Guidelines for the statistical analysis of a collaborative study of a laboratory method for testing disinfectant product performance

Authors: Martin A. Hamilton, Gordon C. Hamilton, Darla M. Goeres, & Albert E. Parker

NOTICE: This is a postprint of an article that originally appeared in Journal of AOAC International on September 2013. DOI: http://dx.doi.org/10.5740/jaoacint.12-217.

Hamilton MA, Hamilton GC, Goeres DM, Parker AE, "Guidelines for the statistical analysis of a collaborative study of a laboratory method for testing disinfectant product performance," J AOAC International. 2013 96(5):1138-1147

Guidelines for the Statistical Analysis of a Collaborative Study of a Laboratory Method for Testing Disinfectant Product Performance

Martin A. Hamilton and Gordon Cord Hamilton
Big Sky Statistical Analysts LLC, 309 South Sixth Ave, Bozeman, MT 59715

Darla M. Goeres and Albert E. Parker
Montana State University, Center for Biofilm Engineering, Bozeman, MT 59717-3980

This paper presents statistical techniques suitable for analyzing a collaborative study (multilaboratory study or ring trial) of a laboratory disinfectant product performance test (DPPT) method. Emphasis is on the assessment of the repeatability, reproducibility, resemblance, and responsiveness of the DPPT method. The suggested statistical techniques are easily modified for application to a single laboratory study. The presentation includes descriptions of the plots and tables that should be constructed during initial examination of the data, including a discussion of outliers and QA checks. The statistical recommendations deal with evaluations of prevailing types of DPPTs, including both quantitative and semiquantitative tests. The presentation emphasizes tests in which the disinfectant treatment is applied to surface-associated microbes and the outcome is a viable cell count; however, the statistical guidelines are appropriate for suspension tests and other test systems. The recommendations also are suitable for disinfectant tests using any microbe (vegetative bacteria, virus, spores, etc.) or any disinfectant treatment. The descriptions of the statistical techniques include either examples of calculations based on published data or citations to published calculations. Computer code is provided in an appendix.

It is imperative that appropriate statistical procedures are used to reveal, summarize, and communicate the results of a multilaboratory collaborative study (ring trial). Various official guidelines are available to steer the statistical analysis of collaborative studies of chemical assay methods (1â€“5). Guidelines for analyzing studies of methods for ascertaining microbial contamination of foods list statistical techniques that are very similar to the official guidance for chemical assays (6, 7). Limited guidance exists, however, for the study of laboratory test methods designed to evaluate the attributes of a disinfectant product performance test (DPPT) method. For studies of chemical assays and microbial contamination of food assays, the statistical analysis guidance documents use generic terminology and conform to established scientific practice; consequently, they are widely applied for method evaluations in other fields. Although they provide some information relevant to DPPT method studies, they are inadequate for analyzing data from a DPPT method collaborative study because DPPT methods differ in many important ways from chemical assays and microbial contamination of food assays (8–10). A DPPT is a microbiological assay that assesses the ability of a disinfectant product to kill or otherwise inactivate living microbes. The DPPTs employ different experimental designs, use different terminology, produce fundamentally different response measurements, have additional goals, and apply the results to different uses than chemical assays or microbial contamination of food assays. Some topics important to DPPTs are not covered in existing guidance documents because those topics are not relevant to chemical assays or microbial contamination of food assays. On the other hand, there is much information in the guidelines for chemical assays and microbial contamination of food assays that does not pertain to DPPTs.

Regulatory authorities and consumers rely on laboratory tests to indicate the performance of disinfectant products when applied in practice. Because disinfection is a critical component of disease prevention strategies, the laboratory DPPT method must be rigorously evaluated so that one can be confident in the test results for any specific disinfectant product. For all these reasons, there is a need to consolidate and present statistical techniques that are suited specifically to collaborative study evaluations of DPPT methods.

The purpose of this paper is to describe statistical techniques for analyzing data from a collaborative study of a standardized DPPT method. The presentation lies at the intersection of microbiological laboratory methods and statistical methods. It organizes the terminology, goals, and statistical techniques into a framework that should facilitate communication between the microbiologist and statistician. Its main goal is to help the statistician by suggesting proven, practical statistical techniques for analyzing DPPT study data.

The terminology and mathematical notation used in this paper are summarized in Tables 1 and 2. The presentation covers prevailing types of DPPTs, including both quantitative and semiquantitative tests, as summarized in Table 1 (11). Until a few years ago, some popular DPPT methods did not require data for untreated control carriers; however, such methods are obsolete and are not discussed here. This paper describes: (a) the basic data collected in collaborative studies

of DPPTs, including the typical study factors and response variables; (b) the plots and tables that have proven helpful during the initial examination of collaborative study data, including a discussion of outlier detection and evaluation; (c) recommended QA checks; and (d) guidance on statistical techniques for assessing the extent to which the DPPT method attains the important attributes of repeatability, reproducibility, resemblance, and responsiveness (12). It is important to remember that a collaborative study is conducted to assess the attributes of the DPPT method, not to evaluate the specifi disinfectant treatments used in the study. Notable topics not covered here include guidelines for the design and management of a collaborative study and serviceable strategies for creating and optimizing a new standardized DPPT method.

Recommendations in this article adapt and extend the repeatability and reproducibility assessment methods described earlier (4, 13), making these recommendations consistent with existing guidelines for collaborative studies of chemical assays or microbial contamination of food assays. For each major analysis step, these guidelines cite a published data analysis and, in an appendix, show how to perform the key calculations using a popular statistical software package. This paper incorporates terminology and methodologies used in recent reports on collaborative studies of DPPT methods (14–16). Because each DPPT has its own special characteristics and a collaborative study may involve alternative goals, one should apply with circumspection the statistical techniques presented in this paper.

Disinfectant Product Performance Tests

A DPPT protocol describes in detail the microbiological techniques used when performing the test. The difficulty of the test method depends on the existent environment being emulated, the means by which a chemical disinfectant is applied (liquid, spray, wet wipe towelette, etc.), and on the test microorganisms (vegetative bacteria, spores, viruses, protozoa, etc.). The microbiological manipulations usually require experienced microbiology laboratory specialists.

The disinfectant may be tested against either planktonic microbes suspended in a liquid or a collection of surface-associated microbes. Various suspension test methods have been evaluated thoroughly and are used widely, for example, to screen chemicals for antimicrobial activity and to assess the efficacy of drinking water disinfectants or recreational water disinfectants (17, 18). Methods for testing a disinfectant against surface-associated microbes typically use easily manipulated carriers, e.g., glass disks. A variety of methods are used to place and hold microbes on the carriers. The examples in this article focus on tests using surface-associated microbes because it is anticipated that most new standardized DPPTs will be surface tests. However, the discussion can be applied to suspension tests if the suspension test beakers or tubes are considered to be carriers.

In a surface DPPT, some microbe-bearing carriers are treated with the disinfectant, and others serve as untreated control carriers. For treated carriers, the disinfectant is neutralized to stop its activity at the end of the designated contact time. Untreated carriers receive the same manipulations as the treated carriers (including neutralization), except that an inactive treatment, such as dilution water, is applied instead of the disinfectant.

A disinfectant is formulated to kill microbes; consequently, for an effective disinfectant, the treated carriers should hold few viable microbes relative to the untreated carriers. Disinfectant efficacy is quantified by comparing the typical number of viable microbes on the treated carriers to the typical number on the untreated carriers. Depending on the microbes and test system, it may be possible to perform a quantitative assessment of the number of viable microbes on a carrier by conventional enumeration methods, e.g., harvesting the microbes from the carrier surface into suspension, disaggregating the microbes, performing a dilution series, plating, incubating, and counting colonies (16, 19).

Collaborative Studies—As Viewed by the Statistical Analyst

A collaborative study of a DPPT method is a multilaboratory assessment of the method conducted by replicate testing of the same “disinfectant treatment” (Table 1) in each of several laboratories. The basic goals of a collaborative study include determining the extent to which the efficacy outcome achieves the desirable attributes of repeatability within a laboratory, reproducibility among laboratories, and responsiveness to increasing disinfectant efficac . An additional goal is determining the extent to which the untreated carriers resemble each other from test to test and from laboratory to laboratory, where resemblance is with respect to the number of microbes on each untreated carrier. Additional goals are not unusual, e.g., to compare a new DPPT method to a similar established DPPT method.

Statistical Analyst’s Preparation

The statistical analyst should study the collaborative study proposal, goals, and protocol, as well as relevant background information about the DPPT method (e.g., research conducted during development of the method). The Study Director should notify the statistical analyst about any deviations from the protocol or the study design and about any unanticipated occurrences that potentially could affect results or the choice of statistical techniques. It is advisable for the statistical analyst to list the study factors, response variables, and any other variables of importance that were recorded in the collaborative study. Note that factors are also called independent variables, predictor variables, or explanatory variables; response variables are also called dependent variables (20). Generally, the other variables of importance are covariates (or concomitant variables), such as ambient humidity and temperature (20). The list of factors and variables will help the analyst understand the scope of the study and choose the statistical summaries and analytical models suitable for attaining the study goals.

Study Factors

In a conventional collaborative study, the study factors will include test date, microbe, disinfectant, efficacy level of the disinfectant, technician (or technician team), and laboratory. The test date is either the day the test began or the day the final response was observed, a choice applied consistently throughout the study. The recorded response for each carrier will be accompanied by a unique code for each test and a code

that indicates whether the response is for an untreated carrier or a treated carrier. If the collaborative study has an expanded list of goals, the data set will include additional factors pertaining to those goals. For example, in a method comparison collaborative study, the data will include a method code.

Response Variables

The response variables are dependent on the type of DPPT being conducted: quantitative, semiquantitative, or alternative (Table 1). For quantitative and semiquantitative studies, the primary response variables for each test are the untreated carrier $\log_{10}$-transformed viable microbe densities and the log reduction (LR) measure of efficacy (formulas given below in subsections "LR for a single quantitative test" and "LR for a single semiquantitative test of type SQ_1" of the *Response Variables* section).

Mathematical formulas are useful when discussing the response variables (notation in Table 2). When viable microbes can be enumerated by counting colony forming units (CFU), let D denote the viable microbe density associated with a carrier, typically expressed in units of CFU/carrier or CFU/cm^2 of carrier surface area. The protocol for the DPPT method should specify the density units, and the same units must be used for both treated and untreated carriers.

The $\log_{10}$-transformed density D is called the log density, denoted by LD. Experience has shown that not only do the LD values for replicate carriers usually follow a symmetric distribution (21), but they also conform to the assumptions required for conventional normal-theory statistical techniques. Therefore, when analyzing DPPT data, it is standard practice to perform statistical calculations on LD values, not on the untransformed densities (13, 22). The mean of LD values, averaged across a set of carriers (formula and notation in the next few paragraphs) is an important summary statistic. When the final results are presented on the density scale, the geometric mean density, which is the antilog of the mean LD, is the preferred indicator of the typical density for the set of carriers (22, 23).

For a single test, let J denote the number of untreated carriers and K denote the number of treated carriers. Let U_j denote the LD for the j^{th} untreated carrier, j = 1, 2, …, J. The mean of untreated carrier LDs in the test is:

$$TestLD = \frac{1}{J}\sum_{j=1}^{J} U_j$$

The geometric mean density for the test is 10^{TestLD}. The $TestLD$ is a standard gauge of the microbial challenge. In addition to being a critical component of the LR, $TestLD$ is the key response variable for a resemblance analysis (details later).

For quantitative and semiquantitative tests, an important response variable is LR, the conventional measure of efficacy for a disinfectant treatment. The LR for a test is calculated differently for quantitative and semiquantitative methods (24). The recommended calculation formulas will now be presented.

LR for a single quantitative test.—For a quantitative test, LR is found by subtracting the mean of log densities for the treated carriers from the mean of log densities for the untreated carriers. Let T_k denote the LD for the k^{th} treated carrier, k = 1, 2, …, K; the mean of the treated carrier log densities is $\overline{T} = \frac{1}{K}\sum_{k=1}^{K} T_k$. Equation 1 presents the formula for the quantitative test LR (25, 26):

$$\mathrm{LR} = TestLD - \overline{T} \quad (1)$$

Let US and TS denote the within-test SD of log densities for untreated carriers and treated carriers in a single test, respectively, as defined in Equation 2

$$US = \sqrt{\frac{\sum_{j=1}^{J}(U_j - TestLD)^2}{J-1}}$$

and

$$TS = \sqrt{\frac{\sum_{k=1}^{K}(T_k - \overline{T})^2}{K-1}} \quad (2)$$

Equation 3 shows the within-test SD of LR (25), which is denoted by S:

$$S = \sqrt{\frac{US^2}{J} + \frac{TS^2}{K}} \quad (3)$$

LR for a single semiquantitative test of type SQ_1.—For a semiquantitative test, some or all of the viable microbe densities are based on positive or negative (P/N) observations (Table 1). For a Type SQ_1 test, which is the most common semiquantitative test, viable microbes are enumerated on untreated carriers, but a P/N outcome is recorded for each treated carrier. Let NP_T denote the number of positive treated carriers among the set of K in the test. Let $\overline{T}_{NP}$ denote the mean LD for treated carriers, a value that can be calculated from the NP_T outcome using the formula shown in Equation 4, where ln(x) is the natural logarithm of the positive number x:

$$\overline{T}_{NP} = \log_{10}\left[-ln\left(\frac{K - NP_T + 0.5}{K + 1.0}\right)\right] \quad (4)$$

The $\overline{T}_{NP}$ value of Equation 4 is the $\log_{10}$ transformation of the single dilution most probable number (MPN), an often-used method for calculating the typical density (27, 28), where the conventional MPN formula has been adjusted slightly by adding 0.5 to the numerator and 1.0 to the denominator when calculating the fraction of carriers that were negative. The adjustment assures that $\overline{T}_{NP}$ can be calculated even when the observed NP_T is one of the extremes, 0 or K. Garthright (29) determined that, as an estimate of the mean LD, the log-transformed MPN is negligibly biased.

Equation 5 provides the formula for LR for a type SQ_1 test (30):

$$\mathrm{LR} = TestLD - \overline{T}_{NP} \quad (5)$$

For semiquantitative tests of types SQ_2 and SQ_3, the LR also can be calculated from log-transformed MPN values for both the treated and untreated carriers.

QA Steps

Before computing analytical results, it is important to check the data for completeness with respect to the list of factors, response variables, and other variables required to achieve the study goals. The analyst should conduct checks on the master data set to discover correctable errors in coding, data entry, basic calculations, etc. It is also important to check the data using tables and plots, which are useful tools for QC checks, for

Table 1. Terminology

Term	Meaning
Alternative test	A DPPT in which measures of viable cells are based on techniques different from conventional culture methods that utilize enumeration or positive/negative outcomes.
Analysis of covariance	An extension of ANOVA to include the mathematical relationship between the response variable and a covariate (e.g., temperature) in the statistical model.
Analysis of variance (ANOVA)	In this context, a statistical methodology for partitioning the observed variance of a response variable into component variances associated with the random effects factors.
Balanced design	All laboratories conduct the same set of tests.
Biofil	A self-organized, cooperative, sessile community of microorganisms, typically embedded in a matrix of extracellular polymeric substances.
Biofilm tes	A DPPT in which the biochallenge is a microbial biofilm on the surface of a carrier; e.g., (31, 32)
Carrier	A small, easily manipulated object with which the microbial biochallenge is associated in biofilm tests or dried surface tests, also called a coupon.
Collaborative study	Evaluation of a standardized DPPT method based on tests of identical disinfectant treatments conducted by a number of laboratories, each following the written protocol for the method.
Density	Number of viable microbes, usually expressed as CFU/unit, e.g., “CFU/carrier” or “CFU/cm^2 of carrier surface area.”
Disinfectant product performance test (DPPT)	A microbiological assay conducted to assess the ability of a disinfectant product to kill, or otherwise inactivate, living microbes.
Disinfectant treatment	Usually a chemical, or a formulation of multiple chemicals, at a specified strength or dilution for a specified contac time, and the means by which the microbes are exposed to the chemical (e.g., immersed or sprayed). Sometimes a physical treatment, such as sonication, or a physicochemical treatment, such as a disinfecting towelette.
Dried surface test	A DPPT in which the biochallenge is an assemblage of microbes in a dried organic film on a carrier surface (33)
Log density	$\log_{10}$-transformed density.
Log reduction (LR)	The mean of $\log_{10}$-transformed densities for the treated carriers subtracted from the mean of $\log_{10}$-transformed densities for the untreated carriers. If the disinfectant inactivates none of the microbes, the expected LR is zero, and if the disinfectant is effective, the expected LR is positive.
Most probable number (MPN)	Based on a serial dilution assay, where a P/N response is observed for each assemblage (well, tube, carrier, etc.) at each dilution; the MPN is the density at which the observed data are most likely (34).
Outlier	Observation that appears to deviate markedly from other members of the data set.
Positive/negative (P/N) outcome	For an assemblage of microbes (on a carrier, in a dilution tube, etc.), the outcome is positive when there is at least one viable microbe in the assemblage and is negative when all microbes in the assemblage are nonviable.
Quantal test	P/N outcome is observed for each treated carrier, and no data are collected for untreated carriers.
Quantitative test	Viable microbes on each carrier are enumerated; the typical result of the enumeration is the density of viable microbes.
Repeatable	A repeatable DPPT produces nearly the same LR values when the same disinfectant treatment is tested on different days within a laboratory.
Reproducible	A reproducible DPPT method will produce nearly the same LR value when the same disinfectant treatment is retested in a different laboratory.
Resemblance	Inoculated carriers are the experimental units in a surface DPPT. It is desirable for the carriers to resemble each other. For quantitative and semiquantitative DPPT methods, resemblance assessment is based on viable microbe densities on control carriers.
Responsive	A responsive DPPT method is sensitive enough that it can detect an important efficacy–response e fect.
Semiquantitative test	Data are collected for both treated and untreated carriers; some or all of the viable microbe densities are based on P/N outcomes.
Suspension test	A DPPT in which the biochallenge is a suspension of planktonic microbes (17).
Type SQ_1 semiquantitative test	For each untreated carrier the viable microbe density is enumerated by a conventional counting technique and for each treated carrier a P/N whole carrier outcome is recorded (35).
Type SQ_2 semiquantitative test	For each carrier, treated or untreated, the microbes are harvested from the carrier into suspension, and a dilution series is formed for each suspension. Instead of using counting techniques for enumerating viable microbes, multiple tubes (or wells) are created from each dilution, a P/N outcome is observed for each tube, and the density of viable microbes for each carrier is found by the MPN method (36).
Type SQ_3 semiquantitative test	A whole-carrier P/N outcome is observed for each treated carrier. For each control carrier, the microbes are harvested, the suspension is serially diluted, multiple tubes (or wells) are created from each dilution, P/N outcomes are observed for each tube, and carrier density is calculated by the MPN method.
Unbalanced design	Some of the laboratories conduct a different set of tests than other laboratories.

Table 2. Notation

Symbol or abbreviation	Denotation or meaning
ANOVA	Analysis of variance
CFU	Colony-forming units (when enumerating viable microbes)
DF	Degrees of freedom
D	Viable microbe density, for which the typical units are CFU/carrier or cfu/cm^2
DPPT	Disinfectant product performance test
higher	When a subscript, indicates the presumed higher efficacy disinfectant treatment when discussing responsivenes
J	Number of untreated carriers/test
K	Number of treated carriers/test
L	Number of participating laboratories
LD	Log density; $\log_{10}$-transformed *D*
lower	When a subscript, indicates the presumed lower efficacy disinfectant treatment when discussing responsiveness
LR	Log reduction
M	Number of replicate tests in each laboratory
MPN	Most probable number method for calculating the density of viable microbes
NP_T	Number of positive carriers among the *K* treated carriers
P/N	Positive or negative; an observation is positive if one or more microbes can produce a growth response and it is negative if all microbes are incapable of a growth response, under the culture conditions of the test
$Resp_{LR}$	Responsiveness measure for LR values; quantitative and semiquantitative tests
S^2_{lab}	Variance among laboratories for the LR
S	Within-test SD of LR
SD	Standard deviation
SEM	Standard error of the mean
SQ_1, SQ_2, SQ_3	Types of semiquantitative test methods (Table 1)
$S_{r\ell}$	Repeatability SD for LR values within laboratory ℓ, ℓ = 1, ..., *L*
S_r	Repeatability SD for the LR, applicable to all laboratories
S_R	Reproducibility SD for the LR
TestLD	Mean log density for the *J* untreated carriers in a test
T	Log density for a treated carrier
$\overline{T}$	Mean log density for the *K* treated carriers in a quantitative DPPT
$\overline{T}_{NP}$	Log density on the typical treated carrier in a semiquantitative DPPT, e.g., Equation 4
T_k	Log density for the k^{th} treated carrier in a DPPT, *k* = 1, 2, …, *K*
TNTC	Colonies on plates or filters are too numerous to coun
TS	Within-test SD for the treated carrier *LDs*
U	Log density for an untreated carrier
U_j	Log density for the j^{th} untreated carrier in a DPPT, *j* = 1, 2, …, *J*
$US_{r\ell}$	Repeatability SD for *TestLD* values within laboratory ℓ, ℓ = 1, ..., *L*
US_r	Repeatability SD for *TestLD*, applicable to all laboratories
US_R	Reproducibility SD for *TestLD*
US^2_{lab}	Variance among laboratories for the untreated carrier *LD*
US^2_{test}	Variance among tests within a laboratory for the untreated carrier *LD*
US	Within-test SD for the untreated carrier *LDs*

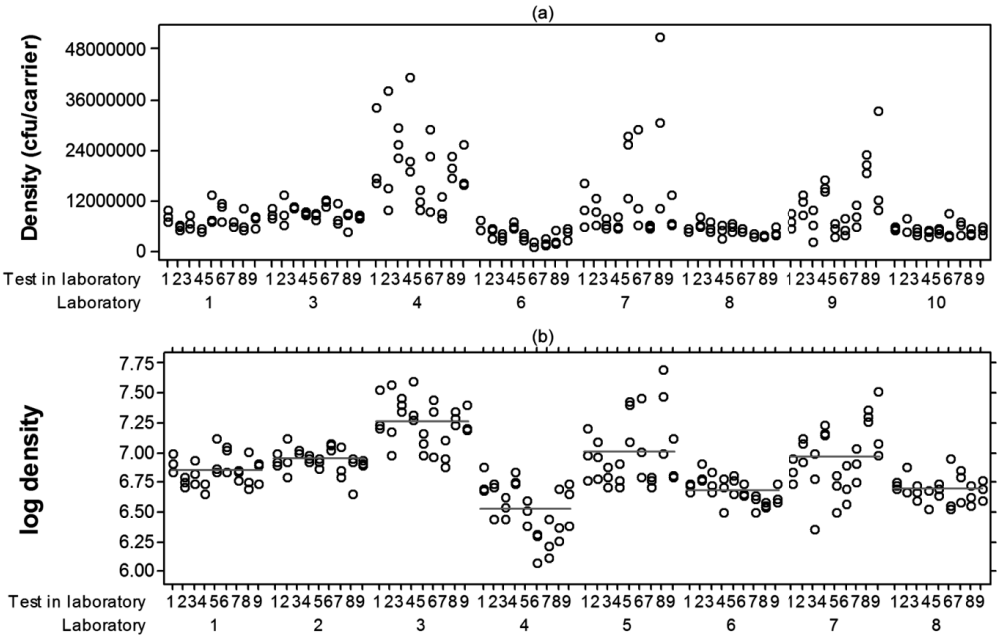

Figure 1. Each symbol is the LD for one untreated carrier. The tests in each laboratory are numbered chronologically. The data are from a collaborative study of a quantitative dried surface test; Appendix B of (ref. 16). The individual responses are aligned vertically for the *J* = 3 untreated carriers/test. Panel (a) shows the density (*D*) and panel (b) shows the log density (*U*). For panel (a), every large deviation from the typical *D* for a test is in the positive direction; that is, the distribution is asymmetric, tailing off gradually for high densities. For panel (b), however, some of the large deviations are high and some are low; the vertical scatter of *U* is generally symmetric. The lines in panel (b) show the mean *U* across all tests within each laboratory.

insights into the appropriate statistical models and techniques, and for visual indications of the results. Whenever anomalies are discovered, the analyst will submit specific questions to the Study Director for resolution. Before responding, the Study Director typically will consult with appropriate laboratory personnel, review original data sheets, or utilize other relevant sources of information. Although this QA discussion focuses on viable cell densities and the LR efficacy measure, the statistical tools can be used for quantities observed in alternative study designs.

Cross-Tabulations

For finding errors in the study factor coding or in executing the study design, it is helpful to construct cross-tabulations where the rows and columns are combinations of factor levels. For example, a cross-tabulation in which each cell in the table provides the number of untreated carriers in each individual DPPT (table columns) in each laboratory (table rows) should show exactly the number *J* specified in the protocol. A value other than *J* in a cell would indicate either a data error or a violation of the study protocol. As another example, a cross-tabulation in which each table cell provides the number of tests for each disinfectant treatment (table columns) conducted by each laboratory (table rows) should match exactly the collaborative study design. Any discrepancies uncovered by the cross-tabulations should be corrected by the analyst if possible (e.g., if there was a correctable laboratory coding error in some spreadsheets); otherwise, they should be resolved by the Study Director. A cross-tabulation can clearly describe the extent of imbalance in the study results. Some imbalance is not unusual, e.g., some carriers may have to be excluded because of contamination during the plate counts, or an isolated replicate test may have to be excluded because of a laboratory error discovered belatedly.

Plots

It is good statistical practice to construct individual value plots for displaying the response for each carrier. Individual value plots show the response on one axis and combinations of study factors on the other axis. For quantitative DPPTs, individual value plots could display the log densities *U* and *T*, or sometimes the density *D* for individual carriers. For type SQ_1 semiquantitative tests, it is informative to plot each of *U,* $\overline{T}_{NP}$, and NP_T. For untreated carriers, the individual value plot of *D* usually shows asymmetric spreads among carriers in the same test (Figure 1a). However, the same plot of *U* is usually relatively symmetric (Figure 1b). These plots and plots for treated carriers usually lead to the conclusion that the log transformation improves symmetry (also see the *Response Variables* section).

The mean log density may vary somewhat among factor levels. It is useful to examine the scatter of log densities around each mean. A scatter that looks about the same for each factor level suggests that the variances (and SDs) are homogeneous.

Individual value plots provide visual guidance for choosing the appropriate way to scale the data for purposes of statistical analysis; in particular, they typically support the convention of performing statistical calculations on *LD* values. For alternative

test methods, individual value plots of the transformed measures provide important insight into the appropriate scale for statistical analysis. Other informative plots are described later (e.g., the *Assessing Repeatability, Reproducibility, Resemblance, and Responsiveness* section).

Data Quality

Although each laboratory may have its own established methods for describing the dilution series, plate counts, and carrier density outcomes, it is imperative that the laboratories use uniform terminology and conduct the calculations the same way. Dilution factors tend to be defined differently in different laboratories, and special attention should be given to dilution factors and scale-up formulas. It is advisable for the statistician to participate in the development of a uniform data entry sheet for use in the study and a spreadsheet for transmitting the data for each test to the Study Director. The participating laboratory specialists must understand the terminology on those sheets. The analyst should review the individual test day spreadsheets that were combined to form the master data set in addition to checking the master data set itself.

The results collected in the laboratory should not be rounded prior to submitting the data for statistical analysis. During the statistical analysis, extra significant digits should be retained for intermediate calculations. Rounding occurs only at the end of the analysis when the final collaborative study results are reported. The appropriate number of significant digits depends on the accuracy with which the original data were recorded and on the statistical uncertainty in the final results

Because the data are often partitioned, reorganized, and recoded during the statistical analysis, data manipulation errors can occur. After each rearrangement of the data set, it is advisable for the analyst to perform cross-checks and to create focused summaries, cross-tabulations, and plots for uncovering any discrepancies between the original data and the reorganized data.

Substitution Rules

When microbes are enumerated on a carrier, occasionally all plate counts for that carrier will be zero, even at the first plated dilution. Such zero counts occur for untreated carriers if the first few dilutions were not plated and, for a quantitative test, occur for treated carriers when the disinfectant is highly effective. A zero count cannot be log-transformed because log(0) is undefined. It would bias the results to drop from the analysis those carriers for which the density was so low that all counts were zero. At the other extreme, if the microbial density is higher than the range covered by the dilution series, the CFUs on some plates at the last dilution may be too numerous to count (TNTC). The density estimate would be underestimated if a carrier was deleted from the data set simply because some or all of the plates were TNTC at the last dilution, unless there were extenuating circumstances, such as a contaminated dilution series.

If these problematic outcomes occur often, the Study Director may decide that some tests should be repeated using a more sensitive CFU count procedure. For example, more dilutions could be plated for untreated carriers. However, if only a few carriers in the collaborative study produce all zeros at the first dilution or TNTC at the last dilution, then the usual approach is to substitute artificial numerical values for the problematic CFU counts, e.g. (16). A variety of substitution rules have been suggested, but there is no obvious best rule to use for collaborative study data. This is an area of ongoing statistical research, and optimum substitution methods may be available in the future.

In the meantime, one could use the following popular substitution rules. When the observed CFU counts are all zeros, find the plate in the data entry form where a count of 1 instead of 0 would produce the smallest possible scaled-up density (*D*). Typically, that critical plate is one plate at the first counted dilution. Replace the observed 0 with ½ and perform the usual scale-up to calculate *D*. Be advised that too many substitutions would lead to downward-biased SD estimates. However, substituting ½ for 0 in a small percentage of tests (e.g., less than 15%) has a negligible effect on the SD (37–39). Alternatives to this substitution rule have been suggested (37–41).

A popular TNTC substitution rule is to insert the highest valid count (e.g., 200 or 300 for the spread plate method, depending on the microbe) for each plate having TNTC at the last dilution, and perform the scale-up as usual. Alternative TNTC rules are available (42, 43).

It is advisable to choose the substitution rules before testing begins and to describe the rules in the collaborative study protocol. Before analyzing the data, the analyst should review those tests where substitutions were necessary and check that the substitution rules were applied correctly. Before proceeding with the statistical analysis, it is advisable to consider whether the DPPT outcomes required an excessive number of substitutions. The analyst could create a summary table showing how many carrier enumerations required a substitution and share the table with the Study Director. If substitutions affected a large percentage of the tests, the Study Director could choose from among alternative actions, such as (*1*) delaying statistical analysis until additional laboratory testing has been performed, (*2*) dropping the case (e.g., all tests of a specific disinfectant treatment) from the study prior to analysis, or (*3*) conducting the analysis with the available data and noting in the report which results are potentially affected by the substitutions.

Outliers

An outlier is an observation that appears to deviate markedly from other members of the data set. Initial examination and plots of the data (also see the *Plots* section above) may uncover outliers. Individual value plots of *LD* and of LR provide convenient, effective tools for discovering outlying carrier responses, tests, or laboratories. For example, a laboratory that produced a response pattern that is systematically quite different from the other laboratories may have submitted invalid data, and the Study Director should investigate.

In DPPT studies, a few unusual values may reflect inherent (microbiological or test system) variability. For example, the microbiologist's experience may show that occasionally a clump of microbes will be inoculated on a carrier's surface, resulting in an exceptionally large viable cell density for the carrier. Or, in a DPPT against biofilm bacteria, experience may indicate that there is an occasional sloughing event when the biofilm is being grown on a carrier's surface, in which case the viable cell density will be exceptionally small. In either

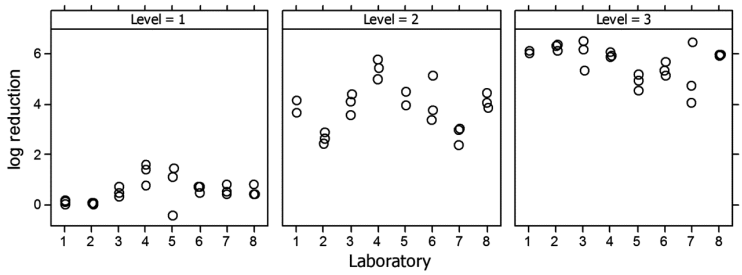

Figure 2. Each symbol is the LR for one test in a collaborative study of a quantitative dried surface test method; *M* = 3 and *L* = 8; NaOCl data in Appendix C of (Ref. 16). The LR values for the three replicate tests are aligned vertically, although the points were jittered horizontally to expose overlying points. Results for the three efficacy levels (1 = lower, 2 = intermediate, and 3 = higher) are displayed in separate panels.

example, the remarkable counts are due to inherent variation in the test method, and it would be a mistake to drop them from the data set.

It is prudent to investigate the documentation for each outlier to see if it is an error that can be corrected (e.g., a data entry error when entering data from the laboratory notebook into the computer spreadsheet). The statistical analyst's professional judgment is required to decide which outliers deserve investigation by the Study Director. The Study Director will determine whether outliers or other anomalies are valid, invalid but correctable, or invalid and should be removed from the data set.

As a general rule, all data except the invalid data should be used when calculating statistical results for a DPPT method. It can happen that the validity of a set of outliers is in doubt, but the Study Director finds no convincing evidence of invalidity. In that case, the analyst can determine the influence of those outliers by conducting each key statistical analysis twice, once with the outliers included in the data set and again with outliers removed. If the results (e.g., the repeatability and reproducibility SDs, defined below in the *Assessing Repeatability, Reproducibility, Resemblance, and Responsiveness* section) are only negligibly different, the outliers might as well be included even though potentially invalid. The Study Director, after considering all pertinent information, might decide there are compelling reasons to remove valid, influential outliers; that decision should be documented in the study report (4).

Assessing Repeatability, Reproducibility, Resemblance, and Responsiveness

The primary goals of a DPPT collaborative study are valid assessments of the repeatability and reproducibility of the disinfectant efficacy outcome (18). These goals are met by calculating the repeatability SD and the reproducibility SD of the LR, where LR is used here as the prototypical efficacy outcome.

Repeatability

A repeatable DPPT method will produce nearly the same LR in independent tests within a laboratory. To produce data for assessing repeatability, the collaborative study of a DPPT requires that the same disinfectant treatment is tested by *M* independent tests in the laboratory, where each test is performed on a different day using fresh microbial preparations, new dilutions of the concentrated disinfectant, etc. The set of factors that cause DPPT results to vary from test to test is different from the set of factors responsible for variability among results from a chemical assay or a microbial contamination of food assay. Therefore, the repeatability conditions (5) for DPPT methods are different from the repeatability conditions that have been officially specified for chemical assays or microbial contamination of food assays (2–4).

For a quantitative or semiquantitative DPPT method, an assessment of repeatability typically begins with an individual value plot of the LR values for each disinfectant treatment, arranged to display the variability among and within laboratories, each plotted point showing the LR for a single test (Figure 2). Initially, for each disinfectant treatment, a separate repeatability SD (denoted by $S_{r\ell}$, $\ell = 1, ..., L$) is calculated for each laboratory by calculating the SD of the *M* individual test LRs. If the $S_{r\ell}$ values for a disinfectant treatment are homogeneous across laboratories, it is appropriate to calculate a single S_r applicable to all laboratories. To estimate the common S_r of LR for a specified disinfectant treatment, use a one-factor, random effects analysis of variance (ANOVA) model to analyze the single test LR values observed across all laboratories; in this approach, laboratory is the random factor (Part A of the Appendix on the *J. AOAC Int.* website, http://aoac.publisher.ingentaconnect.com/content/aoac/jaoac). Good repeatability is indicated by a small S_r.

One might be concerned that $S_{r\ell}$ will vary extensively among laboratories, in which case the data do not conform to the homogeneous variance assumption that underlies the ANOVA

Table 3. Mean, variance components, the repeatability SD (S_r), and reproducibility SD (S_R) of LR for a quantitative dried surface test method, presented separately for each of three efficacy levels of a disinfectant; *M* = 3 and *L* = 8; these are the NaOCl results in (Ref. 16).

		Variance components			
Efficac level	Mean LR	Within labs (S_r^2)	Among labs (S^2_{lab})	S_r	S_R
Low	0.56	0.1641	0.0874	0.41	0.50
Medium	3.92	0.2008	0.7004	0.45	0.95
High	5.71	0.2645	0.1703	0.51	0.66

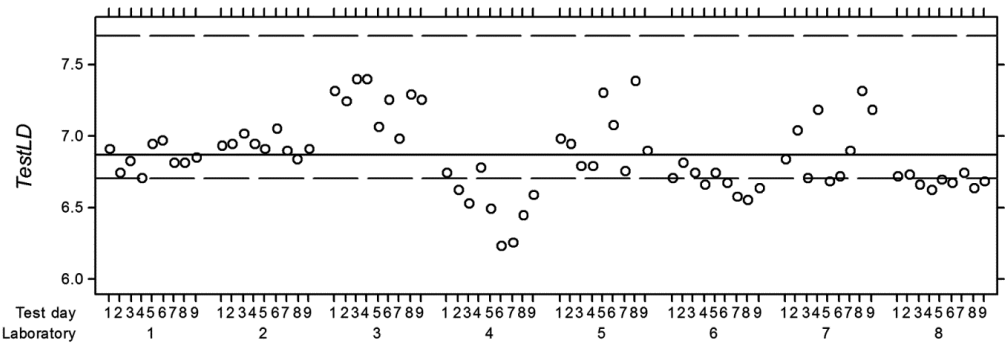

Figure 3. Each symbol in the individual value plot shows the *TestLD* for one test. The tests in each laboratory are numbered chronologically. The data are from the same study as displayed in Figure 1 (Ref. 16). The solid line is at 6.863, the overall mean of *TestLD* values. The dashed lines show the lower limit of $\log_{10}(5 \times 10^6)$ and the upper limit of $\log_{10}(5 \times 10^7)$ that bound the nominal acceptable range for *TestLD*, as suggested by the protocol. No *TestLD* value exceeded the upper limit, but 26% fell below the lower limit.

technique. This concern should have been carefully considered prior to the collaborative study. If there were plausible reasons why the DPPT outcome (e.g., LR) would have different variability in one laboratory than in another, the test protocol should have been revised to control the factors responsible for that difference. To check whether the collaborative study data display homogeneous $S_{r\ell}$ values, these guidelines describe plots, summary statistics, and significance tests. Levene's test can be helpful for detecting heterogeneous $S_{r\ell}$ values (44). However, there seldom are enough data in a collaborative study to show the statistical equivalence of within-laboratory SDs (13). In the event of severe heterogeneity, the study team might consult with the laboratory specialists, attempt to ascertain the cause of the heterogeneity, and, if possible, revise the DPPT protocol accordingly. The main achievement of the collaborative study would then end up being a revised and improved DPPT method, in which case a confirmatory collaborative study would be advisable.

Reproducibility

A reproducible DPPT method will produce nearly the same LR when the disinfectant treatment is tested by a different laboratory. In a collaborative study, the same disinfectant treatment is tested in different laboratories, and the reproducibility across laboratories is indicated by the reproducibility SD for the LR values, denoted by S_R. For a quantitative or semiquantitative test method, the reproducibility of LR is calculated from ANOVA variance component estimates (Appendix–Part A; Ref. 3) using the formula $S_R = \sqrt{S_r^2 + S_{lab}^2}$ (example in Table 3). Good reproducibility is indicated by a small S_R. The ANOVA *F*-test (of the null hypothesis that the true variance among laboratories is zero) is of little value for this application. The ANOVA techniques presented here for assessing repeatability and reproducibility are mathematically similar to the techniques recommended for chemical assay and microbial contamination of food assays (4). *See* Part C of the Appendix on the *J. AOAC Int.* website, (http://aoac.publisher.ingentaconnect.com/content/aoac/jaoac) for computer code that does the ANOVA calculations using the R statistical programming language, along with a numerical example).

The reproducibility variance S_R^2 is also called the "total variance" (20). It is informative to report the percentage of the total variance that is attributable to the variance among laboratories, i.e., $100 \times \frac{S_{lab}^2}{S_R^2}$ %. Although alternatives to the SD of LR have been proposed for semiquantitative tests (45, 46), the SDs of the alternative efficacy measures have neither easily conveyed practical interpretations nor comparative results in the literature.

One might wonder whether the laboratories are statistically representative of all laboratories that will be using the DPPT and whether the reproducibility observed for the collaborating laboratories is too optimistic. In DPPT collaborative studies, the laboratories seldom are a random sample from a relevant population; instead, they usually are self-selected by volunteering to participate. However, there is a practical interpretation of the results that avoids the issue, namely, the observed S_R is a conservatively large estimate of the true SD of LR for a future test of the disinfectant treatment in a single laboratory, if that laboratory is randomly chosen from among those that participated in the collaborative study (13). As is true of all published DPPT collaborative studies, extrapolation of S_R to a wider population of laboratories should be done with circumspection.

A collaborative study typically includes multiple disinfectant treatments of varying efficac . Consequently, the study results list a separate S_R value for each treatment. Historical data consistently show that disinfectant treatments having about the same efficacy also have about the same reproducibility (e.g., 16, 31). Therefore, one can anticipate that the S_R for a new treatment will be similar to the S_R observed for any treatment tested in a collaborative study that has the same efficacy as the new treatment.

Resemblance

An important goal of a DPPT collaborative study is assessing the resemblance of the microbial populations on the untreated carriers by a statistical analysis of the log-transformed viable microbe densities. Each DPPT method includes a protocol for preparing the initial population of microbes and placing them

Table 4. Panel (a) summarizes the ANOVA results for the untreated carrier *LD* values in a series of dried surface tests: $J = 3$, $M = 9$, and $L = 8$; Appendix 3 of (Ref. 16). Based on the ANOVA results in panel (a), panel (b) illustrates the US_r, US_R, and SEM calculations of Equations 6, 7, and 8.

(a)	Source	Variance	Symbol	DF
	Lab	0.04899	US^2_{lab}	7
	Test	0.01607	US^2_{test}	64
	Within-test	0.02097	US^2	144

(b) For the estimated variances in panel a, the repeatability and reproducibility SDs of *TestLD* for $J = 3$ untreated carriers/test are:

$$US_r = [US^2_{test} + (US^2/J)]^{½} = [0.01607 + (0.02097/3)]^{½} = 0.152$$

$$US_R = [US^2_{lab} + US^2_{test} + (US^2/J)]^{½} = [0.04899 + 0.01607 + (0.02097/3)]^{½} = 0.268$$

The overall mean log density (± SEM), averaged across all 216 untreated carriers, was 6.863 (± 0.080). The SEM was calculated as follows (Equation 9):

$$SEM = [(US^2_{lab}/L) + (US^2_{test}/(L \cdot M)) + (US^2/(L \cdot M \cdot J)]^{½} = [(0.04899/8) + (0.01607/72) + (0.02097/216)]^{½} = 0.080$$

on the carriers. This phase of the test method is sometimes called the inoculation or biofilm growth step. A good test protocol will create nearly the same desired microbial population/carrier across replicate tests, even if tests are done in different laboratories. At present, it is not practical to measure fundamental microbial characteristics, such as the genotype/phenotype distribution. Instead, the microbial population is measured by the density of viable microbes on a carrier after it has undergone DPPT manipulations.

Because resemblance pertains to the *LD* on untreated carriers (*U*) or to the mean of the *U* values for a test (*TestLD*), neither treated carrier data nor LR values are used for assessing resemblance. An important indication of good resemblance is little variability of the *TestLD* as measured by the repeatability SD and the reproducibility SD of *TestLD* values. A supplementary indication is little variability of the *U* values as measured by the within-test SD of *U*. In the following discussion, it will be clear from both the context and the notation whether the terms "repeatability" and "reproducibility" pertain to the resemblance analysis of *TestLD* or to the previously discussed repeatability and reproducibility analysis of LR.

To initiate the resemblance analysis, plot the *U* values against time, separately for each laboratory. Time ought to be quantified in a relevant way, e.g., the date of the test, the number of days from the commencement of the study, or an integer representing chronological order (*see* Figure 1b). This chronological plot also is useful for detecting outliers, visualizing the variability of the log densities, and uncovering systematic trends or cycles (cf., *Plots* and *Outliers* sections above).

A trend or cycle evident in the collaborative study would suggest that some underlying factor affects resemblance; discovery of that factor could lead to an improved test method. An analytical assessment of trend often is based on the slope of the least-squares regression line fit to the log densities over time. To estimate the common slope for trend across all laboratories, it is often appropriate to use a mixed effects analysis of covariance, e.g., Chapter 25 of Ref. 20, where the response is the log density *U*, the random factor is laboratory, and the covariate is time. Alternatively, separate trend lines could be calculated for each laboratory. To summarize the trend, it sometimes is appropriate to use a correlation coefficient relating *U* to time. Both the Pearson product-moment correlation coefficient and the Spearman rank correlation coefficie t, along with the associated *P*-value for testing the null hypothesis of no correlation, have been used in resemblance analyses, with the choice depending on the characteristics of the data at hand (20). The trend in time could also be quantified using autocorrelation (20) if there are a sufficient number of data points

It also is informative to construct an individual value plot of *TestLD* values arranged to display the variability among and within the laboratories. The display could show *TestLD* on the vertical axis, with the values arranged according to laboratory and within-laboratory tests that are listed on the horizontal axis (Figure 3).

DPPT protocols often specify an acceptable range (minimum and maximum acceptable values) for the *TestLD*. If the *TestLD* falls within the acceptable range, then the microbial density is considered to be representative of the field conditions where the disinfectant will be applied. The minimum acceptable *TestLD* value is set to guarantee that each test uses a large enough microbial challenge so that an LR value as large as the prevailing performance standard is a possbile outcome. The maximum acceptable *TestLD* may prevent densities so large that the mechanisms of action for the typical disinfectant treatment would be altered, thereby resulting in an irrelevant LR value. If there is an acceptable range, then it is informative to display the minimum and maximum acceptable values on the individual value plot (Figure 3).

Typically, a test having an unacceptable *TestLD* will be judged invalid, and the data for that test will not be included when calculating the collaborative study results. In some studies, however, the acceptable range is tentative, and tests having *TestLD* values outside the nominal range of acceptability are included in the collaborative study analysis. In the latter case, it is informative to report *TestLD* values below the lower acceptability limit and above the upper acceptability limit (e.g., Figure 3).

The prospect of imperfect resemblance is why most DPPT protocols require concurrent untreated carriers in each test. The LR calculation, which takes into account the concurrent *TestLD*, eliminates the potential effect of poor resemblance on the efficacy measure. In general, if the *TestLD* values are within the specified acceptable range, S_r and S_R of LR are unaffected by poor resemblance.

The repeatability (within-laboratory) and reproducibility (among-laboratories) of *TestLD* values are important

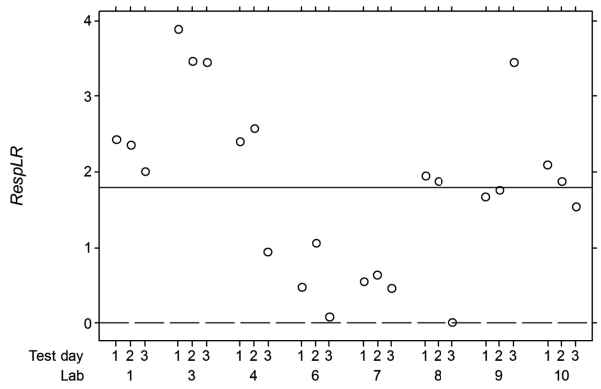

Figure 4. Each symbol is the responsiveness, $Resp_{LR}$, for a test day. The data are from a collaborative study of a quantitative test using spores of *Bacillus subtilis* on glass carriers; high versus medium efficacy levels of NaOCl in (Ref. 16). The dashed horizontal line is at $Resp_{LR} = 0$. The solid horizontal line is at 1.796, the overall mean of $Resp_{LR}$ values, which is statistically significantly greater than zero ($P < 0.001$).

resemblance characteristics. The laboratory ℓ resemblance repeatability SD, denoted by $US_{r\ell}$, is the SD of *TestLD* among all tests within laboratory ℓ. Usually, the $US_{r\ell}$ values are so similar that they can be pooled into a single US_r representative of all laboratories. Levene's test can be helpful in detecting heterogeneous $US_{r\ell}$ values (44).

Pooling is accomplished by running a two-factor, nested, random effects ANOVA of all untreated carrier *LD* values. Here, the random factor for experiment is nested within the random factor for laboratories (Appendix – Part B). See Part C of the Appendix for computer code that does the ANOVA calculations using the R statistical programming language, along with a numerical example.

The ANOVA provides estimates of three variances: within-test (US^2); among tests (US^2_{test}); and among laboratories (US^2_{lab}). The ANOVA pools the within-test variances (*US* values of Equation 2) across all tests in all laboratories. Based on the ANOVA output, the resemblance repeatability SD for the *TestLD* is calculated using Equation 6:

$$US_r = \sqrt{\frac{US^2}{J} + US^2_{test}} \tag{6}$$

Table 4 provides a calculation example. A small US_r indicates good resemblance repeatability for the *TestLD*.

The resemblance reproducibility SD, denoted by US_R, is the typical distance from the true overall mean of *TestLD* values for a *TestLD* in a randomly chosen test in a randomly chosen laboratory. For a test protocol that requires *J* untreated carriers, calculate US_R from the three ANOVA variance components using Equation 7:

$$US_R = \sqrt{\frac{US^2}{J} + US^2_{test} + US^2_{lab}} \tag{7}$$

Table 4a provides an example of the ANOVA results, and Table 4b shows the subsequent calculations. A small US_R indicates good resemblance reproducibility. (*Note*: An alternative way to calculate US_r and US_R is to run the one-factor, random effects model of Part A in the Appendix using *TestLD* as the response variable instead of LR. This one-factor ANOVA of *TestLD* values produces the same US_r and US_R, but does not break out the potentially useful variance components, US^2 and US^2_{test}, that make up US_r.)

The overall mean of *TestLD* values shows the typical *TestLD* for the DPPT method. The overall mean usually is accompanied by an indication of statistical uncertainty, such as the standard error of the mean (SEM) or a confidence interval for the true mean. For a balanced study in which *J* untreated carriers were used in each of *M* tests conducted in each of *L* laboratories, Equation 8 provides the formula for the SEM of the overall mean *TestLD*, averaging across all $L{\cdot}M{\cdot}J$ untreated carrier log densities (47). Table 4b provides an example. A confidence interval for the true overall mean *TestLD* can be calculated using this SEM and the *t*-distribution, with *L*-1 degrees of freedom (20):

$$SEM = \sqrt{\frac{US^2}{LMJ} + \frac{US^2_{test}}{LM} + \frac{US^2_{lab}}{L}} \tag{8}$$

If either US_r or US_R is unacceptably large or a significant number of tests produced *TestLD* values outside the acceptable range, the plots and summary statistics described in the guidelines may help uncover the reasons. Sometimes the resemblance for a DPPT can be improved by altering the protocol to require more untreated carriers/test. To calculate the extent to which averaging over more carriers can improve resemblance, the analyst can use the collaborative study data to find a tolerance interval for subsequent *TestLD* values if based on a larger number of carriers. The tolerance interval is the range expected to contain a specified percentage (usually 90, 95, or 99%) of future *TestLD* values, if one calculates the so-called

Table 5. Responsiveness analysis for high versus medium efficacy levels of a disinfectant tested side-by-side using a quantitative dried surface test; NaOCl data in Appendix C of (Ref. 16). There were $M = 3$ tests in each of $L = 8$ laboratories, each test producing the responsiveness measure $Resp_{LR}$ of Equation 9. Responsiveness was analyzed separately for each laboratory, followed by a responsiveness assessment across all laboratories. The DPPT method exhibited statistically significant responsiveness ($P < 0.05$ in six of eight laboratories; $P < 0.10$ in all eight laboratories). When combined over the eight laboratories using a one-factor, random effects ANOVA, the overall mean responsiveness was highly statistically significant (one-sided *t*-test $P < 0.001$). The *P*-value for each laboratory is for an upper one-sided *t*-test with DF = M-1 = 2. The "All labs" test was an upper one-sided *t*-test with DF = L-1 = 7.

Lab	Mean $Resp_{LR}$	*M*	*P*-value
1	2.268	3	0.002
2	3.605	3	0.001
3	1.974	3	0.031
4	0.541	3	0.098
5	0.555	3	0.005
6	1.282	3	0.091
7	2.300	3	0.029
8	1.841	3	0.004
All labs	1.796	24	<0.001

β-expectation tolerance interval (48). The tolerance interval may show that, by increasing the number of untreated carriers, the *TestLD* acceptable range will be achieved by most tests. Also the analyst can calculate the smaller US_r and US_R values that will occur with a larger number of untreated carriers/test (use Equations 6 and 7 with the observed variance components and the new larger value of *J*). Note that if the collaborative study produced excellent resemblance, the calculations just described can be performed to help decide whether it is possible to reduce the number of untreated carriers in each test while maintaining acceptable resemblance.

Responsiveness

A responsive DPPT method is capable of discriminating between high- and low-efficacy disinfectant treatments. Poor responsiveness means that the DPPT method is too variable or too insensitive for practical use. When evaluating responsiveness, the DPPT is applied at two or more efficacy levels of an established disinfectant. One can adjust the efficacy by altering the concentration of the active ingredient, by changing the contact time, or by altering influential variables such as temperature or pH. The observed efficacy of the presumed higher efficacy treatment should be discernibly greater than the observed efficac of the presumed lower efficacy treatment. A collaborative study can show the extent to which different laboratories observe the same increases in efficacy values when testing the same pair of presumed lower and higher efficacy levels (Figure 2).

For quantitative and semiquantitative test methods, the simplest responsiveness assessment is a comparison of the efficacy responses (usually LR values) observed for two presumed-different efficacy levels of a disinfectant. Generally, a study will have better power for detecting responsiveness if two disinfectant treatments of presumed-different efficacy levels are tested side-by-side (in parallel) on the same test day; however, it is possible to assess responsiveness if the two efficacy levels were tested separately. Estimating and testing responsiveness for each of these two scenarios are discussed next.

For side-by-side testing, the LR values are paired, and the responsiveness ($Resp_{LR}$) for each test day is measured by Equation 9. Good responsiveness is indicated by a large, positive value for $Resp_{LR}$, that is, a value that is both statistically and practically significant

$$Resp_{LR} = \mathrm{LR}_{higher} - \mathrm{LR}_{lower} \quad (9)$$

It is informative to construct an individual value plot of $Resp_{LR}$ values, arranged to display the magnitude and consistency of responsiveness among replicate test days and among laboratories (Figure 4).

To determine whether the mean of $Resp_{LR}$ values is significantly greater than zero, use an upper one-tailed *t*-test, where t = (mean $Resp_{LR}$)/SEM. For a balanced design in which each efficacy level of a disinfectant is evaluated by M independent side-by-side tests in each of L laboratories, the formula for SEM, the standard error of the mean $Resp_{LR}$ having L-1 degrees of freedom, is presented in Equation 10 where the estimated variance components for $Resp_{LR}$ are calculated by ANOVA for the model in Appendix A with $Resp_{LR}$ instead of LR as the response variable (Table 5):

$$SEM = \sqrt{\frac{\text{Repeatability variance of } Resp_{LR}}{L \cdot M} + \frac{\text{Among laboratories variance of } Resp_{LR}}{L}} \quad (10)$$

Now consider the case where the two efficacy levels of a disinfectant were tested in separate tests and the tests of both efficacy levels were replicated. Average the LR values for an efficacy level within a laboratory. The responsiveness for a laboratory ($Resp_{\mathrm{LR}}$) is the difference of those within laboratory means, *higher* minus *lower*. One can then perform an upper one-tailed, one-sample *t*-test based on the L laboratory $Resp_{\mathrm{LR}}$ values to determine whether the overall responsiveness is significantly positive

Discussion

Although a collaborative study of a DPPT method is quite expensive and represents a large cooperative effort, the study produces essential information that justifies the investment. The participating laboratories and stakeholders who will use the results deserve assurance that the study was designed properly, the data were analyzed correctly, and the results were presented clearly. For these reasons, it is highly recommended that the study team includes a professional statistician. It is advisable to use the statistician's expertise when creating a collaborative study design and protocol, as well for analyzing and presenting the study results. The practical and proven statistical techniques recommended in this paper will not be appropriate for every study; therefore, they should be applied only after careful consideration.

To every extent possible, the analyst should use peer-reviewed statistical techniques and avoid ad hoc ones. The analyst should provide evidence that statistical computations were performed correctly. For most collaborative studies, the analyst will utilize reputable statistical computer packages. For a complicated analysis, it is prudent to duplicate the computations using a second statistical software package to confirm the results. It is advisable for the analyst to maintain a running diary to record the statistical decisions that were important to the analysis.

The statistician should watch for outliers during the course of statistical analysis. Most analyses rely on statistical models, especially models that express the response variable as a function of the study factors or of the operational and environmental dynamics of the test method. For a model-based analysis, it is standard practice to consider model diagnostics, e.g., residual plots, normal probability plots, and other standard residual analyses. The diagnostics are effective tools for identifying observations that obviously depart from the model, e.g., Chapter 9 of (Ref. 20). Some statistical computer software packages provide the option of listing residuals that are statistically unusual. The pattern created by large residuals may indicate model misspecification, in which case alternative models should be considered. On the other hand, a large residual may indicate an invalid observation, and that possibility should be investigated.

The resemblance and responsiveness analyses are specific to collaborative studies of DPPTs, and do not appear in guidelines for chemical assays or microbial contamination of food assays. The important aspects of QA and the recommended statistical techniques in this article are determined mainly by the treated versus untreated microbe comparison that is a critical component of DPPT methods (10). Therefore, some parts of the guidance

presented here are not suitable for chemical assays or microbial contamination of food assays.

The threshold for acceptability for S_r or S_R could be established by a stakeholder (e.g., by a company for use in its internal product screening or by a regulatory agency for the acceptability of the method in regulatory decisions). The threshold could be based on historical data (12, 49) or on a specified tolerance (48, 50). Similar reasoning could be used by a stakeholder to choose acceptable thresholds for the resemblance repeatability SD (US_r) and the resemblance reproducibility SD (US_R).

Although this presentation focuses primarily on the LR measure of disinfectant efficac , the statistical techniques can be adapted for studying an alternative test measure. For alternative tests, the analyses described in this article provide reasonable guidance. To use the recommended statistical techniques, it is important that the efficacy measure and the measure of viable cells/carrier are properly transformed so that the probability distributions are symmetric, approximately normal, and have homogenous within-laboratory variances. This technical issue would be resolved during method development work prior to the collaborative study.

The guidance could also be applied to antimicrobial test methods that have different goals than killing microbes, for example, testing a treatment designed to remove microbes from surfaces or testing a treatment that prevents microbes from becoming associated with a surface. In any case, the response measure must be chosen so that the results have a practical interpretation and the statistical modeling assumptions are valid.

The statistical results from a collaborative study often are applied for purposes other than the basic goals discussed in these guidelines. For example, the results may suggest modifications of the DPPT method or may help regulatory authorities craft a performance standard for the DPPT. In some collaborative studies, two or more DPPT methods are used so that the methods can be compared. When a collaborative study is applied to such additional purposes, statistical techniques beyond those in this article may be required [e.g., equivalence testing (14) or pooling data from different treatments to estimate S_r and S_R]. Because it is not possible to anticipate all potential uses for the collaborative study data, publishing the main data matrixes is recommended so that retrospective calculations can be performed.

Important topics, such as guidelines for developing and optimizing a standardized DPPT method or guidelines for designing and managing a collaborative study, are beyond the scope of this paper. In addition to covering the practical aspects of DPPT studies, such guidelines would discuss statistical design principles, e.g., the value of randomization when subjective choices could bias the results and the advantages of blind testing. There is a need for such guidelines applicable specifically to DPP methods.

Acknowledgments

Preparation of this manuscript was supported by the Industrial Associates of the Center for Biofilm Engineering at Montana State University. The recommended statistical techniques evolved from work performed during the past 20 years under cooperative agreements and contracts (including the current Contract No. EP-W-08-014) between the U.S. Environmental Protection Agency and Montana State University.

References

(1) ISO (1994) *5725-1, Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 1: General Principles and Definitions*, International Organization for Standardization, Geneva, Switzerland (Cor. 1:1998)

(2) ISO (1994) *5725-2, Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 2: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method*, International Organization for Standardization, Geneva, Switzerland (Cor. 1:2002)

(3) ISO (1994) *5725-3*, *Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 3: Intermediate Measures of the Precision of a Standard Measurement Method*, International Organization for Standardization, Geneva, Switzerland (Cor. 1:2001)

(4) *Official Methods of Analysis* (2005) 18th Ed., AOAC INTERNATIONAL, Gaithersburg, MD, Appendix D

(5) ASTM International (2009) *Standard E691–09 Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method*, ASTM International, West Conshohocken, PA

(6) ISO (2003) *16140, Microbiology of Food and Animal Feeding Stuffs - Protocol for the Validation of Alternative Methods,* International Organization for Standardization, Geneva, Switzerland

(7) ISO (2006) *19036, Microbiology of Food and Animal Feeding Stuffs -Guidelines for the Estimation of Measurement Uncertainty for Quantitative Determinations,* International Organization for Standardization, Geneva, Switzerland

(8) Niemi, R.M., & Niemela, S.I. (2001) *Accred. Qual. Assur.* **6**, 372–275. http://dx.doi.org/10.1007/s007690100360

(9) Tillett, H.E., & Sartory, D. (2004) *Accred. Qual. Assur.* **9**, 629–632. http://dx.doi.org/10.1007/s00769-004-0798-x

(10) Hamilton, M. (2010) *How the Differences Between Disinfectant Tests and Chemical Assays Affect Method Evaluation Criteria, KSA-SM-05, 2010-08-19*, Center for Biofilm Engineering a Montana State University, Bozeman, MT, http://www.biofilm montana.edu/resources/knowledge_sharing_articles (accessed on July 3, 2013)

(11) Hamilton, M. (2011) *Testing Surface Disinfectants: Quantitative, Semiquantitative, Quantal, and Alternative Methods, KSA-SM-02, 2011-10-25*, Center for Biofil Engineering at Montana State University, Bozeman, MT, http:// www.biofilm.montana.edu/resources/knowledge_sharing articles (accessed on July 3, 2013)

(12) Hamilton, M. (2010) *Testing Surface Disinfectants: Desirable Attributes of a Standardized Method, KSA-SM-03, 2010-06-10*, Center for Biofilm Engineering at Montana State Universit , Bozeman, MT, http://www.biofilm.montana.edu/resources knowledge_sharing_articles (accessed on July 3, 2013)

(13) Youden, W.J., & Steiner, E.H. (1975) *Statistical Manual of the AOAC–Statistical Techniques for Collaborative Tests,* AOAC INTERNATIONAL, Gaithersburg, MD

(14) Tomasino, S.F., & Hamilton, M.A. (2006) *J. AOAC Int.* **89**, 1373–1397

(15) Tomasino, S.F., & Hamilton, M.A. (2007) *J. AOAC Int.* **90**, 456–464

(16) Tomasino, S.F., Pines, R.M., Cottrill, M.P., & Hamilton, M.A. (2008) *J. AOAC Int*. **91**, 833–852

(17) *AOAC Official Method **965.13**, Disinfectants (Water) for Swimming Pools* (1970) AOAC INTERNATIONAL, Gaithersburg, MD

(18) Bloomfield, S. ., & Looney, E. (1992) *J. Appl. Bacteriol.* **73**, 87–93. http://dx.doi.org/10.1111/j.1365-2672.1992.tb04975.x
(19) Hamilton, M., Buckingham-Meyer, K., & Goeres, D. (2009) *J. AOAC Int.* **92**, 1755–1762
(20) Neter, J., Kutner, M.H., Wasserman, W., & Nachtsheim, C.J. (1996) *Applied Linear Statistical Models,* 4th Ed., McGraw-Hill, Boston, MA
(21) Forster, L.I. (2003) *J. AOAC Int.* **86**, 1089–1094
(22) Jarvis, B. (2008) *Statistical Aspects of the Microbiological Examination of Foods,* 2nd Ed., Elsevier, San Diego, CA
(23) Eaton, A.D., Clesceri, L.S., Greenberg, A.E., & Franson, M.A.H.(1995) *Standard Methods for the Examination of Water and Wastewater–Section 9020,* American Public Health Association, Washington, DC
(24) DeVries, T.A., & Hamilton, M.A. (1999) *Quant. Microbiol.* **1**, 29–45. http://dx.doi.org/10.1023/A:1010072226737
(25) Zelver, N., Hamilton, M., Goeres, D., & Heersink, J. (2001) in *Methods Enzymol. Biofilms II*, Vol. 337, R.J. Doyle (Ed.), Academic Press, New York, NY, pp 363–376
(26) Hamilton, M. (2011) *The Log Reduction (LR) Measure of Disinfectant Efficacy, KSA-SM-07, 2011-11-07*, Center for Biofilm Engineering at Montana State Universit , Bozeman, MT, http://www.biofilm.montana.edu/resources/knowledge sharing_articles (accessed on July 13, 2013)
(27) Cochran, W.G. (1950) *Biometrics* **6**, 105–116. http://dx.doi.org/10.2307/3001491
(28) Blodgett, R.J. (2006) in *Bacteriological Analytical Manual,* Appendix 2, FDA. http://www.fda.gov/Food/FoodScienceResearch/LaboratoryMethods/default.htm (accessed on July 3, 2013)
(29) Garthright, W.E. (1993) *Biom. J.* **35**, 299–314. http://dx.doi.org/10.1002/bimj.4710350306
(30) Hamilton, M. (2011) *The P/N Formula for the Log Reduction When Using a Semiquantitative Disinfectant Test of Type SQ1, KSA-SM-08, 2011-02-01*, Center for Biofilm Engineering a Montana State University, Bozeman, MT. http://www.biofilm montana.edu/resources/knowledge_sharing_articles (accessed on July 3, 2013)
(31) ASTM International (2102) *Standard E2799–12 Standard Test Method for Testing Disinfectant Efficacy Against Pseudomonas aeruginosa Biofilm Using the MBEC Assay,* ASTM International, West Conshohocken, PA
(32) ASTM International (2007) *Standard E2562-07 Standard Test Method for Quantification of Pseudomonas aeruginosa Biofilm Grown with High Shear and Continuous Flow Using the CDC Biofilm Reactor*, ASTM International, West Conshohocken, PA
(33) AOAC *Official Method* **2008.05***, Efficacy of Liquid Sporicides Against Spores of Bacillus subtilis on a Hard Nonporous Surface, Quantitative Three-Step Method* (2008) AOAC INTERNATIONAL, Gaithersburg, MD
(34) Garthright, W.E., & Blodgett, R.J. (2003) *Food Microbiol.* **20**, 439–445. http://dx.doi.org/10.1016/S0740-0020(02)00144-2
(35) AOAC INTERNATIONAL (2004) *Official Method* **966.04,** *Sporicidal Activity of Disinfectants*. AOAC INTERNATIONAL, Gaithersburg, MD
(36) EPA (2008) *Protocol for Testing the Efficacy of Disinfectants Used to Inactivate Duck Hepatitis B Virus and to Support Corresponding Label Claims, ver. August 2002*, Developed by MicroBioTest, Inc. and submitted to the U.S. Environmental Protection Agency, http://www.epa.gov/oppad001/pdf_files hbvprotocol.pdf (accessed on July 3, 2013)
(37) Haas, C.N., & Scheff, P.A. (1990) *Environ. Sci. Technol.* **24**, 912–919, http://dx.doi.org/10.1021/es00076a021
(38) EPA (2006) *Data Quality Assessment: Statistical Methods for Practitioners EPA QA/G-9S*, *EPA/240/B-06/003, February 2006*, U.S. Environmental Protection Agency, Office o Environmental Information, Washington, DC. http://www.epa.gov/quality/qs-docs/g9s-final.pd (accessed on July 3, 2013)
(39) Antweiler, R.C., & Taylor, H.E. (2008) *Environ. Sci. Technol.* **42**, 3732–3738. http://dx.doi.org/10.1021/es071301c
(40) Lorimer, M.F., & Kiermeier, A. (2007) *Int. J. Food Microbiol.* **116**, 313–318. http://dx.doi.org/10.1016/j.ijfoodmicro.2007.02.001
(41) Shorten, P.R., Pleasants, A.B., & Soboleva, T.K. (2006) *Int. J. Food Microbiol.* **108**, 369–375
(42) Haas, C.N., & Heller, B. (1988) *Appl. Environ. Microbiol.* **54**, 2069–2072
(43) Blodgett, R.J. (2008) *Food Microbiol.* **25**, 92–98. http://dx.doi.org/10.1016/j.fm.2007.07.006
(44) Gastwirth, J.L., Gel, Y.R., & Miao, W. (2009) *Stat. Sci.* **24**, 343–360. http://dx.doi.org/10.1214/09-STS301
(45) Rubino, J.R., Bauer, J.M., Clarke, P.H., Woodward, B.B., Porter, F.C., & Hilton, H.G. (1992) *J. AOAC Int.* **75**, 635–645
(46) Arlea, C., King, S., Bennie, B., Kemp, K., Mertz, E., & Staub, R. (2008) *J. AOAC Int.* **91**, 152–158
(47) Marcuse, S. (1949) *Biometrics* **5**, 189–206. http://dx.doi.org/10.2307/3001935
(48) Hamilton, M. (2011) *Testing Surface Disinfectants: How to Decide Whether the Reproducibility Standard Deviation is Small Enough, KSA-SM-11, 2011-09-11*, Center for Biofil Engineering at Montana State University, Bozeman, MT, http://www.biofilm.montana.edu/resources/knowledge_sharing articles (accessed on July 3, 2013)
(49) Tilt, N., & Hamilton, M. (1999) *J. AOAC Int.* **82**, 384–389
(50) Hubert, Ph., Nguyen-Huu, J.-J., Boulanger, B., Chapuzet, E., Chiap, P., Cohen, N., Compagnon, P.-A., Dewé, W., Feinberg, M., Lallier, M., Laurentie, M., Mercier, N., Muzard, G., Nivet, C., & Valat, L. (2004) *J. Pharm. Biomed. Anal.* **36**, 579–586