JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 1981, p. 760-768
0095-1137/81/040760-09$02.00/0

Vol. 13, No. 4

Alternative Ways of Estimating Serological Titer Reproducibility

ROSS J. WOOD

Bureau of Laboratories, Centers for Disease Control, Atlanta, Georgia 30333

Received 31 July 1980/Accepted 21 January 1981

A quantitative measure of the reproducibility of serum antibody titers has recently been proposed (R. J. Wood and T. M. Durham, J. Clin. Microbiol. **11**: 541-545, 1980). The measure advocated is "the probability that the maximum ratio of two distinct (integer) titers (obtained in the blind) on the same specimen will not exceed 2." This measure of the reproducibility of serological titers is considered to be a fixed probability for any given specimen and set of test conditions. Although it is a fixed constant during the time period of a study, there are alternative methods one might use to compute an estimate of it, using laboratory data. Four such methods of estimating test reproducibility are discussed and evaluated. The estimates obtained from the two principal methods are evaluated quantitatively by means of Monte Carlo computer simulation. The simulation results show that, from a given sample of replicate integer titers, these two principal methods yield estimates that are highly correlated. In addition, with moderate numbers of replicates (sample size), these methods provide estimates that are on the average properly directed at the true reproducibility values (that is, are essentially unbiased), particularly when the true reproducibility of the test is at least 0.9. The reliability, or stability, of the alternative estimates is studied for selected sample sizes.

A specific quantitative measure of the reproducibility of serum antibody titers has recently been proposed (6). The measure advocated is "the probability that the maximum ratio of two distinct (integer) titers (obtained in the blind) on the same specimen will not exceed 2."

One procedure for computing an estimate of the true reproducibilty (TR) from laboratory data was demonstrated with numerical examples, and a summary algorithm was given to facilitate computerizing the computations

In this paper, some alternative ways of computing an estimate of TR are presented. Also, the validities of the different estimates are studied, along with the reliability or stability of each estimate, based on different sample sizes.

The previously described estimate (6), here designated E1, is "distribution free" or "nonparametric" (3, 4) because the estimating procedure does not depend on the population distribution sampled; E1 is based only on information obtained from the magnitudes of the sample data. The reader is referred to the previous paper (6) for specific details of the E1 estimating procedure.

In contrast, two of the alternative estimating procedures introduced here are not distribution free but are "parametric." They are based on an assumption about the form of the sampled population distribution (see Appendix A) in addition to information from the sample data. The first of these two parametric estimates, designated E2, is obtained as follows. A random sample of n replicate integer titers is obtained in the blind for a chosen specimen. The n titers are converted to logarithms, and their standard deviations (SD) are computed and used with Table 1 to obtain E2, a parametric estimate of the reproducibility of the test system with that specimen (see Appendix B).

A second nonparametric estimate, designated E3, is introduced in this paper. It is based on a pair of replicate integer titers for each of k different sera, all having essentially the same titer. For each pair the maximum ratio is calculated, and the number (x) of pairs with a maximum ratio exceeding 2 is counted. TR is estimated by the equation $E3 = 1 - x/k$.

It is clear that the estimate E3 is easy to compute. The estimate E2 is also fairly easy to obtain because it is read directly from Table 1 after the SD among replicate log (integer titer) values has been computed. A little experience in using laboratory data will reveal that E1 is also readily computed. Consequently, there is little difference between the estimates E1 and E2 on

TABLE 1. *Conversion table for obtaining E2 from the observed SD of the logarithms of integer titers*

E2	Sample SD in logarithms to the base:		
	2	e	10
0.60	0.8402	0.5824	0.2529
0.61	0.8226	0.5702	0.2476
0.62	0.8055	0.5583	0.2425
0.63	0.7888	0.5467	0.2374
0.64	0.7725	0.5354	0.2325
0.65	0.7566	0.5244	0.2278
0.66	0.7411	0.5137	0.2231
0.67	0.7259	0.5032	0.2185
0.68	0.7110	0.4929	0.2140
0.69	0.6965	0.4828	0.2097
0.70	0.6823	0.4729	0.2054
0.71	0.6683	0.4632	0.2012
0.72	0.6545	0.4537	0.1970
0.73	0.6410	0.4443	0.1930
0.74	0.6278	0.4351	0.1890
0.75	0.6147	0.4261	0.1850
0.76	0.6018	0.4171	0.1812
0.77	0.5891	0.4083	0.1773
0.78	0.5765	0.3996	0.1735
0.79	0.5641	0.3910	0.1698
0.80	0.5518	0.3824	0.1661
0.81	0.5395	0.3740	0.1624
0.82	0.5274	0.3656	0.1588
0.83	0.5153	0.3572	0.1551
0.84	0.5033	0.3488	0.1515
0.85	0.4912	0.3405	0.1479
0.86	0.4791	0.3321	0.1442
0.87	0.4670	0.3237	0.1406
0.88	0.4578	0.3152	0.1369
0.89	0.4424	0.3067	0.1332
0.90	0.4299	0.2980	0.1294
0.91	0.4171	0.2891	0.1256
0.92	0.4039	0.2800	0.1216
0.93	0.3903	0.2705	0.1175
0.94	0.3760	0.2606	0.1132
0.95	0.3608	0.2501	0.1086
0.96	0.3443	0.2387	0.1036
0.97	0.3258	0.2259	0.0981
0.98	0.3040	0.2107	0.0915
0.99	0.2745	0.1903	0.0826

the basis of the amount of computation required. Once the practitioner has become familiar with computing E1 and E2, however, the question will arise about which is the better one—better in the sense of representing the actual TR. Here this subject is divided into two parts, the first dealing with how close each estimate averages to TR and the second dealing with the amount of scatter or variation appearing among repeated sample estimates. The first component of the accuracy of an estimate is referred to as "bias," and the second component is termed "reliability" or "precision."

In line with these definitions, a quantitative evaluation of the estimates E1 and E2 is the principal topic of this paper. The evaluation is based on a technique called Monte Carlo distribution sampling. A computer was used to generate repeated samples of log (integer titer) values from population distributions having preselected, and therefore known, TR values. The observed distributions of the repeated sample estimates were studied to assess the magnitude of the errors of estimation one can expect when using the estimators E1 and E2. In contrast, the number (x) of pairs in the estimate E3 can be taken to follow the binomial distribution (5) with sample size k and TR. Therefore, simulation is not required to evaluate E3; it is done directly from knowledge of the binomial distribution.

The terms "integer" and "truncated" titers relate to a concept that is important to the subject of this and related papers. In a titration test on a specimen in practice, the reportable titer is, in general, limited by the discriminating capabilities of the measuring equipment available to the technologist; that is, if better measuring equipment were available, it is conceivable that a titration test could show the endpoint dilution factor (titer) for a specimen to actually be 61.3, for example, if the technologist were sufficiently persistent in pursuing the necessary dilutions to obtain an endpoint at this level of resolution. The more common practice currently is to restrict the search for an endpoint to the examination of twofold dilutions. In this example the titer would therefore be observed and reported as the truncated titer 32, in contrast to the actual integer titer 61.

Some kit tests presently being developed for the market provide integer titers; that is, for the example specimen they would yield a result at or near the integer 61 rather than being restricted to twofold truncated titers. In summary, the titer 32 is simply the integer titer 61 truncated to the next lower integer power of 2. Such truncated titers are in general clearly less accurate than the associated integer titers.

MATERIALS AND METHODS

The IBM-370 computer at the Centers for Disease Control was used to simulate results from five hypothetical serological laboratories making repeated titration runs in the blind on a given specimen (see Appendix A). These five laboratories were made to differ only in their TR when producing replicate integer titers. The different laboratories were assigned TR values of 0.95, 0.90, 0.80, 0.70, and 0.60. In the simulation, the replicates for each laboratory were divided into five groups of 5,000 independent samples of sizes

n = 5, 10, 20, 30, and 40 (see Appendix C).

Within each group of 5,000 samples, the distributions of the resulting 5,000 values of E1 and the 5,000 values of E2 were studied to assess the performance of the estimators at each of the five sample sizes. The same assessments were made for each of the five laboratories.

RESULTS

Correlation of E1 and E2. For computing convenience, each group of 5,000 reproducibility estimates was divided into 10 sets of 500 estimates each, and the Pearson product moment coefficient of correlation between E1 and E2 was ec computed for each set. The mean correlations of these 10 sets are shown in Fig. 1. The figure shows a high linear correlation between E1 and E2 that increases with n over the range of 5 to 40. It also indicates that the correlation coefficient reaches a maximum for TR values in the neighborhood of 0.90.

Percent bias in E1 and E2. The bias of an estimator is defined as the long-term average of the sample estimates minus the population TR value. In this study, the bias was estimated by subtracting TR from the means of the 5,000 E1 and the 5,000 E2 sample estimates. Dividing these differences by TR and multiplying by 100 gave an estimate of the percent bias for E1 and E2 (Fig. 2 and 3).

The figures show that, at a TR of 0.90, the bias in E1 is approximately 1% for a sample size (n) of 30 and that E2 is essentially unbiased. In this situation, high correlation together with the lack of bias indicates that the sample results from E1 are practically the same as those from E2.

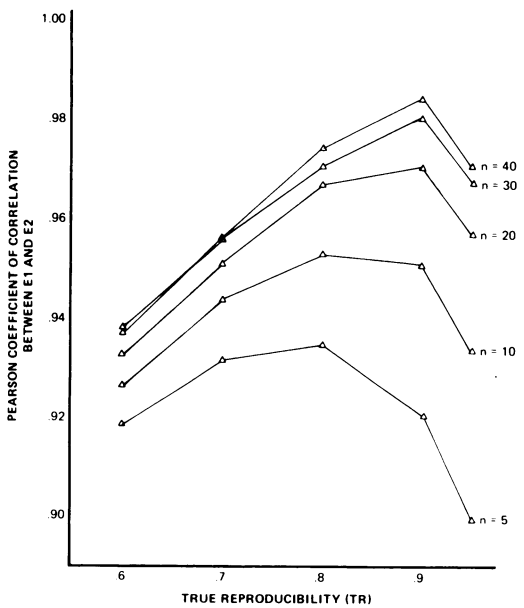

FIG. 1. *Observed correlation between E1 and E2 for selected reproducibilities and sample sizes (based on 10 independent sets of 500 simulated pairs of estimates at each sample size).*

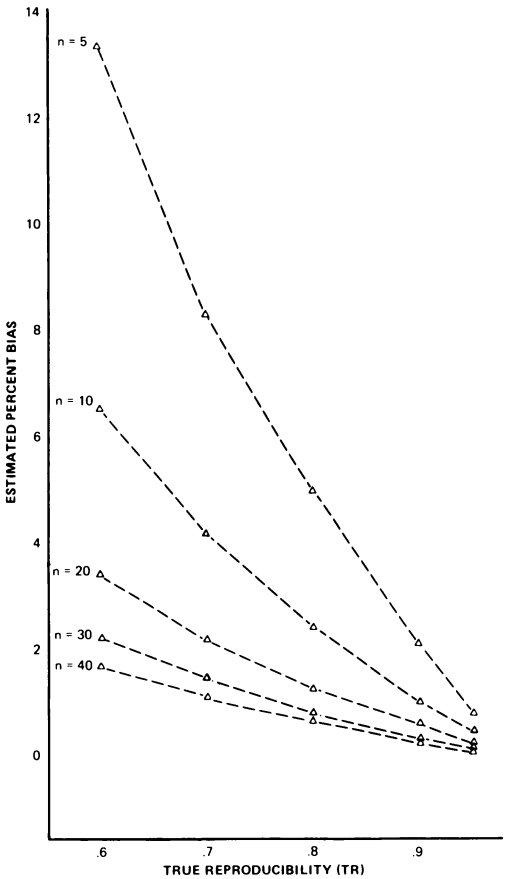

FIG. 2. *Observed percent bias of E1 at selected reproducibilities and sample sizes (based on 5,000 simulated estimates at each sample size).*

Reliability or stability of E1 and E2. When a laboratory sets out to obtain replicate integer titers on a specimen for the purpose of estimating TR, the question of how many replicates are needed arises. If the replicates are properly obtained in the blind, the estimate of TR will, on the average, follow the patterns of bias shown in Fig. 2 or 3. In contrast to bias, however, the question of sample size has its basis in the variability of the estimate. As more replicate titers are used in the estimate it becomes more stable. If an estimate were perfectly stable and also unbiased, it would be TR itself. The question of adequate sample size is addressed here by determining the variability in the estimate for differ-

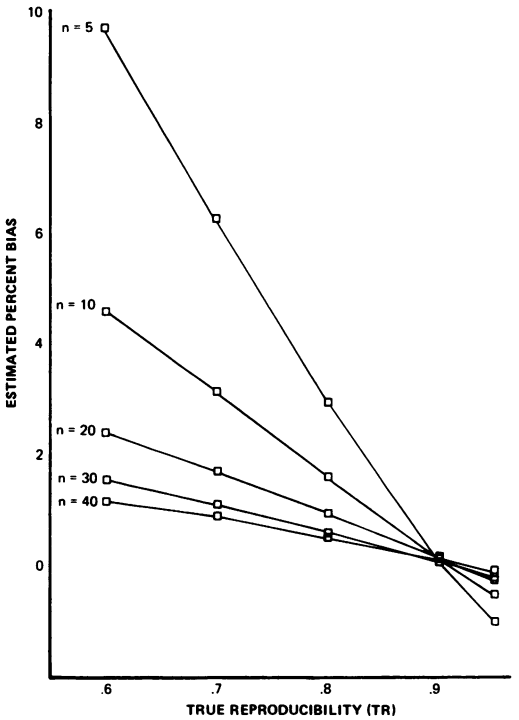

FIG. 3. *Observed percent bias of E2 at selected reproducibilities and sample sizes (based on 5,000 simulated estimates at each sample size).*

ent sample sizes and relating this to levels of possible error in the estimate that can be considered "acceptable" or "unacceptable" for the particular application.

Without substantive experience in applying the proposed quantitative measure of reproducibility to laboratory tests, certain assumptions are necessary. Accordingly, it is assumed that tests having TR values of $0.9 \leq TR < 1.0$ may be considered to be in an acceptable range, that those with TR values of $0.8 \leq TR < 0.9$ may be considered to be in a "marginal" range, and that those with TR values of <0.8 may be considered to be in an unacceptable range. Using these categories, one could obtain a set of replicate integer titers on a chosen specimen, estimate the TR, and classify the test as having a TR that was acceptable, marginal, or unacceptable solely on the basis of the single sample estimate.

Such a procedure for classifying a test would be subject to either of two relatively important errors. The first would occur if the TR were actually in the acceptable range but the sample estimate fell in the unacceptable range (a type I error). The second would occur if the TR were actually in the unacceptable range but the sample estimate fell in the acceptable range (a type II error). In industrial applications, the conditional probabilities of experiencing these errors are referred to as the producer risk and the consumer risk, respectively (5). These risks, that is, the conditional probabiliy of the particular type of error, are directly controllable through the sample size.

The question of adequate sample size was addressed here by studying the maximum probabilities of type I and type II errors as the sample size was increased. These conditional probabilities were obtained from the sample distributions of the 5,000 estimates obtained for the hypothetical laboratories assigned reproducibilities of 0.9 and 0.8 (Fig. 4). The same results are given in Table 2, which is discussed in Appendix D.

Suppose a probability in the range of 0.05 is considered to be acceptable for either type of error. Fig. 4 shows that sample sizes of approximately 16 for E1 and 18 for E2 would be sufficient for the type I error. However, these sample sizes would not be sufficient for the type II error. Sample sizes of 32 and 30 are required (Fig. 4) before the probability of a type II error is reduced to the range of 0.05. With these sample sizes, the probability of a type I error is reduced to less than 0.02.

Returning to the estimate E3 based on k pairs of integer titers, the number (x) of pairs follows the binomial distribution with sample size k and TR. Therefore, the estimate E3 is known to be unbiased. However, it is less reliable than E1 and E2, which are based on independent replicate titers obtained from a single serum. This can be seen by determining the number of pairs

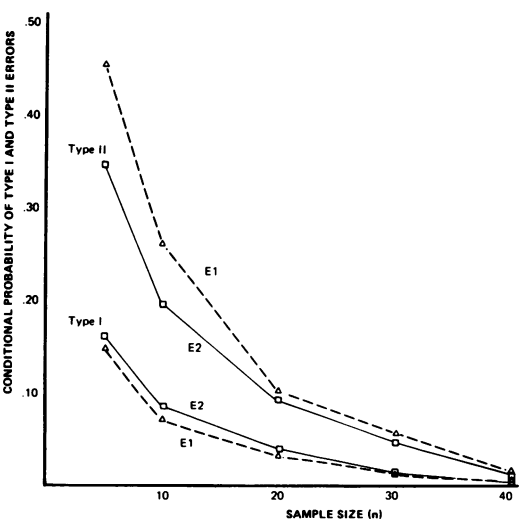

FIG. 4. *Observed conditional probabilities of type I and type II errors for selected sample sizes (based on 5,000 simulated estimates at each sample size).*

TABLE 2. *Conditional probabilities of type I and type II errors associated with E1 and E2, for selected sample sizes*

Sample size	Probability of error			
	Type I, conditional on TR = 0.9		Type II, conditional on TR = 0.8	
	E1	E2	E1	E2
5	0.153[a] (3.3)[b]	0.164 (3.2), 0.159[c]	0.456 (1.5)	0.348 (1.9), 0.342
10	0.074 (5.0)	0.087 (4.6), 0.096	0.265 (2.4)	0.199 (2.8), 0.208
20	0.036 (7.3)	0.040 (6.9), 0.038	0.116 (3.9)	0.097 (4.3), 0.095
30	0.016 (11.1)	0.016 (11.1), 0.016	0.060 (5.6)	0.049 (6.2), 0.048
40	0.006 (18.5)	0.006 (18.5), 0.007	0.034 (7.6)	0.026 (8.7), 0.026

[a] Estimated from simulation study.

[b] Percent CVM values are given within parentheses. CVM = $100 \times \sqrt{P(1 - P)/5{,}000}/P$; in this example, a 95% confidence interval for the long-run average of P is $[1 \pm (2 \times 0.033)]$, $P = (0.143, 0.163)$.

[c] True probability computed from the chi-square distribution, using linear interpolation.

necessary to bring the conditional probability of a type II error with E3 down to 0.05. Using the Gaussian distribution to approximate the sampling distribution of E3, the required number of pairs is found to be 44. Thus, a total of 88 titer measurements are required with E3 to give the same protection against a type II error that 32 or 30 titer measurements will give when E1 or E2 is used. This is a substantial difference in the total number of measurements.

Illustrative example for E1 and E2. The test kit integer titer data in Table 3 were obtained from the 30 morning runs for one specimen in the study previously described (6). From these integer titers, E1 is computed as $1 - (2 \times 40)/(30)^2 = 0.911$. This computation is obtained from Table 3 as follows (6): there are 1×4 possible pairs of integer titers having a maximum ratio exceeding 2 and a least integer titer (LIT) of 10. For an LIT of 11, the number of possible pairs is 1×3; for an LIT of 12, the number is 3×2; for an LIT of 13, the number is 4×2; for an LIT of 14, the number is 5×2; for an LIT of 15, the number is 3×1; and for an LIT of 16, the number is 6×1. The total number of possible pairs of integer titers with a maximum ratio exceeding 2 that could be drawn from Table 3 (drawing with replacement) is therefore $40 = W$. With the formula E1 = $1 - (2 \times W)/N^2$, E1 = 0.911.

TABLE 3. *Test kit integer titers obtained in 30 independent daily runs with a single specimen*

Integer titer	Frequency
10	1
11	1
12	3
13	4
14	5
15	3
16	6
17	2
20	1
22	1
23	1
30	1
33	1

After changing the integer titers in Table 3 to logarithms to the base e, the SD in logarithms to the base e (SDLE) is computed as 0.2568. Table 1 gives E2 as 0.95 when SDLE is 0.2501 and as 0.94 when SDLE is 0.2606. With an SDLE of 0.2568 and linear interpolation, E2 = 0.943.

With $n = 30$ and TR at approximately 0.90, the coefficient of correlation between E1 and E2 is approximatley 0.98 (Fig. 1). The bias in the two estimators is essentially zero (Fig. 2 and 3). The combination of high correlation and low bias is manifested in the agreement of the two estimates in this example, in which E1 = 0.911 and E2 = 0.943. At a sample size of $n = 30$, the likelihood of a type II error is in the range of 5 to 6% (Fig. 4).

Because the log (integer titer) values are Gaussian distributed, a confidence interval for TR can be computed using the sample SDLE. The quantities $[(n - 1) \times \text{SDLE}^2/\text{chi-square}\ (n - 1, 0.025)]^{1/2}$ and $[(n - 1) \times \text{SDLE}^2/\text{chi-square}\ (n - 1, 0.975)]^{1/2}$ provide a 95% confidence interval (5) for the population standard deviation, designated here as SIGMA. From a table of the percentage points of the chi-square distribution (5), one obtains the two values chi-square (29, 0.025) = 45.7 and chi-square (29, 0.975) = 16. The confidence limits for SIGMA are computed as $[29 \times 0.6594624/45.7]^{1/2} = 0.2046$ and $[29 \times 0.06594624/16]^{1/2} = 0.3457$. These two quantities can be used (Table 1) to obtain the 95% confidence interval for TR as (0.843, 0.983).

In summary, the single-point estimate of TR is E2 = 0.943. Furthermore, one has 95% assurance that TR does not reside outside the interval of 0.843 to 0.983.

If the replicate titers are obtained in a single run, E1 and E2 will estimate within-run reproducibility. If they are all obtained in different runs, E1 and E2 will estimate among-run reproducibility. If the replicates are obtained by some other study design, it is possible that one does not have the option of using E1. In these cases, E2 may offer the only means of arriving at a proper estimate of TR (see Appendix E).

Illustrative example for E3. The data in Table 4 were obtained from the same specimen as the data in Table 3. The pairings were made in the blind, subject to the condition that the members of a pair were obtained on different days and therefore in different runs. Among these 44 pairs, there were 3 that had a maximum ratio exceeding 2. This leads to the estimate of TR that is E3 = 1 − 3/44 = 0.931. Note that this estimate is in close agreement with E1 and E2 for the same specimen, but that E3 is here based on 88/30 = 2.9 times as much data.

TABLE 4. *Pairs of test kit integer titers obtained on the specimen used for Table 3*[a]

Pair no.	Titer	
	1st	2nd
1	13	12
2	13	30
3	19	13
4	16	14
5	14	19
6	20	24
7	10	21
8	28	16
9	14	20
10	15	15
11	15	19
12	14	16
13	33	17
14	20	18
15	19	15
16	14	16
17	12	12
18	19	12
19	13	16
20	18	16
21	22	17
22	23	22
23	23	12
24	18	15
25	14	14
26	21	15
27	16	12
28	11	16
29	20	19
30	18	26
31	16	13
32	20	19
33	22	12
34	30	14
35	21	21
36	17	18
37	12	21
38	12	13
39	22	13
40	17	14
41	17	16
42	25	23
43	15	19
44	17	18

[a] Integer titers in each pair were obtained on different days. Maximum ratios were: pair 2, 2.31; pair 7, 2.10; and pair 34, 2.14.

DISCUSSION

Four methods for estimating TR with nontruncated integer titers have been discussed. Two of these, E1 and E3, being nonparametric or distribution-free methods of estimating, can readily be used with either integer or truncated sample titers. The estimators E2 and E4 (the latter discussed in Appendix E), being Gaussian parametric estimators, are properly used only with data that follow the Gaussian distribution. If the log (integer titer) values follow this distribution (Appendix A), then clearly the log (truncated titer) values do not, and therefore the estimators E2 and E4 would not be appropriate with truncated titers. This subject will be pursued in a future report. The estimators E1 and E2 were found to be highly correlated (Fig. 1) and to have negligible biases at moderate sample sizes or moderately high TR levels or both (Fig. 2 and 3). The reliability or stability of the two estimators was studied, using a particular definition of type I and type II errors, for different sample sizes. Approximately 30 replicate integer titers on a chosen specimen were seen to be required for E1 or E2 to provide acceptably reliable estimates of TR (Fig. 4). Approximately this same number (29) of pairs of integer titers was required for the parametric estimator E4 (see Appendix E) to attain the same reliability, whereas the nonparametric estimator E3 required 44 pairs to provide essentially the same degree of reliability in the estimate.

APPENDIX A

The Monte Carlo distribution sampling reported in this paper is based on the Gaussian distribution. The

justification for the assumption that the log (integer titer) values follow this distribution is the subject of this appendix.

A specially designed rubella antibody serum titration study was described previously (6). The study resulted in independent fluorescent-antibody FIAX (International Diagnostic Technology, Santa Clara, Calif.) titer measurements on 14 different hemagglutination inhibition-positive specimens. Each specimen was measured in three separate runs per day for 30 days. In each of the 90 runs, a calibration curve was determined and used to obtain the logarithms of the flourescent-antibody titers for all specimens in that run. Consequently, the resulting titers did not come truncated to integer powers of 2, which are conventional serum dilution titers, but were free to take on any of a continuum of integer values. The resulting 14 geometric mean titers ranged from a low positive of 7.0 to a high positive of 389.3.

Because of the possibility of a day-to-day component of variation, the data for each of the 14 specimens were grouped by the morning run, the midday run, and the afternoon run. This resulted in 30 independent daily log (integer titer) values in each of 14 × 3, or 42, groups. The 30 values in each of the 42 groups were normalized by subtraction of the group mean and division by the group SD.

Within each group of 30, a count was made of the numbers of normalized values that fell in the six successive intervals defined by the 1/6, 2/6, 3/6, 4/6, and 5/6 percentile of the standard normal curve. If the 30 log (integer titer) values did arise from a Gaussian distribution, then on the average a count of 5 was to be expected in each of these six intervals.

The X^2 statistic was computed for each group as $X^2 = \sum_{i=1}^{6} (O_i - 5)^2/5$, where O_i is the observed count for the i-th interval. If the underlying log (integer titer) values are Gaussian, then X^2 is approximately chi-square distributed with 4 degrees of freedom. These values computed for the 42 independent groups were added together to yield an overall X^2 based on 168 degrees of freedom. This overall X^2 value was 162.8, which converts to a normal deviate value of -0.26 through the transformation $z = \sqrt{2X^2} - \sqrt{2df - 1}$.

With the hypothesis that the underlying log (integer titer) distribution is Gaussian, the probability of a poorer fit than the one represented by $z = -0.26$ is found in a table of normal curve values to be $0.79 < P$. From this result, it was concluded that the underlying distribution of log (integer titer) values could be simulated with the Gaussian distribution.

APPENDIX B

The subject measure of reproducibility is "the probability that the maximum ratio of two distinct (integer) titers (obtained in the blind) on the same specimen will not exceed 2." This is equivalent to the logarithms (to the base 2) of the two independent integer titers not differing by more than 1.

The log (integer titer) values for any positive specimen are taken to be Gaussian distributed (see Appendix A) with mean μ and SD σ. Consequently, the difference (d) between two independent log (integer titer) measurements made on the same specimen will also be Gaussian distributed but with a mean of 0 and a standard deviation equal to $\sqrt{2}\sigma$. Then $z = d/\sqrt{2}\sigma$ is a standard normal deviate having a mean of 0 and an SD of 1.

The probability that a randomly drawn d will fall within the interval (± 1) is $P(-1 \leq d \leq 1) = 1 - 2\,P(1 < d) = 1 - 2[1 - P(d < 1)] = 2P(d < 1) - 1$. Writing this in terms of z, the subject reproducibility is $R = 2P(z < 1/\sqrt{2}\sigma) - 1 = 2F(1/\sqrt{2}\sigma - 1$, where F is the cumulative distribution function of z. This relationship between the SD of the log (integer titer) population distribution and the subject reproducibility offers a second means of estimating that reproducibility. It only remains to obtain a random sample estimate of σ (SD) and compute $E2 = 2\,F(1/\sqrt{2}SD) - 1$. This was made possible when the function F was specified as a result of assuming a population distribution for the log (integer titer) values (see Appendix A).

To facilitate the use of this parametric estimating procedure, a conversion table of SD and E2 values was computed by using a power series representation (1) (Table 1).

APPENDIX C

The IBM VSPC-APL interactive software system was used to generate random numbers with the IBM-370/158 computer at the Centers for Disease Control. These uniform numbers were used pairwise in the Box-Muller transformation (2) to yield independent Gaussian random numbers from a population with 0 mean and unit variance. Independent studies were carried out for the five sample sizes n = 5, 10, 20, 30, and 40.

For each sample size, 5,000 independent samples of size n were generated from the standard (0, 1) Gaussian population. For each of the 5,000 samples of size n, the sample mean and SD were computed. Each of the n log (integer titer) values, together with the sample SD, was multipled by a conversion number corresponding to a preselected reproducibility value for the sampled population (see Appendix B and Table 1). This caused the sample of n to appear as if it had been drawn from a population having the preselected reproducibility parameter. The sample estimates E1 and E2 were then computed.

Each of the n original log (integer titer) values in the sample, together with the sample SD, was then multipled by a second conversion number corresponding to a second preselected reproducibility parameter value. A second E1 and E2 were computed. This procedure was repeated for the same sample of n log (integer titer) values a total of five times. In this manner, each sample of size n yielded both a nonparametric estimate (E1) and a parametric estimate (E2) of the system reproducibility when the latter was set in turn to 0.95, 0.90, 0.80, 0.70, and 0.60. In each case, the 5,000 estimates of reproducibility were taken to approximate the corresponding population distrubution of estimates, and selected statistics were computed. The descriptive capability of each of these selected statistics, when viewed across the range of preselected TR values, was enhanced through the correlation induced by using the same sample of ele-

ments with the five different selected reproducibility values.

APPENDIX D

For the nonparametric estimator E1, the simulation-based estimates of the conditional probabilities of type I and type II errors are shown in Fig. 4. They are also given in Table 2, together with a measure called here the "coefficient of variation of the mean" (CVM). This useful measure is 100 times the standard error of the mean divided by the mean, or the conventional coefficient of variation divided by the square root of the sample size. The CVM readily leads to a useful confidence interval for the true mean. This is obtained as 100% ± 2 × percent CVM to yield a 95% confidence interval. Consequently, one has assurance at approximately the 95% level that the observed sample mean is within 2 × percent CVM of the population mean. From Table 2, the conditional probability of a type I error with E1 is estimated to be 0.153 when $n = 5$. The estimated CVM of 3.3% indicates that we have approximately 95% assurance that this sample estimate, 0.153, is within 2 × 3.3% = 6.6% of the true value in the population. A 95% confidence interval for the true conditional probability is given by (1.000 ± 0.066) × 0.153, or 0.143 to 0.163.

For the parametric estimator E2, the conditional probabilites of type I and type II errors can be computed directly from the chi-square distribution. This is because a constant times the square of the sample SD from Gaussian log (integer titer) values follows the chi-square distribution. These true probabilities were computed and are given in Table 2, together with the corresponding estimates obtained from the simulation. Comparison of the computed true probabilites with those obtained from the simulation for E2 offers an overall check on the accuracy of the simulation.

TABLE 5. *Differences of the logarithms to the base e for the first 29 pairs of integer titers in Table 4*

Pair no.	Difference of natural logarithms	Difference squared[a]
1	0.800 427	0.006 406 835
2	0.836 248	0.699 310 758
3	0.379 490	0.144 012 373
4	0.133 531	0.017 830 633
5	0.305 382	0.093 257 952
6	0.182 322	0.033 241 150
7	0.741 937	0.550 471 024
8	0.599 616	0.313 169 830
9	0.356 675	0.127 217 016
10	0.064 539	0.093 257 952
11	0.236 389	0.055 879 654
12	0.133 531	0.017 830 633
13	0.633 294	0.439 959 219
14	0.105 361	0.011 100 838
15	0.236 389	0.055 879 654
16	0.133 531	0.017 830 633
17	0	0
18	0.459 532	0.211 169 962
19	0.207 639	0.431 141 058
20	0.177 783	0.013 872 843
21	0.257 829	0.066 475 850
22	0.044 452	0.001 975 959
23	0.650 588	0.423 264 181
24	0.182 322	0.033 241 150
25	0	0
26	0.336 472	0.113 213 566
27	0.287 682	0.082 760 975
28	0.374 693	0.140 395 181
29	0.051 293	0.002 631 002

[a] Total differences squared, 4.107 705 150. SDLE = $\sqrt{4.107\ 705\ 150/58}$ = 0.2661. E4 = 0.934.

For the parametric estimator E2, the conditional probabilites of type I and type II errors can be computed directly from the chi-square distribution. This is because a constant times the square of the sample SD from Gaussian log (integer titer) values follows the chi-square distribution. These true probabilities were computed and are given in Table 2, together with the corresponding estimates obtained from the simulation. Comparison of the computed true probabilites with those obtained from the simulation for E2 offers an overall check on the accuracy of the simulation.

APPENDIX E

It has been emphasized that E1, E2, and E3 require data having a particular structure; that is, each of the replicate integer titers must be from an independent run of the test system. Such data contain a single measure of within-run variation and a single measure of among-run variation. This is so because within-run measurement error, or variation, is introduced at the point of measurement within the particular run, and a different among-run component of variation is an integral part of each run. When the difference of logarithms is formed for such measurements, it contains the difference between two independent within-run measurement components as well as the difference between the two independent among-run components. The key point is that all such differences between any two log (integer titer) values contain these same basic components in equal quantities, that is, two of each. An SD computed from such integer titers is based on these differences and, therefore, reflects an equal weighting of the within-run component and the among-run component. Since E2 is obtained directly from this standard deviation, it too reflects an equal weighting of the within-run and among-run components. Clearly then, the basic measure of overall reproducibility considered here rests on an equal weighting of these two components of measurement variation.

The essence of the above statements is that the underlying restriction does not actually apply to the structure of the data itself but rather to the SD that is computed. The data may have any structure whatever, so long as the proper SD estimate can be extracted from it. The data collection study design discussed in this paper is such that the straightforward SD of the log (integer titer) values contains the two subject components of variation automatically weighted equally. Other study designs may be used, however, so long as the proper sample SD can be synthesized from the data. With certain designs, this can readily be done by means of the analysis of variance procedure to estimate singly the within-run component of variance and the among-run component of variance. These can be combined to yield an overall SD estimate. These two SD estimates can themselves be used in Table 1 to obtain estimates of the within-run reproducibility and the overall reproducibility of the subject test system.

In the same vein, the analyst restricted to paired data is not limited to the nonparametic estimate E3.

The analysis of variance procedure can be used to obtain a parametraic estimate in the manner just described. For this estimate, called E4, the k differences between log (integer titer) values within each pair would be squared, summed, and divided by $2k$. The square root of this quantity is the SD to be used in Table 1. In this case, the effective sample size is $k + 1$ relative to the estimate E2. That is, k pairs of titer measurements lead to E4, a parametric estimate of TR that has no more reliability than E2 has with only $k + 1$ replicates. In an example given above, it was seen that E3 requires 44 pairs to provide the same reliability obtainable with E2 based on 30 replicate titer measurements on the same specimen. This same reliability is obtainable with E4 by using 29 pairs of titer measurements. In this example, the reliability of the parametric estimate E4 is equivalent to that of the nonparametric estimate E3, when E4 is based on only 29/44, or two-thirds, as much data. The data in Table 5 are the differences of the logarithms to the base e for the first 29 pairs of integer titers shown in Table 4. The sum of the squared differences divided by $2k = 58$ is 0.070823. The square root of this quantity is 0.2661 which, in Table 1, gives the estimate of TR that is E4 = 0.934.

LITERATURE CITED

1. **Abramowitz, M., and I. A. Stegun.** 1965. Handbook of mathematical functions. National Bureau of Standards, Washington, D.C.
2. **Box, G. E. P., and M. E. Muller.** 1958. A note on the generation of random normal deviates. Ann. Math. Stat. **29**:610-611.
3. **Bradley, J. V.** 1968. Distribution-free statistical tests. Prentice-Hall, Inc. Englewood Cliffs, N.J.
4. **Hollander, M., and D. A. Wolfe.** 1973. Nonparametric statistical methods. John Wiley & Sons, Inc. New York.
5. **Ostle, B.** 1963. Statistics in research. The Iowa State University Press, Ames.
6. **Wood, R. J., and T. M. Durham** 1980. Reproducibility of serological titers. J. Clin. Microbiol. **11**:541-545.