Analyzing the power and error of *Listeria monocytogenes* growth challenge studies

Mark R. Powell *

United States Department of Agriculture, Office of Risk Assessment and Cost Benefit Analysis, 1400 Independence Ave., SW, Rm. 4032 SAC, Washington, DC 20250, USA

**Abstract**

Domestic and international food safety policy developments have spurred interest in the design and interpretation of experimental growth challenge studies to determine whether ready-to-eat (RTE) foods are able to support growth of *Listeria monocytogenes*. Existing challenge study protocols and those under development differ markedly in terms of experimental design and the acceptance criteria under which a RTE food is determined not to support *L. monocytogenes* growth. Consequently, the protocols differ substantially with respect to the probability of incorrectly determining that growth occurs and the statistical power to detect growth if it does occur. Applying a fixed acceptance criteria exceedance value (e.g., less than a 0.5 log<sub>10</sub> or 1 log<sub>10</sub> increase) to distinguish real growth from qualitative measurement uncertainty over different experimental designs and/or measurement uncertainty values implies highly inconsistent type I error (α) probabilities. None of the *L. monocytogenes* growth challenge study designs currently being considered are likely to provide an F-test with α = 0.05 and power ≥ 0.8 to detect a 1 log<sub>10</sub> increase in mean concentration over the entire range of measurement uncertainty values for enumeration of *L. monocytogenes* reported in food samples in a validation study of ISO Method 11290-2.

1. Introduction

Domestic and international food safety policy developments have spurred interest in the design and interpretation of experimental growth challenge studies to determine whether ready-to-eat (RTE) foods are able to support growth of *Listeria monocytogenes*. In 2005, the European Commission (EC) defined a food safety criteria limit of 100 colony forming units (CFU)/g for RTE foods “unable to support the growth of *L. monocytogenes*” (European Commission, 2005). The EC regulation also states that as necessary, food business operators shall conduct studies to evaluate the growth of *L. monocytogenes* that may be present in the product during the shelf-life under reasonably foreseeable conditions of storage, distribution, and use. In January 2008, the EU Community Reference Laboratory for *L. monocytogenes* issued a draft guidance document to operationally define acceptance criteria under which a RTE food is determined unable to support *L. monocytogenes* growth on the basis of shelf-life study results and to describe procedures for conducting shelf-life studies to determine compliance with the EC regulatory criteria (EUCRL, 2008a). In November 2008, the laboratory issued a revised working document providing technical guidance on shelf-life studies for *L. monocytogenes* in RTE foods. Under EUCRL (2008b), a RTE product is determined unable to support growth of *L. monocytogenes* if the difference between the initial and final sample median concentrations is less than 0.5 log<sub>10</sub> CFU/g for all batches tested.

In February 2008, the U.S. Food and Drug Administration (FDA) issued a draft compliance policy guide stating that “FDA may regard a RTE food that does not support the growth of *L. monocytogenes* to be adulterated … when *L. monocytogenes* is present at or above 100 [CFU/g] of food” (Food and Drug Administration, 2008a). Food and Drug Administration (2008a) states that a “listeristatic control measure is generally considered to be effective if growth studies show less than a one log increase in the number of *L. monocytogenes* during replicate trials with the food of interest.” Food and Drug Administration (2008b) cites Scott et al. (2005) as an example of guidance for conducting *L. monocytogenes* growth challenge studies. In 2007, the National Advisory Committee on Microbiological Criteria for Foods (NACMCF) was charged with providing guidance to U.S. food safety agencies on inoculated pack and challenge study protocols (NACMCF, 2008). The scope of this charge includes, but is not limited to, *L. monocytogenes* growth challenge studies.

The Codex Committee on Food Hygiene (CCFH) proposed draft microbiological criteria that would establish a limit of 100 CFU/g for RTE foods “in which growth of *L. monocytogenes* will not occur” (CCFH, 2009). The proposed draft states that “a food in which growth of *L. monocytogenes* will not occur will not have an observable increase in *L. monocytogenes* levels greater than (on average) 0.5 log<sub>10</sub> CFU/g for at least the expected shelf life.” In July 2009, the Codex Alimentarius Commission adopted the proposed draft microbiological criteria for *L. monocytogenes* in RTE foods (CAC, 2009).
This paper considers several inter-related issues regarding the design of *L. monocytogenes* growth challenge studies: the acceptance criteria for distinguishing real growth from quantitative measurement uncertainty and the false positive error probability and the statistical power of a study in the context of the acceptance criteria and uncertainty.

2. Materials and methods

2.1. Definitions and assumptions

A type I (false positive) error is the rejection of a true null hypothesis (H0). The probability of a type I error (α) is often called the test level of significance. A type II (false negative) error occurs if a false H0 is not rejected. The probability of a type II error is denoted p(type II error) = β. The power of a hypothesis test (1 − β) is often called the producer’s risk, and β is often called the consumer’s risk (Montgomery, 2005). By convention, α is set at 0.05, and statistical power is considered adequate if (1 − β) ≥ 0.8, although this customary experimental design practice does not consider the severity of type I and type II errors in the context of specific decisions (Di Stefano, 2003).

Let \( y(t) = CFU/g \) at time \( t \) and \( x = \log_{10}(y) \). Assume that \( x \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \), denoted \( x \sim Normal(\mu, \sigma^2) \). Unless stated otherwise, the analysis is simplified by ignoring variability in growth response and assuming that \( x \) is subject only to quantitative measurement uncertainty, which includes measurement error as well as the inherent variability (e.g., in subsamples or dilutions) of a strictly independent contrast (C) that has a total number of treatments \( T \) with coefficients that sum to zero (e.g., \( C_1 = -T_1 + T_2 \)). Multiple contrasts are orthogonal if the products of corresponding coefficients sum to zero (Hicks, 1982). (For example, if \( C_2 = -T_3 + T_4 \) and \( C_2 = -T_5 + T_6 \), then \( C_1C_2 \) and \( C_2C_3 \) are orthogonal.)

2.2. Type I error probability and power for a fixed exceedance value

Under current protocols for *L. monocytogenes* growth challenge studies, the objective of distinguishing real growth from quantitative measurement uncertainty is satisfied by setting a fixed exceedance value, or upper limit, that is intended to account for measurement uncertainty. In risk analysis, a probability of exceedance value (M) is defined for a random variable (y) by: \( p(y \geq M) = \alpha \) (National Research Council, 2000). In contrast, a fixed exceedance value is defined without specifying an allowable \( \alpha \). Based on current *L. monocytogenes* protocols, two forms of a fixed exceedance value are considered. An exceedance value for a difference in two sample medians is denoted by \( (m_1 - m_2) < M_m \), where \( M_m \) is the 50th percentile of the \( x_q \) sample values. An exceedance value for a difference in two sample means is denoted by \( (x_1 - x_2) < M_x \), where \( M_x = \sum x_q / n \).

The type I error probability for a fixed exceedance limit depends not only on its statistical form (median or mean) and value (0.5 \( \log_{10} \) or 1.0 \( \log_{10} \) but also on the challenge study design specifications (e.g., the enumeration method and sampling plan) and the acceptance criteria for determining whether a RTE product may support growth of *L. monocytogenes* (e.g., number of sample values allowed to exceed the specified upper limit). Under EUCL (2008b), a RTE product is determined unable to support growth of *L. monocytogenes* if the difference between the initial and final sample median concentrations \( (m_i(t_i) - m_i(t_0)) < 0.5 \log_{10} \) CFU/g for all batches tested. That is, \( M_m = 0.5 \log_{10} \) with zero allowable exceedances of \( M_m \). The protocol calls for testing \( b \geq 3 \) different batches to account for variability of the RTE food product. (Note that batches may represent a random effect under the protocol.) The protocol results in a test with \( c \geq 3 \) independent, pair-wise comparisons: \( m_i(t_i) - m_i(t_0) \) for \( i = 1, \ldots, b \geq 3 \). (Note that the comparisons are orthogonal.) The median concentration in the \( k \)-th batch at each of \( k = 2 \) sampling times is based on a sample of \( n_{(k)} = 3 \). The median sample is insensitive to outliers and can be calculated if one of the three results is below the limit of enumeration; however, it is a less efficient estimator of \( \mu \) than the sample mean. (See discussion regarding Eq. (7) below.)

As noted in Sec. 1, the acceptance criteria under Food and Drug Administration (2008a) include “a less than one \( \log_{10} \) increase in the number of *L. monocytogenes* during replicate trials.” Similarly, Scott et al. (2005) concludes that “\( a < 1 \log_{10} \) increase above the initial inoculum level throughout the shelf-life of the product and across replicate trials would be an appropriate acceptance criterion” due to the “inherent variation that exists with enumeration of microorganisms.” Food and Drug Administration (2008a) and Scott et al. (2005) do not specify whether the nominal exceedance value of \( 1 \log_{10} \) refers to a difference in means of the log-transformed enumeration data, as recommended by NSF International (2000). For the purposes of this analysis, assume the criteria are operationally defined as a difference in means, with \( (x_1(t_i) - x_1(t_0)) < M_x = 1 \log_{10} \) CFU/g for \( i > 1 \) over all \( i \). (Note that \( t = b \) for \( i = 1 \).)

Scott et al. (2005) recommends a minimum of \( k = 5-7 \) sampling times and \( n = 2-3 \) samples per sampling time. Food and Drug Administration (2008a) and Scott et al. (2005) do not set a minimum number of replicate trials to account for variability of the RTE food product. (For example, conditions for *L. monocytogenes* growth may be more favorable in a single, specifically formulated batch than in a small random sample of batches.) Because any one \( \log_{10} \) increase above the initial level throughout the study violates the acceptance criteria, analyzing one growth trial involves a test with \( c = k - 1 \) dependent, pair-wise comparisons and zero allowable exceedances. (Note that the comparisons represent many-to-one, non-orthogonal contrasts. For example, Dunnett’s \( t \)-test is used for multiple comparisons of treatments with a control rather than the standard \( t \)-test to account for the dependency among comparisons (Dunnett, 1964).)

As discussed below, this dependency complicates evaluation of \( \alpha \) for a fixed exceedance value.

As noted in Sec. 1, CCFH (2009) proposed that “a food in which growth of *L. monocytogenes* will not occur will not have an observable increase in *L. monocytogenes* levels greater than (on average) 0.5 \( \log_{10} \) CFU/g for at least the expected shelf life.” That is, \( M_m = 0.5 \log_{10} \) with zero allowable exceedances of \( M_m \). CCFH (2009) specifies no experimental design parameters for *L. monocytogenes* growth challenge studies.

The stated basis for the proposed exceedance value is that “0.5 log is two times of the estimated standard deviation (i.e., 0.25 log) associated with the experimental enumeration of viable counting/plate counts” (CCFH, 2009). However, this calculation refers to an approximate upper limit for a two-tailed, 95% confidence interval for the random variable \( x \) (\( \log_{10} CFU/g \) subject to known measurement uncertainty (\( \sigma_x \)) (Montgomery, 2005):

\[
\bar{x} \pm z_{(1-\alpha/2)} \sigma_x
\]

where \( \alpha = 0.05, z_{(0.975)} = 1.96, \) and \( \sigma_x = 0.25 \log_{10} \) CFU/g. This interval is equivalent to the “expanded uncertainty” about a measurement
result \((x \pm u, \text{ where } U=2\alpha_x)\), within which the true value of the measurand can be asserted to lie with approximately 95% confidence (Corry et al., 2007; Lombard, 2006).

In general, however, the variance of the difference between two independent random variables \((A \text{ and } B)\) is \((\text{Rice, 1988})\):

\[
\sigma_{A-B}^2 = \sigma_A^2 + \sigma_B^2
\]  

(2)

Consequently, because both the initial and subsequent pathogen concentrations are subject to measurement uncertainty, the variance of the difference of the means (assuming independence) is calculated as:

\[
\sigma_{x(t_i)-x(t_0)}^2 = \sigma^2_{x(t_i)} + \sigma^2_{x(t_0)}
\]  

(3)

where \(\sigma^2 = \sigma^2/n\) (Rice, 1988).

Assuming equal sample sizes and variance:

\[
\sigma_{x(t_i)-x(t_0)}^2 = \frac{2\sigma^2}{n}
\]  

(4)

Therefore, \(\sigma_{x(t_i)-x(t_0)} = \sigma_x\) only if \(n = 2\).

Consequently, \(\alpha\) for a test based on a fixed \(M_x\) is calculated as follows. Assuming equal variances and sample sizes, if \(H_0: \mu(t) - \mu(t_0) \leq 0\) is true, then for a single pair-wise comparison:

\[
p(x(t) - x(t_0)) > M_x \leq 1 - \Phi\left(\frac{\mu_x + \sigma}{\sqrt{\frac{2\sigma^2}{n}}}\right) = \alpha'
\]  

(5)

where \(\Phi = \text{normal cumulative distribution function. (Note that because } \mu \leq 0 \text{ also satisfies } H_0, \text{ this calculation provides an upper bound on } \alpha.\)\) For a test with \(c\) independent comparisons, the experiment-wise type I error probability \((\alpha)\) is \((\text{Sidak, 1967})\):

\[
\alpha = 1 - \left(1 - \alpha'/c\right)^c
\]  

(6)

(No\te that for \(c = 1, \alpha = \alpha'\).)

Assuming \(x \sim \text{Normal}(\mu, \sigma^2)\), both \(\bar{x}\) and \(n\) are consistent estimators of \(\mu\), although the sampling distributions of the mean and median differ. Asymptotically, the mean is more efficient by a factor of \((\text{Greene, 1997})\):

\[
r = \frac{\sigma_m}{\sigma_y} \sqrt{\frac{n}{2}} = 1.25
\]  

(7)

However, the value of \(r\) depends on \(n\). For \(n = 2\) and \(3, r = 1.00\) and \(1.16\), respectively. (The ratios were calculated using the \text{SAS}\text{® STMED}(n) function.)

Therefore, for a single pair-wise comparison based on a fixed \(M_m\), \(\alpha\) is calculated as above (Eqs. (5) and (6)), substituting:

\[
p(m_x(t) - m_x(t_0)) > M_m \leq 1 - \Phi\left(\frac{\mu_m + \sigma}{\sqrt{\frac{2\sigma^2}{n}}}\right) = \alpha'
\]  

(8)

The straightforward \((\text{Sidak, 1967})\) adjustment provides control of experiment-wise \(\alpha\) only when multiple comparisons are independent. For multiple dependent comparisons \(x(t) - x(t_0), \alpha\) is estimated by Monte Carlo simulation. Assuming \(x \sim \text{Normal}(\mu, \sigma^2_x)\), the true mean concentration is held constant over time \((\mu(t) = \mu) \log_{10} \text{CFU/g}\), and simulated sample means \(\bar{x}(t)\) vary randomly only due to measurement uncertainty \((\sigma_x)\) in the observations \(x(t)\). The proportion of simulated growth study replications resulting in an exceedance of \(M_m\) (and therefore a decision to falsely reject \(H_0: \mu(t) - \mu(t_0) = 0\)) provides an upper bound estimate of \(\alpha\) (for \(H_0: \mu(t) - \mu(t_0) \leq 0\)).

Monte Carlo simulations were performed with Latin hypercube sampling using Palisades® @Risk™, Ver. 4.5.2, an add-on to Microsoft® Excel™ 2000 (Palisade Corporation, Ithaca, NY, USA). For each value of \(\sigma_x\), a simulation with 10,000 iterations was performed.

European Commission (2005, Food and Drug Administration (2008a), and CCFH (2009) identify ISO Method 11290-2 (ISO, 2004) as the reference method to enumerate \(L_{\text{monocytogenes}}\) in RTE foods. In a multi-laboratory validation study of ISO Method 11290-2, Scotter et al. (2001) reported standard deviation of reproducibility \((s_k, \text{ an intra-laboratory measure of quantitative measurement uncertainty (Lombard, 2006)})\) values in food samples with a range from 0.17 log_{10} CFU/g to 0.45 log_{10} CFU/g. Therefore, consider \(\sigma_x\) values in the range from 0.15 log_{10} CFU/g to 0.50 log_{10} CFU/g by increments of 0.05. (Note that under EUCLRL (2008b), if at \(t_0\) the sample standard deviation within a batch \(\geq 0.3 \log_{10} \text{CFU/g},\) then the trial for the batch is unacceptable and must be repeated to achieve acceptable contamination homogeneity and enumeration precision.) In calculating \(\alpha\) for a fixed exceedance value, consider challenge studies with \(c = 1\)–6 comparisons and \(n = 2\)–3 samples per sampling time. Consistent with EUCLRL (2008b), consider only independent comparisons for the case of \(M_m = 0.5 \log_{10} \text{CFU/g}\).

In general, the power for a test depends on the magnitude and pattern of growth assumed under \(H_0\). To illustrate the power of a test based on a fixed exceedance value, consider the simplest case involving a single comparison. For \(c = 1,\) the power of a test with a fixed \(M_m\) is calculated as follows. Assuming equal variances and sample sizes, if \(H_0: \mu(t) - \mu(t_0) = 0\), then:

\[
p(x(t) - x(t_0)) > M_m \leq 1 - \Phi\left(\frac{\mu_m + \sigma}{\sqrt{\frac{2\sigma^2}{n}}}\right) = (1-\beta).
\]  

(9)

Similarly, the power of a single comparison test based on a fixed exceedance value \(M_m\) for a difference in medians is calculated as:

\[
p(m_x(t) - m_x(t_0)) > M_m \leq 1 - \Phi\left(\frac{\mu_m + \sigma}{\sqrt{\frac{2\sigma^2}{n}}}\right) = (1-\beta).
\]  

(10)

For the power of a test based on fixed exceedance values, consider:

- \(M_m = 0.5 \log_{10} \text{CFU/g}\) and \(1.0 \log_{10} \text{CFU/g}\) and \(M_m = 0.5 \log_{10} \text{CFU/g}\); \(\alpha_m = 0.15 \log_{10} \text{CFU/g}\) to 0.50 \log_{10} \text{CFU/g} (by increments of 0.05); \(n = 2\)–3; and \(\delta = 0.5 \log_{10} \text{CFU/g}\) and 1.0 \log_{10} \text{CFU/g}.

2.3. Power of \(F\)-test in one-way ANOVA

Analysis of variance (ANOVA) is commonly used in the peer-reviewed literature for statistical analysis of \(L_{\text{monocytogenes}}\) growth challenge studies (Pal et al., 2008; Thompson et al., 2008). One-way ANOVA applies to an experiment with a single treatment factor and a continuous response variable. The model for a simple growth trial is:

\[
x_{ij} = \mu + \tau_i + \epsilon_{ij}
\]  

(11)

where \(\tau_i\) is a fixed treatment effect for time with \(k\) discrete levels and \(\epsilon \sim \text{Normal}(0, \sigma^2_{\epsilon})\). In contrast to regression methods, ANOVA requires no assumptions about growth model form and may be applied to data that exhibit a lag phase or no lag phase, linear or non-linear growth, and monotonic or non-monotonic growth patterns (e.g., adapted cells may resume growth after an initial, partial die-off).

The \(F\)-test in one-way ANOVA assumes that the observations are samples from \(k\) normally distributed populations with the same variance \((\sigma^2_{\epsilon})\) but possibly different means and tests \(H_0: \mu_i = \mu\) for all \(i\).

The alternative hypothesis is that at least one \(\mu_i\) is different, i.e. \(H_1: \) at
least one $\tau_i \neq 0$. (Note that the F-test is inherently two-sided.) For a balanced one-way ANOVA, the F-test statistic is (Hicks, 1982):

$$F = \frac{MS_k}{MS_e}$$

(12)

where $MS_r$ = mean square for sampling times (between groups), $MS_e$ = mean square residual (within groups). The F-test is a generalization of the two-sample F-test for simultaneously comparing more than two sample means. In contrast to the z-test where the variance is considered known (and assumed here to be limited to measurement uncertainty ($\sigma_x^2$)), the F-test applies where the unknown variance ($\sigma_x^2$) is estimated from the sample data.

The power of the F-test is given by (Taylor and Muller, 1995):

$$1 - \beta = 1 - F_{\alpha} \left[F_{\text{crit}}(1 - \alpha)\right] \left| df_1, df_2, \alpha \right|$$

(13)

where: $F_{\alpha} = \text{non-central F distribution, } F_{\text{crit}}(1 - \alpha) = \text{critical F value for rejecting } H_0 = (1 - \alpha)$ percentile of the F distribution with $df_1$ (numerator degrees of freedom) = $k - 1$ and $df_2$ (denominator degrees of freedom) = $k(n - 1)$ (assuming equal sample size), and $\alpha = \text{non-centrality parameter}$. For balanced one-way ANOVA:

$$\omega = \frac{n}{\sum_{i=1}^{k} (\mu_i - \mu)^2} \sigma^2$$

(14)

where $\mu_i = i$th group mean and $\mu = \text{overall mean specified under } H_0$, and $\sigma^2 = \text{within-group variance}$. The power of the F-test increases with $\omega$. For any scenario in which two elements on the right hand side of Eq. (14) are fixed, we can solve for the third to obtain the desired power. Note that for a given level of measurement uncertainty, assuming that the variance within groups reflects no heterogeneity in growth response at time $t$ provides an upper bound on the power of the F-test.

In analyzing the power of the overall F-test for growth challenge studies, denote the ordered group means by $\mu(1) \leq \ldots \leq \mu(k)$, and let $\delta = (\mu_k - \mu_1)$. For $k > 2$ sampling times, the power of the F-test depends on the specific configuration of the $k$ group means (i.e., the pattern of growth) postulated under $H_2$ (George, 2001). For a given $\delta$, power is minimized (maximized) by minimizing (maximizing) the sum of squares between groups (the second term in the numerator of Eq. (14)). The configuration that minimizes power is $\mu(i) = (\mu(1) + \mu(k))/2$ for $i \neq 1, k$ (e.g., growth under $H_2$ has a plateau at $\mu$ between $t_0$ and $t_2$). The configuration that maximizes power is two polarized clusters of maximally equal size with one cluster mean = $\mu(1)$, and the second cluster mean = $\mu(k)$ (e.g., growth under $H_2$ is a step function with lag time $= t_0/2$). These two cases represent atypical growth patterns but provide upper and lower bounds on the power of the F-test for a given $\delta$.

A series of power curves is constructed to plot the probability of rejecting $H_0$: $\mu = \mu$ at $\alpha = 0.05$ for $\delta = 1 \log_{10}$ CFU/g against $\sigma_x$ values by combinations of $k$ and $n$. Consider $\sigma_x = 0.15 \log_{10}$ CFU/g to 0.50 $\log_{10}$ CFU/g; $k = 2$–7 sampling times; and $n = 2$–3 samples per sampling time. Calculations were performed with SAS® PROC POWER.

### 2.4. Power of a planned comparison

If the F-test in ANOVA rejects the null hypothesis, the result does not indicate which group means are significantly different or the direction of the difference (increase or decrease). A variety of post hoc tests are available for conducting multiple comparisons while maintaining the experiment-wise $\alpha$ level. For example, Dunnett’s test evaluates pair-wise comparisons of $k - 1$ treatments with a control, Tukey’s test evaluates all $k(k - 1)/2$ pair-wise comparisons, and Scheffé’s test allows comparison of all possible (not only pair-wise) combinations of $k$ treatments (Shavelson, 1981). After accounting for a large number of comparisons to maintain the experiment-wise $\alpha$ level, however, the power of such tests may be insufficient.

Alternatively, a planned (a priori) comparison may provide more power than the overall F-test and can be applied irrespective of the omnibus test result without impacting $\alpha$ because the hypotheses tested are not equivalent (Rutherford, 2001). A logical planned comparison for growth challenge studies is to test whether the final concentration is significantly greater than the initial concentration. For comparing two sample means, the ANOVA F-test is equivalent to the two-sided t-test, with $F(1 - \alpha) = F(1 - \alpha/2)$ (Evans et al., 1993). For foodborne pathogens, however, the concern is with processes leading to unacceptable high values, and this implies a one-sided limit (AOAC International, 2006). An advantage of a one-sided limit or hypothesis test is that for a given sample size, it has more power than its two-sided counterpart (Rosner, 2005).

Therefore, consider a one-sided t-test of $H_0$: $\mu(t) - \mu(0) \leq 0$. The power of the test is calculated for $\alpha = 0.05$; $\bar{\delta} = 0.5 \log_{10}$ CFU/g and $1 \log_{10}$ CFU/g; $\sigma_x = 0.15 \log_{10}$ CFU/g to 0.50 $\log_{10}$ CFU/g (by increments of 0.05); and $n = 2$–3 samples per sampling time. In addition, the required sample size ($n^*$) for $(1 - \beta) \geq 0.8$ is obtained. (Note that $n \geq 2$ is required to calculate the sample mean.) Calculations were performed with SAS® PROC POWER.

### 2.5. Power of F-test in two-way mixed model

Growth challenge studies may be designed to include factors such as batches or pathogen strains to account for potential differences in L. monocytogenes growth response across such factors. The ANOVA F-test and its power depend on whether such factors are considered to have fixed or random effects. The duration of a growth challenge study is not randomly determined but related to the intended shelf-life of the RTE product. Time represents a fixed effect. If the levels of other factors (e.g., batches) are selected at random and inferences from the challenge study are to be generalized to the underlying population of a RTE food product, then these factors are random effects. Models that contain fixed and random effects are mixed models (Hicks, 1982).

Consider the following mixed model:

$$x_{ij} = \mu + \tau_i + B_j + \epsilon_{ij}$$

(15)

where $\tau_i$ is a fixed effect for time with $k$ levels, $B_j$ is a random effect for batch with $b$ levels, and $j = 1, \ldots, n_{ij}$. Based on the expected mean square (EMS), the F-test statistic for time (the fixed effect) in the mixed model is (Hicks, 1982):

$$F_{\bar{df}, df} = \frac{MS_r}{MS_{\tau B}}$$

(16)

where $MS_r$ = mean square between sampling times with $df = k - 1$ and $MS_{\tau B}$ = mean square of the $\tau B$ interaction with $df = (k - 1)(b - 1)$. By comparison, the denominator for a fixed two-way ANOVA model would be $MS_r$ (mean square residual) with $df = (k - 1)b$ (Hicks, 1982). As the interaction mean square is generally larger than $MS_r$ and has fewer degrees of freedom, the F-test for the mixed model is generally less powerful than the F-test for the fixed model. This reduction in power illustrates the inherently greater challenge of expanding inferences to the entire population rather than just the specific treatments contained in a given study (Chapman and Seidel, 2008). (Batches may be considered to have fixed effects if the challenge study is specifically designed to address key formulation factors known to affect L. monocytogenes growth, as recommended by Scott et al. (2005).)

SAS® PROC POWER does not contain a feature for power analysis of mixed models; however, SAS® PROC MIXED performs ANOVA for mixed models. Therefore the power of the F-test in the two-way
mixed model was estimated by simulating data from multiple replications of the experiment with known inputs and calculating the proportion of F-tests rejected at α using SAS© PROC MIXED. The simulations were performed by parametric bootstrapping (Vose, 2000) using a SAS© DATA step. Based on the EMS for the fixed effect F-test, the denominator degrees of freedom were calculated using the containment method in SAS© PROC MIXED.

Consistent with EUCRL (2008b), consider a two-way mixed model with k = 2 sampling times (t0 and t1), b = 3 batches, and n(t0) = 3. Bootstrap samples were generated assuming for fi: x(t0) ~ Normal(μ, σ2) and for fi: x(t1) ~ Normal(μ + δi, σ2). Under H0, the difference in means varies among batches, with δi ~ Normal(μδ, σδ2). In all simulations, the average difference is fixed at $μδ = 1.0 \log_{10} \text{CFU/g}$. To assess the sensitivity of power to the degree of variability in growth among batches, the standard deviation of δi (δi) was increased across simulations from 0.1 log10 CFU/g to 0.50 log10 CFU/g by increments of 0.1. (For example, specifying $δ_1 = 0.5 \log_{10} \text{CFU/g}$, $δ_2 = 1.0 \log_{10} \text{CFU/g}$, and $δ_3 = 1.5 \log_{10} \text{CFU/g}$ obtains $μδ = 1.0 \log_{10} \text{CFU/g}$ and $α_δ = 0.50 \log_{10} \text{CFU/g}$. Regarding the lower bound for αδ, the difference in power between $αδ = 0.01$ (results not shown) and $αδ = 0.1$ is negligible. Thus, the coefficient of variation for batch differences (CV = $αδ/μδ$) in the sensitivity analysis ranged from 10% to 50% across simulations. (Regarding the upper bound $αδ = 0.5$ in the sensitivity analysis, by comparison, Cornu et al. (2006) reported L. monocytogenes growth rates among batches of cold-smoked salmon with cv $\approx 60%$.) Consider $αδ = 0.50 \log_{10} \text{CFU/g}$ to 0.50 log10 CFU/g (by increments of 0.05). For each of the 40 combinations of αδ and $αδ$, a simulation with 10,000 bootstrap iterations was performed. For each simulation, the power of the fixed effect F-test at $α = 0.05$ under the alternative growth hypothesis was estimated as the proportion of the 10,000 iterations with $p(\text{greater } F) \leq 0.05$.

### 3. Results

#### 3.1. Type I error probability and power for a fixed exceedance value

Tables 1–3 present the $α$ levels for fixed exceedance values of $M_0 = 0.5 \log_{10} \text{CFU/g}$, $M_0 = 0.5 \log_{10} \text{CFU/g}$, and $M_0 = 1.0 \log_{10} \text{CFU/g}$, respectively. The results demonstrate that applying any fixed exceedance value to distinguish real growth from quantitative measurement uncertainty over different experimental designs and/or measurement uncertainty values implies highly inconsistent allowable type I error probabilities. In general, α increases with measurement uncertainty ($σε$, std. dev.) and number of comparisons (c) and decreases with increasing sample size (n) and exceedance value (M). The $M_0$ level is lower for dependent relative to independent comparisons and for $M_0$ relative to $M_0$.

Table 4 presents the power of a single comparison test for fixed exceedance values of $M_0 = 0.5 \log_{10} \text{CFU/g}$, $M_0 = 0.5 \log_{10} \text{CFU/g}$, and $M_0 = 1.0 \log_{10} \text{CFU/g}$.

### Table 2

<table>
<thead>
<tr>
<th>Type I error probability for difference in means fixed exceedance value ($M_0 = 0.5 \log_{10} \text{CFU/g}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Std. Dev. (log10 CFU/g)</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$p(\text{type I error}) \leq α$</td>
</tr>
<tr>
<td>0.15</td>
</tr>
<tr>
<td>0.20</td>
</tr>
<tr>
<td>0.25</td>
</tr>
<tr>
<td>0.30</td>
</tr>
<tr>
<td>0.35</td>
</tr>
<tr>
<td>0.40</td>
</tr>
<tr>
<td>0.45</td>
</tr>
<tr>
<td>0.50</td>
</tr>
</tbody>
</table>

**$p < 0.01$.**
where growth is log-linear with no lag time falls closer to the minimum power curve. Depending on the specific alternative, the power curve for the case where there is a substantial lag before log-linear growth can fall closer to the maximum power curve.

None of the challenge study designs considered provide an F-test with \( \alpha = 0.05 \) and minimum power \( \geq 0.8 \) to detect a 1 log difference in the mean over the entire range of measurement uncertainty values reported in food samples by Scotter et al. (2001). The challenge study designs considered, with \( n = 3 \), do not have a sample size sufficient to provide power \( \geq 0.8 \) over the reported range of measurement uncertainty.

### Table 4: Power of a single comparison test for fixed exceedance values.

<table>
<thead>
<tr>
<th>Std. Dev.</th>
<th>Sample size ((n) = 2)</th>
<th>Sample size ((n) = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.50 &gt; 0.99</td>
<td>0.50 &gt; 0.99</td>
</tr>
<tr>
<td>0.20</td>
<td>0.50 &gt; 0.99</td>
<td>0.50 &gt; 0.99</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50 &gt; 0.99</td>
<td>0.50 &gt; 0.99</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50 &gt; 0.95</td>
<td>0.50 &gt; 0.95</td>
</tr>
<tr>
<td>0.35</td>
<td>0.50 &gt; 0.92</td>
<td>0.50 &gt; 0.92</td>
</tr>
<tr>
<td>0.40</td>
<td>0.50 &gt; 0.89</td>
<td>0.50 &gt; 0.91</td>
</tr>
<tr>
<td>0.45</td>
<td>0.50 &gt; 0.87</td>
<td>0.50 &gt; 0.88</td>
</tr>
<tr>
<td>0.50</td>
<td>0.50 &gt; 0.84</td>
<td>0.50 &gt; 0.85</td>
</tr>
</tbody>
</table>

### Table 5: Power analysis of a one-sided t-test \((\alpha = 0.05, \delta = 10^{log_{10} CFU/g})\) with sample size \(n = 3\) and sampling times \(k = 2-7\).

<table>
<thead>
<tr>
<th>Std. Dev. ((log_{10} CFU/g))</th>
<th>True increase ((\delta) = 0.5)</th>
<th>True increase ((\delta) = 1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size ((n))</td>
<td>Sample size ((n*))</td>
<td>Sample size ((n*))</td>
</tr>
<tr>
<td>0.15</td>
<td>0.69</td>
<td>0.95</td>
</tr>
<tr>
<td>0.20</td>
<td>0.50</td>
<td>0.81</td>
</tr>
<tr>
<td>0.25</td>
<td>0.38</td>
<td>0.65</td>
</tr>
<tr>
<td>0.30</td>
<td>0.31</td>
<td>0.52</td>
</tr>
<tr>
<td>0.35</td>
<td>0.26</td>
<td>0.42</td>
</tr>
<tr>
<td>0.40</td>
<td>0.22</td>
<td>0.36</td>
</tr>
<tr>
<td>0.45</td>
<td>0.19</td>
<td>0.31</td>
</tr>
<tr>
<td>0.50</td>
<td>0.17</td>
<td>0.27</td>
</tr>
</tbody>
</table>

\(n*\) = Required sample size for power \(\geq 0.8\).
Table 6
Power analysis of F-test in two-way mixed model (α = 0.005).

<table>
<thead>
<tr>
<th>Std. Dev. (Root mean square error) (log_{10} CFU/g)</th>
<th>True mean increase (μ) = 1.0 log_{10} CFU/g</th>
<th>Std. Dev. growth among batches (σ_u) (log_{10} CFU/g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>&gt;0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>0.20</td>
<td>&gt;0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>0.25</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>0.30</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>0.35</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>0.40</td>
<td>0.73</td>
<td>0.66</td>
</tr>
<tr>
<td>0.45</td>
<td>0.58</td>
<td>0.53</td>
</tr>
<tr>
<td>0.50</td>
<td>0.46</td>
<td>0.42</td>
</tr>
</tbody>
</table>

4. Discussion

To the extent that *L. monocytogenes* growth response is heterogeneous among samples, the observed variance in data from challenge studies performed using ISO Method 11290-2 may exceed the upper end of the range of measurement uncertainty values reported in food samples by Scotter et al. (2001). Holding other factors constant, as variance increases, the statistical power of a hypothesis test decreases. Recognizing this limitation of the analysis reinforces the central conclusion that none of the *L. monocytogenes* growth challenge study designs currently being considered are likely to support statistical tests that provide α = 0.05 and power ≥0.8 to detect a 1 log_{10} increase in mean pathogen concentration over the range of measurement uncertainty associated with the enumeration method referenced by European Commission (2005), Food and Drug Administration (2008a), and CCFH (2009). Satisfying conventional experimental design criteria would require larger sample sizes, lower measurement uncertainty, or both.

With regard to growth challenge studies designed to account for random differences in *L. monocytogenes* growth response among batches of RTE food, an investigator may select multiple samples per batch due to an interest in estimating the magnitude of the batch effect. However, for a fixed total sample size, power is maximized by selecting a single sample from multiple batches rather than drawing multiple samples from a small number of batches. The stronger the batch effect, the greater the gain in power by sampling more batches (Holson et al., 2008; National Toxicology Program, 2001).

Alternatively, power may be gained by not seeking to generalize from the challenge study to the underlying population of a RTE food product. Preliminary sampling of RTE food batches can serve to characterize the variability in pH, water activity, and other physicochemical properties known to effect *L. monocytogenes* growth. Batches specifically formulated to provide favorable conditions for *L. monocytogenes* growth (as discussed by Scotter et al. (2005)) may be considered to have fixed treatment effects, resulting in more power than a mixed model. This presumes, however, that the conditions favoring *L. monocytogenes* growth, including interactions among multiple physicochemical parameters, are well understood in the RTE product under study and that the variability of the key physicochemical factors in the RTE product is well characterized. When confronted with uncertainty, it may be more practical and expedient in some circumstances to simply conduct a larger growth challenge study using randomly selected batches.

Finally, sampling plans and methods commonly developed as part of microbiological criteria are intended for lot-by-lot acceptance decisions (ICMSF, 2002). In contrast, a determination of whether a RTE food product supports growth of *L. monocytogenes* is a non-recurring, one-off decision that may establish the regulatory standard applied to all lots produced henceforth. Practices associated with routine acceptance sampling where microbiological testing is ongoing are not necessarily suitable for generating data for standard-setting purposes. Therefore, protocols for conducting and interpreting *L. monocytogenes* growth challenge studies to establish regulatory criteria should strive for consistency in application and give careful consideration to experimental design criteria, including the importance and magnitude of producer’s and consumer’s risks.

References


