# Using Computer Simulated Results of a Bulk Drug Substance Assay to Determine Acceptance Criteria for Method Validation

**Richard D. Bunnell**[1,2]

*Received July 31, 1996; accepted December 4, 1996*

***Purpose.*** To determine the statistical variability expected for a well designed HPLC assay of a bulk drug substance (BDS). The results are used to develop appropriate acceptance criteria for a method validation protocol as well as to evaluate the level of uncertainty expected for assay results using a variety of sampling/injection schemes.

***Methods.*** Computer simulation was used to generate a large quantity of data and the variability of the mock results was evaluated. Error propagation was also calculated, whenever possible, to confirm results obtained from the simulations.

***Results.*** Protocol acceptance criteria were developed that were consistent with the expected variability for data resulting from the execution of the validation protocol. In certain cases simulations provided the only avenue of obtaining results that could not otherwise be readily determined.

***Conclusions.*** Computer modeling can be used to obtain suitable acceptance criteria for validation results which are consistent with method variability. This is particularly significant in the case of linearity where it has been difficult to develop acceptance criteria based on anything other than analyst intuition and experience. Assay simulations clearly demonstrated that the variability expected for a typical BDS assay is large relative to the average specification range and therefore little insight about relative purity can be gained comparing individual passing assay results.

**KEY WORDS:** method validation; HPLC assay; statistics; computer simulation.

## INTRODUCTION

Validation of analytical methods is a vital part of today's regulatory submission. The United States Pharmacopeia (USP) includes a general chapter on analytical methods validation (1) and the International Conference on Harmonization (ICH) published a guideline on the subject that appeared in the Federal Register (2). Both identify analytical parameters that should be addressed as part of the validation process but neither gives much detail on how the parameters are to be evaluated. The ICH recently initiated a draft guideline (3) intended to address some general aspects of how evaluations are to be conducted, but this document does not and cannot cover all the details required for a method validation.

This paper presents a protocol intended for use in the validation of a conventional HPLC assay for a BDS. The protocol avoids falling into the trap of attempting to generate sufficient data to prove beyond a statistical shadow of a doubt that Beer's law is reliable or that high performance liquid chromatography is based on sound scientific principles. Such approaches ultimately lead to enormous amounts of experiments of dubious value. Instead we started by asking the pragmatic question: "What will a manageable number of typical validation results look like when the method performs as intended?". Validation protocol acceptance criteria were based on the answer to this question. Next we asked: "What will typical validation results look like when the method does not perform as intended?". The answers to this second question were used to probe the statistical power of the protocol.

Three fundamental assumptions served as the statistical starting points upon which the acceptance criteria were based:

1. All sample and standard solutions are prepared in such a way that the uncertainties in concentrations (mg/mL) will be limited to a relative standard deviation (rsd) of 0.3% or less;
2. The rsd for HPLC injection volume will be limited to 0.5% or less;
3. HPLC peak detection/integration will have an rsd of 0.3% or less.

How these starting values were arrived at is discussed in detail.

Computer simulations were used to model thousands of experimental trials of the proposed validation protocol. Random error, derived from the three fundamental assumptions for anticipated experimental error, was introduced into each of the simulated experiments. The variability observed between protocol simulations was used to establish the acceptance criteria by defining the ranges of acceptable protocol results.

Computer modeling was also used to investigate the question: "Just how accurate will a typical assay result be?" Various sampling schemes are discussed in terms of the uncertainties of the results associated with each. Views are also presented on the importance of assay results for use in characterizing a BDS.

## EXPERIMENTAL

Statistical calculations and simulations were performed using StatView 4.5 (FPU) statistical software (Abacus Concepts, Inc., Berkeley, CA) installed on an Apple Power Macintosh 7100/80 with 16 Mb RAM and 700 Mb hard drive.

## VALIDATION PROTOCOL

The following protocol addresses the Analytical Performance Parameters listed in USP 23 (1) for a Category I Assay. With the exception of ruggedness, each Performance Parameter is evaluated as a single set of experiments using the same HPLC equipment, batch of mobile phase, balance, etc. throughout the set.

Standard and sample weighings must be kept within ±5% of the nominal target. Unless otherwise indicated, the same source of high purity BDS should be used to generate all validation data.

All weighings must meet the USP definition of accurate such that three times the standard deviation of the reproducibility of the measurement should not exceed 0.1% of the weight. Serial dilutions of stock solutions are limited to one. Only class-

[1] Preformulation Department, Pharmaceutical Sciences, Allergan, 2525 Dupont Dr., Irvine, California 92713.
[2] To whom correspondence should be addressed.

0724-8741/97/0200-0156$12.50/0 

A volumetric flasks of 25.0 mL or larger volume and Class-A pipets 2.0 mL or larger are to be used.

The manufacturers performance specifications for any HPLC injector employed must stipulate an injection precision ≤0.5% for relative standard deviation (rsd) at the injection volume used.

## Precision

*Total Method Precision*

Independently prepare six sample solutions at 100% of the nominal assay concentration and label the solutions #1 through #6. Make three consecutive injections of each solution and take the average of the detector responses for the triplicate injections. Normalize the average of the three injections by dividing it by the solution concentration (mg/mL). Calculate the %rsd ($S_t$) for the six normalized results (4). *Acceptance Criterion*: $S_t \leq 0.68\%$.

*Injection Precision*

Make nine additional successive injections of solution #1. Calculate the %rsd for detector response of the nine injections (4). *Acceptance Criterion*: $S_a \leq 0.82\%$.

## Linearity and Range

Independently prepare one solutions each at 25%, 50%, 75%, 100%, and 125% of the nominal assay concentration. For each solution take the average detector response for three injections. Perform linear regression analysis on the results using the average detector responses as the ordinate (y axis) and the calculated % of nominal assay concentration as the abscissa (x axis). Calculate the y-intercept (b), the slope (m), and ($S_m$) the percent standard error of the slope (4). *Acceptance Criteria*: $S_m \leq 0.9\%$. $|b| \leq 0.9\%$ of the theoretical 100% response ($y_{100\%}$) where: $y_{100\%} = m(100\%) + b$.

## Specificity

Prepare a drug substance solution at 100% the nominal assay concentration that also contains as many known or potential impurities as possible. If authentic samples of impurities are available, spike the solution to achieve either 1% by area or a level of twice the specification for the impurity (whichever is larger). Evaluate the drug substance HPLC peak for purity using diode array or other suitable detector. Calculate the resolution between the sample and each impurity detected. Visually inspect the chromatographic trace and note any abnormalities. *Acceptance Criterion*: Specificity ≥ 1.2 between the drug substance peak and each impurity detected. Peak purity analysis of the sample peak must be indicative of a pure peak. Visual inspection of the chromatogram must not reveal any significant interference or abnormalities.

## Ruggedness

Testing is to be carried out in two laboratories designated Lab #1 and Lab #2. Lab #1 will usually be the same laboratory where the method is validated for precision, linearity, and specificity. At minimum the HPLC column, HPLC hardware, solutions, and analyst should differ between laboratories.

*Inter Laboratory Assay Comparison (Lab #1 & 2)*

Independently assay three typical lots of drug substance (Sample N, where N = 1 through 3) in both laboratories. For each assay, three preparations each of standard and sample are to be used. The same three standard preparations can be used for all three assays. The bracketing scheme outlined in Table 1 must be followed to generate the assay results. Results for all injections of standard and sample should be normalized and averaged so as to arrive at a single assay value for each lot of drug assayed. *Acceptance Criterion*: The absolute value of the difference in assay results between lab #1 and lab #2 should not exceed 1.4% of the result from lab #1.

*Sample Solution Stability (Lab #1)*

Following the completion of the assay, allow a time to elapse that is at least equal to the expiration age for a sample solution as designated in the method. Prepare three fresh reference standard solutions and re-assay sample #1 using the injection scheme outlined in Table 1. *Acceptance Criterion*: Result for the re-assay should not differ from the original result by more than 1.4%.

*Injection Precision (Lab #2)*

Make nine successive injections of the standard solution used in the assay. Calculate the %rsd ($S_{a'}$) for detector response of the nine injections (4). *Acceptance Criterion*: $S_{a'} \leq 0.82\%$.

*Linearity (Lab #2)*

Prepare solutions at 75%, 100%, and 125% of the nominal assay concentration. Make three injections per solution and take the average for each. Perform linear regression analysis on the results using mean detector response as the ordinate (y axis) and calculated % of nominal assay concentration as the abscissa (x axis). Calculate the y-intercept (b'), the slope (m'), and ($S_{m'}$) the % standard error of the slope (4). *Acceptance Criteria*: $S_{m'} \leq 2.5\%$ and $|b'| \leq 4.1\%$ of the theoretical 100% response.

**Table 1.** Injection Scheme to be Used in Inter Laboratory Assay Comparison as Part of Ruggedness Portion of the Validation Protocol

| Solution | Number of Injections |
|---|---|
| Standard Prep. 1 | 1 |
| Standard Prep. 2 | 1 |
| Standard Prep. 3 | 1 |
| Sample N, Prep. 1 | 2 |
| Sample N, Prep. 2 | 2 |
| Sample N, Prep. 3 | 2 |
| Standard Prep. 1 | 1 |
| Standard Prep. 2 | 1 |
| Standard Prep. 3 | 1 |

*Miscellaneous (Lab #2)*

Evaluate the method for clarity, completeness, and any other appropriate characteristic. *Acceptance Criteria*: The method must be found to adequately exhibit those characteristics associated with a valid, well written, and rugged method.

### Accuracy

Review the results for Precession, Linearity, Specificity and Ruggedness testing. *Acceptance Criteria*: The method is considered sufficiently accurate when the acceptance criteria for Precession, Linearity, Specificity and Ruggedness have been met.

## RESULTS

One can't be absolutely sure about anything, consequently no amount of testing will "prove" a method is valid. Validation can only provide data that is consistent with the assertion that a method is accurate, precise, linear, etc. to some statistical level of confidence. The acceptance criteria presented in this paper were chosen to be "consistent" for a valid method within the confines of the experiments that were described in the preceding protocol.

The validation approach is predicated on three fundamental assumptions regarding variability. First, the rsd for any solution concentration ($\sigma_p$, where concentration is in units of [mg sample]/[mL solution]) is described by a population with rsd limited to $\sigma_p \leq 0.3\%$. Second, HPLC injectors used have precision for injection volume ($\sigma_i$) of $\sigma_i \leq 0.5\%$. Finally, the precision for peak area quantitation ($\sigma_d$) of the detector/integrator used will be capable of $\sigma_d \leq 0.3\%$. These fundamental assumptions were based on the physical limitations of the analyst and the laboratory equipment. From these starting points it was possible to predict the degree to which experimental results would be expected to vary given no other significant factors effecting the analysis (i.e. the method was valid and there were no additional significant sources of error). Tables 2 and 3 contain definitions for the population and sample statistics used to describe the various aspects of the expected method variability. These variabilities were then used to set acceptance criteria for the elements of the validation protocol (Table 4). Also contained in Table 4 are acceptance criteria that would result from other possible starting values of $\sigma_p$, $\sigma_i$, and $\sigma_d$ when the same set of validation experiments outlined in the protocol are used.

**Table 2.** Definitions of Population Statistics

| Population Statistic | Definition |
|---|---|
| $\sigma_d$ | Standard deviation for the detection/integration of sample injections. |
| $\sigma_i$ | Standard deviation for the HPLC injection volumes delivered. |
| $\sigma_p$ | Standard deviation for the concentrations of solutions. |
| $\sigma_a$ | Standard deviation for individual HPLC measurements $= (\sigma_d^2 + \sigma_i^2)^{0.5}$. |
| $\sigma_t$ | Standard deviation for normalized total HPLC measurements of a given sample $= (\sigma_p^2 + \sigma_a^2/N)^{0.5}$; where N = the number of injections. |
| $\sigma_{\%purity}$ | Standard deviation for assay results using a given assay scheme. |

**Table 3.** Definitions of Sample Statistics

| Sample Statistic | Definition |
|---|---|
| $S_a$ | Standard deviation for individual HPLC measurements made in Lab #1 calculated from nine injections of a solution. |
| $S_{a'}$ | Standard deviation for individual HPLC measurements made in Lab #2 calculated from nine injections of a solution. |
| $S_t$ | Standard deviation calculated from the six normalized results of the six individual preparations of a homogenous sample. |
| $S_m$ | Standard error in the slope calculated from the five point linearity experiment. |
| $S_{m'}$ | Standard error in the slope calculated from the three point linearity experiment in Lab #2. |

### Primary Sources of Method Variability ($\sigma_p$, $\sigma_i$, and $\sigma_d$)

*Solution Concentration ($\sigma_p$)*

A worst case uncertainty of $\sigma_p = 0.3\%$ for solution concentration was one of the pivotal assumptions upon which the acceptance criteria were based. This uncertainty in concentration can be further broken down into three independent components; sample weighing, deviation of glassware from nominal volumes, and analyst variability.

Sample weights conforming to the USP definition of "accurate" will be controlled so that three times the rsd of the measurement will not exceed 0.1% of the weight. For a typical five place analytical balance the standard deviation for a measurement is 0.02 mg and this makes 60 mg the minimum weight that can be measured "accurately" on such a balance.

Class-A glassware is allowed to deviate from nominal volumes within defined tolerances (5). Tolerance is not a statistical concept and therefore many labs ignore this potential source of uncertainty. When very accurate volumes are required glassware manufacturers can provide (for a fee) certificates documenting volume to a high degree of precision. The approach taken in this paper was to account for this source of uncertainty by considering all possible volumes within a tolerance range to be equally probable. This represents a compromise between simply ignoring this potential source of uncertainty and virtually eliminating it through the exclusive use of certified glassware.

Analyst variability typically constitutes the single largest source of error in solution concentration. Reading of a meniscus or delivering volumes have variabilities that can be approximated by normal distributions. Precision in using volumetric glassware has been investigated by gravimetric means. A rsd ~0.1% represents a realistic (although conservative) estimate for both transfer pipets (2 mL and up) and volumetric flasks (25 mL and up) in the hands of a competent analyst.

In order to develop an estimate for the uncertainty in solution concentration ($\sigma_p$) one must combine the errors for sample weighing, deviation of glassware from nominal volumes, and analyst variability. It is not readily apparent how to

**Table 4.** Maximum Protocol Acceptance Criteria Resulting for Various Starting Assumptions of Intrinsic Method Variability

| Validation Parameter | $\sigma_d$ = 0.3%<br>$\sigma_i$ = 0.5%<br>$\sigma_p$ = 0.3%<br>($\sigma_t$ = 0.45%) | $\sigma_d$ = 0.3%<br>$\sigma_i$ = 1.0%<br>$\sigma_p$ = 0.3%<br>($\sigma_t$ = 0.67%) | $\sigma_d$ = 0.1%<br>$\sigma_i$ = 0.5%<br>$\sigma_p$ = 0.1%<br>($\sigma_t$ = 0.31%) | $\sigma_d$ = 0.1%<br>$\sigma_i$ = 0.1%<br>$\sigma_p$ = 0.1%<br>($\sigma_t$ = 0.13%) |
|---|---|---|---|---|
| Precision: | | | | |
| $S_t$ | 0.68% | 0.97% | 0.46% | 0.19% |
| $S_a$ | 0.82% | 1.45% | 0.71% | 0.20% |
| Linearity & Range: | | | | |
| $S_m$ | 0.9% | 1.4% | 0.6% | 0.3% |
| \|b\| | 0.9% | 1.4% | 0.6% | 0.3% |
| (corr. coefficient) | (0.9998) | (0.9995) | (0.9999) | (1.0000) |
| Ruggedness: | | | | |
| $S_{a'}$ | 0.82% | 1.45% | 0.71% | 0.20% |
| $S_{m'}$ | 2.5% | 3.7% | 1.7% | 0.7% |
| \|b'\| | 4.1% | 6.2% | 2.9% | 1.2% |
| (corr. coefficient) | (0.9990) | (0.9979) | (0.9995) | (0.9999) |
| Absolute Value in Assay Results<br>(Lab #1 vs Lab #2)<br>(Old vs Fresh Std.) | 1.4% | 2.2% | 1.0% | 0.4% |

combine these flat and normal distributions using conventional error propagation. The approach taken was to simulate a worst case solution preparations using a computer model. The solution preparation simulated was as follows: "Accurately weigh 60 mg of sample into a 25 mL Class-A volumetric flask, q.s. (Solution #1). Using a 2 mL Class-A transfer pipet, deliver 2 mL of Solution #1 into a 25 mL Class-A volumetric flask, q.s. (Solution #2). The nominal concentration of solution #2 = 192 μg/mL."

Ten thousand solution preparations were simulated in this fashion with the appropriate random uncertainty introduced into each variable. The resulting variability in Solution #2 closely approximated a normal distribution with a mean equal to the 192 μg/mL and a rsd = 0.27%. Based on the simulations it was concluded that uncertainty in solution concentrations ($\sigma_p$) could be limited to 0.3% or less by restricting serial dilution's to no more than one and using the appropriate types of glassware and analytical balance as specified in the protocol.

*Injection Volume ($\sigma_i$)*

The variability experienced in measuring a solution using a conventional HPLC system will in part depend on the precision with which a volume can be injected ($\sigma_i$). A review of various manufacturers specifications for autoinjectors indicates that a precision of $\sigma_i \leq 0.5\%$ is the norm. Consequently, for the purpose of deriving protocol acceptance criteria, an injector precision of $\sigma_i$ = 0.5% was assumed.

*Peak Detection/Integration ($\sigma_d$)*

The last major source of uncertainty in an HPLC measurement is attributable to the precision of the detection/integration measurement ($\sigma_d$). A well developed HPLC method, utilizing a conventional UV detector, will have adequate peak shape, signal to noise, and specificity to minimize this source of error. This source of variability was estimated to be $\sigma_d \sim 0.3\%$ by evaluating multiple injections of a well resolved two component HPLC solution. The uncertainty in the peak ratio for a two component solution is independent of injection volume and is equal to the square root of the sum of the squares of each peaks detection/integration uncertainty. Chromatograms where peaks were well resolved, symmetrical in shape, and of roughly equal areas were found to have ratio variabilities of about rsd = 0.4% over a wide range of injection volumes. Assuming equal contributions from each peak, the uncertainty in each individual peak was estimated at $\sigma_d$ = 0.3%.

**Rationale for Precision Acceptance Criteria**

The total repeatability ($\sigma_t$) for a normalized HPLC measurement (peak area per unit concentration of solution) can be broken down into two independent components; the precision of the solution preparation ($\sigma_p$), and precision of the measurement ($\sigma_a$) of that solution. The individual terms combine to yield the value of $\sigma_t$ shown in equation 1, where N is the total number of injections made.

$$\sigma_t = \sqrt{\sigma_p^2 + \sigma_a^2/N} \quad (1)$$

Equation 1 states that the total uncertainty is equal to the square root of the sum of the squares of the individual sources of variability. Note that as N gets large, the uncertainty due to the measurement ($\sigma_a^2/N$) disappears and the precision of solution preparation ($\sigma_p$) becomes the only source of variability. By evaluating results from multiple injections of a single solution, $\sigma_a$ can be directly probed. By testing multiple preparations of a homogenous sample, $\sigma_t$ can be probed.

Just as the total precision can be divided into two parts so too can the precision of an individual HPLC measurement ($\sigma_a$) be considered as resulting from two independent sources, the uncertainty due to the repeatability of the HPLC injector ($\sigma_i$) and the uncertainty in detecting and integrating the signal resulting from a given injection ($\sigma_d$). Given the primary values for injector precision of $\sigma_i$ = 0.5% and detection/integration

precision of $\sigma_d = 0.3\%$ the precision for a single HPLC peak measurement was calculated as; $\sigma_a = (\sigma_i^2 + \sigma_d^2)^{1/2} = (0.5^2 + 0.3^2)^{1/2} = 0.58\%$.

This calculation of $\sigma_a = 0.58\%$ confirmed the results obtained by computer simulation as follows. Areas for ten thousand simulated injections ($Inj_i$ for i = 1 to 10,000) were produced using the random number generating function (RandomNormal) in StatView 4.5 (FPU). Simulated injections were based on a nominal injection volume of 100%, a rsd for injection volume of 0.5%, a nominal peak area of 1, and a rsd in peak area of 0.003. The algorithm used for the simulation appears in equation 2. Simulated values for $Inj_i$ resulted in a population with a mean of 100.0011% and a rsd = 0.5826%.

$$Inj_i = \text{RandomNormal}(100.0, 0.5) \times \text{RandomNormal}(1, 0.003) \quad (2)$$

In the laboratory, validation experiments for injection precision yield a value for the sample statistic ($S_a$) which is an estimate of $\sigma_a$. The sample statistic for precision in the measurement of a solution is calculated using equation 3 where $Inj_{ave}$ is the average area for the N injections.

$$S_a = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Inj_i - Inj_{ave})^2} \quad (3)$$

Simulated experimental results for nine injections were generated using the random number function of StatView 4.5 from a parent injection population with mean = 100% and $\sigma_a = 0.58\%$. The nine simulated injection values were then used to calculate a value for $S_a$ using equation 3. This process was repeated 10,000 times. In 95% of the cases, $S_a \leq 0.82\%$ and this value was designated as the acceptance criterion for the protocol. This value sets the probability of committing a Type I error at $\alpha = 0.05$ (i.e. at $\alpha = 0.05$, there is a 5% probability of obtaining a value for $S_a > 0.82\%$ by random chance alone when $\sigma_a = 0.58\%$).

A similar result was obtained with the *F* Statistic (6) using equation 4, where 1.94 is the $F_{.05}$ statistic and $\sigma_a = 0.58$.

$$\text{Acceptable Values of } S_a \leq \sqrt{1.94 \times \sigma_a^2} \quad (4)$$

The *F* statistic 1.94 is the ratio between the variances that would be calculated 95% of the time for a sample with eight degrees of freedom (N − 1) in the numerator and a sample size with an infinite number of degrees of freedom in the denominator when both samples are taken from the same parent population. The result arrived at by equation 4 was $S_a \leq 0.81\%$.

The statistical power of the protocol was determined by computer simulation through an iterative process. An initial guess was made at a value of $\sigma_a$ that would result in a Type II error at $\beta = 0.05$ (i.e. at $\beta = 0.05$, there is a 5% probability of obtaining a value for $S_a \leq 0.82\%$ by random chance alone when $\sigma_a$ is much larger than our estimate of 0.58%). Simulated data were then generated following the same process described for determining the protocol acceptance criteria value for $S_a \leq 0.82\%$. Several iterations of this process were carried out until it was found that at $\sigma_a = 1.4\%$ only 5% of the resulting $S_a$ values met acceptance criterion.

Presumably the *F* statistic could also be used to determine the power. An equation similar to equation 4 would be required except that an $F_{.95}$ statistic would be substituted for the $F_{.05}$ value and 0.82 would be substituted for 0.58.

The second requirement of the precision protocol is to determine a value for the sample statistic $S_t$ (total repeatability for a normalized measurement) from triplicate injections of each of six sample solutions. Equation 1 predicts a value of $\sigma_t = 0.45\%$ when $\sigma_a = 0.58\%$, $\sigma_p = 0.3\%$, and N = 3 injections per solution. This result verified that obtained by computer simulation. Following the approach used to determine $S_a$, a value of $S_t \leq 0.68\%$ at $\alpha = 0.05$ was determined.

Power was calculated using the same iterative technique described for the injection solution experiments. For acceptance criterion $S_t \leq 0.68\%$ and $\beta = 0.05$ the statistical power is sufficient to distinguish a value $\sigma_t \geq 1.3\%$.

### Rationale for Linearity and Range Acceptance Criteria

The USP suggests using variance of the slope ($S_m^2$) as a measure of linearity, but it has been our practice to use $S_m$, the square root of the variance, for this purpose. Either can be used to probe linearity.

Expected slope variability was determined using StatView 4.5. Sixty thousand linearity experiments were simulated resulting in a population of $S_m$ values that, for all practical purposes, describes the theoretical parent population for this sample statistic. It was not possible to confirm by a calculation the variability in slopes determined by the simulation.

The first step in the simulation was to generate five solution simulations corresponding to the five targeted concentrations expressed as percentages of the nominal assay concentration ($C_{X\%}$). Simulated solutions were allowed to vary up to ±5% relative to the nominal targeted concentration. Simulated solution concentrations were generated using equation 5, where X is the target concentration expressed as a percentage of the nominal assay concentration.

$$C_{X\%} = \text{RandomUniform}(0.95X, 1.05X) \quad (5)$$

RandomUniform(0.95X, 1.05X) is the StatView expression to generate a random number between 0.95X and 1.05X. from a uniform (i.e. flat) distribution such that all values within the range are equally probable. For example a solution targeting 25% of the nominal was simulated by: $C_{25\%}$ = RandomUniform(23.75, 26.25).

Y coordinates ($R_{Y\%}$) of the linearity simulations correspond to normalized HPLC measurements of each hypothetical solution with rsd = 0.45% (equation 1). $R_{Y\%}$ values were simulated by equation 6 which mimics the variability expected for the average of three injections of a single solution.

$$R_{Y\%} = \text{RandomNormal}(C_{X\%}, 0.0045C_{X\%}) \quad (6)$$

Results from equations 5 and 6 were used to calculate the linear regression in the usual manner (4).

In the model just described a constant relative variance is used to model simulated data even though constant absolute variance is assumed in the derivation of linear regression equations (6). This apparent contradiction does not constitute a problem. The simulation generates data typical of expected results and subjects this data to linear regression analysis. Many such simulations are used to determine how the results will vary from experiment to experiment and acceptance criteria are based on this variability. The fact that the linear regression

derivation makes assumptions that are not accurate is irrelevant to the determination of the expected variability for regression results.

Sixty thousand linear regression simulations resulted in a population of values 99.9% of which yielded $S_m \leq 0.9\%$. The results also produced a population of y-intercepts (b), approximately 99.9% of which had absolute values $|b| \leq 0.9\%$ relative to the $R_{100\%}$ value calculated from the linear regression line.

The Type I error in $S_m$ acceptance criterion was controlled at $\alpha = 0.001$ instead of the more conventional level $\alpha = 0.05$ because variability in simulated $S_m$ values, for a truly linear method, is a function of data precision. Controlling $S_m$ at $\alpha = 0.05$ is needlessly excessive considering precision acceptance criteria already incurs a 5% probability of data rejection from this same source of variability. It is worth noting that even at $\alpha = 0.001$ the correlation coefficients (r) for simulations that met the criteria were such that $r \geq 0.9998$.

As a measure of possible systematic bias, the intercept has more to do with accuracy than linearity. Nevertheless, since the y-intercept is a direct result of linear regression analysis, it was included in linearity criteria. An intercept of |0.9%| or less should occur about 99.9% of the time when method precision is at least that expected and the true intercept is zero. For $b = \pm 1.3\%$, $\beta = 0.05$ for the acceptance criteria. A true 1.3% intercept would have a marginal impact on the typical drug substance assays since this systematic error will occur in both standard and sample and will therefore tend to cancel out.

## Rationale for Specificity Acceptance Criteria

Specificity is not evaluated in a statistical fashion in the protocol. Resolution is measured between the main analyte peak and potential sources of interference. Lack of evidence of interference with the sample peak is taken as conformation of the absence of a bias as described in the USP definition of specificity. Resolution (R) acceptance criteria was set at $R \geq 1.2$ since this degree of resolution is indicative of virtually complete baseline resolution. Specificity will decrease as impurity levels increase (assuming sample peak size is constant). Protocol impurity levels were chosen to surpass worst case levels for typical samples while still not resulting in overly restrictive specificity requirements. Diode array and visual inspection criteria address the potential co-elution of sample with an unanticipated impurity.

## Rationale for Ruggedness Acceptance Criteria

The protocol employs a holistic approach to ruggedness by showing the method performs as intended in different laboratories. Robustness is the measure of a methods capacity to remain unaffected by small but deliberate variations in parameters and is not addressed in the protocol. Robustness should be limited to those rare occasions where parameters are identified as having a reasonable probability of a significant impact on performance under normal conditions. Most individual aspects of a method are not individually challenged using our approach, but every experiment that yields a result consistent with expectations increases the confidence in the overall method and the robustness of all it's aspects. When a particular parameter is identified as a potential source of trouble, the protocol should be amended accordingly.

*Injection precision and Linearity* were singled out for ruggedness testing because they provide a rigorous system suitability check in lab #2. Acceptance criteria was set using methodology already described for precision and linearity. Each criterion was controlled at $\alpha = 0.01$ because this testing was performed as a secondary exercise supporting the inter laboratory assay comparison and as such a larger $\alpha$ would incur needlessly high additional occurrences of a Type I error.

*Inter laboratory Assay Comparison* is the heart of ruggedness testing. Results from this testing demonstrate that the method yields statistically indistinguishable results between labs. Determination of an acceptance criterion (yielding the value $\alpha = 0.05$) was dependent on the sampling/injection scheme specified. How the sampling/injection scheme specified led to the acceptance criterion is discussed in the section of this paper titled "Reliability of Assay Results".

*Sample Solution Stability* was the final variable chosen for regular investigation. Stability of sample solutions must always be considered during method development. The acceptance criterion standardizes the minimum solution stability acceptable for an HPLC assay of this type.

*Miscellaneous* acceptance criteria allows for a subjective non statistical critique of the method from a fresh perspective. The value of this criteria is self evident.

## Rationale for Accuracy Acceptance Criteria

Method accuracy can be defined as the closeness of the hypothetical mean purity, derived from an infinite number of assays, to the "true purity". Method accuracy is a function of systematic errors. On the other hand the accuracy of an individual test result will be a function of both method precision and method accuracy. Method accuracy is a property intrinsic to a method and is addressed during method validation. The accuracy that can be expected for an individual result is another matter and will be discussed later.

Method accuracy is commonly evaluated using recovery experiments, but in the context of a bulk drug assay "percent recovery" is a meaningless concept. Recovery experiments are usually conducted by comparing results from a placebo spiked with sample against those for the same sample in solvent only. For a typical bulk drug assay the placebo and the pure solvent are one in the same.

Bulk drug standard and sample are almost always dissolved in the same solvent. This is not surprising considering that standard is generally just another lot of the bulk drug (reference standard purity determination is a topic for another paper). Unless a sample is grossly out of specification, standard and sample solutions will have very similar compositions. The statistical distributions describing HPLC measurements of standard and sample will be virtually identical except for differences in the means of the two distributions corresponding to the true difference in purity between the two. Even in cases where there exists a hidden built in method bias, its effect on accuracy should be mitigated since the bias will be present in both sample and standard and thus will tend to cancel out.

Based on this reasoning it was concluded that under normal conditions a bulk drug assays will be intrinsically accurate. Meeting acceptance criteria for precision, linearity, specificity and ruggedness qualifies as acceptance criteria for accuracy

because these support the assertion that the method is performing as expected.

## DISCUSSION

### Reliability of Assay Results

It can be a daunting task explaining to the uninitiated how, on the one hand a method can be very accurate, while on the other hand an individual result obtained from this "accurate" method may not be as reliable as desired. For a perfectly accurate assay, the reliability of a result will be a function of the precision of the individual measurements and the number of replicates of these measurements performed. Since assay schemes often vary between laboratories, the reliability of test results can vary even when all other variables are equal. The following is a discussion of some of the more simple assay schemes possible and how these schemes effect the reliability of the results obtained.

Probably the simplest scheme compares a single injection of a standard solution against a single injection of a sample solution. For this scheme the percent purity of the sample can then be calculated using equation 7 where $A_{spl}$ and $A_{std}$ are the detector responses for the sample and standard respectively, $C_{spl}$ and $C_{std}$ are the calculated concentrations (mg/mL) for the sample and standard preparations respectively, and $\%purity_{std}$ is the %purity of the standard.

$$\%purity_{spl} = \%purity_{std} \frac{A_{spl}C_{std}}{A_{std}C_{spl}} \quad (7)$$

Uncertainty in the results from equation 7 was determined by multiple computer simulations and was confirmed through the calculation of the square root of the sum of the squares of uncertainty using equation 8.

$$\sigma_{\%purity} = \sqrt{2\sigma_p^2 + 2\sigma_a^2} \quad (8)$$

Given the values for $\sigma_p$ and $\sigma_a$ that were discussed previously, one can calculate the uncertainty resulting from this scheme to be $\sigma_{\%purity} = 0.92\%$. This is not a very satisfactory level of precision for a bulk drug assay result. For example, a typical bulk drug might have a purity specification of 98.5 to 101.0% by weight. Using this scheme one can calculate the probability of obtaining a failing result to be 19% ($\alpha = 0.19$) for those cases where the sample is in fact 100% pure (note that since this specification is asymmetric about 100%, the probability of a obtaining a failing assay will be due to a result above 101.0% in the majority of cases). Chances of a Type II error occurring when the sample is less pure than 98.5% will be such that $\beta = 0.05$ for a purity of 97.0%.

A more precise value results using the scheme designated for ruggedness testing given in Table 1. The normal uncertainty associated with this approach was determined by computer simulation and the result of the simulation was confirmed using equation 9 to be $\sigma_{\%purity} = 0.41\%$.

$$\sigma_{\%purity} = \sqrt{\frac{2}{3}\sigma_p^2 + \frac{2}{6}\sigma_a^2} \quad (9)$$

For a bulk drug specification of 98.5 to 101.0%, $\alpha = 0.009$ for obtaining a failing assay result for a bulk drug that is exactly 100% pure. The Type II error of identifying a sample as meeting minimum specification when in fact it is less than 98.5% pure will be such that at a purity of 97.8%, $\beta = 0.05$. This is far superior to the first scheme but still implies that at a 95% confidence level that an assay result is only good to about ±0.8%.

At this point a further discussion of the ruggedness criterion for *Inter laboratory Assay Comparison (Lab #1 & 2)* is in order. Since the assay result expected from the sampling/injection scheme in Table 1 yields $\sigma_{\%purity} = 0.41\%$, the probability of assay results from the two labs agreeing to within 1.4% of each other was calculated to be about 98.4%. The chances that three out of three separate assays will agree to within 1.4% is about 95%. This calculation was the basis for the acceptance criterion specified in the protocol.

Neither of the schemes described so far are representative of common day to day practice. A typical scheme might go as follows; Prepare three independent solutions of a sample and two independent solutions of a standard (STD-1 and STD-2). Assay one standard against the other making triplicate injections of each. If normalized average peak areas for the two standards are within some predetermined limit of each other (±1 or 2% typically) then use STD-1 for the assay. Make triplicate injections of the standard at the start of the assay, then make triplicate injections of each of the three sample preparations, followed by three further injections of the standard. Calculate a single assay value using an average of the six standard injections and all of the three sample preparations and injections. The uncertainty expected was calculated using equation 10, confirming the result obtained by computer simulation.

$$\sigma_{\%purity} = \sqrt{\frac{4}{3}\sigma_p^2 + \frac{5}{18}\sigma_a^2} \quad (10)$$

The uncertainty was determined to be $\sigma_{\%purity} = 0.46\%$. For a bulk drug specification of 98.5% to 101.0% one can calculate that $\alpha = 0.02$ for the assay of a sample that is 100% pure and $\beta = 0.05$ for a sample of purity equal to 97.7%.

How should the assay result obtained from a valid method be viewed in light of results from the preceding three schemes? It seems clear that when a reasonable scheme is employed the expected uncertainty of the result approaches the magnitude of the entire span of the typical BDS specification, and it would be a mistake to rely too heavily on an assay result alone as proof of bulk drug purity. A pure drug will routinely assay within specification but where an assay value falls within the specification range will not be very informative. It is virtually impossible to distinguish between assay results that differ by less than 1%. Assay results are best viewed as just one in a number of factors that need to be considered when accessing the acceptability of a BDS. A passing assay alone is a very weak evidence of purity.

It has been suggested that to comply with the spirit of the Barr decision every injection of a BDS solution should yield a result that meets specification. This is clearly unrealistic based on the preceding analysis. It was already shown that for a single injection of standard and sample that a failing assay would result for a 100% pure BDS in 19% of the cases when the specification is 98.5% to 101.0%. If three injections of sample (and standard) were used to calculate three values for the same sample in this way, the chances that all three would meet specification is only 53%. Chances of a passing result could be

improved by using the average of the three standard injections to calculate a value for each of three sample injections but even then all three results would be expected to meet specification only 66% of the time.

A similar situation exists when any given sample solution preparation is evaluated against the BDS specification. If for example, each of three sample solutions is evaluated separately against the specification then the probability of producing a failing result is almost triple that for the single sample preparation. BDS assays are not intended as content uniformity tests and should not be treated as such. During a BDS assay, multiple sample solutions should be prepared for the express purpose of producing a single, statistically refined, assay result. If each preparation must be evaluated separately, most labs would probably opt for assaying a single sample solution and would thus produce a less reliable value. The chances of obtaining a failing result (given the specification of 98.5% to 101.0%) for an assay where triplicate injections of standard and sample are used to produce a single assay result is 7%. The chances that at least one of a group of any three such assays would fail is 19%. There is only a 4% chance that the average value would fail. The situation would improve if $\sigma_a$ and/or $\sigma_p$ were less than the magnitudes assumed in this paper, but it is difficult to imagine how that could be the case. Any refinements to $\sigma_a$ and/or $\sigma_p$ would almost certainly result in increases in these values thus painting an even bleaker picture.

### Other Assay Refinements

Throughout this paper the assumption is made that most of the variability describing sample preparations and measurements are normally distributed. This is a necessary and, for the most part, reasonable assumption. Nevertheless, in the real world, sample can be lost during weighing, or a drop of solution can fall from a pipet tip during the transfer of an aliquot, etc.; these are occurrences that are not well accounted for by the approximation of a normally distributed experimental variability. The conventional approach to mitigate such possibilities is to test for outliers. It is beyond the scope of this paper to review the merits of outlier testing, but the effects of some types of outlier testing is evaluated.

For example; a method may require that results for three preparations of a homogenous sample all be within 1.0% of each other (average normalized area of 3 injections per preparation). A method which exhibits the level of precision upon which the validation protocol acceptance criteria were based will meet this criterion 74% of the time. The cases that survive this criteria are described by a parent population with a statistical distribution that is slightly narrower than that describing the data before the criteria was imposed ($\sigma_t$ = 0.37% vs 0.45%). This criterion will detect major deviations due to errors not accounted for by normal variability but little is gained in terms of reducing normal variabilities even at a 26% rejection level. When a 2% criterion is imposed, over 99% of the normally distributed trios will meet the criteria and $\sigma_t$ is virtually unchanged.

When a very precise assay value is needed, it is best to use an internal standard. When properly designed, an internal standard method can avoid a great deal of the sources of uncertainty already discussed. Using the scheme in the previous example where three sample solutions are bracketed by a single reference standard (three injections from each HPLC vial) one can calculate $\sigma_{\%purity}$ = 0.19%. This is a considerable improvement over the absence of an internal standard where $\sigma_{\%purity}$ = 0.46%.

## CONCLUSIONS

An approach has been presented that allows acceptance criteria for validation experiments to be chosen based on performance expectations for a valid method. Computer modeling allowed for the prediction of expected variability of the validation process elements. This was true even in the case of linearity experiments where a conventional error propagation calculation could not be readily determined. The approach is versatile and can be customized for most Class I and Class II quantitative assays by designing an appropriate validation protocol and choosing reasonable values for the primary sources of method variability ($\sigma_d$, $\sigma_i$, and $\sigma_p$).

The reliability of results obtained from a valid method was considered in terms of various assay schemes. It was concluded that in order to obtain an assay result precise enough to ensure an acceptable degree of confidence that a truly pure BDS could be relied upon to meet the typical specification, multiple injections of multiple sample preparations are required to generate a single assay value. Assay results derived from individual injections and/or individual sample solutions will typically not have adequate precision, given current HPLC technology and the typical BDS specification. While results calculated from multiple injections of multiple sample preparations of a pure sample can be relied upon to meet specification, the uncertainty of such results will be such that no meaningful comparisons are possible between assay results that differ by about 1% or less.

## REFERENCES

1. *USP 23*, General Chapter 1225, pp. 1982–1984.
2. *Federal Register*, Vol. 60, No. 40, Wednesday, March 1, 1995.
3. ICH3 Announcement, *Validation of Analytical Procedures: Methodology*, Step 2 of the ICH Process, 29 November 1995.
4. Hugh D. Young. *Statistical Treatment of Experimental Data*, McGraw Hill, Inc., 1962.
5. *USP 23*, General Chapter 31, p. 1680.
6. Sanford Bolton. *Pharmaceutical Statistics*, Marcel Dekker, Inc., 2nd edition, 1990, Drugs and the Pharmaceutical Sciences Volume 44.