(66) C. G. Moertel, A. J. Schutt, R. J. Reitemeier, and R. G. Hahn, *Cancer Chemother. Rep.*, **56**, 95(1972).
(67) A. S. Kende, T. J. Bentley, R. W. Draper, J. K. Jenkins, M. Joyeux, and I. Kubo, *Tetrahedron Lett.*, **1973**, 1307.
(68) A. I. Meyers, R. L. Nolen, E. W. Collington, T. A. Narwid, and R. C. Strickland, *J. Org. Chem.*, **38**, 1974(1973).
(69) S. Danishefsky and R. Volkmann, *Tetrahedron Lett.*, **1973**, 2521.
(70) S. Danishefsky, J. Quick, and S. B. Horwitz, *ibid.*, **1973**, 2525.
(71) A. G. Schultz, *Chem. Rev.*, **73**, 385(1973).

ACKNOWLEDGMENTS AND ADDRESSES

Received from the *Department of Chemistry, Pennsylvania State University, University Park, PA 16802*

The authors are grateful to the National Institutes of Health for Grant CA-11450.

*To whom inquiries should be directed.

RESEARCH ARTICLES

Comparison of Criteria for Content Uniformity

BRUCE FLANN

Abstract □ The criterion of lot quality for pharmaceutical unit doses based on the percentage of defectives has been generally used in the evaluation of tests determining content uniformity. This criterion is rejected in favor of joint criteria based on the deviation of the lot mean content from label claim and on the standard deviation of the unit content within the lot. By utilizing methods of simulation by computer, several tests by attributes, including those tests currently in the USP and the NF, are compared with selected tests by variables for reliability, flexibility, and simplicity. (Robustness with respect to type of distribution is rejected as a criterion because it is dependent on the definition of lot quality.) Several tests by variables are quite superior to tests by attributes with respect to reliability and flexibility. Tests for mean content and for weight variation are also examined.

Keyphrases □ Content uniformity—comparison of criteria, percentage of defectives compared to deviation of lot mean content from label claim and standard deviation of within lot unit content, computer simulations □ Statistical evaluation, content uniformity criteria—percentage of defectives method compared to joint use of deviation of lot mean content from label claim and standard deviation of within lot unit content, computer simulations □ Tablets and capsules, content uniformity—comparison of criteria for determination, computer simulations, statistical evaluations

The Pharmaceutical Manufacturers Association of the United States, the United States Pharmacopeial Convention, and the National Formulary Board have pioneered in the field of quality control of the finished pharmaceutical with official recognition of the need for a criterion of content uniformity among single-dosage units. Since 1965, the USP (1) and NF (2) have described, for some tableted pharmaceuticals, a test of this uniformity. Subsequently, they have included a special variation for application to capsules.

A review by Olson and Lee (3) mentioned that the statistics of acceptance sampling was developed during World War II for the armament industry, and shortly after the technique was applied by the pharmaceutical industry to the control of tablet weight (4). In 1960, F. Wiley of the Food and Drug Administration presented data (5) showing that the test controlling tablet weight variation was not controlling content variation. He suggested the replacement of this test with a test for content uniformity. Both old and new tests required counting the number of units beyond an acceptance range (for weight or assay, respectively) and the number beyond double-the-range where the midpoint of the acceptance range was a function either of average tablet weight or average assay.

The tests subsequently accepted by the NF and the USP are of the sequential type requiring 10 or 30 assays. Decisions are based on the number of assays beyond the range 85–115% of label claim. Also, in some cases a sample is automatically rejected if any assays are beyond the range 75–125% of label claim.

Those tests, which are based on the number of measurements (made on single units from a sample) outside an acceptance range, are called tests by attributes. Examples are described in military specifications (6). Other tests, which are based on the magnitude of one or more variables describing the sample, are called tests by variables. A well-known variation of this class of test is based on the magnitude of "the absolute deviation of the mean from the target value plus a multiple of the standard deviation," as described in another set of military specifications (7) and by Lieberman and Resnikoff (8).

Generally, it is accepted that the tests by variables are the more reliable tests of dispersion; *i.e.*, conclusions are less subject to the vagaries of random sampling, at least where the variable is normally distributed. However, there is ample evidence that the content of tablets is not always normally distributed (4, 9–11). Papers comparing the two types of tests can be divided into two groups: those that suggest that most pharmaceutical lots have an approximately

normal distribution so that the more reliable test by variables should be preferred (4, 12, 13) and those that suggest that sufficient pharmaceutical lots have a nonnormal (and unknown) distribution so that a distribution-independent test by attributes should be preferred (14, 15).

Dunnett and Crisafio (4) compared the performance of several possible tests for weight variation by generating 200 groups of pseudoweights from numbers in a table of random normal deviates. To obtain the weights, they combined the gaussian random numbers with a preselected standard deviation for the lot (by definition, lot mean weight equals 100%). They assessed lot quality by computing the percentage of defective tablets in the lot, *a defective tablet being defined as one that differs in weight from the lot mean by more than 5% of the lot mean.* Then they plotted, for each test, the percentage of samples passing the test against the percentage of defective tablets in the lot. By repeating this process for a series of standard deviations, the *OC* (operating characteristic) curve was estimated.

Subsequently, other authors (14, 16), utilizing computer simulation, obtained estimated *OC* curves for several tests for content uniformity. The method was analogous to that of Dunnett and Crisafio (4).

In the above, lot quality has been equated with percentage defective; in the following article, lot quality is equated jointly with both the deviation of the lot mean content from label claim and the standard deviation of the content of individual units about the lot mean content. Hence, the *OC* curves currently in the literature are two dimensional, giving the probability of a sample from a lot passing the test as a function of the percent defective, *i.e.*, the percentage of the units in a lot with an assay beyond an acceptance range. In contrast, the following curves are three dimensional, giving the probability of a sample from a lot passing as a function of the mean assay of the lot and of the standard deviation of the assay of the units of the lot.

In this article, *OC* curves for a selection of reasonable tests, initially restricted to those requiring 20 assays, are estimated by computer simulation, with the restriction that the simulated assays have a normal distribution about the lot mean. Both the lot mean and the standard deviation are varied systematically. Subsequently, the curves for the more promising tests are obtained for four nonnormal distributions already suggested in the literature (9, 14). Then the effect of varying the number of assays per sample is studied. And, finally, sequential variations of these tests are compared.

The general use of the normal distribution in this work does not imply any assumption of normality in the content distribution in pharmaceutical lots. There is ample evidence (4, 11) that an appreciable proportion of the many distributions of drug content are significantly nonnormal. Even an extreme binodal distribution is possible, although presumably very rare. This same evidence shows that the normal distribution is a reasonable distribution to be used initially in the simulation of pharmaceutical assays.

For practical reasons, a criterion for content uniformity can only be applied in conjunction with a criterion for mean content. These criteria can be applied as a single merged criterion, as two criteria combined in a single expression, or as separate criteria.

EXPERIMENTAL

The simulations were carried out in FORTRAN IV G on a large computer[1]. Storage of the program required 50K bytes, and additional storage of approximately 2–6K bytes was required for the results of each test. Up to 11 min of computing time was required for a run. The results, stored on magnetic tape, were then used in conjunction with other programs to generate instruction tapes for an incremental plotter[2] which, in turn, produced the appended graphs.

Estimation of the operating characteristic curves was made as follows:

1. Twenty-five values of δ, generally increasing from 0.0 in steps of 0.5, were selected to represent the percentage deviation of the lot mean from label claim. All values of δ were taken as positive, since the *OC* curves for lots with means below label claim will be mirror images of those with means above label claim.
2. Twenty-four or more values of σ, generally increasing from 0.5 in steps of 0.5, were selected to represent the standard deviation of the lot about its mean, expressed as a percentage of label claim.
3. Values of δ and σ were taken systematically to form 600 or more pairs of lot parameters.
4. Thirty numbers were randomly generated from a population of numbers with mean 0.0, standard deviation of 1.0, and a selected probability distribution, either normal or nonnormal, as discussed below.
5. By using the first or next pair of lot parameters, 30 assays, expressed as deviations from label claim, were derived by multiplying each of the random numbers by σ and adding to δ.
6. Then using the numbers of assays specified by the test, each sample was evaluated by each test. The result was stored according to the values of the parameters and test number. If the test was of the sequential type, the number of assays required was also stored.
7. Then the program returned to the next pair of parameters under Step 5 until all pairs of parameters had been used. Then the program returned to the generation of the next set of 30 assays under Step 4 and so on until a thousand sets of assays had been generated.
8. The results were expressed as percentages of samples passing according to the values of the parameters and of the test numbers. For sequential tests, the average numbers of assays required were similarly calculated.
9. The *OC* curves were drawn.
10. The *ASN* (average sample number) curves for sequential tests were also drawn in cross section or in a three-dimensional representation.

The following scheme was used to generate pseudorandom numbers of mean 0.0, standard deviation 1.0, and the selected probability distribution.

A library function, RANDU, is available. Upon input of a suitable random integer, it generates both a new integer for the next cycle and a number between 0.0 and 1.0, which has uniform probability of being equal to any value in this range. To generate a normally distributed random number, another library function, GAUSS, calls RANDU 12 times and adds the 12 numbers to obtain a new number, which belongs to a population with mean 6.0 and a virtually normal distribution. Subtracting 6.0 from this number and multiplying by a normalizing factor yield the desired gaussian random number. (GAUSS cannot generate numbers more than 5 standard deviations from the mean, and presumably the frequency will be low before this limit is reached.)

The subroutine GAUSS cannot be used directly when nonnormal distributions are desired. Instead, the desired distribution is

[1] IBM 360/85 computer (System-Dimensions Ltd., Ottawa, Canada).
[2] Calcomp 663 plotter.

Table I—Percentage[a] of Single-Dose Units more than 15.05% from Label Claim

Standard Deviation about Mean Assay, %	Deviation of Population Mean Assay from Label Claim, %												
	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0
1.0														1.9
2.0											0.5	2.0	6.2	15.2
3.0								0.3	0.9	2.1	4.5	8.5	15.3	
4.0					0.2	0.5	1.1	2.1	3.9	6.3	10.2	15.4		
5.0	0.2	0.2	0.4	0.7	1.3	2.1	3.5	5.2	7.6	11.1	15.5			
6.0	1.1	1.2	1.6	2.2	3.2	4.6	6.4	8.7	11.8	15.5	20.1			
7.0	3.1	3.2	3.8	4.6	5.9	7.5	9.6	12.4	15.6	19.4				
8.0	6.1	6.2	6.8	7.7	8.9	10.9	13.1	15.9	19.1					
9.0	9.4	9.7	10.1	11.0	12.6	14.4	16.6	19.3						
10.0	13.2	13.2	13.9	14.9	16.2	17.9	20.1							
11.0	16.9	17.2	17.7	18.6	19.9									
12.0														

[a] Only percentages between 0.15 and 20.55% are included in Table I. The sample size is 30,000.

expressed as the sum of n components, all with normal distributions, where the ith component is present in proportion P_i and has a mean of $\bar{\epsilon}_i$ and a standard deviation of σ_i'.

Since the sum of the proportions must equal 1.0 and, by choice, the mean and standard deviation of the random numbers are to be 0.0 and 1.0, respectively, the values of P_i, $\bar{\epsilon}_i$, and σ_i' must be found so that:

$$\sum^{n} P_i = 1.0 \qquad \text{(Eq. 1)}$$

$$\sum^{n} P_i \bar{\epsilon}_i = 0.0 \qquad \text{(Eq. 2)}$$

$$\sum^{n} P_i(\bar{\epsilon}_i^2 + \sigma_i'^2) = (1.0)^2 \qquad \text{(Eq. 3)}^3$$

Now a scale of unit length can be subdivided into n regions such that the boundaries of the ith subregion are located at $\Sigma^{n-i-1} P_i$ and $\Sigma^{n-i} P_i$. The program calls the next random number from RANDU and determines the subregion in which it falls. If this is the ith subregion, the next gaussian number to be generated is assigned to the ith component, *etc.*

As a crosscheck, the program[4] obtains the mean, standard deviation, coefficient of skewness, and coefficient of kurtosis of the population of numbers generated.

To compare different tests, certain arbitrary numerical limits are required. The value 15.05, found in the description of the tests in several tables, reflects the 85.0–115.0% acceptance range for single-dosage units in the current pharmacopeial tests. Similarly, the value 25.05, required only for a few tests, reflects the rejection range for single-dosage units. The value 7.55 was chosen as the limiting value for the standard deviation and for the deviation of the mean from label claim, both expressed as a percentage of label claim. By use of these arbitrary limits, the tests were described so that the specifications implied by the different tests were as nearly equivalent as possible, thus simplifying the intercomparison.

Also, to compare tests, the four following criteria were developed: reliability, robustness, flexibility, and simplicity.

Ideally a test for content uniformity will always pass or always fail a particular lot of tablets according to its degree of uniformity. In practice, as can be seen from the figures, there is a region of limited uniformity in which a lot sometimes passes and sometimes fails, depending on the vagaries of random sampling. The smaller this region, the more reliable is the test.

An *ad hoc* estimator of the reliability of the various tests is given by Eq. 4, in which "area" refers to the area between that contour indicated by the subscript and the axes in the *OC* curves found in Figs. 2–5:

$$\text{reliability} = \left(1.0 - \frac{\text{area}_{5\%} - \text{area}_{95\%}}{2 \times \text{area}_{50\%}}\right) \times 100\% \qquad \text{(Eq. 4)}$$

This definition arbitrarily equates the case, where the area between the 5 and 95% contours is twice that between the 50% contour and the axes, with 0% reliability.

Robustness of a statistical test refers to its ability to come to the same conclusion, on the average, for populations of a given quality, regardless of the violation of assumptions. Here an arbitrary measure of the robustness of selected tests with respect to type of distribution was obtained by estimating the deviation of the *OC* curve (still in three dimensions) for a test as applied to a set of four populations, each with a particular nonnormal distribution from the corresponding curve for normal distributions. This was carried out by computer simulation as already described except that the output was in the form of tables, the 12 columns of which correspond to the 12 integral values of the deviation of the mean from label claim from 0.0 to 11.0 inclusive and the 14 rows of which correspond to the standard deviation of units about population mean from 1.0 to 14.0 inclusive. The values in the body of the tables give the percentage of times for which the samples from respective populations passed the particular test. After tables were generated for four nonnormal and the normal distribution, four tables of differences were obtained by subtracting the normal table from the four nonnormal tables. Then the absolute values of all the numbers in each table of differences were added. This sum is proportional to the volume between the curves, and the larger it is the less robust is the test. This sum was arbitrarily divided by 4000 and subtracted from 1.0. As shown by Eq. 5, this difference multiplied by 100% was taken as a measure of robustness:

$$\text{robustness} = \left[1.0 - \frac{\sum^{n=4} \begin{pmatrix}\text{volume}\\ \text{between}\\ \text{curves}\end{pmatrix}}{4000}\right] \times 100\% \qquad \text{(Eq. 5)}$$

As with the coefficient of reliability, a value of 100% would indicate a completely robust test while unsatisfactory tests would have a low or even negative value for this coefficient.

[3] Define f_i as the frequency of the ith component in an unlimited number, N, of the random numbers. Then:

$$P_i = f_i/N$$

By the definition of variance:

$$(\sigma_i')^2 = [\Sigma(\epsilon_i - \bar{\epsilon}_i)^2]/f_i$$
$$= [\Sigma(\epsilon_i)^2]/f_i] - (\bar{\epsilon}_i)^2$$

Therefore:

$$\sum^{n} P_i[(\bar{\epsilon}_i^2) + (\sigma_i')^2] = \sum^{n} P_i[\Sigma(\epsilon_i)^2]/f_i$$
$$= \left\{\sum^{n}[\Sigma(\epsilon_i)^2]\right\}/N$$

Since, by choice, the mean of all the numbers is 0.0, the latter quantity is the variance of these numbers.

[4] The details of the program and the flowsheet are available on request.

Table II—Comparison of the Distribution of Simulated Assays with That Predicted for a Normal Population

Standard Deviation about Population Mean[a], %	Percentage of Assays beyond the Range 84.95–115.05%		Standard Deviation of Experimental Results, Calculated[b]
	Experimental	Calculated	
5.0	0.20	0.26	0.03
6.0	1.10	1.21	0.06
7.0	3.14	3.16	0.10
8.0	6.06	6.04	0.14
10.0	13.15	13.23	0.20
12.0	21.03	20.98	0.24

[a] In this case, the population mean is the label claim (100%). [b] Calculated through Eq. 6.

Flexibility, the adaptability of the basic test criteria to different test formats without, in effect, altering the definition of acceptable lot quality, was not quantified.

And simplicity, the ease with which the basic test criteria or the format are understood, remembered, and applied, was not quantified either.

TESTS EXAMINED AND RESULTS

The normality of the pseudorandom numbers can be tested when the computer program includes the requirement that the number of defective units in each group of 30 assays for each combination of lot mean and standard deviation be stored by addition, and then the sum for 1000 groups of 30 can be converted to a percentage. Representative values of those obtained are given in Table I. Certain of these values, restricted to those for populations with mean equal to the label claim, are compared with calculated values in Table II. The calculated values were obtained by interpolation in a table of the normal distribution (17), where z equals the ratio of 15.05% to the standard deviation of the lot about the population mean. The standard deviation to be expected in the proportion of experimental, *i.e.*, computer-generated, assays beyond the acceptance range can be calculated with the aid of Eq. 6:

$$\sigma'^2 = P(1 - P)/N \qquad \text{(Eq. 6)}$$

as derived in Ref. 18, for binominal distributions. The experimental and calculated values are in good agreement. There is an indication of a low frequency in the tails of the distribution, but this will have a trivial effect on the *OC* curves.

For all computer simulations of 1000 groups of 30 numbers from a given distribution, the same set of random numbers was used for practical reasons. However, for different distributions, different sets of numbers were generated.

Test of Mean Content—Very frequently, pharmacopeial monographs specify that one weigh, powder, and mix 20 dose units and assay an aliquot. The lot is accepted if the assay is not beyond certain limits. Although the variance is larger, this assay has the same expected value as the mean of the assays of 20 individual dose units. By computer simulation, the probability of acceptance by this latter test for lots of various means and standard deviations is shown graphically in Fig. 1. In this and other cases, the normal distribution of content within lots is used unless the nonnormal is specified.

Tests for Content Uniformity—The numbering system for distinguishing between tests uses a three-component number with components joined by hyphens. The first component, the type number, lies between 100 and 199 if the test is a test by variables; otherwise, it lies between 0 and 99. The second component indicates the number of possible conclusions. And the third number indicates the number of assays required by the test. For sequential tests, this number can take the form 10/20, 10/30, or 10/20/30. Variations of sequential tests are distinguished by a letter following the number of assays. Generally, tests are described and discussed in the order in which the results appear in the tables. Comparison of types of tests are generally made with variations

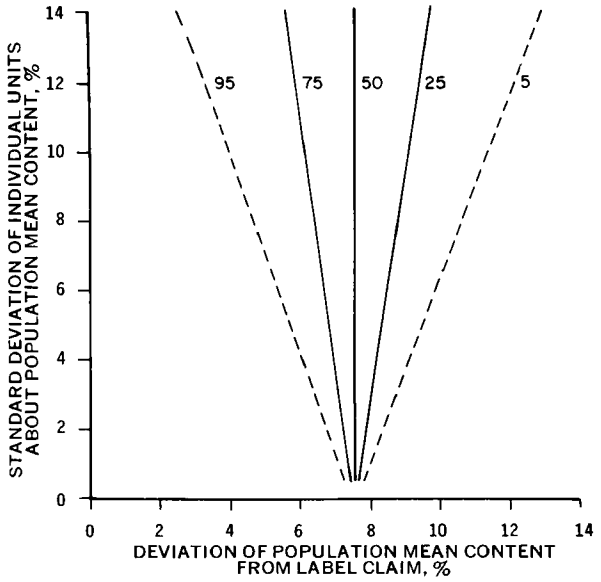

Figure 1—*The* OC *curve of an idealized form of the pharmacopeial test for mean content. The sample passes if the mean of 20 individual assays is not greater than 7.55% from label claim. The five contour lines, labeled 95, 75, 50, 25, and 5, respectively, indicate the probability of samples passing the test, expressed as a percentage.*

requiring a fixed number of assays, generally 20 assays. All tests are described in terms of percentages of label claim.

Comparison of Nonsequential Tests for Reliability and Robustness—

Test 10-2-20: If the mean assay for 20 dose units is beyond the range 92.45–107.55% of label claim or if two or more individual assays are beyond the range 84.95–115.05%, the lot fails.

Test 80-2-20: If seven or more individual assays from 20 dose units are beyond the range 92.45–107.55% of label claim, the lot fails.

Test 90-2-20: If the mean assay for 20 dose units is beyond the range 92.45–107.55% of label claim, or if three or more individual assays are beyond the range 87.74–112.26%, or if any individual assays are beyond the range 81.88–118.12%, the lot fails.

Test 100-2-20: If the mean assay for 20 dose units is beyond the range 92.45–107.55% of label claim or if the standard deviation of these assays is greater than 7.55%, the lot fails.

Test 120-2-20: If the mean assay for 20 dose units is beyond the range 92.45–107.55% or if the sum of the (positive) deviation of the mean assay from label claim (100%) plus 2.0 times the standard deviation of the same assays is greater than 15.05%, the lot fails.

Test 121-2-20: This test is the same as 120-2-20, except the factor 1.8 is used instead of 2.0.

Test 124-2-20: This test is the same as 121-2-20, except 2.4 times the mean deviation is used instead of 1.8 times the standard deviation.

Test 130-2-20: If the sum of the (positive) deviation of the mean assay of 20 dose units from label claim (100%) plus 0.90 times the standard deviation of these assays is greater than 7.55%, the lot fails.

Test 134-2-20: This test is the same as 130-2-20 except 1.2 times the mean deviation is used instead of 0.90 times the standard deviation.

Test 140-2-20: If the sum of the square of the deviation of the mean assay of 20 dose units from label claim (100%) plus the variance, *i.e.*, the square of the standard deviation of these assays, is greater than $(7.55\%)^2$, the lot fails.

Test 150-2-20: If the mean of the (absolute) deviations of the assays of 20 dose units from label claim (100%) is greater than 7.55%, the lot fails.

Test 160-2-20: If the mean square of the deviations of the assays of 20 dose units from label claim (100%) is greater than $(7.55\%)^2$, the lot fails.

Table III—Comparison of Basically Different Tests for Content Uniformity (All Tests Require 20 Assays)

Test	Class of Test	Abbreviated Description of Test[a], Sample Passes If:	Estimate of Reliability[b], %	Estimate of Robustness[c], %	*OC* Curve, Figure
10-2-20	By attributes	$\lvert\bar{X}\rvert \leq 7.55$, $(No._{20} > 15.05\%) \leq 1$	49	46	2A
80-2-20	By attributes	$(No._{20} > 7.55\%) \leq 6$	41	−3	2B
90-2-20	By attributes	$\lvert\bar{X}\rvert \leq 7.55$, $(No._{20} > 12.26\%) \leq 2$ and $(No._{20} > 18.11\%) = 0$	55		2C
100-2-20	By variables	$\lvert\bar{X}\rvert \leq 7.55$, $S \leq 7.55$	53	73	3A
120-2-20	By variables	$\lvert\bar{X}\rvert \leq 7.55$, $(\lvert\bar{X}\rvert + 2S) \leq 15.05$	60		See 3B[d]
121-2-20	By variables	$\lvert\bar{X}\rvert \leq 7.55$, $(\lvert\bar{X}\rvert + 1.8S) \leq 15.05$	59	73	3B
124-2-20	By variables	$\lvert\bar{X}\rvert \leq 7.55$, $(\lvert\bar{X}\rvert + 2.4M) \leq 15.05$	58		See 3B[e]
130-2-20	By variables	$\lvert\bar{X}\rvert + 0.9S \leq 7.55$	52		3C
134-2-20	By variables	$\lvert\bar{X}\rvert + 1.2M \leq 7.55$	52		See 3C[e]
140-2-20	By variables	$(\bar{X})^2 + S^2 \leq (7.55)^2$	58		See 3E[e]
150-2-20	By variables	$\Sigma\lvert\bar{X}\rvert/N \leq 7.55$	51	40	3D
160-2-20	By variables	$(\Sigma X^2)/N \leq (7.55)^2$	59	73	3E
170-2-20	By variables	$(\Sigma\lvert X^3\rvert)/N \leq (7.55)^3$	60		3F

[a] The symbols used in this column have the following meanings: X = deviation of individual assays from label claim expressed as a percent; $\bar{X}$ = mean value of X; S = standard deviation of X about $\bar{X}$; M = mean deviation of X about $\bar{X}$; N = number of values of X involved, *i.e.*, 10, 20, or 30, in the generalized format of certain tests by variables. Similarly, superscript or subscript 10, 20, or 30 indicates the total number of assays in question. [b] Reliability is defined by Eq. 4 in the text. [c] Robustness is defined by Eq. 5 in the text. The estimates of robustness have been brought forward from Table V. [d] Similar to Fig. 3B; however, the contour lines are closer to the origin. [e] Virtually the same as the quoted figure.

Table IV—Characterization of Normal and Nonnormal[a] Populations Utilized in Evaluating the Robustness of Certain Tests for Content Uniformity

	Description of Components				Description of Population Generated[b]			
					Moments about Zero		Coefficient of	
Type of Distribution	Number of Components	Proportion of Population	Standard Deviation (Relative)[c]	Mean (Relative)[c]	First	Second	Skewness	Kurtosis
Normal	1	1.000	1.000	0.000	0.003	1.008	−0.004	2.886
Platykurtic	2	0.500 0.500	0.830 0.830	1.000 −1.000	0.003	1.000	−0.005	2.280
Leptokurtic	2	0.500 0.500	0.500 2.000	0.000 0.000	−0.007	0.992	−0.059	5.149
Leptokurtic skewed	2	0.910 0.090	1.000 1.000	−0.317 3.205	0.011	1.023	1.037	4.486
Platykurtic	10	0.100 (all)	1.000 (all)	$[-2.250 + (n - 1) \times 0.500]$, where $n = 1, 2, \ldots, 10$	−0.008	0.999	−0.009	2.456

[a] The nonnormal populations correspond to those suggested by Haynes *et al.* (9, 14). [b] Based on a sample of 30,000 numbers. [c] The relative values must be divided by a normalizing factor before use so that Eq. 3 of the text is not violated.

Table V—Estimation of Robustness of Selected Tests for Content Uniformity (All Tests Require 20 Assays)

			Scores[b] for Nonnormal Distributions[c]					Estimate of Relative Robustness,[d] %
Test	Class of Test	Abbreviated Description of Test[a], Sample Passes If:	Platykurtic	Leptokurtic	Leptokurtic Skewed	Platykurtic (10 Components)	Total Score	
10-2-20	By attributes	$\lvert\bar{X}\rvert \leq 7.55$, $(No._{20} > 15.05\%) \leq 1$	279	645	1015	202	2141	46
80-2-20	By attributes	$(No._{20} > 7.55\%) \leq 6$	407	1853	1686	173	4119	−3
100-2-20	By variables	$\lvert\bar{X}\rvert \leq 7.55$, $S \leq 7.55$	131	423	428	104	1086	73
121-2-20	By variables	$\lvert\bar{X}\rvert \leq 7.55$, $(\lvert\bar{X}\rvert + 1.8S) \leq 15.05$	124	390	453	92	1059	73
150-2-20	By variables	$(\Sigma\lvert\bar{X}\rvert)/N \leq 7.55$	269	949	1004	161	2383	40
160-2-20	By variables	$(\Sigma X^2)/N \leq (7.55)^2$	108	423	434	120	1085	73

[a] Table III, Footnote *a*, defines the symbols required by this column. [b] The "score" is proportional to the scalar volume between *OC* curves of the nonnormal distribution and the normal distribution. [c] The distributions are described in Table IV. [d] Robustness is defined by Eq. 5 in the text.

Test 170-2-20: If the mean cube of the (absolute) deviations of the assays of 20 dose units from label claim (100%) is greater than $(7.55\%)^3$, the lot fails.

The numerical results obtained for these tests are summarized in Table III. Figures 2 and 3 compare the application of the tests to samples from normal populations. The *ad hoc* estimates of reliability are derived from these figures. By using the four nonnormal populations described in Table IV, the estimates of robust-

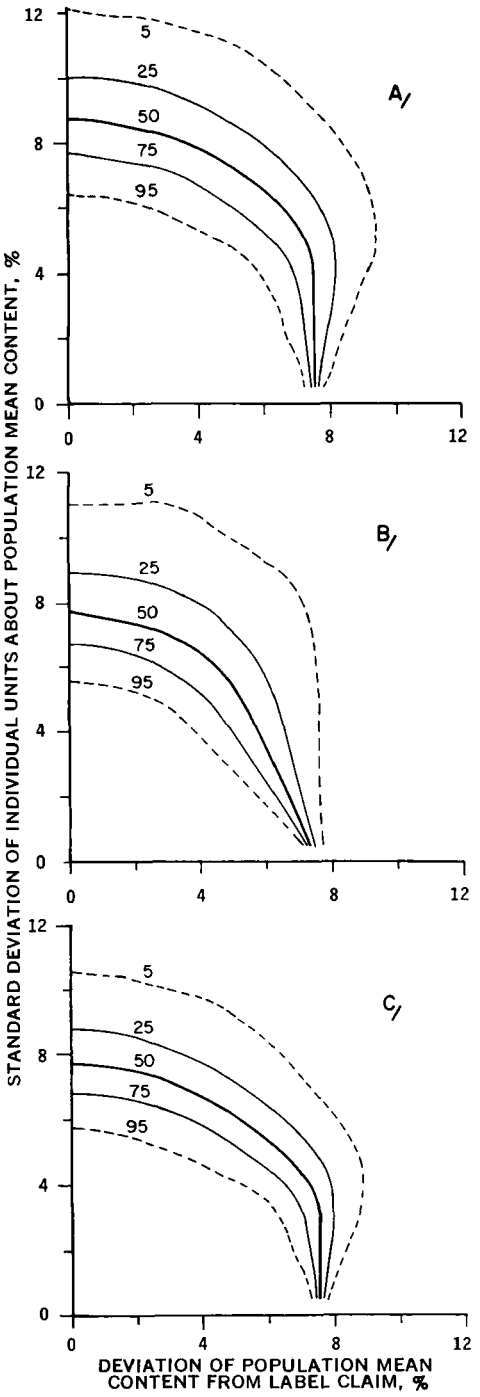

Figure 2—*The* OC *curves for three tests by attributes, each requiring 20 assays, for content uniformity. The tests are: (A) No. 10-2-20,* $|\bar{X}| \leq 7.55$, $(No._{20} > 15.05\%) \leq 1$; *(B) No. 80-2-20,* $(No._{20} > 7.55\%) \leq 6$; *and (C) No. 90-2-20,* $|\bar{X}| \leq 7.55$, $(No._{20} > 12.26\%) \leq 2$, $(No._{20} > 18.11\%) = 0$. *The symbols are defined in Footnote* a, *Table III.*

ness are derived within Table V which, in turn, is derived from tables giving in detail the differences between the *OC* curves obtained with normal and nonnormal distributions. Since the significance of the estimates of robustness with respect to pharmaceutical quality is questioned, only a few values were derived for comparison with earlier work by Haynes *et al.* (14).

The results of Table III clearly show that the reliabilities of the tests by attributes are lower than most of the tests by variables. Nevertheless, sequential variations of these tests were examined to facilitate comparison with the work by Haynes *et al.* (14). On the other hand, Test 170-2-20 is not considered further, despite the high efficiency, because it seems unlikely that the adverse effects of unit doses varying from label claim will vary as the third power of this variation. Tests 120-2-20 and 124-2-20 also are not considered further because of their resemblance to Test 121-2-20. Test 90-2-20, suggested by Haynes *et al.* (14), is not considered further because it is not fundamentally different from Test 10-2-20 when lot quality is measured in terms of mean content and standard deviation. Also, those tests with relatively low reliabilities (<56%) are not considered further.

This latter decision is somewhat arbitrary in view of the differences in shape and position of the 50% contour line for the various tests. Test 100-2-20 probably has a moderately low reliability simply because it encloses a relatively high proportion of points representing high standard deviations. However, the premise of the test, that the acceptable limits for the deviation of the mean and for the standard deviation are mutually independent, seems unlikely. (In contrast, Tests 130-2-20 and 134-2-20 seem to be unduly harsh with respect to lots with both a moderately high deviation of the mean and a moderately high standard deviation, say 5%.)

Comparison of Basically Different Types of Sequential Tests for Reliability and Robustness—For practical reasons, only those tests requiring 10 assays with an optional 20 additional assays are considered. In the following discussion, *acceptance range* means the range 92.45–107.55% of label claim.

Test 11-2-10/30: If the mean assay for 10 dose units is beyond the acceptance range, or if two or more individual assays are beyond the range 84.95–115.05%, or if any assay is beyond 74.95–125.05%, the lot fails. If the mean assay is not beyond the acceptance range and no individuals are beyond the range 84.95–115.05%, the lot passes. Otherwise, a total of 30 assays is required. If the mean assay of the 30 is beyond the acceptance range or two or more individuals are beyond the range 84.95–115.05%, the lot fails. Otherwise it passes.

Test 80-2-10/30: If more than five of the first 10 individual assays are beyond the acceptance range, *i.e.*, 92.45–107.55%, the lot fails. If none is beyond the acceptance range, the lot passes. Otherwise, a total of 30 assays is utilized. If more than nine of the 30 assays are beyond the acceptance range, the lot fails. Otherwise it passes.

Test 121-2-10/30: If the mean assay for 10 dose units is beyond 1.1 times the acceptance range or if the absolute value of the mean assay plus 1.8 times the standard deviation of the individuals is greater than 1.3 times 15.05%, the lots fails. If the mean assay is less than 0.9 times the acceptance range and if the absolute value of the mean assay plus 1.8 times the standard deviation of the individuals is less than 0.6 times 15.05%, the lot passes. Otherwise, a total of 30 assays is utilized. If the mean assay of the 30 is beyond the acceptance range or if the absolute value of the mean assay plus 1.8 times the standard deviation of the individuals is greater than 15.05%, the lot fails. Otherwise it passes.

Test 160-2-10/30A: If the mean of the squares of the deviations of the first 10 individual assays from label claim (100%) is greater than 2.0 times $(7.55\%)^2$, the lot fails. If this mean is greater than 1.2 times $(7.55\%)^2$ and the standard deviation of the individual assays is less than 2.0%, the lot fails. If this mean is less than 0.33 times $(7.55\%)^2$, the lot passes. If this mean is less than 0.8 times $(7.55\%)^2$ and the standard deviation of the individual assays is less than 2.0%, the lot passes. Otherwise, a total of 30 assays is utilized. If the corresponding mean of the squares of the deviations of the 30 assays is greater than $(7.55\%)^2$, the lot fails. Otherwise it passes.

The above four tests and the corresponding numerical results are summarized in Table VI and Fig. 4. The *ad hoc* estimates of reliability are derived from Figs. 4A–4D. The quantity, the maximum average sample size, is derived from Figs. 4E–4H and refers only to the special case where the lot standard deviation equals the deviation of the lot mean from label claim. Tables giving the average sample size, for the sequential tests, as a function of lot mean and standard deviation are available on request.

Reliability of Typical Tests as a Function of Number of Assays Required—Table VII compares the reliabilities of five types of tests for 10, 20, and 30 assays. Figure 5 indicates the location of the 5, 50, and 95% contours for each test. The table includes abbreviated descriptions of the tests, all of which have been described before in principle.

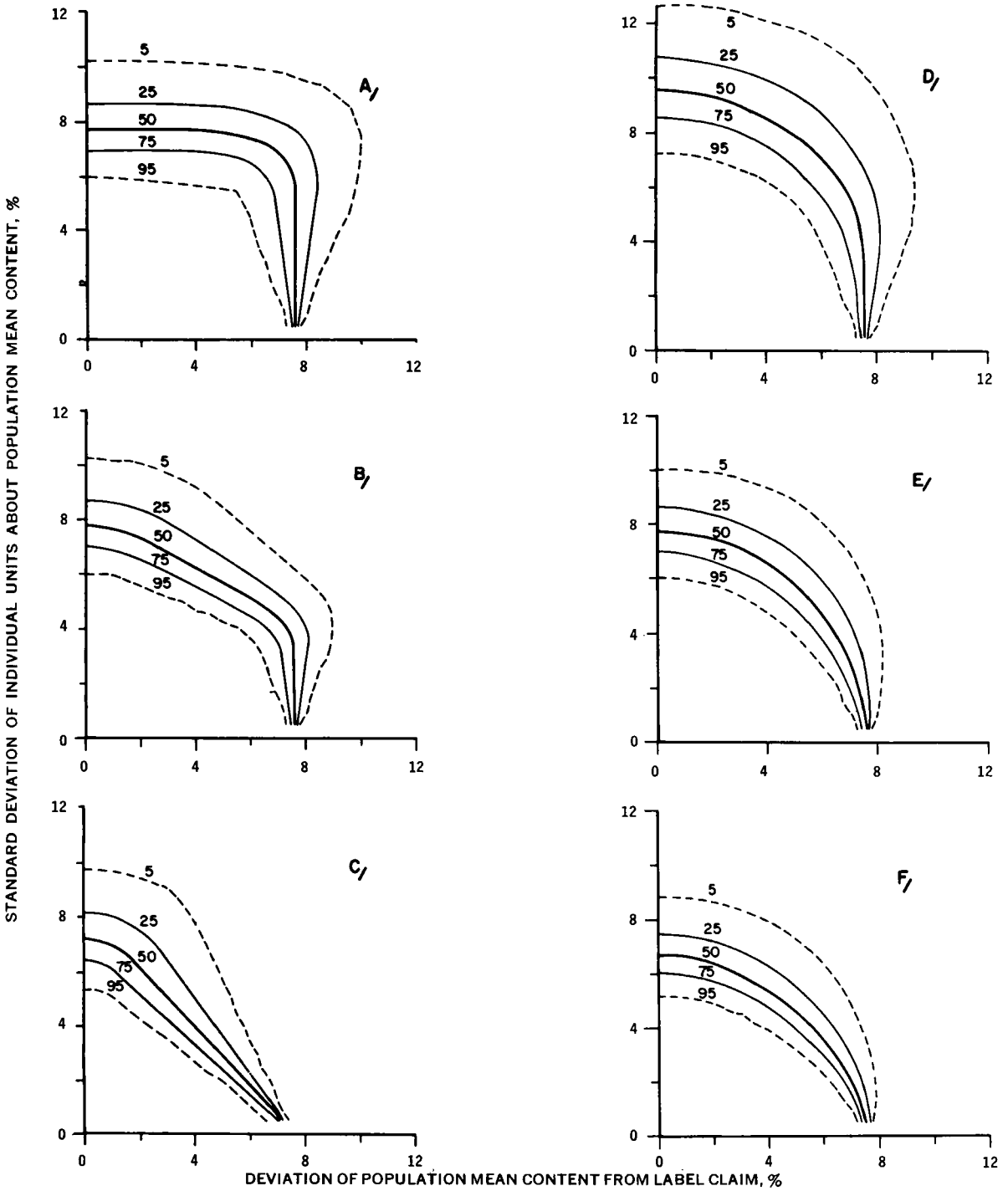

Figure 3—*The* OC *curves for six tests by variables, each requiring 20 assays, for content uniformity. The tests are: (A) No. 100-2-20,* $|\bar{X}| \leq 7.55$, $S \leq 7.55$; *(B) No. 121-2-20,* $|\bar{X}| \leq 7.55$, $(|\bar{X}| + 1.8S) \leq 15.05$; *(C) No. 130-2-20,* $|\bar{X}| + 0.9S \leq 7.55$; *(D) No. 150-2-20,* $\Sigma|\bar{X}|/N \leq 7.55$; *(E) No. 160-2-20,* $(\Sigma X^2)/N \leq (7.55)^2$; *and (F) No. 170-2-20,* $(\Sigma|X^3|)/N \leq (7.55)^3$.

Flexibility of Different Types of Tests for Content Uniformity—Besides the simple format of a fixed number of assays with a simple pass-fail conclusion, sequential formats and formats with three or more graded conclusions may be used advantageously on occasion. An acceptable type of test must not only permit these other formats without introducing too much complexity but must do so without appreciably changing the definition of an acceptable lot, *i.e.*, without changing appreciably the location of the 50% contour. As discussed later in connection with Tables VI and VII, it is already apparent that tests by attributes do not have this flexibility. Only those tests by variables that still appear promising are considered further.

The sequential tests studied were limited, for practical reasons, to those requiring a minimum of 10 assays and a maximum of either 20 or 30, with the number of assays required increasing in steps of 10 or a single step of 20. Graded-conclusion tests were limited to those requiring 10 or 20 assays and having four possible conclusions.

Representative sequential tests are listed in Table VI. These tests can always be designed to have virtually the same reliability

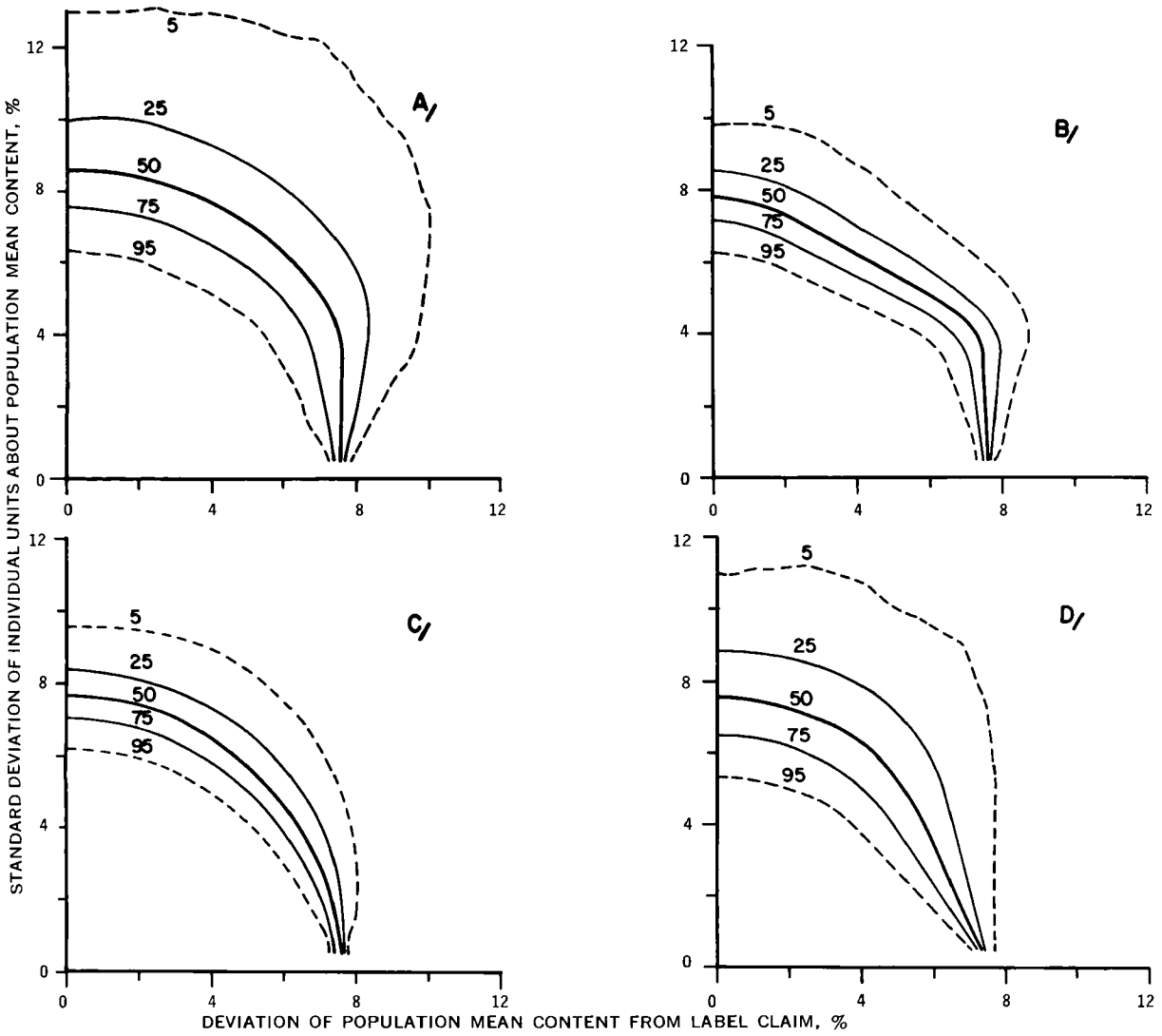

Figure 4A–D—*The* OC *curves for the four sequential tests for content uniformity as completely described in Table VI. The tests and basic criteria are: (A) No. 11-2-10/30,* $|\bar{X}| \leq 7.55$, $(No._n > 15.05\%) \leq R$; *(B) No. 121-2-10/30,* $|\bar{X}| \leq 7.55$, $(|\bar{X}| + 1.8S) \leq 15.05$; *(C) No. 160-2-10/30A,* $(\Sigma X^2)/N \leq (7.55)^2$; *and (D) No. 80-2-10/30,* $(No._n > 7.55) \leq (3n/10)$.

as the corresponding test that has the number of assays equal to the maximum possible number in the sequential version, but there remains the question of whether this design is wasteful by requiring redundant assays. In designing and evaluating sequential tests, it appeared likely that the width of the zone in which additional assays should be required (column 4 of Table VI) should be a function of the magnitude of the standard deviation of the units in the lot. This latter quantity is, in practice, unknown but can be estimated with the standard deviation of the sample. In Table VIII, three types of sequential designs are evaluated, their limitations are observed, and then a modification is suggested.

Three graded-conclusion tests, corresponding to the sequential tests listed as the second variation in Table VIII, are listed in Table IX, and their curves that correspond to *OC* curves are given in Fig. 6.

Comparison of Pharmacopeial Tests for Content Uniformity—The following pharmacopeial tests for content uniformity were altered by the inclusion of the restriction on the magnitude of the mean assay. In contrast with the above sequential tests, the magnitude of the mean assay cannot directly influence the number of assays required.

Test 12-2-10/30: This is the test for tablets in USP XVII (1) and NF XII and XIII (2). This test is identical to Test 11-2-10/30 except that there is no special restriction with respect to assays beyond the range 74.95–125.05%.

Test 13-2-10/30: This is the test for tablets in USP XVIII (1). If the mean assay for 10 tablets is beyond the acceptance range, if any individual assays are beyond the range 74.95–125.05%, or if three or more assays are beyond the range 84.95–115.05%, the lot fails. If the mean assay is not beyond the acceptance range and no more than one is beyond the range 84.95–115.05%, the lot passes. Otherwise, a total of 30 assays is required. If the mean assay of the 30 is beyond the acceptance range or three or more are beyond the range 84.95–115.05%, the lot fails. Otherwise it passes. (Test 13-2-10/30 allows one more defective unit than either Test 11- or 12-.)

Test 14-2-10/30: This is the test for capsules in USP XVIII (1) and NF XIII (2). If the mean assay of 10 capsules is beyond the acceptance range, if any individual assays are beyond the range 74.95–125.05%, or if four or more are beyond 84.95–115.05%, the lot fails. If the mean assay is not beyond the acceptance range and not more than one individual assay is beyond the range 84.95–115.05% (but not beyond 74.95–125.05%), the lot passes. Otherwise, a total of 30 assays is required. If the mean assay of the 30 is beyond the acceptance range, if any assays are beyond the range 74.95–125.05%, or if four or more are beyond the range 84.95–115.05%, the lot fails. Otherwise it passes.

The *OC* curves and the *ASN* curves for these three pharmacopeial tests are shown, in a three-dimensional representation, in Figs. 7 and 8 with, for contrast, the curves of the sequential tests by variables listed in Table VIII. When using Eq. 4, the values for the respective reliabilities of the pharmacopeial tests are 28, 15, and 19% while those of the tests by variables are 64 or 65%.

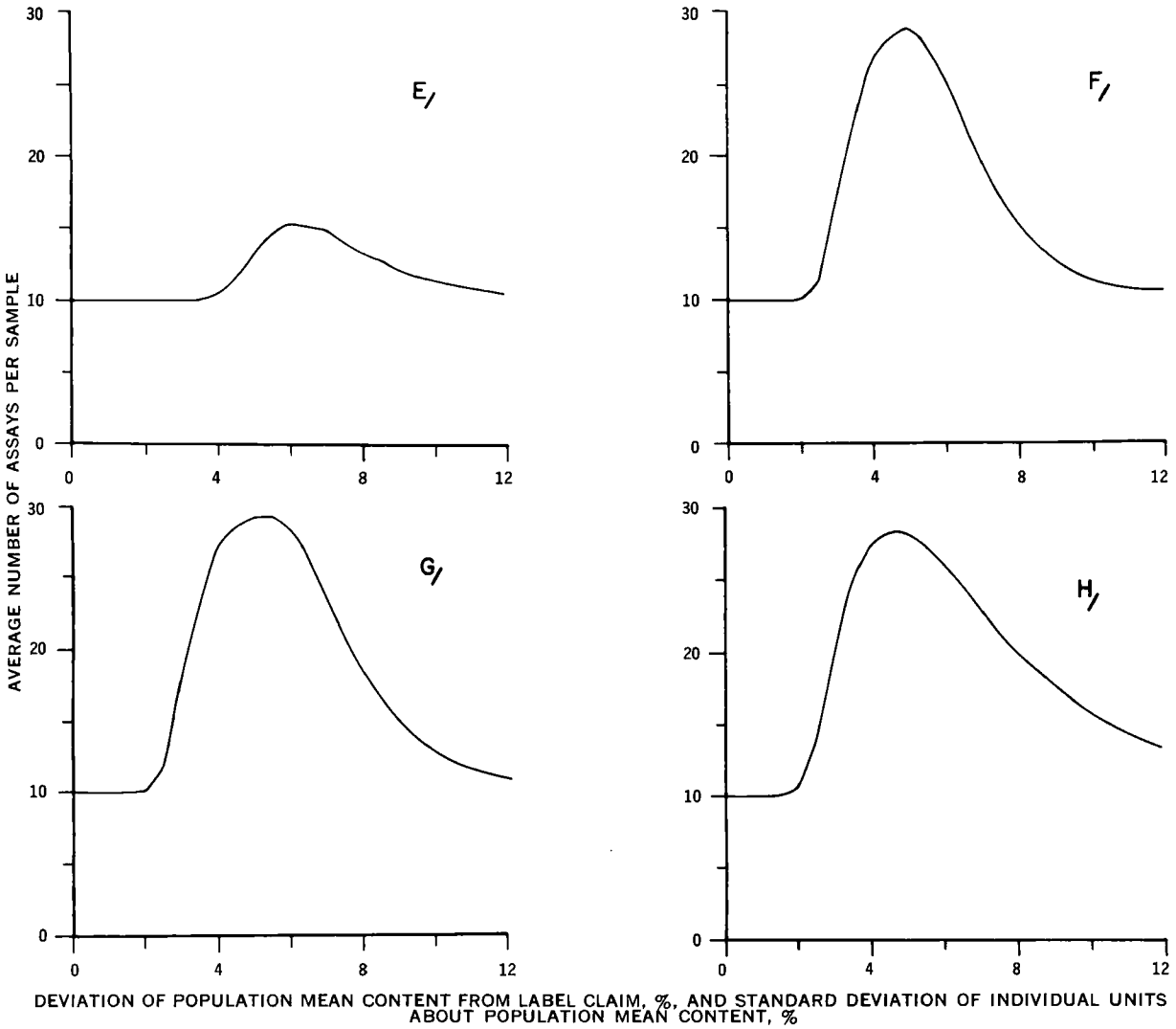

Figure 4E–H—*The* ASN *curves corresponding to the diagonals (starting at the origin) of the graphs in Figs. 4A–4D, taken in the same order. Note that the mean content and the standard deviations will be equal.*

The inclusion, in Test 11-2-10/30, of the restriction that no assays be beyond the range 74.95–125.05% produces a maximum increase in the probability of failing the test of 0.5% when the probability is between 25 and 75% and the lot mean equals label claim. The effect is, therefore, of marginal significance. However, other tests (not listed) show that this restriction increases the probability of failing by approximately 5% for two of the current pharmacopeial tests, 13-2-10/30 and 14-2-10/30. This double restriction, in which the principal criterion corresponds to the definition of defectiveness, is different in principle from Test 90-2-20, suggested by Haynes *et al.* (14), in which two restrictions were evenly balanced about the definition of defectiveness.

DISCUSSION

Error—The location of the contour lines in the *OC* curves is subject to an uncertainty originating in the pseudorandom nature of the simulated assays used in the evaluation of the performance of the test in question. This uncertainty is essentially experimental error. The standard deviation of this error, σ'', for an individual point can be evaluated with Eq. 7, a modified form of Eq. 6. For this equation, N samples, randomly chosen from a population with a given mean and standard deviation, each has a probability, P, of passing the test. The letter, G, refers to the gradient (at the corresponding point) of the appropriate three-dimensional surface. The reciprocal of the gradient, G, converts the standard deviation in percent passing to a unit of length both in the plane perpendicular to the *percent passing* axis and in the direction perpendicular to the contour line:

$$\sigma''^2 = [P(1 - P)/N]/G^2 \qquad \text{(Eq. 7)}$$

As already mentioned, in this work N equals 1000. For the 5, 25, 50, 75, and 95% contours, σ'' equals (0.7%), (1.4%), (1.6%), (1.4%), and (0.7%)/G, respectively. By inspection of the figures, one can see that this quotient is roughly constant for corresponding points on the different contours of a figure and that its magnitude could be used as a measure of reliability.

This value for σ'' applies only to a single point based on its own series of random numbers. Nevertheless, one might surmise, since all the derived points on the *OC* curves are based on the same series of random numbers, that experimental error will affect neighboring portions of the *OC* curve in a corresponding manner, especially for tests by variables.

The location of contours within the figures has been made by linear interpolation within the tables. Presumably the linear aspect of this interpolation has been the source of some of the irregularities in the curves in certain figures.

Criteria of Lot Quality—The definition of lot quality, ideally, requires a pharmacological basis, but for practical reasons pharmaceutical chemists have, for years, temporized. As early as 1951, Smith (19) indicated that the generally accepted criterion of quality with respect to tablet weight was based on the percentage of defectives, *i.e.*, the percentage beyond certain limits in weight. Dunnett and Crisafio (4) used this criterion in the comparison of

Table VI—Estimation of Robustness and Reliability of Selected Tests for Content Uniformity (All Tests Are of the Sequential Type Requiring 10 or 30 Assays)

Test	Class of Test	Basic Test Criteria[a]	Sequential Criteria[a], 30 Assays Required If:	Estimated Robustness[b], %	Estimated Reliability[c], %	Maximum Average Sample Size[d]
11-2-10/30	By attributes	$\lvert\bar{X}\rvert \leq 7.55$ and number of deviations greater than 15.05%	$\lvert\bar{X}_{10}\rvert \leq 7.55$, $(No._{10} > 15.05\%) = 1$, and $(No._{10} > 25.05\%) = 0$	45	28	15
80-2-10/30	By attributes	Number of deviations greater than 7.55%	$(No._{10} > 7.55\%) =$ 1, 2, 3, 4, or 5	−3	37	28
121-2-10/30	By variables	$\lvert\bar{X}\rvert \leq 7.55$, and $(\lvert\bar{X}\rvert + 1.8S) \leq 15.05$	$(\lvert\bar{X}_{10}\rvert \leq 1.1 \times 7.55)$ and $(\lvert\bar{X}_{10}\rvert + 1.8S \leq 1.3 \times 15.05)$ while either $(\lvert\bar{X}_{10}\rvert \leq 0.9 \times 7.55)$ or $(\lvert\bar{X}_{10}\rvert + 1.8S \geq 0.6 \times 15.05)$	69	64	29
160-2-10/30A	By variables	$(\Sigma X^2)/N \leq (7.55)^2$	Case I: $-S_{10} < 2.0$; then $0.8 \times (7.55)^2 \leq (\Sigma^{10} X^2)/10 \leq 1.2 \times (7.55)^2$ Case II: $-S_{10} \geq 2.0$; then $0.33 \times (7.55)^2 \leq (\Sigma^{10} X^2)/10 \leq 2.0 \times (7.55)^2$	78	65	29

[a] Table III, Footnote *a*, defines the symbols required by these columns. [b] Robustness is defined by Eq. 5 in the text. It is determined in an analogous manner to that of Table V. [c] Reliability, defined by Eq. 4 in the text, is derived from Fig. 4. [d] Maximum average sample size is derived from the original data of those tables represented in Figs. 4E–H.

Table VII—Reliability *versus* Number of Assays, and Intercept of 50% Contour on the Standard Deviation Axis *versus* Number of Assays

Test	Class of Test	Abbreviated Description of Test[a]	Estimated Reliability[b], %			Intercept of 50% Contour with Standard Deviation Axis, %		
			10 Assays	20 Assays	30 Assays	10 Assays	20 Assays	30 Assays
10-2-10	By attributes	$\lvert\bar{X}_{10}\rvert \leq 7.55$, $(No._{10} > 15.05\%) = 0$	19	—	—	8.2	—	—
11-2-20[c]	By attributes	$\lvert\bar{X}_{20}\rvert \leq 7.55$, $(No._{20} > 15.05\%) = 0$	—	48	—	—	7.2	—
10-2-30	By attributes	$\lvert\bar{X}_{30}\rvert \leq 7.55$, $(No._{30} > 15.05\%) \leq 1$	—	—	58	—	—	7.9
121-2-...	By variables	$\lvert\bar{X}\rvert \leq 7.55$, $(\lvert\bar{X}\rvert + 1.8S) \leq 15.05$	31	59	65	7.6	7.8	7.9
140-2-...	By variables	$(\bar{X})^2 + S^2 \leq (7.55)^2$	32	58	65	7.4	7.5	7.5
160-2-...	By variables	$(\Sigma X^2)/N \leq (7.55)^2$	30	59	65	7.8	7.7	7.7
80-2-10	By attributes	$(No._{10} > 7.55\%) \leq 3$	1	—	—	8.1	—	—
80-2-20	By attributes	$(No._{20} > 7.55\%) \leq 6$	—	41	—	—	7.8	—
80-2-30	By attributes	$(No._{30} > 7.55\%) \leq 9$	—	—	37	—	—	7.6

[a] Table III, Footnote *a*, defines the symbols required by this column. [b] Reliability, defined by Eq. 4 in the text, is derived from Fig. 5 except for the 140 series. [c] Test 10-2-20, as listed in Table III, is similar except a single assay beyond 15.05% is permitted. The reliability is 49%, and the intercept of the 50% contour on the standard deviation axis is 8.8%.

the reliabilities of several tests for weight variation. Subsequently, authors employed the analogous criterion, *i.e.*, the percentage of assays beyond certain limits, in the evaluation of tests for content uniformity.

Regardless of whether units with assays beyond these limits are called defectives or outsiders, this classification implies that all units within the range are pharmacologically equally satisfactory and those outside the range are equally unsatisfactory. However, data, as recently reviewed by Ritschel (20), indicate that variations in drug content, modified by the accumulative effect of a succession of units administered at intervals, are reflected by variations in blood levels which, in turn, are reflected by variations in pharmacological activity. While these relationships will not necessarily be linearly proportional, they will be generally continuous functions, increasing and decreasing in unison.

This reasoning eliminates the use of an *acceptance range* as an *intrinsic* measure of unit quality and *percentage beyond* as an *intrinsic* measure of lot quality. Rather, it suggests the use of a simple function of the deviation of unit content from label claim as a measure of unit quality and the pooled magnitude as a measure of lot quality.

Of the various possible functions, the second moment of the deviations about label claim has a certain elegance and reasonableness, but it is an unfamiliar and unproven function. Mean drug content and standard deviation are familiar quantities which can be independently related to different aspects of the manufacturing process. Also, without any restriction on the nature of the distribution, they are intrinsic descriptive parameters for each pharmaceutical lot (although they do not completely characterize it). In addition, the mean blood level will be generally proportional to the mean unit content; the standard deviation of the blood level, especially where there is an accumulative effect, will be generally proportional to the standard deviation of the unit content. Therefore, mean drug content and standard deviation are used here jointly as indexes of lot quality (although they do not necessarily completely characterize lot quality in the pharmacological sense).

If mean content and standard deviation did completely characterize lot quality in the pharmacological sense, then the charac-

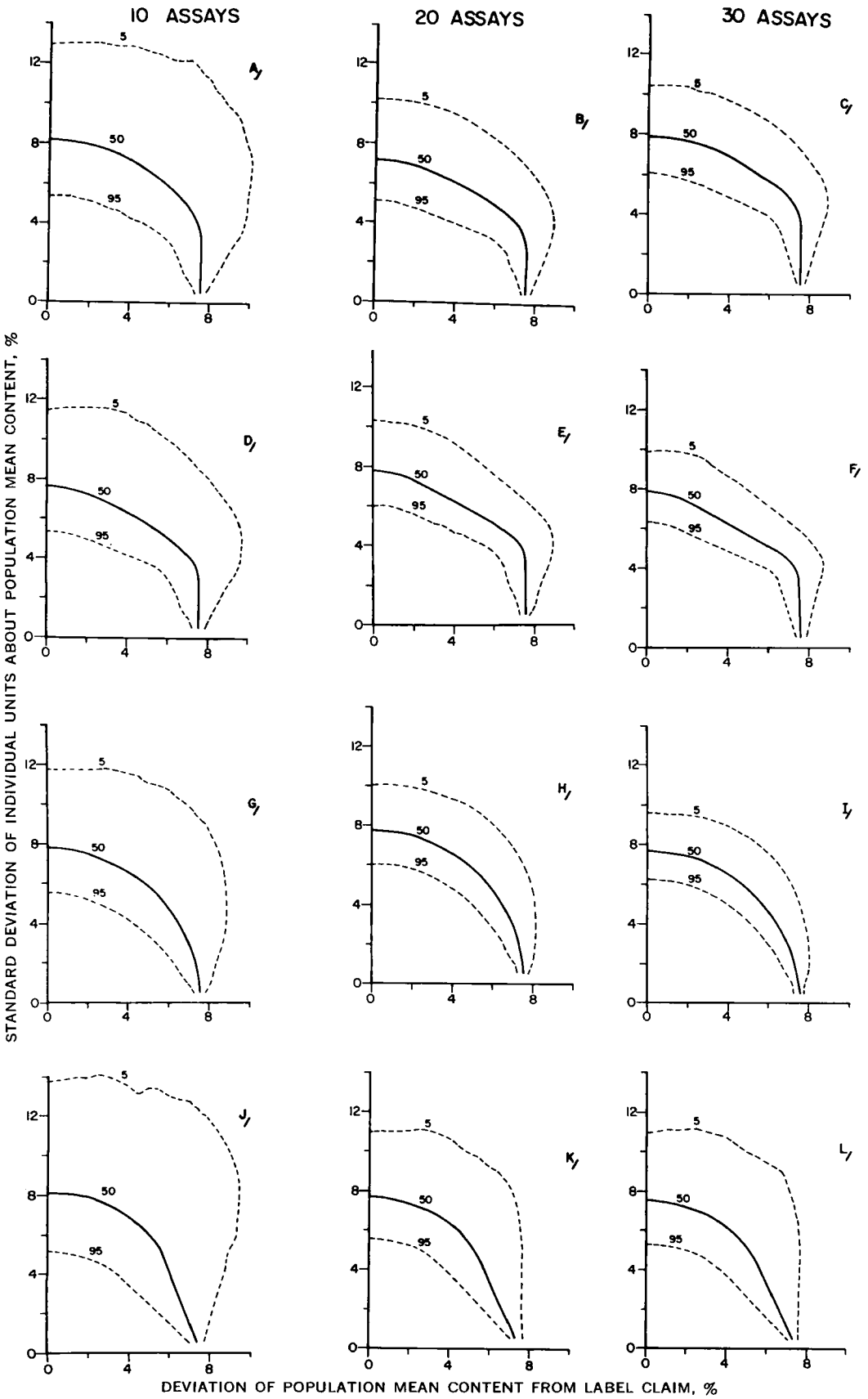

Figure 5—*The* OC *curves for four types of tests for content uniformity, showing the variation with 10, 20, and 30 assays. The tests and criteria are: (A) No. 10-2-10,* $|\bar{X}_{10}| \leq 7.55$, $(No._{10} > 15.05\%) = 0$; *(B) No. 11-2-20,* $|\bar{X}_{20}| \leq 7.55$, $(No._{20} > 15.05\%) = 0$; *(C) No. 10-2-30,* $|\bar{X}_{30}| \leq 7.55$, $(No._{30} > 15.05\%) \leq 1$; *(D, E, F) No. 121-2-. . .,* $|\bar{X}| \leq 7.55$, $(|\bar{X}| + 1.8S) \leq 15.05$; *(G, H, I) No. 160-2-. . .,* $(\Sigma X^2)/N \leq (7.55)^2$; *(J) No. 80-2-10,* $(No._{10} > 7.55\%) \leq 3$; *(K) No. 80-2-20,* $(No._{20} > 7.55\%) \leq 6$; *and (L) No. 80-2-30,* $(No._{30} > 7.55\%) \leq 9$.

Table VIII—Flexibility of Sequential Format[a]

Primary Test Criteria	First Variation		Second Variation	
	Test	30 Assays Required If:	Test	30 Assays Required If:
$\lvert\bar{X}\rvert \lesssim 7.55$ and $(\lvert\bar{X}\rvert + 1.8S) \lesssim 15.05$	121-2-10/30A	$(\lvert\bar{X}_{10}\rvert \lesssim 7.55 + 0.63S)$ and $(\lvert\bar{X}_{10}\rvert + 1.8S \lesssim 15.05 + 0.45S)$ while either $(\lvert\bar{X}_{10}\rvert \gtrsim 7.55 - 0.63S)$ or $(\lvert\bar{X}_{10}\rvert + 1.8S \gtrsim 15.05 - 1.65S)$	121-2-10/30B	$(\lvert\bar{X}_{10}\rvert \lesssim 1.1 \times 7.55)$ and $(\lvert\bar{X}_{10}\rvert + 1.8S \lesssim 1.4 \times 15.05)$ while either $(\lvert\bar{X}_{10}\rvert \gtrsim 0.9 \times 7.55)$ or $(\lvert\bar{X}_{10}\rvert + 1.8S \gtrsim 0.6 \times 15.05)$
$(\Sigma X^2)/N \lesssim (7.55)^2$	160-2-10/30	$0.33(7.55)^2 \lesssim (\Sigma^{10}X^2)/10$ and $(\Sigma^{10}X^2)/10 \lesssim 2.0(7.55)^2$	160-2-10/30C	$(1.0 - 0.15S) \times 0.9 \times (7.55)^2 \lesssim (\Sigma^{10}X^2)/10$ and $(\Sigma^{10}X^2)/10 \lesssim (1.0 + 0.15S) \times 1.1 \times (7.55)^2$
$(\bar{X})^2 + K \cdot S^2 \lesssim (7.55)^2$, where $K = 1.0$ generally	140-2-10/30	$0.33(7.55)^2 \lesssim (\bar{X}_{10})^2 + S^2$ and $(\bar{X}_{10})^2 + S^2 \lesssim 2.0(7.55)^2$	140-2-10/30B	$0.9 \times (7.55)^2 \lesssim (\bar{X}_{10})^2 + 3.0 \times S^2$ and $(\bar{X}_{10})^2 + 0.5 \times S^2 \lesssim 1.1 \times (7.55)^2$

[a] For the six variations listed in the table, the reliabilities are 64 or 65%. Of the first three variations, Test 121-2-10/30A is considered too complicated for routine use while the other two call for redundant assays when S is small. The second three variations are considered satisfactory.

Table IX—Several Multiconclusion Tests for Content Uniformity (All Tests Require 10 Assays)

Test	Sample Passes If[a]	Basic Test Criteria	Sample Fails If[b]	Results and Comments
121-4-10	$\lvert\bar{X}_{10}\rvert < 0.9 \times 7.55$ and $\lvert\bar{X}_{10}\rvert + 1.8S < 0.6 \times 15.05$	$\lvert\bar{X}\rvert \lesssim 7.55$ and $\lvert\bar{X}\rvert + 1.8S \lesssim 15.05$	$\lvert\bar{X}_{10}\rvert > 1.1 \times 7.55$ or $\lvert X_{10}\rvert + 1.8S > 1.4 \times 15.05$	See Fig. 6A; the acceptance and the rejection regions overlap each other equally at the 5% contours
160-4-10B	$(\Sigma^{10}X^2)/N < (1.0 - 0.15S) \times 0.9 \times (7.55)^2$	$(\Sigma X^2)/N \lesssim (7.55)^2$	$(\Sigma^{10}X^2)/10 > (1.0 + 0.15S) \times 1.1 \times (7.55)^2$	See Fig. 6B; similar to above except at the 2% contour; the rejection region is poorly defined
140-4-10	$(\bar{X}_{10})^2 + 3.0S^2 < 0.9 \times (7.55)^2$	$(\bar{X})^2 + S^2 \lesssim (7.55)^2$	$(\bar{X}_{10})^2 + 0.5S^2 > 1.1 \times (7.55)^2$	See Fig. 6C; the acceptance and the rejection regions overlap each other equally at the 15% contours

[a] Samples that do not pass but are within the basic test criteria are classed as *probably acceptable.* [b] Samples that are beyond the basic test criteria but do not fail are classed as *probably unacceptable.*

terization would be distribution free; however, experimental difficulties preclude the determination of the completeness of characterization. (The question whether lot quality as characterized by certain parameters is distribution free should not be confused with the question whether an acceptance test is distribution free, *i.e.*, robust.)

Recently, Oie *et al.* (21) tested, with aspirin (acetylsalicylic acid) tablets, for a correlation between a controlled interdose variation and plasma levels, but the results were not conclusive.

Criteria for Selecting Tests for Quality Control—The primary criterion must be reliability. Due to the random variation in sampling, tests on successive samples from the same lot do not always provide the same conclusion. The risks this involves for both the manufacturer and the regulatory agency are apparent, and the risks must be minimized, at least, to the point where the improvement in the definition of quality becomes trivial in terms of the additional work involved. The estimates of reliability show that some tests are much more reliable than others. Where several tests have the required degree of reliability, the most economical test is to be preferred.

The second criterion must be flexibility. The individual test cannot be flexible in itself; rather the principle on which the test is based must be adaptable to other formats without substantially changing the definition of acceptability, *i.e.*, without substantially changing the location of the 50% contour. Furthermore, since lot quality is taken as a function both of dispersion and of deviation of the mean, the principle of a test should permit the variation of the relative influence of the dispersion and the deviation of the mean. The following two minor aspects of flexibility should also be considered: *(a)* the test, or at least the basic principle of the test, should be applicable to other measurements besides content uniformity, in particular, to weight variation; and *(b)* the test should be directly applicable to the situation where the analytical methods with the necessary precision for single-tablet assays do not have the necessary accuracy for the determination of the mean content. While the criterion of flexibility is difficult to quantify, its use should permit the ranking of tests with comparable reliabilities.

The third criterion is simplicity with respect both to appreciation by persons with quite limited training in statistics and to calculation manually or by desk calculators and computers. This criterion underlies the initial selection of the tests to be examined.

In more specific terms than those preceding Eq. 5, robustness of a test refers to independence, for a given lot quality, from the assumption of normality. Unfortunately, in the development (14, 15) of the current pharmacopeial tests (1, 2), considerable weight was attached to the relative robustness where lot quality was expressed as percentage defective. Until either the relative robustness has been shown to be independent of the definition of lot

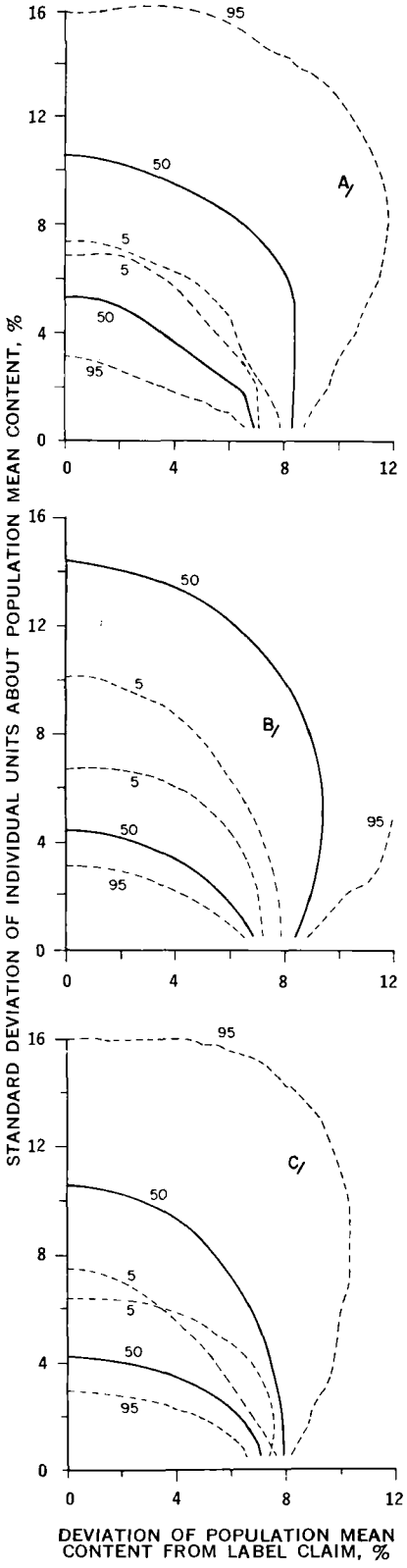

Figure 6—*The analog of* OC *curves for three multigrade tests by variables, each requiring 10 assays, for content uniformity, as completely described in Table IX. The tests and basic criteria are: (A) No. 121-4-10,* $|\bar{X}| \leq 7.55$, $(|\bar{X}| + 1.8S) \leq 15.05$; *(B) No. 160-4-10,* $(\Sigma X^2)/N \leq (7.55)^2$; *and (C) No. 140-4-10,* $(\bar{X})^2 + S^2 \leq (7.55)^2$. *The three contours that intersect the axes closer to the origin indicate the probability of passing the test. The other three contours indicate the probability of failing the test.*

quality or the definition of lot quality is established in pharmacological terms, robustness should not be used as a criterion for selecting tests.

Test of Mean Content—The *OC* curve for the test of mean content (based on the mean of 20 assays) is given in Fig. 1 for comparison with those of the various tests for content uniformity. The figure includes lots with high (14%) standard deviations, not because they are common but because they occasionally do occur, usually where the active ingredient is in a macroscopically nonhomogeneous phase.

It is, perhaps, of greater importance that the figure illustrates the limitations of the usual approach to determining mean content as described in Refs. 1 and 2. The *OC* curve actually depicts the optimum performance of the test. If the mixture from the 20 units is not homogeneous with respect to the size of the weighed portion or if the analytical method (for two or three assays) has a relatively high coefficient of variation, the uncertainty is increased.

While tests for mean content have not been systematically considered, one is suggested here for comparison with the tests for content uniformity. If the absolute deviation from label claim of the mean assay for 10 dose units is greater than 7.55% plus twice the "standard deviation of the ten assays divided by the square root of 10," the lot fails. If the absolute deviation is less than 7.55% minus twice the "standard deviation divided by the square root of 10," the lot passes. Otherwise, 30 additional dose units are assayed (singly or in aliquots of powdered and mixed units). If the absolute deviation is greater than 7.55%, the lot fails. Otherwise it passes.

This test can only be used without modification if the analytical method with the necessary precision for determining the dispersion is sufficiently accurate for the determination of the mean.

Tests for Content Uniformity—As shown by Table III and Figs. 2 and 3, the conclusions reached, based on 20 assays, by the various classes of tests were not radically different. In most of the graphs, the 50% contour forms a rough quarter-circle about the origin. To some degree, this was intentional because it simplifies the intercomparison.

In two cases, for reasons inherent in the test definition, the 50% contours do not form a rough quarter-circle. As shown by Fig. 3A, the 50% contour for Test 100-2-20 is composed basically of a straight line reflecting the condition limiting the mean assay and a second straight line reflecting the condition limiting the standard deviation. In contrast, as shown by Fig. 3C, the 50% contour for Test 130-2-20 (and, by analogy, Test 134-2-20) is a single diagonal line reflecting the single condition limiting the sum of two linear components. For subjective reasons mentioned earlier, these tests are not considered further.

The data of Table III show that the three tests by attributes (Fig. 2) are generally not as reliable as the tests by variables (Fig. 3). This is undoubtedly a consequence of the loss of information in conversion of assay values to single binary digits, *i.e.*, 0 or 1, in the cases of Test 10-2-20 (a modified form of the current pharmacopeial tests) and Test 80-2-20. Test 90-2-20 (Fig. 2C), as suggested by Haynes *et al.* (14), loses much less information with respect to the distribution of assay values and has a correspondingly higher reliability.

Haynes *et al.* (14) showed that tests by attributes (based on the number of defectives) were more robust than a test by variables based on standard deviation, while the data of Table III indicate the converse. This, apparently, reflects the change in criteria of lot quality.

Table VI and Figs. 4A–4D show that sequential tests by variables are decidedly more reliable than the two sequential tests by attributes, one of which, Test 11-2-10/30, is very similar to the current pharmacopeial tests.

By comparison of the maximum average sample size as listed in Table VI or as shown by Figs. 4E–4H, it is apparent that the sequential mechanism of Test 11-2-10/30 fails in its purpose, *i.e.*, to require the maximum number of assays in borderline cases. Apparently, this defect can also be traced back to the loss of information in conversion of assay values to single binary digits.

The values for robustness for sequential tests are virtually the same as for the corresponding values for the tests requiring 20 assays.

Table VII and Fig. 5 show that the reliabilities of the sequential tests just discussed closely approach those for the corresponding

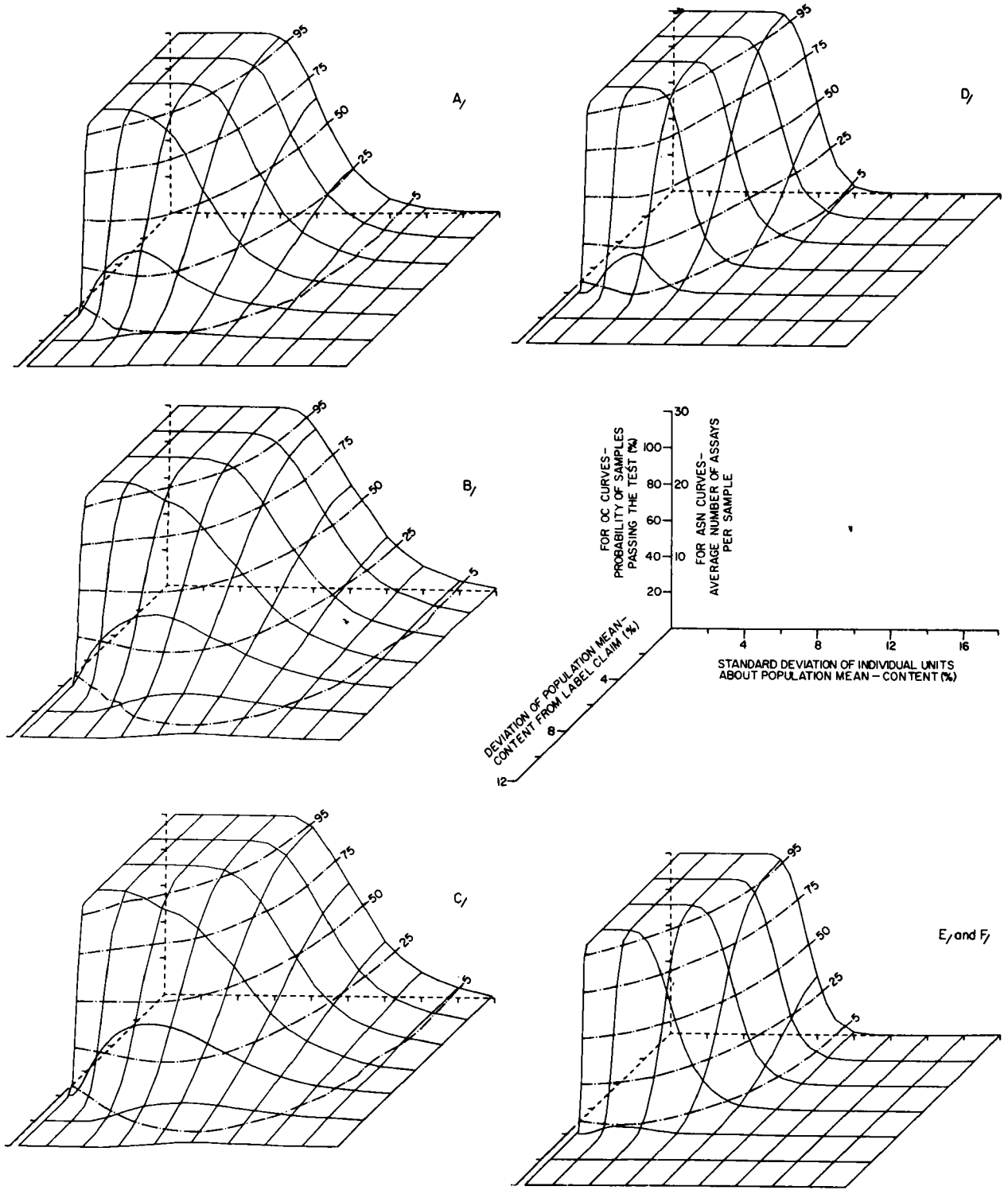

Figure 7—*The* OC *curves for the three current pharmacopeial tests and the three preferred tests by variables for content uniformity. The tests and their source or their basic criteria are: (A) No. 12-2-10/30, NF XIII; (B) No. 13-2-10/30, USP XVIII;* (C) *No. 14-2-10/30, NF and USP method for capsules; (D) No. 121-2-10/30B,* $|\bar{X}| \leq 7.55$, $(|\bar{X}| + 1.8S) \leq 15.05$; *(E) No. 160-2-10/30C,* $(\Sigma X^2)/N \leq (7.55)^2$; *and (F) No. 140-2-10/30B,* $(\bar{X})^2 + S^2 \leq (7.55)^2$. *Since E is nearly identical to F, it has not been included as such.*

tests requiring 30 assays, except for Test 10-2-10/30. This is further evidence that the "sequential mechanism" of this test is unsatisfactory.

Figure 5 also shows the relatively large improvement in reliability obtained in going from 10 to 20 assays compared with going from 20 to 30 assays.

Table VIII contains the essence of attempts to improve the sequential format of the more promising classes of tests. If a class has the desired flexibility, it is possible to design the sequential format so that, for lots with parameters in the vicinity of those of the 50% contours, the number of assays called for is equal or close to the maximum while for other lots the required number of assays is rapidly decreased to the minimum as the lot parameters move away from this contour. The *ASN* curves of Figs. 8D–8F show that the preferred tests by variables do have this flexibility, in contrast with the current pharmacopeial tests (Figs. 8A–8C).

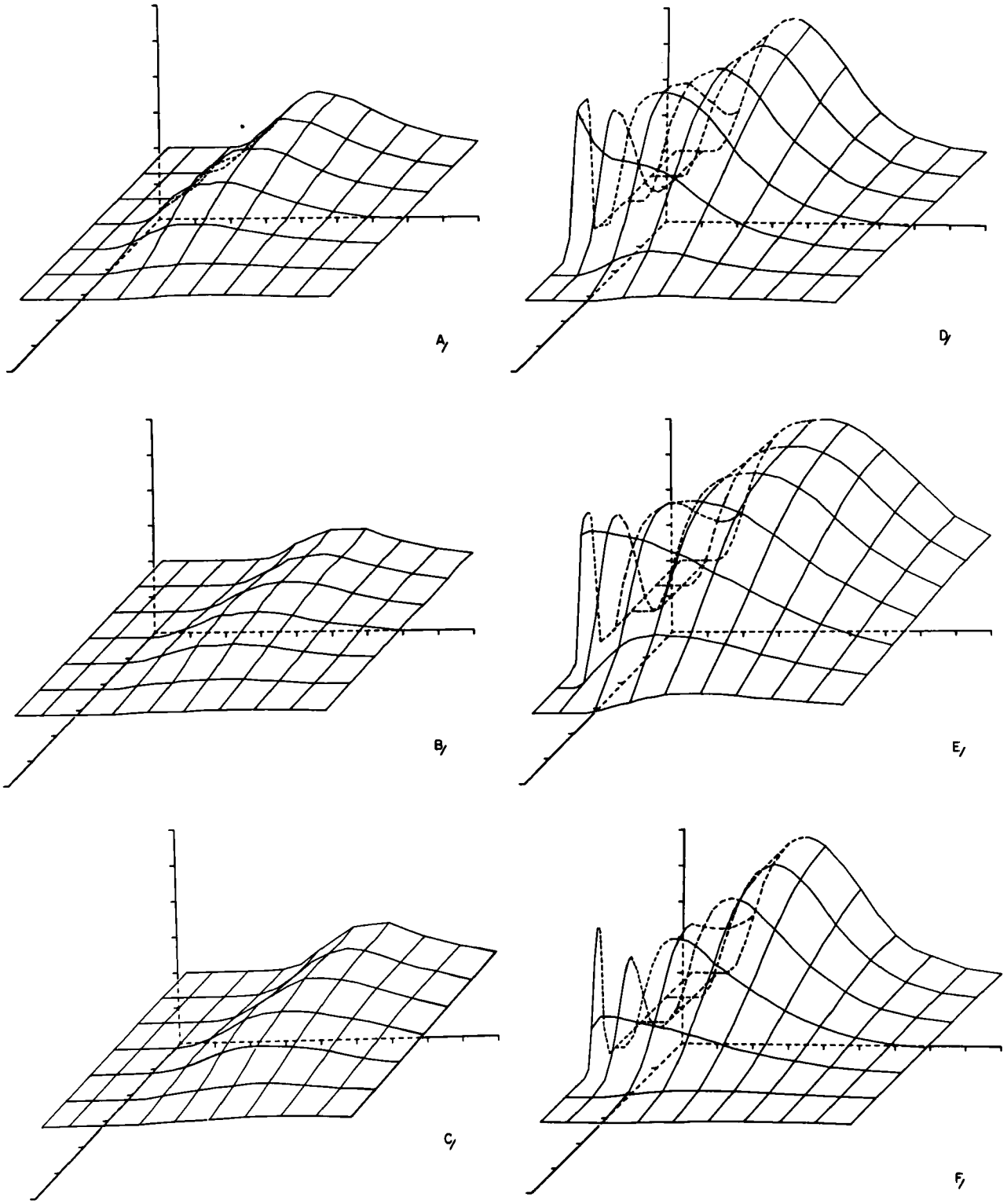

Figure 8—*The* ASN *curves corresponding to the* OC *curves of Fig. 7.*

Presumably this sequential format could be improved further, but these improvements must not be made at the expense of simplicity.

Perhaps a finer test of the flexibility of the various classes of tests is the adaptability to the graded-conclusion format. Arbitrarily, the format is considered satisfactory if the 5% contour of the acceptance region is within and parallel to the 50% contour of the corresponding sequential test while the 5% contour of the rejection region is beyond and parallel to the same 50% contour. Table IX and Fig. 6 show that Test 121-4-10 is successful in this respect. Presumably with further adjustment of parameters, Tests 140-4-10 and 160-4-10 would be successful also.

Since those tests by attributes that classify lots according to the number of assays beyond a single limit have relative difficulty in classifying lots into even two groups, they have not been considered for graded-conclusion formats which would divide lots into three or more groups. More complex tests by attributes, similar to the one suggested by Haynes *et al.* (14), might be convertible to the graded-conclusion format; but due to inherent inflexibility, it is doubtful that they could fulfill the above criterion.

The primary comparison to be made in this work is between the current pharmacopeial tests for content uniformity and the best of the known potential alternatives. The criteria for the comparison are reliability, flexibility, simplicity, and reasonableness.

The reliabilities of the current pharmacopeial tests are in the range 15–30%, while those of the preferred tests by variables are approximately 65%.

The criterion of flexibility requires the following: *(a)* the definition of acceptability, *i.e.*, the location of the reference contour be independent of sample size; *(b)* the sequential format and graded-conclusion formats be successful; and *(c)* the relative importance of mean assay and dispersion be variable. Table VII and Figs. 4A–4D and 5 show that the location of the 50% contour for tests similar in principle to the current pharmacopeial tests varies considerably. In contrast, the preferred tests by variables form a homologous series with virtually the same 50% contour. With respect to the sequential format, for lots of intermediate quality, a sequential test should consistently call for the full complement of assays to estimate reliably the lot quality. The *ASN* curves of Fig. 8 show that the tests by variables successfully meet this condition while, in contrast, the tests by attributes call for the full complement, at best, only 30% of the time. Figure 6 shows that the preferred tests can be used in a graded-conclusion format; but the corresponding format for the tests by attributes, where the four-way classification would depend on whether there was zero, one, two, or three defectives in 10 assays, was not considered sufficiently promising for investigation. The relative importance of the mean assay and the relative importance of content uniformity in the pharmacopeial tests and in Test 121-2-10/30 are independently variable since they are controlled by completely separate criteria. Also, in one preferred test by variables, Test 140-2-10/30, the relative importance is varied by changing the proportionality constant k. However, in the test based on the mean-square deviation, Test 160-2-10/30, the relative importance of mean assay and content uniformity cannot be varied. (In some cases, single-tablet assay procedures can yield relative assays only, and the mean assay is determined by a separate procedure. In this circumstance, it becomes arithmetically simpler to use either Test 121-2-10/30 or 140-2-10/30, which are based on the mean and the standard deviation, rather than the other tests which are based on individual assay values.)

The emphasis of the criterion of simplicity has perhaps changed in the last decade, at least where the calculations are handled by a large or small computer. The current pharmacopeial tests for content uniformity are manually simpler than the tests by variables, but it is doubtful whether simplification ever justified the increase in likelihood of an erroneous conclusion. The above tests are intended to be sufficiently simple for routine use.

In summary, the tests based on $|\bar{X}| + 1.8S$ and on $(\bar{X})^2 + S^2$ are decidedly more reliable and more flexible than the current pharmacopeial tests but suffer a minor loss of simplicity with respect to manual calculation.

New Tests for Weight Variation—As already mentioned above, Wiley (5), in 1960, suggested replacement of the weight variation test with a test for content uniformity. This replacement has been carried out in the USP and NF, where the unit dosage is not more than 50 mg and where there is a suitable analytical method. The weight variation test (22, 23) generally applies to the others. This is consistent with the view that the weight variation test is a crude but simple measure of content variation.

Comer *et al.* (24) suggested, as an alternative approach, that one assume that unit weight and the proportion of drug in a unit vary *independently* from unit to unit, then estimate the variance in weight with 100 units and the variance in proportion with nine units, and then combine these values to obtain a relatively accurate estimate of the variance of the drug content. This latter value can then be employed in one of several tests for content and content uniformity. However, the validity of the assumption that unit weight and proportion of drug vary independently is not yet established. Where the density of the drug is different from that of the other components or where there is an interaction between the drug and other components on mixing, unit weight may be a dependent function of proportion.

A third approach is to use the weight variation test in parallel to, but independently of, the test for content uniformity as a test for *good manufacturing practices.*

In the references cited and in general, the test of weight variation is a particular variation of: "Weigh n groups of 10 units. The sample passes if not more than n units are more than x% by weight from the sample mean and none are more than $2x$%." This is a test by attributes, the same in principle as the pharmacopeial test for content uniformity except for one aspect. The calculation of the deviations is based on the sample mean instead of the theoretical or *target* weight, *i.e.*, the weight corresponding to 100% assay. The target weight is readily calculated since the mean sample weight and mean assay are known (subject to experimental error). Perhaps it is time to bring the older test for weight variation onto the same footing as the test for content uniformity.

The search for an improved test for weight variation would seem to parallel that for content uniformity, except that presumably more weights would be required. For reasons discussed under *Tests for Content Uniformity*, tests based on $|\bar{X}| + 1.8S$ and $(\bar{X})^2 + S^2$ (where $\bar{X}$ is the deviation of the mean weight from the target weight and S is the standard deviation of the weights of the unit dose) are indicated.

Extensions—This work is meant only as a comparative study of selected basic types of tests. The optimum number of assays, the most efficient sequential arrangement, and the maximum practical number of grades were not investigated. Indeed, the computer program requires modification before certain variations, such as a 10/20/40 format for the sequential tests, can be studied.

Neither the question of the optimum values of the acceptance and rejection ranges nor the question of the desirability of more tolerant standards for pharmaceuticals with inherently high standard deviation of the unit content has been considered because these are pharmacological questions.

In this discussion, the distinction between the variance of the assay values and the variance of the content of the units assayed has been ignored. Similarly, the distinction between the mean of the assay values and the mean content of the units assayed has been ignored. Generally, the accuracy and precision of the analytical method are such that the distinction is secondary. Also, USP XVIII (25) states, with respect to tolerances: "These limits allow for assay error, for unavoidable variations in manufacturing and compounding, and for deterioration to an extent considered insignificant under practical conditions." In any event, the curves would seem to be valid for either case provided that one is consistent, *i.e.*, provided both lot and sample are described in terms of either absolute or assay values. In actual practice, the true mean and variance are not known so that the assay values must be used. Recently, the effect of analytical error on the conclusion of the USP XVII test for content uniformity was determined (26, 27).

There remains the question as to whether more reliable tests, for a given number of assays, can be designed. In converting the assay values into the numbers utilized in the tests, there is a loss of information. This becomes apparent if one tries to use these numbers to regenerate the original assay values. Presumably, the most reliable tests will be those utilizing, to the greatest degree, all available information. Perhaps such tests, besides involving the mean and standard deviation, will utilize the coefficients of skewness and kurtosis. However, due to nonrandom variations in the actual mean and variance from one portion of a lot to another, the limitations in obtaining a representative sample may mask small improvements in the test for content uniformity.

REFERENCES

(1) "The United States Pharmacopeia," 17th rev., Mack Publishing Co., Easton, Pa., 1965, p. 905; *ibid.*, 18th rev., 1970, p. 930.

(2) "The National Formulary," 12th ed., Mack Publishing Co., Easton, Pa., 1965, p. 449; *ibid.*, 13th ed., 1970, p. 798.

(3) T. N. T. Olson and I. Lee, *J. Pharm. Sci.*, **55**, 1(1966).

(4) C. W. Dunnett and R. Crisafio, *J. Pharm. Pharmacol.*, **7**, 314(1955).

(5) "P.M.A. Year-Book, 1961-1962," Pharmaceutical Manufacturers Association, Washington, D.C., 1962, pp. 375, 395–401.

(6) "Sampling Procedures and Tables for Inspection by Attributes (Military Standard MIL-STD-105D)," U.S. Government Printing Office, Washington, D.C., 1963.

(7) "Sampling Procedures and Tables for Inspection by Variables for Percent Defective (Military Standard MIL-STD-414),"

U.S. Government Printing Office, Washington, D.C., 1957.
(8) G. J. Lieberman and G. J. Resnikoff, *J. Amer. Stat. Ass.*, **50**, 457(1955).
(9) J. D. Haynes, M. Schnall, and A. S. Doniger, "Types of Unit-to-Unit Variation and Their Implications," presented to the Industrial Pharmacy Section, APhA Academy of Pharmaceutical Science, New York meeting, Nov. 1963. (Copy received from authors at Lederle Laboratories, Pearl River, N.Y.)
(10) J. F. Paul, presented to the Quality Control Section, PMA section meetings, unpublished report; through Ref. 3.
(11) C. D. Smith, T. P. Michaels, M. J. Chertkoff, and L. P. Sinotte, *J. Pharm. Sci.*, **52**, 1183(1963).
(12) H. L. Breunig and E. P. King, *ibid.*, **51**, 1187(1962).
(13) I. Setnikar and F. Fontani, *ibid.*, **59**, 1319(1970).
(14) J. D. Haynes, M. Schnall, and R. A. Lamm, "Variability Tests in Acceptance Sampling from Non-Normal Populations," presented to the American Statistical Association, Cleveland, Ohio, Sept. 1963.
(15) "PMA Year-Book, 1963–1964," Pharmaceutical Manufacturers Association, Washington, D.C., 1965, pp. 458–462.
(16) C. B. Sampson, "Statistical Procedures for the Estimation of Content Uniformity," presented to the APhA Academy of Pharmaceutical Sciences, Washington meeting, Nov. 1969.
(17) W. J. Dixon and F. J. Massey, "Introduction to Statistical Analysis," 2nd ed., McGraw-Hill, New York, N.Y., 1957, p. 381.
(18) *Ibid.*, p. 349.
(19) A. N. Smith, *Pharm. J.*, **167**, 143, 270, 323(1951); through Ref. 4.
(20) W. A. Ritschel, *Drug Intel. Clin. Pharm.*, **6**, 246(1972) (especially p. 248, starting with fourth paragraph).
(21) S. Oie, K. Frislid, T. Waaler, E. Arnesen, and E. Enger, *Pharm. Acta Helv.*, **46**, 702(1971).
(22) "The United States Pharmacopeia," 18th rev., Mack Publishing Co., Easton, Pa., 1970, p. 950.
(23) "The National Formulary," 13th ed., Mack Publishing Co., Easton, Pa., 1970, p. 902.
(24) J. P. Comer, H. L. Breunig, D. E. Broadlick, and C. B. Sampson, *J. Pharm. Sci.*, **59**, 210(1970).
(25) "The United States Pharmacopeia," 18th rev., Mack Publishing Co., Easton, Pa., 1970, p. 2.
(26) A. O. Pedersen, Y. Torud, and T. Waaler, *Pharm. Acta Helv.*, **46**, 21(1971).
(27) A. O. Pedersen and Y. Torud, *ibid.*, **46**, 114(1971).

ACKNOWLEDGMENTS AND ADDRESSES

Received July 20, 1972, from the *Drug Research Laboratories, Health Protection Branch, Department of National Health and Welfare, Ottawa, Canada.*

Accepted for publication September 27, 1973.

The author acknowledges the valuable discussions with Dr. W. N. French of these laboratories and the assistance in programming of Mr. H. Yule and Mr. R. Mackey of the Division of Statistics and Information Science.

Figures 1–4 were included, by permission of the *Journal of Pharmaceutical Sciences*, in material presented to the 1973 Land O'Lakes Conference on Pharmaceutical Analysis, Land O'Lakes, Wis.

Use of 2-(4′-Hydroxybenzeneazo)benzoic Acid to Study the Binding of L-Thyroxine to Serum Albumins

RALPH I. NAZARETH, THEODORE D. SOKOLOSKI[x], DONALD T. WITIAK, and ALLEN T. HOPPER*

Abstract □ A study was made to investigate the use of the dye 2-(4′-hydroxybenzeneazo)benzoic acid as a potential agent that would reflect the binding of L-thyroxine to serum albumins. The dye is a strong visible absorbing material which interacts with serum albumins to give characteristic spectrophotometric peaks and, as such, provides the basis for an extremely convenient assay to measure amounts, free and bound, after their separation in the presence of serum albumin and potential competitive inhibitors. The results obtained showed that the dye and L-thyroxine compete for the same binding site on bovine and rat serum albumins; thus the dye can be used to gauge the displacement of L-thyroxine from serum albumin by potential competitive inhibitors. Additionally, a method was introduced to evaluate statistically differences, between data sets, in y-intercepts obtained by linear regression. The method used should have general application.

Keyphrases □ L-Thyroxine binding to serum albumins—measured using 2-(4′-hydroxybenzeneazo)benzoic acid, spectrophotometry □ Serum albumin binding of L-thyroxine—measured using 2-(4′-hydroxybenzeneazo)benzoic acid, spectrophotometry □ 2-(4′-Hydroxybenzeneazo)benzoic acid—used to measure L-thyroxine binding to serum albumins

2-(4′-Hydroxybenzeneazo)benzoic acid (I) is a dye that interacts with serum albumins to give characteristic spectrophotometric peaks (1). The intensity of these peaks can be related to the extent of binding of the dye to serum albumin and, therefore, could represent an extremely convenient means of studying displacement reactions. The binding of I to serum albumins and its displacement from serum albumin by chlorophenoxyacetic acids were studied previously (2, 3). Moriguchi *et al.* (4) also studied the binding of I to bovine serum albumin and described a procedure that utilized I in determining the binding of acidic drugs. Drugs studied with bovine serum albumin were substituted benzoic acids (5) and sulfonamides (6). This study was made to investigate the use of I as a potential agent that would reflect the binding of L-thyroxine to serum albumins. The objective was to determine if I and L-thyroxine are bound to the same site on serum albumin so that the dye could be used to gauge the displacement of L-thyroxine from serum albumin. The results using bovine and rat serum albumins are presented here.