Use and Interpretation of Common Statistical Tests in Method-Comparison Studies

James O. Westgard and Marian R. Hunt

We have studied the usefulness of common statistical tests as applied to method comparison studies. We simulated different types of errors in test sets of data to determine the sensitivity of different statistical parameters. Least-squares parameters (slope of least-squares line, its *y* intercept, and the standard error of estimate in the *y* direction) provide specific estimates of proportional, constant, and random errors, but comparison data must be presented graphically to detect limitations caused by nonlinearity and errant points. *t*-test parameters (bias, standard deviation of difference) provide estimates of constant and random errors, but only when proportional error is absent. Least-squares analysis can estimate proportional error and should be considered a prerequisite to *t*-test analysis. The correlation coefficient *(r)* is sensitive only to random error, but is not easily interpreted. Values for *r*, *t*, and *F* are not useful in making decisions on the acceptability of performance. These decisions should be judgments on the errors that are tolerable. Statistical tests can be applied in a manner that provides specific estimates of these errors.

Additional Keyphrases: *Student's t-test · least-squares parameters · proportional, constant, and random error · limitations of statistical tests · correlation coefficient · F test · decision-making on method acceptability · glucose determination as an example*

New methods of analysis must be studied to determine their precision and accuracy in order objectively to judge their acceptability for daily clinical use. In practice, one approach is to compare the new method with a "reference" method. If the "test" method compares favorably, it is judged to be acceptable. Detailed schemes for performance evaluation via comparison studies have been presented, with Barnett's scheme *(1, 2)* being most widely accepted and referred to in published evaluation studies. Others *(3, 4)* are referenced in the "Information for Authors" sections of journals such as this *(5)*.

These evaluation schemes specifically recommend analyzing comparison data by statistical techniques such as the F test, *t*-test, least-squares analysis, and correlation coefficients. Improper use or faulty interpretation of the statistical parameters may result in invalid judgments on the acceptability of methods. This report clarifies the use of common statistical tests in making decisions on performance of new methods.

Methods and Materials

Perspective on Evaluation Studies

Customarily we evaluate a method in terms of precision and accuracy, though this is a division based only on types of errors. Precision refers to random (indeterminate) errors whereas accuracy refers to systematic (determinate) errors. Systematic errors may be constant or proportional; constant error, as used here, refers to a systematic error in concentration units and proportional error to a systematic error in percentage units. Random errors are usually studied first because systematic errors are difficult to evaluate when large random errors are present. Hence, we experimentally study precision and then accuracy.

It is useful to know both the types and magnitudes of errors because different errors have different causes and affect results in different ways. In evaluating the performance of methods, random error is estimated by "precision" studies, proportional error by "recovery" studies, and constant error by "interference" studies. Because error studies need to be extensive to demonstrate accuracy, it is generally more efficient to evaluate a method by comparison with a method already well characterized by error studies. A series of patient samples are analyzed by both methods and the comparison data are subjected to statistical analysis. The statistical calculations *cannot* provide yes or no answers on the acceptability of a method; however, they can provide specific estimates of the types and magnitudes of errors. This is

From the Department of Medicine and the Clinical Laboratories, University of Wisconsin Center for Health Sciences, Madison, Wis. 53706

Presented in part at the 40th National Meeting of the ASMT, Minneapolis, Minn., June 11-16, 1972, and at the 164th National Meeting of the ACS, New York City, August 28–Sept. 1, 1972.

Received Sept. 5, 1972; accepted Oct. 7, 1972.

[Reprinted from CLINICAL CHEMISTRY, 19, 49 (1973).]

the essential information for deciding whether a method is accepable. The absence (or rarity) of errors substantiates acceptable performance. Decisions on acceptability are judgments as to what amount of error is tolerable.

Data Simulation and Processing

To demonstrate how various errors affect the common statistical tests, we generated an arbitrary set of reference values for glucose and systematically introduced different types and magnitudes of errors in the "test" set of data. Simulation approaches have also been used by Reed *(6)* to study the influence of statistical methods on the estimation of normal range, and by Amador *(7, 8)* to study the sensitivity of various quality control systems to specific errors. The reference set of 41 values is shown in Table 1.

The "test" set of data was formed by mathematical manipulation of the reference values. Random error was introduced by alternately adding and subtracting a constant concentration value. While this is not a true random variation, the simulation is adequate for demonstration purposes. Determinate errors were introduced by *systematically* adding or subtracting a constant concentration value or a constant percentage of the reference value.

After introduction of specific errors, the test and reference data were analyzed statistically and plotted with the ELLA (*E*xperimental *L*inc *L*aboratory *A*nalysis) system *(9)*. The least-squares parameters (slope, m; y intercept, b; standard error of estimate, S_y)[1] were calculated by the usual equations *(10)*. Paired Student t-test parameters (bias; SD of difference, SD_d; t) and the F test were calculated as suggested by Barnett *(1)*, and the Pearson product moment correlation coefficient (r) by equation 8.2 in McNear *(11)*.

Laboratory Methods

Example sets of comparison data were obtained for several glucose methods. The standard neocuproine method was initially used on the "SMA 12/60" (Technicon Instruments Corp., Tarrytown, N.Y. 10591). Later we studied a glucose oxidase procedure, which involved the use of the chromagen "ABTS" [2,2-azinodiethylbenzthiazoline sulfonic acid; *(12)*] and commercial reagents (Boehringer Mannheim Corp., 219 East 44th St., New York, N.Y. 10017). The flow system differed only slightly from that recently reported by Bigat and Saifer *(13)*. Other specific methods included a manual hexokinase determination as performed with the "ESKA-LAB" system (Smith Kline Laboratories, 3400 Hillview Ave., Palo Alto, Calif. 94306) and an *o*-toluidine method as performed with a single channel Auto-Analyzer [Technicon; *(14)*]. All methods were compared to the hexokinase method on the "Du Pont ACA" (E. I. Du Pont de Nemours and Co., Inc., Wilmington, Del. 19898).

Table 1. Reference Set of Glucose Values for Simulation Sudies[a]

1. 20	12. 82	23. 104	34. 160
2. 30	13. 84	24. 106	35. 170
3. 40	14. 86	25. 109	36. 180
4. 50	15. 88	26. 112	37. 200
5. 55	16. 90	27. 115	38. 220
6. 60	17. 92	28. 120	39. 240
7. 65	18. 94	29. 125	40. 260
8. 70	19. 96	30. 130	41. 280
9. 73	20. 98	31. 135	
10. 76	21. 100	32. 140	
11. 79	22. 102	33. 150	

[a] mg/dl.

Results and Discussion

Sensitivity to Types of Errors

No errors. Ideal comparison data would have exactly the same values for test and reference; all the points would fall exactly on a line making a 45° angle and intersecting the axes at the origin. Line *a* of Table 2 shows the statistical results. Ideal statistical values are 1.000 for slope and correlation coefficient and zero for all others except t. The t-value is undefined in this case because it is a ratio of two zero terms (0/0 = undefined). The usefulness of t will be discussed later.

Random error. The effects of random error are illustrated in Figure 1, in which 5 mg/dl has been alternately added and subtracted from the reference set of values. Random error shows up in the plot as scatter in the points around the least-squares line.

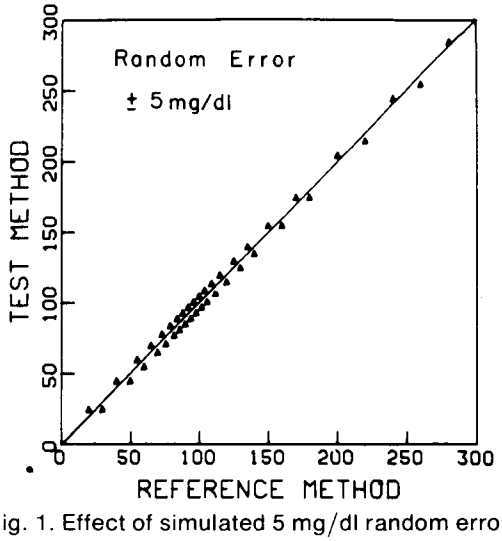


Fig. 1. Effect of simulated 5 mg/dl random error

[1] Nonstandard abbreviations used: m, slope of the least-squares line; b, the y intercept of the least-squares line; S_y, standard error of estimate in the y direction; SD_d, standard deviation of differences; t, Student's t-value; and r, correlation coefficient. S_y is the standard deviation of the differences of the actual Y value from the Y value calculated from the least-squares equation ($Y = mX + b$). Other authors may refer to this as the standard deviation of the residuals and use the abbreviations $S_{y/x}$ or S_r.

Table 2. Effects of Various Errors on Statistical Results, with Glucose Determination as an Example

Type of error(s)			Statistical parameters						
Random mg/dl	Constant mg/dl	Proportional percent	*m*	*b* mg/dl	Sy mg/dl	Bias mg/dl	SD_d mg/dl	*t*	*r*
(*a*) ...	...	...	1.000	0.00	0.00	0.00	0.00	Undefined	1.000
(*b*) ±2	...	...	1.001	−0.02	2.00	0.05	2.00	0.16	0.999
5	...	...	1.001	−0.04	5.00	0.12	5.00	0.16	0.993
10	...	...	1.003	−0.08	10.00	0.24	10.00	0.16	0.986
(*c*) ...	2	...	1.000	2.00	0.00	2.00	0.00	∞	1.000
...	5	...	1.000	5.00	0.00	5.00	0.00	∞	1.000
...	10	...	1.000	10.00	0.00	10.00	0.00	∞	1.000
(*d*) ...	...	2	0.980	0.00	0.08	2.29	1.18	12.43	1.000
...	...	5	0.950	0.00	0.08	5.72	2.95	12.42	1.000
...	...	10	0.900	−0.04	0.14	11.45	5.88	12.47	1.000
(*e*) ±5	2	...	1.001	1.96	5.00	2.12	5.00	2.72	0.996
5	5	...	1.002	4.92	5.02	5.10	5.03	6.50	0.996
5	10	...	1.002	9.95	5.02	10.10	5.03	12.86	0.996
±10	2	...	1.003	1.92	10.00	2.24	10.00	1.44	0.986
10	5	...	1.003	4.92	10.00	5.24	10.00	3.36	0.986
10	10	...	1.003	9.92	10.00	10.24	10.00	6.56	0.986
(*f*) ±2	...	5	0.951	−0.02	2.00	5.67	3.53	10.27	0.999
5	...	5	0.951	−0.04	5.00	5.59	5.76	6.27	0.996
10	...	5	0.953	−0.08	10.00	5.47	10.38	3.38	0.985
±5	...	10	0.900	0.23	5.01	11.28	7.85	9.09	0.996
5	...	25	0.751	−0.02	5.00	28.45	15.68	11.62	0.994
5	...	50	0.501	−0.04	5.00	57.02	30.17	12.10	0.986
5	...	75	0.251	−0.04	5.00	85.60	44.94	12.20	0.948
(*g*) ...	2	5	0.950	2.00	0.08	3.72	2.95	8.07	1.000
...	5	5	0.950	5.00	0.08	0.72	2.95	1.55	1.000
...	10	5	0.950	10.00	0.08	4.28	2.95	9.31	1.000
(*h*) ±2	5	10	0.905	4.99	2.03	6.37	6.28	6.50	0.999
10	2	5	0.953	1.89	10.01	3.48	10.51	2.12	0.984
5	10	2	0.981	9.96	4.99	7.84	5.18	9.69	0.996

Statistical data are shown for errors of ±2, ±5, and ±10 mg/dl (Table 2, line *b*). There are no changes in *m*, *b*, and bias, but S_y and SD_d reflect the magnitude of the random error. The correlation coefficient decreases as random error increases, but the changes in *r* are small.

Constant error. The effect of constant error is shown in Figure 2, where the line does not go through the origin. In this case, 10 mg/dl has been added to each value to form the test set of data. Statistical results are shown for errors of 2, 5, and 10 mg/dl (Table 2, line *c*). From the Table, we see that *m*, S_y, SD_d, and *r* are *not sensitive* to constant error, but that *b* and bias reflect it exactly.

Proportional error. The effect of proportional error is shown in Figure 3, in which the glucose test results are 2%, 5%, and 10% lower than the reference results. Proportional error changes the steepness of the line and the exact magnitude of the error is quantitated by the changes in *m* (Table 2, line *d*). Proportional error does not affect *b*, S_y, and *r*, but both bias and SD_d increase.

Combinations of errors were also studied and statistical results are included in Table 2 (lines *e*–*h*). The sensitivities of the individual statistical parameters are the same as when single errors are introduced, and are summarized in Table 3.

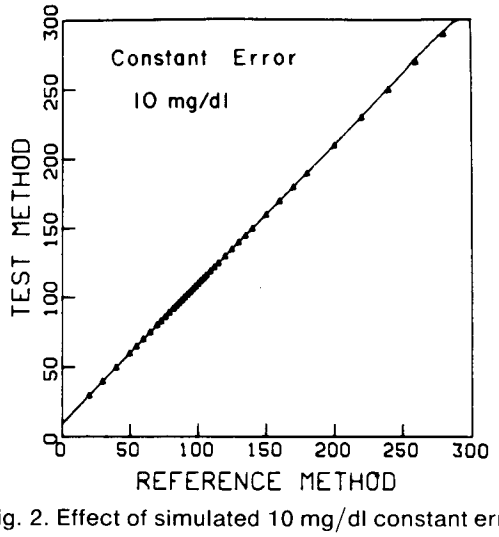


Fig. 2. Effect of simulated 10 mg/dl constant error

Specific Estimates of Errors

The error-simulation study shows that different statistical parameters are sensitive to different types of errors, and in some cases to more than one error.

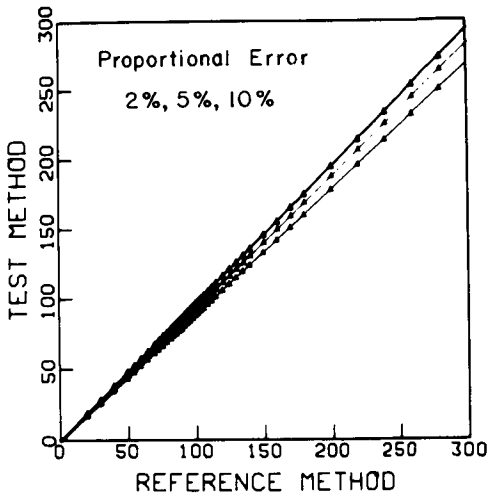


Fig. 3. Effect of simulated 2%, 5%, and 10% proportional error (lines from top to bottom)

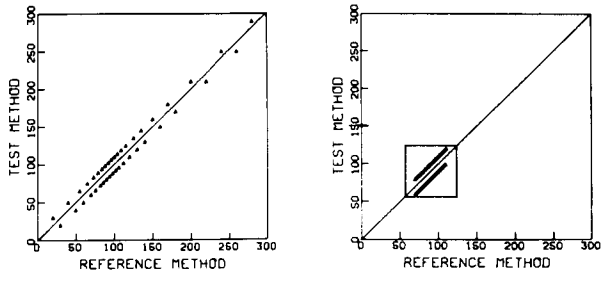
Our purpose is to apply the statistical tests in a manner that provides specific estimates of random, constant, and proportional errors.

Random error. From Table 3, it is apparent that S_y, SD_d, and r all respond to random error. SD_d is also influenced by proportional error, which means that this parameter does not provide a specific estimate of error when proportional error is present. S_y and r are sensitive only to random error, but they differ both in units and numerical values. S_y is in units of concentration and is interpreted as a standard deviation, thus a value of 5.0 mg/dl means that values will agree within ±10 mg/dl for 95% of the samples (±2 SD limits). The r term is unitless and differences from 1.000 indicate the magnitude of the error. But what does 0.996 mean in terms of the actual random error between the two methods? For a given value by the reference method, what range of values would be expected by the test method? These questions are answered very simply when we know S_y, but not when we know only r.

A further deficiency of r is that it depends on the range covered by the data. For example, in the two plots shown in Figure 4, the random error is the same, ±10 mg/dl, yet the values for r are very different (0.986 vs. 0.764; Table 4, line *a*). The plot on the left has a wide range and simulates data from low abnormal to high abnormal concentrations. The scatter is small compared to the range of the data and the correlation coefficient is high. The plot on the right has the same number of data points, but all are in the normal range. The amount of scatter is large with respect to the range covered, or to the area of the small box drawn around normal range. The low r value of 0.764 increases to 0.849 by addition of one point at 20 mg/dl and to 0.953 by addition of a point at 280 mg/dl. Because of this dependence on range, different investigators could get different correlation coefficients simply because their ranges of values were different. For example, compare correlation coefficients reported in various evaluations of the ACA glucose method. Perry et al. *(15)* quote 0.995 *vs.* *o*-toluidine, Westgard and Lahmeyer *(16)* report 0.993 *vs.* *o*-toluidine, 0.997 *vs.* manual hexokinase, and 0.996 *vs.* neocuproine, and Speicher et al. *(17)* report 0.909 *vs.* neocuproine. Speicher's value is low because his study included only samples in the normal range, whereas the others included elevations up to 300 to 400 mg/dl. Because of the difficulty of interpreting r, S_y is a more useful parameter for quantitating random error.

Table 3. Sensitivity of Statistical Parameters to Different Types of Errors

	Type of error		
	Random	Constant	Proportional
Least squares			
Slope, *m*	No	No	Yes
y intercept, *b*	No	Yes	No
Std. error, Sy	Yes	No	No
t-test			
Bias	No	Yes	Yes
SD_d	Yes	No	Yes
Correl. coeff., r	Yes	No	No



Range	0 to 300	70 to 110
Random Error	10 mg/dl	10 mg/dl
Corr Coef	0.986	0.764

Fig. 4. Effect of range on correlation coefficient. Simulated random error of 10 mg/dl for both sets of comparison data

Constant error. Table 3 shows that b and bias are both sensitive to constant error. Both estimate the error in units of concentration and give similar values when proportional error is absent. However, proportional error does affect the bias term; therefore, bias cannot be interpreted as a specific estimate of constant error unless proportional error is absent. Because b can estimate constant error in the presence of proportional error, b is a more useful parameter for quantitating constant error.

Proportional error. Table 3 shows that m, bias, and SD_d are all sensitive to proportional error. The difference of m from 1.000 provides an exact estimate of the magnitude of the proportional error. The bias and SD_d terms do not reflect the nature of the error, nor do they provide a useful estimate of its magnitude. The *t*-test parameters therefore are not useful

Table 4. Effects of Range and Nonlinearity on Statistical Results

	N	m	b	Sy	Bias	SD_d	t	r
			mg/dl					
(a) Simulated data; range								
0–300	41	1.003	−0.08	10.00	0.24	10.00	0.16	0.986
70–110	41	1.000	0.24	10.00	0.24	10.12	0.15	0.764
70–110 + (20,20)	42	1.002	0.10	9.86	0.24	10.00	0.15	0.849
70–110 + (280,280)	42	0.999	0.34	9.88	0.24	10.00	0.15	0.953
70–110 + (20 & 280 pts.)	43	0.999	0.29	9.76	0.23	9.88	0.15	0.958
(b) Simulated data; nonlinearity								
Above 199	41	0.958	3.74	1.97	1.06	3.16	2.15	0.999
149	41	0.910	7.76	3.42	2.52	6.31	2.56	0.998
124	41	0.818	15.39	6.51	5.37	12.53	2.74	0.991
(c) Real data; range								
0–400	126	0.984	6.93	5.95	4.88	6.08	9.01	0.996
0–300	120	0.984	6.95	5.52	5.04	5.61	9.85	0.995
0–200	105	0.977	7.63	5.36	5.22	5.43	9.84	0.984
0–150	93	0.976	7.72	5.44	5.40	5.49	9.48	0.965
70–110	60	0.904	13.81	5.65	5.45	5.76	7.32	0.807

in estimating proportional error and, furthermore, should not be used when proportional error is present because they do not provide specific estimates of random and constant error in this situation. Proportional error is best quantitated by *m*.

Applicability of Statistical Tests

The correlation coefficient provides information about random error, but *r* cannot be easily interpreted and therefore is of no practical use in the statistical analysis of comparison data. Analysis by *t*-test can provide specific estimates of random and constant errors, but only when proportional error is absent. In practice, when the *t*-test is applied, least-squares analysis should also be performed to determine whether proportional error is present and whether the *t*-test results do represent specific estimates of errors. Least-squares analysis is potentially the most useful statistical technique, because it provides specific estimates for all types of errors.

However, least-squares parameters will not provide accurate estimates of errors unless the comparison data show a linear relationship between methods. In Figure 5, test values above 200 mg/dl are low, simulating the nonlinear response that can occur when reagents are depleted. All three parameters change (Table 4, line *b*), which shows that estimates of error are not specific when nonlinearity is present. To obtain specific estimates, linearity must be ensured by *(a)* initial linearity studies on both methods and *(b)* presentation of comparison data graphically to permit visual detection of nonlinear relationships. Least-squares can similarly be invalidated by one or a few errant points near the upper or lower end of the line. Again, visual observation of graphical data will permit detection of these situations.

Least-square results may also be inaccurate when

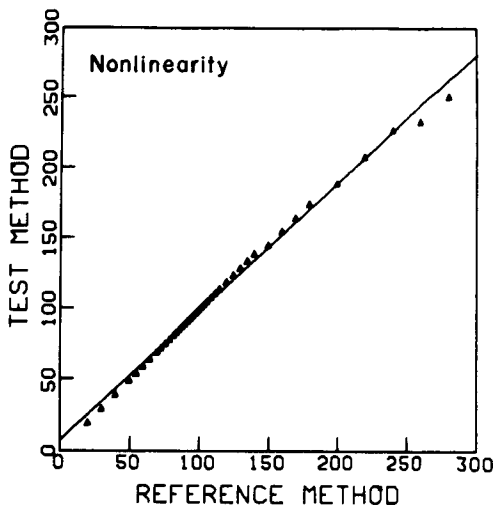
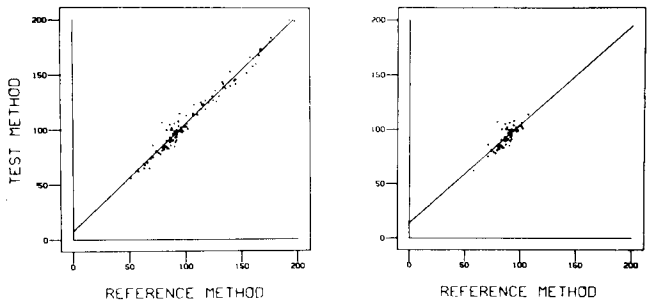


Fig. 5. Effect of nonlinearity

random error is large and the range of the data is small. For example, in Table 4 (line *c*), the statistical results are shown for a set of real data where range has been reduced by eliminating all values above chosen limits. S_y is relatively constant as the range decreases, but *m* and *b* show large changes once the data points are restricted to the normal range. S_y shows that the range or random variation in the *y* direction is about 22 mg/dl (4 SD), whereas the range of concentration variation in the *x* direction is only 40 mg/dl. As seen in the right plot in Figure 6, we are trying to draw a straight line through a set of points whose random variation is too large to define the location precisely. The line is better defined by including samples over a wide range of concentration, as shown in the left plot of Figure 6. Again, inspection of graphical results will provide a means of detecting the appropriateness of the least-squares application.



Range	0 to 200	70 to 110
N	105	60
Slope	0.977	0.904
Y intercept	7.63	13.81
Std Error	5.36	5.65

Fig. 6. Effect of range on least-squares results. Real data with random error of approximately 5.5 mg/dl

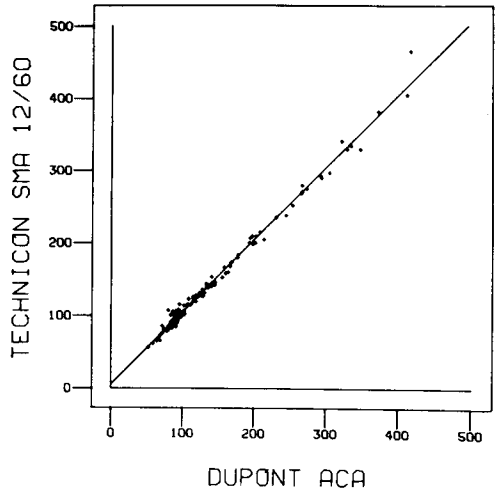


Fig. 7. Glucose by neocuproine as determined on the Technicon SMA 12/60 vs. glucose by hexokinase, as determined on the Du Pont ACA

Examples of Error Estimates

A series of comparison studies on glucose illustrate the interpretation of statistical data as estimates of errors.

1. Neocuproine vs. hexokinase (Table 5, line *a;* Figure 7): The slope is 0.999, which indicates a proportional error of only 0.1%. Constant error is estimated at 5.23 mg/dl by the *y* intercept and 5.13 mg/dl by the bias term. Both the standard error and standard deviation terms estimate random error at 7.23 mg/dl. Results of estimates by the *t*-test agree well with the least-squares estimates because proportional error is small. The correlation coefficient is 0.996, or nearly ideal. For a glucose value of 100 mg/dl by hexokinase, the neocuproine method will on the average give a result of 105 mg/dl, and we are 95% sure that the value will be between 90 and 120 mg/dl (±2 SD, ±15 mg/dl).

2. Manual hexokinase vs. automated hexokinase (Table 5, line *b;* Figure 8): Proportional error is small (0.7%) and constant error is −0.38 mg/dl by the intercept and −1.38 mg/dl by bias. Random error is 7.18 and 7.21 mg/dl by the two estimates and the correlation coefficient is high (0.993). For a value of 100 mg/dl by the automated hexokinase method, the manual method should give 99 ± 14 mg/dl.

3. o-Toluidine vs. hexokinase (Table 5, line *c;* Figure 9): Proportional error is 0.8%, constant error is −0.9 mg/dl by b and −1.7 mg/dl by bias, and random error is 5.85 mg/dl by S_y and 5.90 mg/dl by SD_d. The correlation coefficient is 0.994. A value of 100 mg/dl by hexokinase would be 99 ± 12 mg/dl.

4. Neocuproine vs. hexokinase, uremic samples (Table 5, line *d;* Figure 10): Proportional error is only 0.6%. Constant error is 17.4 mg/dl by the *y* intercept and 16.6 mg/dl by bias, thus the neocuproine method averages about 17 mg/dl higher when uremic samples are analyzed. Both estimates of random error are 10.5 mg/dl. The errors observed here are larger than for non-uremic samples (Table 5, line *a*), where constant error was 5 mg/dl and random error 7 mg/dl. The increased errors are due to the interferences in uremic samples and differences in the specificity of the methods. The interferences are reflected as a constant error because we have selected a group of samples that all have interferences, and they are also reflected as random error because the amount of interference varies from sample to sample. The changes in least-squares and *t*-test parameters are quite marked, but note that the changes in correlation coefficients are small and do not suggest any significant errors (0.996 to 0.990).

5. Glucose oxidase vs. hexokinase (Table 5, line *e;* Figure 11): Least-squares parameters suggest a proportional error of 5.1%, a constant error of −7.4 mg/dl, and a random error of 4.9 mg/dl; *t*-test esti-

Table 5. Statistical Results for Glucose Comparison Studies

Method vs. ACA hexokinase	N	*m*	*b*	Sy	Bias	SD_d	*t*	*r*
				mg/dl				
(a) Neocuproine	128	0.999	5.23	7.23	5.13	7.23	8.03	0.996
(b) Manual hexokinase	96	0.993	−0.38	7.18	−1.38	7.21	1.87	0.993
(c) o-Toluidine	81	0.992	−0.87	5.85	−1.74	5.90	2.65	0.994
(d) Neocuproine (uremic samples)	32	0.994	17.4	10.5	16.6	10.5	9.08	0.990
(e) Glucose oxidase	61	1.051	−7.44	4.90	−2.07	5.34	3.02	0.993
(f) Glucose oxidase	59	1.006	−3.12	4.65	−2.54	4.70	4.16	0.986

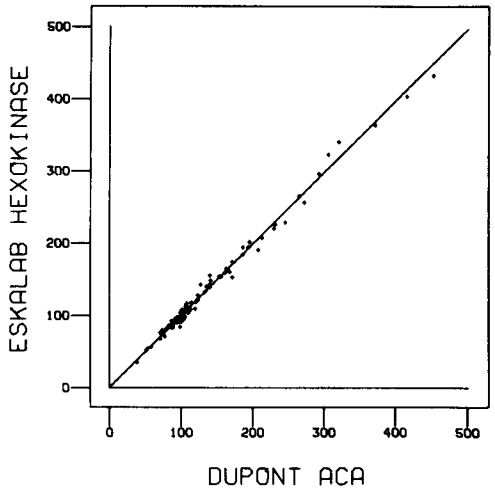


Fig. 8. Glucose by manual hexokinase, as on the ESKALAB vs. glucose by hexokinase, as determined on the Du Pont ACA

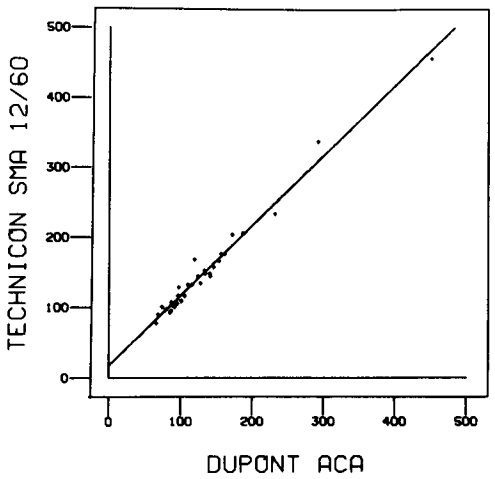


Fig. 10. Glucose for uremic patients as determined by two procedures, neocuproine method on the Technicon SMA 12/60 vs. hexokinase method on the Du Pont ACA

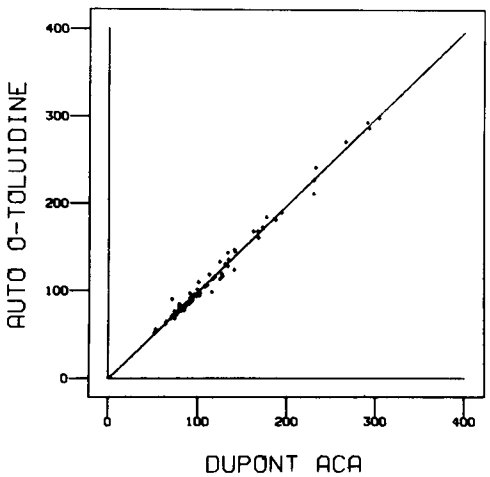
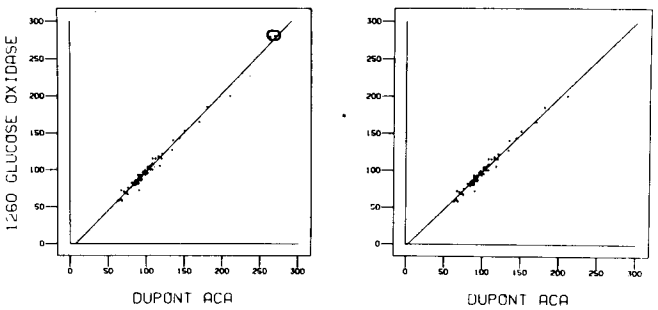


Fig. 9. Glucose by *o*-toluidine, as determined on AutoAnalyzer vs. glucose by hexokinase, as determined on the Du Pont ACA



N	61	59
Slope	1.051	1.006
Y intercept	-7.44	-3.12
Std Error	4.90	4.65

Fig. 11. Glucose by glucose oxidase, as determined on the SMA 12/60 vs. glucose by hexokinase, as determined on the Du Pont ACA: Plot on *left* includes two high points that resulted from nonlinearity. Plot on *right* shows statistical data when these two points are eliminated

mates of errors do not agree well with those from least squares. From the plot on the left side of Figure 11, it appears that two points at the upper end of the least-squares line may be too high. When these two are eliminated (right plot), the least-squares results (Table 5, line *f*) show essentially no proportional error (0.6%), a constant error of −3.1 mg/dl, and a random error of 4.6 mg/dl. Parameters for the *t*-test now give similar estimates for constant and random error. The differences between the two sets of least-squares results show the influence of the two errant points. Further study revealed a definite nonlinear response for the glucose oxidase method, with values being too high at elevated concentrations.

Acceptability

Criteria for acceptability. Acceptability of a method depends on both applicability and performance. Applicability encompasses factors such as sample size, types of samples usable, speed of analysis, equipment needed, personnel requirements, cost, and the like. Performance considers the type and magnitude of errors. Applicability and performance together define the criteria for acceptability. These criteria originate in the laboratory and in the clinical situation where the values from the method are used. Statistical tests do not provide the criteria for acceptability.

Decisions on acceptability. This discussion concerns performance rather than applicability, although both types of criteria must be met for the method to be acceptable. Decisions on acceptability of performance should be based on judgments on tolerable limits of error. Unfortunately, values for *t* and F have often been interpreted as indicators of acceptability for accuracy and precision, respectively, even though they are intended to tell only whether differences between methods are statistically signi-

ficant, not whether the performance of either method is acceptable.

t-Test

This statistical test is usually interpreted by comparing the calculated value with the "critical" value, which is found in a statistics table. When the calculated value is larger than the critical value (2.021, N = 40, $P = 0.05$ or 95% confidence limits), it is generally concluded that the difference between methods is large and that the performance of the test method is not acceptable. When smaller, it is generally concluded that the methods agree well and performance of the test method is acceptable.

Such judgments may be erroneous when based on the *t*-value alone because *t* is a ratio of constant and random error terms, not a measure of total error [$t = (\text{bias}/SD_d)\sqrt{N}$]. This is analogous to blood pH being determined by the ratio of bicarbonate to P_{CO_2}. A low or acid pH does not tell whether the bicarbonate is low or P_{CO_2} is high. Interpretation of the pH or treatment of the acidosis requires assessment of the individual metabolic and respiratory factors. Similarly, interpretation of a *t*-value requires the individual assessment of the constant and random error terms. At least four situations can cause erroneous judgments if only *t* is considered:

- *t* may be small when random error is large. For example, a small constant error of 1 mg/dl and a large random error of 20 mg/dl would give a *t*-value of 0.32 (n = 41).
- *t* may be small when both constant and random error are large. A bias of 10 mg/dl and SD_d of 40 mg/dl give a *t*-value of 1.60 (n = 41).
- *t* may be large when both errors are small. A bias of 2 mg/dl and SD_d of 5 mg/dl give a *t*-value of 2.56 (n = 41).
- *t* may give different values for given error levels if the number of samples varies. For a bias of 1 mg/dl and SD_d of 5 mg/dl, *t*-values are 1.28, 1.81, 2.22, and 2.56, respectively, for N = 41, 82, 123, and 164. If only *t* were considered, small values may result in acceptance in the first two situations, even though the individual error levels may not be acceptable. In the last two situations, large *t*-values may result in rejection of the test method, even though error levels may be acceptable.

The *t*-value provides information only on the relative magnitudes of the constant and random error terms. The important information for judging performance is the individual terms, not the value of *t*. Proper use of the *t*-test requires that all parameters be presented, not just the *t*-value.

F-Test

The F value is calculated from the *individual* standard deviations of the test and reference methods by squaring each and dividing the larger variance by the smaller variance: $F = (SD_A)^2/(SD_B)^2$. When the calculated value is larger than the tabulated "critical value," the difference in precision between the methods is real, i.e., statistically significant. Like the *t*-test, this is a comparison of error levels, not an indicator of the acceptability of errors.

Tolerable Error Levels

Random error. Judgments on acceptability should compare the actual standard deviation with maximum acceptable standard deviations. For glucose, Barnett *(18)* has recommended a "medically significant standard deviation" of 5 mg/dl, which represents the performance necessary for adequate medical care as judged by a group of physicians and laboratory scientists. Vanko *(19)* suggested standard deviations of 2.4 to 4.4 mg/dl as acceptable, and Cotlove et al. *(20)* recommended 4.5 mg/dl as the "tolerable analytical variation." Judgments on the acceptability of day-to-day precision should compare the calculated standard deviation to these performance standards, or to those needed in the particular application of the method.

We must also distinguish between the random error of an individual method (SD, considered above) and the random error between methods (SD_d), which is larger because the errors of both methods are included: $SD_d = \sqrt{SD_{test}^2 + SD_{ref}^2}$. For laboratories that use two different methods for the same constituent (perhaps one for routine and one for emergency determinations), the standard deviation between methods represents the overall performance of the laboratory. If both methods had the maximal acceptable SD, the maximal SD_d would be 1.4 times larger, $SD_d = \sqrt{2\,SD_{max}^2}$.

Systematic errors. In principle, only random error need be tolerated. Systematic errors can be eliminated by appropriate improvements in methodology. For example, the presence of proportional error suggests that standardization and calibration procedures be examined, and the presence of constant error suggests that the specificity of the method be studied. In practice, however, small systematic errors as well as small random errors may be tolerable.

Acceptability depends on whether the errors limit the clinical usefulness of the method. This requires consideration of the exact clinical situations in which the method would be used and where the interpretation is most critical. Further definition and clarification of the decision-making process is needed, but one possible approach is suggested here. Critical reference values can be assumed and the values by the test method calculated from the least-squares estimates of slope and intercepts. The 95% ranges for random variation can be calculated for both the test and reference methods using the day-to-day precision data for the individual methods. These ranges can then be compared to determine whether the clinical interpretation would change if the test method were used. If it does change, the errors are not acceptable. If it does not change, the errors are tolerable.

Summarizing Comments

In characterizing performance, we should characterize errors in a manner that is useful to others who must judge acceptability in their laboratory situations. The criteria for methods will differ in different laboratories; thus, acceptability will depend on the particular application. Least-squares analysis is most useful for estimating errors, but we must be conscious of the limitations caused by nonlinearity, errant points, and a small range of values. Comparison results must be presented graphically to judge these limitations. Analysis by *t*-test is next in usefulness, but will not provide specific estimates of errors when proportional error is present. The calculations can be performed manually and therefore will be used frequently when calculators and computers are not available. When used, it is important to estimate proportional error, at least by manually graphing the comparison values and observing the slope of the best line, and preferably by estimating the slope by least squares. Interpretation must consider the individual parameters rather than the *t*-value itself. *t*, F, and *r*, though often used, have no practical value in characterizing errors, and they should not be used as indicators of acceptability. Statistical tests can provide specific estimates of errors upon which judgments can be made, but they are not a substitute for judgments.

This material was prepared for instructional use in the Medical Technology Program. We thank Miss Alice Thorngate for her encouragement and support, and W. J. Blaedel, I. H. Carlson, M. A. Evenson, and F. C. Larson for their helpful comments on the manuscript. Statistical programs were provided by G. Cembrowski, E. C. Toren, Jr., and A. A. Eggert.

References

1. Barnett, R. N., A scheme for the comparison of quantitative methods. *Amer. J. Clin. Pathol.* **43,** 562 (1965).

2. Barnett, R. N., and Youden, W. J., A revised scheme for the comparison of quantitative methods. *Amer. J. Clin. Pathol.* **54,** 454 (1970).

3. Henry, J. B., Beeler, M. F., Copeland, B. E., and Wert, E. B., A format for description of methods in clinical pathology. *Amer. J. Clin. Pathol.* **52,** 296 (1969).

4. Broughton, P. M. G., Buttolph, M. A., Gowenlock, A. H., Neill, D. W., and Skentelbery, R. G., Recommended scheme for the evaluation of instruments for automatic analysis in the clinical biochemistry laboratory. *J. Clin. Pathol.* **22,** 278 (1969).

5. Information for Authors. *Clin. Chem.* **19,** 1 (1973).

6. Reed, A. H., Henry, R. J., and Mason, W. B., Influence of statistical method used on the resulting estimate of normal range. *Clin. Chem.* **17,** 275 (1971).

7. Amador, E., Quality control by the reference sample method: Error detection as a function of the variability of the control data. *Amer. J. Clin. Pathol.* **50,** 360 (1968).

8. Amador, E., Bartholomew, P. H., and Massod, M. F., An evaluation of the "Average of Normals" and related methods of quality control. *Amer. J. Clin. Pathol.* **50,** 369 (1968).

9. Hicks, G. P., Eggert, A. A., and Toren, E. C., Jr., Application of an on-line computer to the automation of analytical experiments. *Anal. Chem.* **42,** 729 (1970).

10. *Handbook of Chemistry and Physics,* 47th ed., R. C. Weast, and S. M. Selby, Eds. Chemical Rubber Co., Cleveland, Ohio, 1966, p A-244.

11. McNear, Q., *Psychological Statistics,* 3rd ed., John Wiley & Sons, Inc., New York, N.Y., 1962.

12. Werner, W., Ray, H. G., and Wielinger, H., Uber die Eigenschaften eines neuen Chromogens fur die Blutzuckerbestimmung nach der GOD/POD Methode. *Z. Anal. Chem.* **252,** 224 (1970).

13. Bigat, T. K., and Saifer, A., Some methodological modifications of the Technicon "SMA 12/60 AutoAnalyzer" system. *Clin. Chem.* **18,** 630 (1972).

14. Sudduth, M. C., Widish, J. R., and Moore, J. L., Automation of glucose measurement using *o*-toluidine reagent. *Amer. J. Clin. Pathol.* **53,** 181 (1970).

15. Perry, B. W., Hosty, T. A., Coker, J. G., Doumas, B., Straumfjord, J. W., *A Field Evaluation of the DuPont Automatic Clinical Analyzer,* E. I. Du Pont de Nemours and Co., Inc., Wilmington, Del., 1970.

16. Westgard, J. O., and Lahmeyer, B. L., Comparison of results from the Du Pont ACA and Technicon SMA 12/60. *Clin. Chem.* **18,** 340 (1972).

17. Speicher, C. E., Fetrat, M. E., Fiske, M. L., and Henry, J. B., An automatic clinical analyzer: A critical evaluation. *Amer. J. Clin. Pathol.* **50,** 671 (1968).

18. Barnett, R. N., Medical significance of laboratory results. *Amer. J. Clin. Pathol.* **50,** 671 (1968).

19. Vanko, M., Selected factors which influence the design of a quality control program. In *Advances in Automated Analyses, Technicon International Congress 1970,* **1.** E. C. Barton et al., Eds. Thurman Associates, Miami, Fla. 33132, p 159.

20. Cotlove, E., Harris, E. K., and Williams, G. Q., Biological and analytic components of variation in long-term studies of serum constituents in normal subjects; III. Physiological and medical implications. *Clin. Chem.* **16,** 1028 (1970).