regardless of the design and model, however, system suitability parameters should be directly related to the quality of the bioassay. In parallel-line assays low values of the Standard slope typically yield estimates of potency with low precision. Rather than reject assays with low slope, analysts may find it more effective to temporarily use additional replicate assays until the assay system can be improved to consistently yield higher-precision estimates of potency. It may be particularly relevant to monitor the range of response levels and location of asymptotes associated with controls or Standard sample to ensure appropriate levels of response. A drift or a trend in some of the criteria may indicate a systematic degradation of a critical reagent or Standard material. Statistical process control (SPC) methods should be implemented to detect systematic trends in system suitability parameters.

Two common measures of system suitability are assessment of the adequacy of the model (goodness of fit) and of precision. With replicates, a pure error term may be separated from assessment of lack of fit. Care should be taken in deriving a criterion for lack of fit. The use of the wrong error term may result in an artificial assessment. A laboratory may choose alternative measures of system suitability or add others if deemed necessary or useful to maintain control of the assay.

For adequacy of the model, a sound measure is the lack of fit sum of squares from the model for the Standard alone. This value should be compared to the distribution of such values obtained historically. If the data are greater than some threshold percentile (e.g., 99th) of the historical distribution, then the data are not suitable. Note that the Test data are not used here. Adequacy of the model for the Test is implicitly tested with the parallelism assessment.

For assessment of precision, two alternatives can be considered: One follows an approach similar to the assessment of model adequacy and uses the mean squared error (residual variance) from the model fit to the Standard. This requires appropriate replication. However, this may be associated with a small number of degrees of freedom for the variance estimate. A second alternative is to use the mean squared error from the separate model fits to Standard and Test. Once the measure is selected, use historical data to determine a threshold for acceptance.

Sample suitability. Sample suitability or sample acceptance in bioassay consists of two major components: assessment of similarity and assessment of range. Bioassays can report relative potency only from samples that are similar to Standard and are within the range of the assay system. Samples for which a relative potency estimate is within range and the samples are not similar to the Standard should be reported as nonsimilar to Standard without a quantitative potency estimate.

Similarity. For the assessment of similarity (parallelism), an equivalence testing approach offers advantages. The equivalence test criterion may be formulated as a hypothesis test where the null hypothesis of nonsimilarity is that a similarity measure has an absolute value less than a critical value required to demonstrate similarity. When the resulting *P* value is below a suitable threshold, such as 0.05, there is sufficient evidence to reject the null hypothesis of nonsimilarity. Also, when a (given $P = 0.05$) 90% confidence interval on the similarity measure is entirely between the equivalence bounds, there is sufficient evidence to reject the null hypothesis of nonsimilarity.

Equivalence testing is contrasted with classical hypothesis testing (also known as difference testing) in which the null hypothesis is that there is no difference between Standard and Test values of a similarity parameter, because an equivalence testing approach states that sufficiently similar (at most an unimportant departure from similarity) is the alternative statistical hypothesis, and nonsimilarity is the null hypothesis. In this manner, equivalence testing controls (via alpha, the type I error rate) the risk of falsely declaring samples to be equivalent. With difference testing the risk of falsely declaring samples to be equivalent is a type II error. Without careful attention to

and monitoring of the variation of the similarity measure, the sample size, and the critical value of the similarity measure, analysts run the risk of falsely declaring samples similar. With equivalence testing, less information and monitoring are needed to control the critical risk of falsely declaring a sample to be similar to Standard. However, appropriately justified critical values of the similarity measure are needed. “Sufficiently similar” must exclude important differences.

More simply expressed: traditionally, classical hypothesis testing (difference testing)—not equivalence testing—has been used to establish parallelism between a Test sample and the Standard sample. From such an approach the laboratory cannot conclude that the slopes are equal. The data may be too variable, or the assay design may be too weak to establish a difference. The laboratory can conclude that the slopes are sufficiently similar using the equivalence testing approach.

Equivalence testing has practical advantages compared to difference testing, including that improved assay replication or precision will increase the chances that samples will pass the similarity criteria, that decreased assay replication or precision will decrease the chances that samples will pass the similarity criteria, and that sound approaches to combining data from multiple assays of the same sample to better understand whether a sample is truly similar to Standard or not are obtained.

The challenge in implementing equivalence testing is in setting appropriate equivalence bounds for the similarity measures. Ideally, information is available to link departures from similarity to relevant outcomes in clinical or animal data. Information may be available from evaluation of similar assays. Because different assays have responses and doses measured on different scales, it is challenging to compare changes in slope, intercepts (in slope-ratio assays), or asymptotes across assays. For perspective it can be useful to consider slope differences as a percentage of the Standard slope. Similarly, it can be useful to consider asymptote (or intercept) differences as a percentage of the range of the Standard asymptotes. Though these may not be universal experiences, some have observed that changes in slopes smaller than 30% and changes in asymptotes smaller than 5% are rarely meaningful. Setting critical values for changes in slopes > 50% or changes in asymptotes > 15% should be done with special care to ensure that there is a solid scientific case that these limits are reasonable. In routine use of the assay, the critical values of similarity parameters should not be on percentage scale because these scales have more uncertainty (wider confidence limits).

Because of the advantages associated with the use of equivalence testing in the assessment of similarity, analysts can transition existing assays to equivalence testing or can implement equivalence testing methods when changes are made to existing assays. It is informative to examine the risk that the assay will fail good samples. This risk depends on the precision of the assay system, the amount of replication in the assay system, and the critical values of the similarity parameters (a process capability analysis is one approach to this risk analysis). One approach to transitioning an established assay from difference testing to equivalence testing (for similarity) is to use the process capability of the assay to set critical values for similarity parameters. This

approach is reasonable for an established assay because the risks (of falsely declaring samples similar and falsely declaring samples nonsimilar) are implicitly acceptable because of the assay's history of successful use. Equivalence testing improves the relationships between these risks with changes in assay precision or assay sample size. In contrast, a new assay's capacity for keeping the risks at acceptable levels is unknown, and using a process capability analysis to set critical values of similarity measures would yield an assay system with unknown and unmanaged risks.

The similarity measures have, in many cases, interpretable, practical meaning in the assay (these measures are based on the parameters of the dose–response curve and include, e.g., slope for a straight parallel-line assay, intercept for a slope-ratio assay, slope and asymptotes for a four-parameter logistic parallel-line assay, and slope, asymptotes, and nonsymmetry parameter in a five-parameter sigmoid model). In some cases these changes in curve shape predict certain classes of interaction between the product and the bioassay. When possible, discussion of these interactions and their likely impacts is a valuable part of setting appropriate equivalence boundaries.

Implementing equivalence testing for similarity (parallelism). Many statistical procedures for assessing similarity are based on a statistical null hypothesis consistent with similarity and an alternative hypothesis of some degree of nonsimilarity. Failure to find statistically significant nonsimilarity is then taken as a conclusion of similarity. However, failure to detect nonsimilarity does not prove similarity. Equivalence testing provides a method for the analyst to proceed to a conclusion (if warranted by the data) of sufficiently similar while controlling the risk of doing so inappropriately. Following is a sequence for this process:

(1) Choose a measure of nonsimilarity. For the parallel-line case, this could be the difference or ratio of slopes. (The ratio of slopes can be less sensitive to the value of the slope. Also, framing the slope difference as a proportional change from Standard rather than in absolute slope units has an advantage because it is invariant to the units on the concentration and response axes.) For a slope-ratio assay, the measure of nonsimilarity involves the difference in y-intercepts between Test and Standard samples. Again, it can be advantageous to frame this difference as a proportion of the (possibly transformed) response range of Standard to make the measure invariant to the units of the response. For the four-parameter logistic model, similarity between Standard and Test samples must be assessed on the basis of three parameters: the upper asymptote, the slope, and the lower asymptote. If sigmoid curves with additional parameters are used to fit bioassay data, it is also important to consider addressing similarity between Standard and Test preparations of the additional curve parameters (e.g., asymmetry parameter of the five-parameter model). The comparison could be based on the individual parameters, one at a time, or some single composite measure of nonparallelism. One such composite measure is the parallelism sum of squares. This is found as the difference in residual sum of squares (RSSE) between the value obtained from fitting the Standard and Test curves separately and the value obtained from imposing parallelism:

$$\text{Parallelism sum of squares} = RSSE_P - RSSE_S - RSSE_T$$

where the subscripts P, S, and T denote Parallel model, Standard model, and Test model, respectively. With any composite measure, the analyst must consider the implicit relative weighting of the importance of the three (or more) parameters and whether the weighting is appropriate for the problem at hand. For the parallelism sum of squares, for example, with nonlinear models, the weighting given to the comparison of the asymptotes depends on the amount of data in the current assay on and near the asymptotes.

(2) Specify a range of acceptable values, typically termed an equivalence interval or "indifference zone," for the measure of nonsimilarity. When a ratio of slopes is used as a measure of nonsimilarity (nonparallelism), that measure is free to vary above or below a ratio of 1.0 (a ratio of 1.0 indicates perfect parallelism). The acceptable values for nonparallelism then constitute an interval. Following are four approaches that can be used to determine this interval. If pharmacopeial limits have been specified for a defined measure of nonsimilarity, then the assay should satisfy these requirements.

a. The first approach is to compile historical data that compare the Standard to itself and to determine, from the historical data, the equivalence interval as a tolerance interval for the measure of nonparallelism. The advantage of using historical data is that they give the laboratory control of the false-failure rate (the rate of failing an assay that is in fact acceptable). The disadvantage is that there is no control of the false pass rate (the rate of passing an assay that may have an unacceptable difference in performance relative to the Standard). The equivalence interval specification is driven solely by assay capability. Laboratories that use this approach should take care that an assay in need of improvement is not driving overly wide equivalence intervals. Also note that any change in assay capability means changing the equivalence interval—the interval must be relevant to the assay as it is currently conducted. Because of its limitations, this approach is best suited only for early stages in assay development where historical data may be the only information available.

b. The approach of (a) is simple to implement in routine use and can be used with assay designs that do not provide reliable estimates of within-assay variation and hence confidence intervals. However, there is a risk that assays with larger than usual amounts of within-assay variation can pass inappropriately. The preferable alternative to (a) is therefore to determine a tolerance interval for the confidence interval for the measure of nonparallelism. The following is particularly appropriate to transition an existing assay with a substantial body of historical data on both Standard and Test samples from a difference testing approach to an equivalence approach:

i. For each value of the measure of nonparallelism from the historical data, determine a 95% confidence interval, (a, b).
ii. For each confidence interval, determine its maximum departure from perfect parallelism. This is max(|a|, |b|) for differences, max(1/a, b) for ratios, and simply b for quantities that must be positive, such as a sum of squares.
iii. Determine a tolerance interval for the maximum departures obtained in (ii). This will be a two-sided tolerance interval for differences and ratios and a one-sided tolerance interval for necessarily positive quantities. A nonparametric tolerance interval approach is preferred.
iv. "Sufficiently parallel" is concluded for new data if the confidence interval for the measure of nonparallelism falls completely within the interval determined in (iii).

As with (a), this approach is best suited for early stages in assay development.

c. The third approach starts with historical data comparing the Standard to itself and adds data comparing the Standard to known failures, e.g., to degraded samples. Compare values of the measure of nonparallelism for data for which a conclusion of parallelism is appropriate (Standard against itself) and data for which a conclusion of parallelism is not appropriate, e.g., degraded samples. Based on this comparison, determine a value of the measure of nonparallelism that discriminates between the two cases. If using this approach, take care to include more than extreme examples. The work should include a range of samples for which a conclusion of parallelism is not appropriate, and be sure to cover samples with the minimal important change. For nonlinear models, this comparison also can be used to determine which parameters should be assessed. Some may not be sensitive to the failures that can occur with the specific assay.

d. The fourth approach is based on what is known of the product and the assay, much as (0.80, 1.25) is used for bioequivalence of generic products, or based on general experience. For example, when the ratio of slopes is not more than 1.3 and not less than 0.77, departure from parallelism is difficult to discern. In contrast, ratios greater than 1.5 or less than 0.67 are typically substantial. A related consideration could be the therapeutic index of the drug

For approaches (a) and (b) of Step 2, compare the obtained value of the measure of nonparallelism (a) or its confidence interval (b) to the interval obtained in Step 2. The value must be within the limits if one uses (a), or the confidence interval must be completely within the limits if one uses (b).

For approach (c) of Step 2, an alternative approach can be used. This approach essentially treats the parallelism as a discrimination problem. The choice of the cut point in (c) should take into account the rates of false positive and false negative decisions (and the acceptable risks to the laboratory) and should reflect the between-assay variability in precision. Thus it is reasonable to compare the point estimate of the measure of nonparallelism to the cut point and to not use confidence intervals. This approach is simpler to implement in routine use and can be used with assay designs that cannot provide reliable estimates of within-assay variation.

For approach (d) of Step 2, demonstrate that the measure of nonparallelism is significantly greater than the lower endpoint of the acceptance interval and significantly less than the upper endpoint. (If the acceptance interval is one-sided, then apply only the single applicable test.) This is use of the TOST procedure. For most situations, TOST can be most simply implemented by calculating a 90% two-sided confidence interval, which corresponds to a 5% equivalence test. If this confidence interval lies entirely within the equivalence interval specified in Step 2, then similarity is sufficiently demonstrated. For parallel-line models, one can use the confidence interval based on value $\pm$ k times the standard error of the value for the difference of slopes or Fieller's Theorem for the ratio of slopes. (This also applies to the difference of intercepts in slope-ratio models.) For nonlinear models, there is evidence that these simple confidence interval methods do not attain the stated level of confidence, and methods based on likelihood profile or resampling are more appropriate.

An alternative to the approach described above is to use an average (historical) value for the variance of the ratio or difference in a similarity parameter—obtained from some number of individual assays—to compute an acceptance interval for a point estimate of the similarity parameter. This approach is simpler to implement in routine use and can be used with assay designs that are unable to provide reliable estimates of within-assay variation. However, there is a price. The equivalence testing approach that relies on assay-specific (within-assay) measure(s) of variation (i.e., the confidence intervals) is conservative in the sense that it will fail to pass similarity for samples from assays that have larger than usual amounts of within-assay variation. Using an acceptance region for a similarity parameter—rather than an acceptance region for confidence intervals for the similarity parameter—loses this conservative property and hence is not preferred where alternatives exist.

***Range*.** The range for a relative potency bioassay is the interval between the upper and lower relative potencies for which the bioassay is demonstrated to have a suitable level of similarity, precision, accuracy, and assay linearity. It is straightforward to determine whether or not a sample that is similar to Standard has a relative potency within the (validated) range of the assay system. For samples that are apparently not similar according to standard assay methodology, it is more challenging to determine whether a relative potency estimate for the sample might be obtained. In a nonlinear parallel-line assay a sample that does not have data on one asymptote can usually be assumed to

be out of the potency range of the assay. In a parallel straight-line assay a sample that does not have three or more points on the steep portion of the response curve can also be assumed to be out of the potency range of the assay. In some assay systems fit with nonlinear models, it may be reasonable to declare a sample to be out of range based on an estimate of relative potency that is not predicated on a demonstration of similarity. In some assay systems fit with parallel straight-line models, it may be reasonable to declare a sample out of range using a crude potency estimate based on a subset of the Test sample that contains only two doses.

4.8 Outliers

Bioassay data should be screened for outliers before relative potency analysis. Outliers may be a simple random event or a signal of a systematic problem in the bioassay. Systematic error that generates outliers may be due to a dilution error at one or more concentrations of a Test article or the Standard or due to a mechanical error (e.g., system malfunction). Several approaches for outlier detection can be considered. Visual inspection is frequently utilized but should be augmented with a more objective approach to avoid potential bias.

An outlier is a datum that appears not to belong among the other data present. An outlier may have a distinct, identifiable cause, such as a mistake in the bench work, equipment malfunction, or a data recording error, or it could just be an unusual value relative to the variability typically seen and may appear without an identifiable cause. The essential question pertaining to an outlier becomes: Is the apparent outlier sampled from the same population as the other, less discordant, data, or is it from another population? If it comes from the same population and the datum is, therefore, an unusual (yet still legitimate) value obtained by chance, then the datum should stand. If it comes from another population and the datum's excursive value is due to human error or instrument malfunction, then the datum should be omitted from calculations. In practice, the answer to this essential question is often unknown, and investigations into causes are often inconclusive. Outlier management relies on procedures and practices to yield the best answer possible to that essential question and to guide response accordingly.

General Chapter *Analytical Data—Interpretation and Treatment* <1010>, addresses outlier labeling, identification, and rejection, including statistical methods, and provides material that the bioassay practitioner will find useful. General Chapter <1010> also lists additional sources of information that can provide a comprehensive review of the relevant statistical methodology. General Chapter <1010> makes no explicit remarks regarding outlier analysis in linear or nonlinear regression. Outlier analysis techniques appropriate for data obtained from regression of response on concentration can be used. Some remarks about outliers are provided here in the context of bioassays to emphasize or complement the information in <1010>.

Of the procedures employed for analysis of drug compounds and biological drugs, the bioassay can be expected to be the most prone to outlying data. The management of

outliers is appropriate with bioassay data on at least two levels: where an individual datum or a group of data (e.g., data at a concentration) can be checked against expected responses for the sample and concentration; and, separately, when estimates of relative potency from an assay can be checked for consistency with other independent estimates of the potency of the same material.

Three important aspects of outlier management are prevention, labeling, and identification.

Outlier prevention is obviously preferred and is facilitated by procedures that are less subject to error and by checks that are sensitive to the sorts of errors that, given the experience gained in assay development, may be expected to occur. In effect, the error never becomes an outlier because it is prevented from occurring.

Good practice calls for the examination of data for outliers and labeling ("flagging") of the apparently outlying observation(s) for investigation. If investigation finds a cause, then the outlying datum may be excluded from analysis. Because of the ordinary occurrence of substantial variability in bioassay response, a laboratory's investigation into the outlying observation is likely to yield no determinable cause. However, the lack of evidence regarding an outlier's cause is not a clear indication that statistical outlier testing is warranted. Knowledge of the typical range of assay response variability should be the justification for the use of statistical outlier tests.

Outlier identification is the use of rules to confirm that the values are inconsistent with the known or assumed statistical model. For outliers with no determined cause, it is tempting to use statistical outlier identification procedures to discard unusual values. Discarding data solely because of statistical considerations should be a rare event. Falsely discarding data leads to overly optimistic estimates of variability and can bias potency estimates. The laboratory should monitor the failure rate for its outlier procedure and should take action when this is significantly higher than expected.

Statistical procedures for outlier identification depend on assumptions about the distribution of the data without outliers. Identification of data as outliers may mean only that the assumption about distribution is not correct. If dropping outliers because of statistical considerations is common, particularly if outliers tend to occur more often at high values or at high responses, then this may be an indication that the data require some adjustment, such as log transformation, as part of the assay procedure. Following are presentations of two approaches to statistical assessment of outlying data: replication based and model based.

Replication-based approaches. When replicates are performed at concentrations of a Test article and the Standard, an "extra variability" (EV) criterion may be employed to detect outliers. Historical data can be analyzed to determine the range in variability commonly observed among replicates, and this distribution of ranges can be used to establish an extreme in the range that might signal an outlier. Metrics that can be utilized are the simple range (maximum replicate minus minimum replicate), the

standard deviation, or the CV or RSD among replicates. However, if the bioassay exhibits heterogeneity of variability, assumptions about uniform scatter of data are unsupported. Analysts can use a variable criterion across levels in the bioassay, or they can perform a transformation of the data to a scale that yields homogeneity of variability. Transformation can be performed with a POM approach as discussed previously. Where heterogeneity exists nonnormality is likely present, and the range rather than standard deviation or RSD should be used.

The actions taken upon detection of a potential outlier depend in part on the number of replicates. If EV is detected within a pair (n = 2) at a concentration of a Test article or the Standard, it will not always be clear which of the replicates is aberrant, and the laboratory should eliminate the concentration from further processing. If more than two replicates are performed at each dilution the laboratory may choose to adopt a strategy that identifies which of the extremes may be the outlier. Alternatively the laboratory may choose to eliminate the dilution from further processing.

Outliers may also be detected among replicate potency measurements on a Test article from a series of runs. Methods for outlier detection are described in <1010>, or analysts may use methods described above for replicates within a dilution of a Test article or the Standard.

Model-based approaches. Model-based approaches may be used to detect outliers within bioassay data. These approaches use the residuals from the fit of the appropriate bioassay model (the difference between the observed response and the response predicted by the model). Statistical methods are available for assessing outliers among residuals in this framework. Statistical support should be obtained in order to identify a method that is suitable for a particular application.

Lastly, an alternative to discarding outlying data is to use robust methods that are less sensitive to influence by outlying observations. Use of the median rather than the mean to describe the data's center exemplifies a robust perspective. Also, regression using the method of least squares, which underlies many of the methods in this chapter, is not robust in the presence of outliers. The use of contemporary methods of robust regression may be appropriate. This is left for investigation beyond this chapter.

4.9 Fixed and Random Effects in Models of Bioassay Response

The choice of treating design factors as fixed or random is important both to the design and the statistical analysis of the assay. Fixed effects are factors for which all levels, or all levels of interest, are discretely present, e.g., concentration. Fixed effects, such as temperature and duration of thaw, are expected to cause a consistent shift in responses. Analysts study fixed effects by controlling them in the design and examining changes in means across levels of the factor. In a bioassay, sample and concentration are fixed effects. Studies of other fixed effects are typically performed early in assay development as part of the process of optimizing assay performance.

Random effects are factors whose levels in a particular run of an assay are considered representative of levels that could have been present. That is, there is no expectation of a particular value for a random effect. Rather, that value may vary subject to some expected distribution of values and thus may be a source of variability. Examples of random effects include reagent lot, operator, or date of assay if there is no interest in specific reagent lots, operators, or dates as sources of variability. Analysts study random effects by measuring the variance components corresponding to each random effect. As with fixed effects, these are profitably studied during assay development.

The choice of treating a factor as fixed or random is important to the design of the assay and to proper reporting of its precision. Treating all factors as fixed, for example, leads to an understatement of assay variability because it ignores all sources of variability other than replication. The goal is to identify specific sources of variability that can be controlled, to properly include those factors in the design, and then to include other factors as random.

Additionally, if the factor may switch from random to fixed effect or vice versa, the factor should be modeled as a random effect. For example, reagent lots cannot be controlled, so different lots are typically considered to cause variability, and reagent lot would be considered a random effect. However, if a large shift in response values has been traced to a particular lot, a comparison among a set of lots could be performed using the predicted levels of each lot's fixed effect. Similarly, within-assay location (e.g., block, plate, plate row, plate column, or well) or sequence may be considered a source of random variation or a source of a consistent (fixed) effect.

Assay designs that consist of multiple factors are efficient, but they require corresponding statistical techniques that incorporate the factors as fixed or random effects in the analysis. If all factors are fixed, the statistical model is termed a fixed-effects model. If all are random, it is termed a random-effects model. If some factors are fixed and some random, the model is a mixed-effects model. Note that the concepts of fixed and random effects apply to models for qualitative and integer responses as well as for quantitative responses.

For assay designs that include multiple experimental units (e.g., samples assigned to sets of tubes and concentrations assigned to preplate tubes) a mixed-effects model in which the experimental units are treated as random effects is particularly effective.

5. Stages in the Bioassay Process

5.1 Design: assay layout, replication strategy, blocking, and randomization

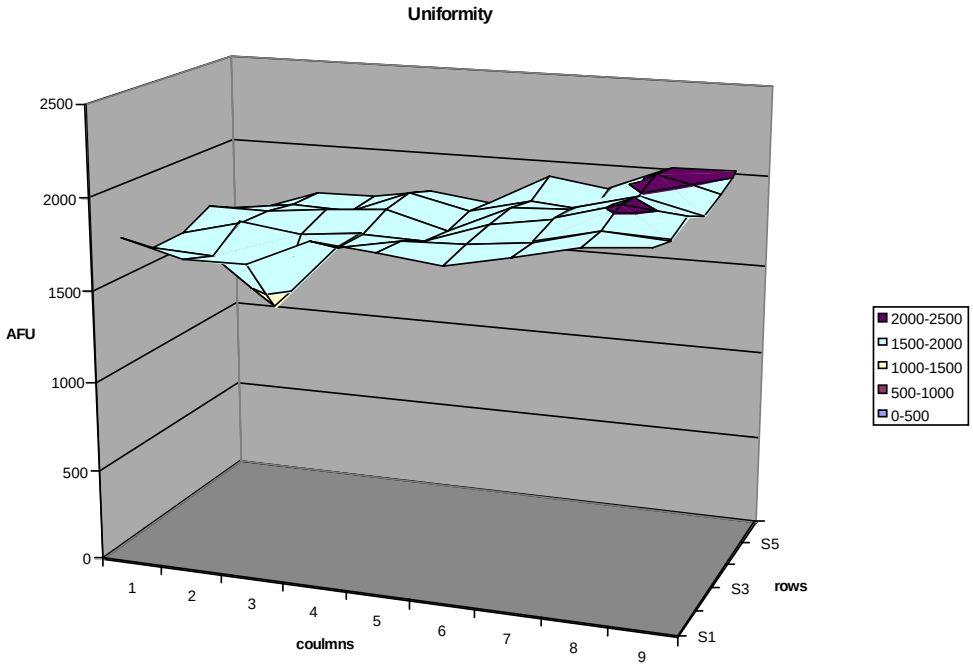
Assay layout. The assay layout depends on the type of bioassay being performed. In general all bioassays include operational units such as wells of a plate or individual animals to which one of a series of concentrations of a Test article or a Standard is administered. The numbers and levels of concentrations, as well as numbers and origin of the replicates of each concentration constitute a portion of the assay layout. Layout

considerations should be weighed against the potential for bias and their impact on bioassay variability. To the extent possible, operations that potentially can lead to bias should be managed so that the bias is translated into random variability. This is because although random variability can be effectively managed by replication, bias cannot.

Most cell-based assays are performed using a cell culture plate (6-, 12-, 96-, or 384-well microtiter plate) that not only must support cell culture in a uniform fashion from well to well but may also undergo a series of wash steps and incubations that may induce different cell responses in certain regions of the plate. One goal during development is to find operating conditions for the assay so that the cell responses to the analyte are free of systematic effects across the plate, across multiple plates, and across assays. Consistent performance is an important component of assay capability. Even assay conditions intended to minimize the potential for systematic bias (e.g., good analyst technique, careful calibration of pipettes, control of incubation time or temperature) often yield systematic gradients on the plate. These gradients may occur across rows, across columns, or from the edge to the center of the plate and are termed *plate effects*. Even modest-sized or inconsistent gradients (across plates, days, or analysts) should be addressed during assay development (plate layout, randomization, and replication).

Plate effects can be evaluated in a *uniformity trial* in which a single assay dose (often chosen from the portion of the assay curve that represents the greatest sensitivity to change in dose) is used across the entire plate. In the 3D plot seen in Figure 1, a trend of decreasing signal is evident from right to left. In this instance, the wash program for the platewasher had to be adjusted because plates were being washed harder on the left side of the plate. Another common plate effect is a differential cell-growth pattern in which the outer wells of the plate grow cells in such a way that the assay signal is attenuated. When this occurs, it is best to not utilize the outer wells of the assay plate.

Figure 1. 3D plot of change in sensitivity across a plate (note decreasing sensitivity from right to left).



Complete randomization of sample treatments and dilutions has been advocated as the best approach to minimizing assay bias or, more accurately, to protecting the assay results from known and unknown sources of bias by converting the latter into variance. Randomization constitutes a good strategy because it addresses systematic bias. Increased replication alone is likely to improve precision, but it cannot address systematic bias without some degree of randomization. Even when complete randomization is not practical, a plate layout can be designed to minimize plate effects by alternating sample positions across plates and the pattern of dilutions within and across plates. When no plate effects are evident, the plate layout design should, at minimum, alternate sample positions across plates within an assay run to accommodate possible bias introduced by the analyst or equipment on a given day. It is prudent to use a balanced rotation of layouts on plates so that the collection of replicates (each of which uses a different layout) provides some protection against likely sources of bias.

Figure 2 illustrates the design of a single assay plate wherein the samples are titrated in two different directions. This can help protect against plate column effects. Figure 3 illustrates a simple alternation of Test (Test sample 1 = “1”; Test sample 2 = “2”) and Standard positions across plates (across a single assay run). If the samples are alternated to different positions across an assay run, the design can help protect against plate row effects. Combining the two methods can effectively help convert plate bias into assay variance, which can be addressed by increased assay replication (increased number of plates in an assay).

For additional discussion of randomization, see Blocking and randomization, below.

Figure 2. A single assay plate in which samples are titrated in two different directions.

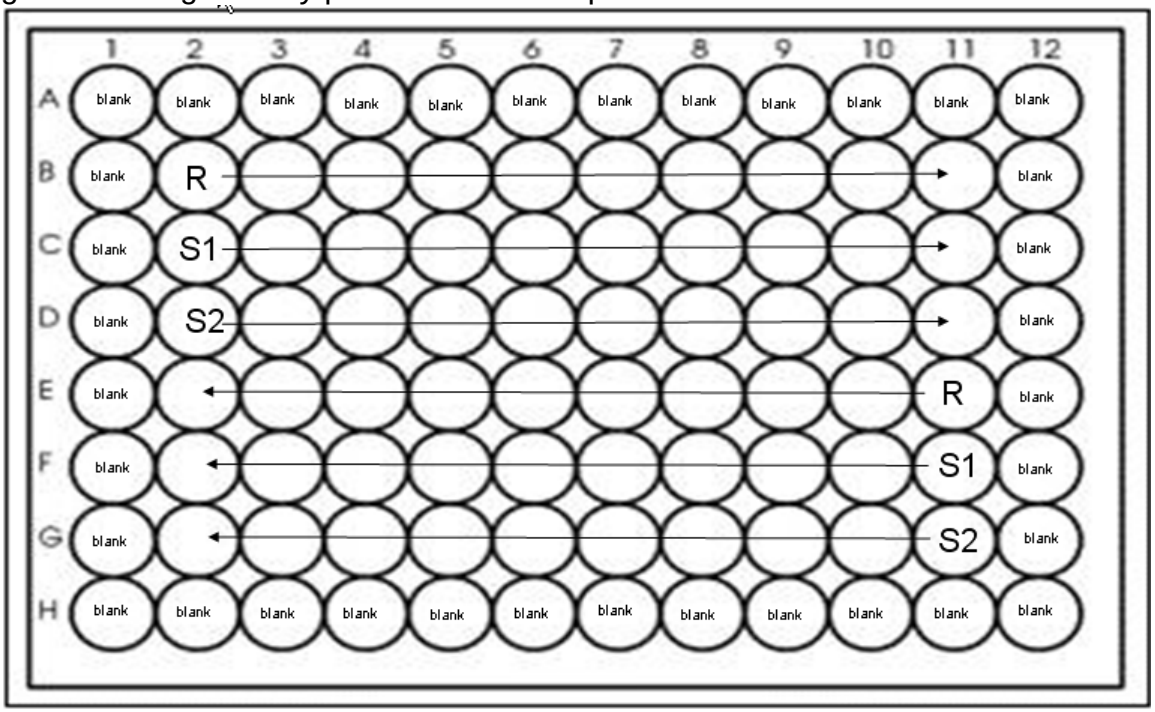


Figure 3. Schematic of a single assay plate with simple alternation of Test samples' and Standard positions.

Plate Row	Plate 1	Plate 2	Plate 3
B	R	2	1
C	1	R	2
D	2	1	R
E	R	2	1
F	1	R	2
G	2	1	R

Replication strategy. Concentrations of a Test article and the Standard can be derived in multiple ways. Laboratories that use manual multichannel pipettes often perform serial dilutions wherein each dilution is prepared from one or more previous dilutions of the material. Alternatively the laboratory may prepare independent dilutions from the Test article and Standard to obtain the concentration series. These two strategies result in the same nominal concentrations, but they have very different properties related to bias and noise. Serial dilutions are subject to propagation of error across the dilution series, and a dilution error made at an early dilution will result in a shift in the dilution

profile throughout the series. Independent dilutions help mitigate the bias resulting from such a dilution error. Although serial dilutions are easier to perform in manual operations, independent dilutions can be readily obtained with robotics.

Replicates at each concentration similarly may be achieved in several ways. When multiple units of volume from a single point in a dilution series are dispensed into multiple wells, pseudoreplicates are present. Other strategies that provide more useful information are using independent dilution series of the Test article and Standard or making independent dilution replicates at each dilution point from each material. The reward for independent replicates is less potential bias and greater precision.

The number and levels of concentrations depend on desired throughput and the model that will be used to process bioassay data. The number of concentrations should increase as the number of model parameters increases. Linear modeling requires fewer concentrations than do nonlinear logistic regression models (see Section 3). Additionally, the levels of concentrations should be selected to support the model. For a linear model at least four concentrations are desirable to assess goodness of fit and obtain reliable estimates of relative potency. For four- or five-parameter logistic models it is useful to have at least four concentrations in the “linear” region of the curve and at least two concentrations in the minimum and maximum regions (asymptotes).

Blocking and randomization. Blocking and randomization are used to transform bias into random variability. Bias is introduced into bioassay measurement by operational factors related to instrumentation and the bioassay environment. Operational factors that may induce bias are edge effects in plate-based bioassays, reaction across a plate in a time-sensitive bioassay, and cage location in animal-based bioassays. The influence of these effects is mitigated by blocking and randomization, either separately or in combination. Blocking is the grouping of related experimental units in experimental designs, such as a dilution series of a Test article or the Standard. Randomization is the process of assigning treatment to experimental units based on chance so that all equal-sized groups of units have an equal chance of receiving a given treatment, e.g., the placement of a bioassay unit or a block in random configuration in a bioassay.

Different blocking and randomization strategies can be utilized by a bioassay laboratory. Figure 4 illustrates a poor plate-based assay design. Dilutions and replicates of the Test preparations (A and B) and the Standard (R) are placed together on the plate. Bias due to a plate or incubator effect can influence part or all the concentrations of one of the samples.

Figure 4. Schematic of a poorly designed plate (note that dilutions and replicates of the Test preparations and the Standard are placed together on the plate).

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	
	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	

A split-plot design, an alternative that randomizes samples across plate rows and dilutions (concentrations) within each row, is seen in Figure 5. Such a strategy is tedious and requires robotics for routine reliability.

Figure 5. Schematic of a split-plot design that randomizes samples across plate rows and dilutions (concentrations).

	A3	A6	A10	A2	A5	A1	A4	A9	A8	A7	
	R6	R9	R1	R2	R5	R10	R3	R8	R7	R4	
	B1	B5	B2	B8	B3	B10	B7	B9	B4	B6	
	R5	R1	R9	R4	R2	R8	R3	R10	R6	R3	
	A4	A2	A8	A7	A1	A9	A3	A10	A5	A6	
	B10	B3	B7	B2	B4	B5	B9	B6	B1	B10	

A layout that provides some protection from plate effects and can be performed manually is a strip-plot design, shown in Figure 6.

Figure 6. Schematic of a layout that provides some protection from plate effects and can be performed manually.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	
	R10	R9	R8	R7	R6	R5	R4	R3	R2	R1	
	A10	A9	A8	A7	A6	A5	A4	A3	A2	A1	
	B10	B9	B8	B7	B6	B5	B4	B3	B2	B1	

Here samples are again randomized to rows of a plate, but dilution series are performed in different directions in different locations on the plate to mitigate bias across columns of the plate. An added advantage of the strip-plot design is the ability to detect a plate effect by the interaction of sample and dilution direction (left-to-right or right-to-left).

For related discussion pertaining to plate effects in cell-based assays, see *Assay layout*, above.

5.2 Development

Assay development ends with the completion of the standard operating procedure (SOP) for the bioassay. The SOP is a documented protocol for bioassay implementation. The SOP should include enough detail so that a qualified laboratory with a trained analyst can perform the procedure in a routine manner. A strategic part of development is a look forward toward performance maintenance. Standard operating procedures for reagent and technician qualification, as well as for calibration of the working Standard, help complete the bioassay development package.

One factor at a time vs design of experiments. Bioassay development proceeds through a series of experiments in which conditions and levels of assay factors are varied to identify those that support a reliable and robust bioassay qualified for routine use. Those experiments may be conducted one factor at a time (OFAT), studying each parameter separately to identify ideal conditions, or through the use of multi-factor design of experiments (DOE). DOE is an efficient and effective strategy for developing a bioassay and optimizing bioassay performance, thus helping to obtain a measurement system that meets its requirements. In comparison to OFAT, DOE generally requires fewer experiments and also provides insight into interactions of factors that affect bioassay performance. Assay development and optimization using DOE proceeds through a series of steps: process mapping and risk analysis; screening; response optimization; and confirmation.

Process mapping and risk analysis. Bioassay optimization begins with a systematic examination of bioassay factors and a risk assessment to identify those factors that may influence performance. It is useful to visualize bioassay factors using a bioassay process map such as a cause-and-effect or fishbone diagram. Using the process map as a guide, the laboratory can examine assay factors that might affect assay performance, such as buffer pH, incubation temperature, and incubation time. Historical experience with one or several of the bioassay steps, along with sound scientific judgment, can identify key factors that require further evaluation. One tool that may be used to prioritize factors is a failure mode and effects analysis. Factors are typically scored by the combination of their potential severity and the likelihood that they will occur. The laboratory must be careful to recognize potential interactions between assay parameters.

Screening. Once potential key factors have been identified from process mapping and risk analysis, the laboratory should conduct an initial screening experiment to probe for effects that may require control. Screening designs such as factorial and fractional factorial designs are commonly used for this purpose. Software is available to assist the practitioner in the selection of the design and in subsequent analysis. Analysts should take care, however, to understand their assumptions about design selection and analysis to ensure accurate identification of experimental factors. Care should also be

taken in defining design variables, which are measurements made from design runs and should be directly or indirectly related to bioassay performance.

Response optimization. A screening design will usually detect a few important factors from among those studied. Such factors can be further studied in a response-optimization design. Response-optimization designs such as central composite designs are performed to determine optimal settings for combinations of bioassay factors for achieving desired response. The information obtained from response optimization is typically depicted as a response surface and can be used to establish ranges that yield acceptable assay performance. With the advent of Quality by Design, that region is called the design space for the bioassay. This is also the set of ranges that will be incorporated into the bioassay SOP.

Confirmation. The mathematical model depicting assay performance as a function of changes in key assay factors is an approximation. Thus it is customary to confirm performance at the ideal settings of the bioassay. Confirmation can take the form of a qualification trial in which the assay is performed, preferably multiple independent times using optimal values for factors. Or the laboratory may determine that the bioassay has been adequately developed and may move to validation. Qualification is a good practice, not a regulatory requirement. The decision to perform confirmatory qualifying runs or to engage in validation is a business decision and depends upon the strength of the accumulated information obtained throughout development.

5.3 Data Analysis during Assay Development

Analysis of bioassay data during assay development enables analysts to make decisions regarding the statistical model that will be used for routine analysis, including transformation and/or weighting of data. The analysis also provides information regarding which elements of design structure should be used during outlier detection and the fitting of a full model. This may also include a plan for choosing subsets of data, such as a linear portion, for analysis or, for nonlinear bioassays, a model reduction strategy for samples similar to Standard. Once these decisions are made and checked during validation, they are not revisited for each bioassay analysis. These goals may be pursued in a stepwise manner, as follows:

(1) Choose an appropriate statistical model (see Section 4.6). Many considerations are involved in making this determination. First, the model should be appropriate for the type of assay endpoint—continuous, a count, or present/absent. Second, the model should incorporate the structure of the assay design. For any design other than completely randomized, there will be terms in the model for the structural elements. These could be, for example, within-plate blocking, location of cage in the animal facility, day, etc. A third consideration is applicable to continuous endpoints and involves whether to use a regression model or a means model (an analysis of variance model that fits a separate mean at each dilution level of each sample tested), with appropriate error terms.

(2) Fit the chosen statistical model to the data without the assumption of parallelism, and then assess the distribution of the residuals (differences between the data and the model), specifically examining them for departures from normality and constant variance. Transform the data as necessary or, if needed, choose a weighting scheme (Section 4.3). Use as large a body of assay data as possible for this step—at least several, and preferably dozens, of independent assays. The primary goal is to address any departure from constant variance of responses across the range of concentrations in the assay. Step 2 will likely alternate between imposing a transformation and assessing the distribution of the residuals.

(3) Screen for outliers. This step normally follows the initial choice of a suitable transformation and/or weighting method. Outlier analysis should be conducted in a predefined and consistent manner and is best done on the residuals from a model that is fit to all the data within an assay. Ideally the model used for outlier detection contains the important elements of the assay design structure, allows nonsimilar curves, and makes fewer assumptions about the functional shape of the concentration–response curve than did the model used to assess similarity. Outlier analysis on a small number of replicates or pseudoreplicates is inadvisable. See Section 4.8 and General Chapter <1010>. In some cases outliers may be so severe that a reasonable model cannot be fit, and thus residuals will not be available. In such cases, it is necessary to screen the raw data for outliers before attempting to fit the model.

(4) Remove outliers as appropriate. See Section 4.8 for a discussion of methods. Routinely, before an observation is declared an outlier, an investigation should be conducted to determine whether a cause can be identified.

During development, prepare instructions for the investigation and treatment of an outlier observation, including any limits on how many outliers are acceptable. Include these instructions in the assay procedure. Good practice includes recording the result of this investigation, outlier test(s) applied, and results from the latter. Note that outlier procedures must be considered apart from the investigation and treatment of an out-of-specification (OOS) result (reportable value). Decisions to remove an outlier from data analysis should not be made on the basis of how the reportable value will be affected in terms of a potential OOS result.

Removing data as outliers should be rare. If many values from a run are removed as outliers, that run should be considered suspect.

(5) After outlier removal, refit the model with the transformation and/or weighting previously imposed (Step 2) and re-assess the appropriateness of the model.

(6) If necessary or desired, choose a scheme for identifying subsets of data to use for potency estimation, whether the model is linear or nonlinear (see 4.5).

5.4 Bioassay Validation

The bioassay validation is a protocol-driven study that demonstrates that the procedure is fit for use. The validation should not be carried out until there are no expectations of further changes in the procedure and an assay SOP has been developed. Preliminary system and sample suitability controls should be established and clearly written into the assay SOP. These may be finalized based on additional experience gained from the validation exercise. USP Chapter *Biological Assay Validation* <1033> provides details regarding the design and analysis of the validation.

5.5 Bioassay Maintenance

The development and validation of a bioassay, though discrete operations, lead to ongoing activities. Assay improvements may be implemented as technologies change and as the laboratory becomes more skilled with the procedure, and changes to bioassay methodology require re-evaluation of bioassay performance. Some of these changes may be responses to unexpected performance during routine processing. Corrective action should be monitored using routine control procedures. Substantial changes may require a study verifying that the bioassay remains fit for use. An equivalence testing approach can be used to show that the change has resulted in acceptable performance. A statistically oriented study can be performed to demonstrate that the change does not compromise the previously acceptable performance characteristics of the assay.

Assay transfer. Assay transfer assumes both a known intended use of the bioassay in the recipient lab and the associated required capability for the assay system. These implicitly, though perhaps not precisely, inform limits on the amount of bias and precision degradation allowed between labs. Consider, for example, an assay with an established potency range from 0.5 to 2.0 with validation data demonstrating bias less than 5% and intermediate precision from a single assay of 9% geometric relative standard deviation. If this assay is used to support a product with a potency specification of 0.71 to 1.41, replicate assays will be required on each lot to achieve a process capability of 1.3. Using two laboratories interchangeably to support one product will require considering variation between labs in addition to intermediate precision for sample size requirements to determine process capability. With only two laboratories it is not sensible to estimate a variance, and so an estimate of reproducibility is not available. It is reasonable to determine a 90% confidence interval on the bias between labs and to add the most extreme endpoint of this bias interval to the assay bias in the process capability calculation. Working through these calculations for various magnitudes of assumed bias will demonstrate the impact of various assay transfer bias acceptance limits on the future replication requirements of the assay system. For a discussion and example pertaining to the interrelationship of bias, process capability,

and validation, see A Bioassay Validation Example in *Biological Assay Validation* <1033>.

Improving or updating a bioassay system. If a new version of a bioassay improves the bias, precision, range, robustness, or specificity, the new version may lower the operating costs or offer other compelling advantages. When improving or updating a bioassay system analysts can use a bridging study to compare the performance of the new vs established versions. A wide variety of samples (e.g., lot release, stability, stressed, critical isoforms) should be used for demonstrating equivalence of estimated potencies. Compare the performance of the new and existing assays using the method described above under assay transfer. If the assay systems are very different (e.g., an animal bioassay vs a cell-based bioassay) but use the same Standard and mechanism of action, one can reasonably expect comparable potencies. If the new assay uses a different Standard, the minimum requirement for a comparison is a unit slope of the log linear relationship between the estimated potencies. An important implication of this recommendation is that poor precision or biased assays used early can have lasting impact on the replication requirements, even if the assay is later replaced by an improved assay.