



Address 100 Barr Harbor Drive
PO Box C700
W. Conshohocken, PA
19428-2959 | USA

Phone 610.832.9500
Fax 610.832.9666
Web www.astm.org

COMMITTEE D02 ON PETROLEUM PRODUCTS, LIQUID FUELS, AND LUBRICANTS

CHAIRMAN: RANDY F JENNINGS, TENNESSEE DEPT OF AGRIC, P O BOX 40627, NASHVILLE, TN 37204, UNITED STATES (615) 837-5327, FAX: (615) 837-5335, E-MAIL: RANDY.JENNINGS@TN.GOV
FIRST VICE CHAIRMAN: JAMES J SIMNICK, BP AMERICA, 150 W WARRENVILLE RD, NAPERVILLE, IL 60563, UNITED STATES (630) 420-5936, FAX: (630) 420-4831, E-MAIL: SIMNICJJ@BP.COM
SECOND VICE CHAIRMAN: MICHAEL A COLLIER, PETROLEUM ANALYZER CO LP, 21114 HWY 113, CUSTER PARK, IL 60481, UNITED STATES (815) 458-0216, FAX: (815) 458-0217, E-MAIL: MICHAEL.COLLIER@PACLP.COM
SECOND SECRETARY: HIND M ABI-AKAR, CATERPILLAR INC, BLDG H3000, OLD GALENA ROAD, MOSSVILLE, IL 61552, UNITED STATES (309) 578-9553, E-MAIL: ABI-AKAR_HIND@CAT.COM
SECRETARY: SCOTT FENWICK, NATIONAL BIODIESEL BOARD, PO BOX 104848, JEFFERSON CITY, MO 65110-4898, UNITED STATES (800) 841-5849, FAX: (537) 635-7913, E-MAIL: SFENWICK@BIODIESEL.ORG
STAFF MANAGER: ALYSON FICK, (610) 832-9681, FAX: (610) 832-9668, E-MAIL: AFICK@ASTM.ORG

Issued: 04.19.2019
Reply to: Dan Worcester
Southwest Research Institute
6220 Culebra Rd.
San Antonio, TX 78238
Phone: 210.522.2405
Email: dworcester@swri.org

These are the unapproved minutes of the 04.17.2019 Sequence VI Conference Call.

This document is not an ASTM standard; it is under consideration within an ASTM technical committee but has not received all approvals required to become an ASTM standard. It shall not be reproduced or circulated or quoted, in whole or in part, outside of ASTM committee activities except with the approval of the chairman of the committee having jurisdiction and the president of the society. Copyright ASTM, 100 Barr Harbor Drive, West Conshohocken, PA 19428-2959.

The meeting was called to order at 1:04 PM Central Time by Chair Andrew Stevens.

Agenda

- 1.0 Roll Call: The Attendance list is **Attachment 1**.
- 1.1 There were no membership changes.
- 1.2 The Agenda is **Attachment 2**.
- 1.3 The minutes for the 03.22.2019 call were not reviewed.

2.0 Old Business

- 2.1 There was an issue with wording on Break-In hours. The wording was slightly different in the VIE and VIF procedures.
- 2.2 This issue was resolved with the release of Information Letter 19.2.

3.0 New Business

- 3.1 Discussion about standards for introducing a new fuel supplier.
 - 3.1.1 Prasad presented some options to consider. His presentations are included as **Attachments 3 and 4**. Doyle brought up that there might be a response difference for FEI1 and FEI2. Tracey noted there currently is no additional testing as new batches are produced. There was a lot of discussion on the level of testing needed.

- Bob: The candidates should be run the same as the reference if a particular fuel is used for that reference test.
- Rich: Would both fuels be available? There are procedures on how to move to different fuels. For the VH, a new fuel can be added to 10% of the previous batch.
- Bob: The alternate supplier should provide some data for comparison. He also asked if there was a cost benefit for an alternate supplier.
- Jim C: 2-4 tests would begin the review process.
- Mike: Noted the T13 bearings were a different supplier and material and approved for use with 2 tests.

3.1.2 What is the process to approve another fuel supplier?

- Adrian: Any testing should compare back to the Precision Matrix.
- Mike: Asked whether a matrix was required?
- Jeff: Suggested an ABBA matrix on one stand as a minimum.
- Bob: The Surveillance Panel should provide the number of tests needed.
- Pat: He noted this option may set a pattern moving forward for alternate suppliers. This is a recommendation for the TGC. The Stats Group would be involved.
- Travis: There should be different matrices for different options.
- Andrew: A Fuel Task Force will be created. He will send recommendations and start calls. There would be representatives for Stats and Test Engineers.

ACTION: Create a Task Force. The goals would be to define single or multiple suppliers, how to determine equivalence, and whether tests would be on first runs only, or engines could be calibrate on the second run, and the first run count as part of the matrix data.

The meeting adjourned at 2:20 PM Central Time

ASTM SEQUENCE VI

Name	Email/Phone	Company	Attend
Adrian Alfonso Voting Member	Phone: (210) 838-0431 Adrian.Alfonso@intertek.com	Intertek	ATTEND
Jason Bowden Voting Member	Phone: (440) 354-7007 jhbowden@ohtech.com	OHT	ATTEND
Kevin Brodwater Voting Member	Phone: (510) 242-2291 KBrodwater@chevron.com	Chevron	
Aleise Gauer Voting Member	Phone: (248) 296-0761 aleise.gauer@gm.com	GM	ATTEND
Rich Grundza Voting Member	Phone: (412) 365-1034 reg@astmtmc.cmu.edu	TMC	ATTEND
Jeff Hsu Voting Member	Phone: (832) 419-3482 j.hsu@shell.com	Shell	ATTEND
Teri Kowalski Voting Member	Phone: (734) 995-4032 Teri.Kowalski@tema.toyota.com	Toyota	
Dan Lanctot Voting Member	Phone: (210) 690-1958 dlanctot@tei-net.com	TEI	ATTEND
Ben Maddock Voting Member	Phone: (804) 788-5743 Ben.Maddock@AftonChemical.com	Afton	ATTEND
Brienne Pentz Voting Member	Phone: (973) 317-6364 Brienne.Pentz@bp.com	BP	
Andy Ritchie Voting Member	Phone: (908) 474-2097 Andrew.Ritchie@infineum.com	Infineum	
Ron Romano Voting Member	Phone: (313) 845-4068 rromano@ford.com	Ford	
Clifford Salvesen Voting Member	Phone: (856) 224-2954 Clifford.R.Salvesen@exxonmobil.com	ExxonMobil	
Amol Savant Voting Member	Phone: (606) 585-8982 acsavant@valvoline.com	Valvoline	
Andrew Stevens Voting Member	Phone: (440) 347-4020 andrew.stevens@Lubrizol.com	Lubrizol	
Haiying Tang Voting Member	Phone: (248) 512-0593 HT146@Chrysler.com	Chrysler	
Prasad Tumati Voting Member	Phone: (313) 300-8300 ptumati@jhaltermann.com	Haltermann	ATTEND
Dan Worcester Voting Member	Phone: (210) 522-2405 Dan.Worcester@swri.org	SwRI	ATTEND

ASTM SEQUENCE VI

Name	Email/Phone	Company	Attend
Ed Altman	Ed.Altman@aftonchemical.com	Afton	
Bill Anderson	Bill.anderson@aftonchemical.com	Afton	
Bob Campbell	Bob.Campbell@aftonchemical.com	Afton	
Lisa Dingwell	Lisa.Dingwell@AftonChemical.com	Afton	
Todd Dvorak	Todd.Dvorak@aftonchemical.com	Afton	
Greg Guinther	Greg.Guinther@aftonchemical.com	Afton	
Terry Hoffman	Terry.Hoffman@aftonchemical.com	Afton	
Christian Porter	Christian.Porter@aftonchemical.com	Afton	
Jeremy Styer	Jeremy.Styer@aftonchemical.com	Afton	
Tisha Joy	Tisha.Joy@bp.com	BP	
Michael Blumenfeld	Michael.I.Blumenfeld@exxonmobil.com Phone: (856) 224.2865	EM	
Jim Carter	jcarter@gageproducts.com	Gage	ATTEND
Andy Buczynsky	Andrew.Buczynsky@gm.com	GM	
Meryn Hopp	Meryn.Hopp@GM.com	GM	
Michael Raney	Michael.p.Raney@gm.com	GM	
Doyle Boese	Doyle.Boese@infineum.com Phone: (908) 474-3176	Infineum	ATTEND
Gordon Farnsworth	Gordon.Farnsworth@infineum.com	Infineum	
Charlie Leverett	Charlie.Leverett@yahoo.com Phone: (210) 414-5448	Infineum	
Mike McMillan	mmcmillan123@comcast.net	Infineum	
Jordan Pastor	Jordan.Pastor@Infineum.com Phone: (313) 348-3120	Infineum	
William Buscher	William.Buscher@intertek.com	Intertek	
Martin Chadwick	Martin.Chadwick@intertek.com	Intertek	
Al Lopez	Al.Lopez@intertek.com	Intertek	
Mike Noriega	Mike.Noriega@intertek.com	Intertek	
Addison Schweitzer	Addison.Schweitzer@intertek.com	Intertek	
Scott Rajala	srajala@ILAcorp.com	Idemitsu	
Dave Passmore	dpassmore@imtsind.com	IMTS	
Jerry Brys	Jerome.Brys@lubrizol.com Phone: (440) 347.2631	Lubrizol	
Tony Jang	Tony.Jang@Lubrizol.com	Lubrizol	
Joe Gleason	Jog1@lubrizol.com	Lubrizol	
James Matasik	James.Matasik@lubrizol.com	Lubrizol	
Greg Miranda	Greg.Miranda@Lubrizol.com	Lubrizol	
Kevin O'Malley	Kevin.OMalley@lubrizol.com	Lubrizol	
Chris Castanien	Chris.Castanien@neste.com Phone: (440) 290-9766	Neste	

ASTM SEQUENCE VI

Name	Email/Phone	Company	Attend
Dwight Bowden	dhbowden@ohtech.com	OHT	
Matt Bowden	mjbowden@ohtech.com	OHT	
Ricardo Affinito	affinito@chevron.com Phone: (510) 242-4625	Oronite	
Ian Elliot	IanElliott@chevron.com	Oronite	
Jo Martinez	jogm@chevron.com	Oronite	ATTEND
Robert Stockwell	rsto@chevron.com	Oronite	ATTEND
Don Smolenski	Donald.Smolenski@gmail.com	Strategic	
Dan Engstrom	daniel.engstrom@swRI.org Phone: (210) 522-3472	SwRI	ATTEND
Travis Kostan	Travis.Kostan@swRI.org Phone: (210) 522-2407	SwRI	ATTEND
Patrick Lang	Patrick.Lang@swRI.org Phone: (210) 522-2820	SwRI	ATTEND
Michael Lochte	mlochte@swri.org Phone: (210) 522-5430	SwRI	ATTEND
Khalid Rais	Khalid.Rais@swri.org	SwRI	
Karen Haumann	Karen.Haumann@shell.com	Shell	
Scott Stap	Scott.Stap@tgdirect.com	TG Direct	
Jeff Clark	jac@astmtmc.cmu.edu	TMC	
Hirano Satoshi	Satoshi_Hirano_aa@mail.toyota.co.jp	Toyota	
Jim Linden	lindenjim@jlindenconsulting.com Phone: (248) 321-5343	Consultant	
Mark Adams	mark@tribologytesting.com	Tribology Testing	
Timothy Caudill	Tlcaudill@valvoline.com	Valvoline	
Thom Smith	trsmith@valvoline.com	Valvoline	
Hap Thompson	Hapjthom@aol.com	Vix Facilitator	
Chris Taylor	Chris.Taylor@vpracingfuels.com	VP Racing Fuels	
Tracey King			ATTEND
Nick Jansen			ATTEND
	1		

ASTM SEQUENCE VI

Name	Email/Phone	Company	Attend
MOTION:			
Adrian Alfonso Voting Member			
Jason Bowden Voting Member			
Kevin Brodwater Voting Member			
Aleise Gauer Voting Member			
Rich Grundza Voting Member			
Jeff Hsu Voting Member			
Teri Kowalski Voting Member			
Dan Lanctot Voting Member			
Ben Maddock Voting Member			
Brienne Pentz Voting Member			
Andy Ritchie Voting Member			
Ron Romano Voting Member			
Clifford Salvesen Voting Member			
Amol Savant Voting Member			
Andrew Stevens Voting Member			
Haiying Tang Voting Member			
Prasad Tumati Voting Member			
Dan Worcester Voting Member			

ASTM SEQUENCE VI

Name	Email/Phone	Company	Attend
MOTION:			
Adrian Alfonso Voting Member			
Jason Bowden Voting Member			
Kevin Brodwater Voting Member			
Aleise Gauer Voting Member			
Rich Grundza Voting Member			
Jeff Hsu Voting Member			
Teri Kowalski Voting Member			
Dan Lanctot Voting Member			
Ben Maddock Voting Member			
Brienne Pentz Voting Member			
Mike Raney Voting Member			
Andy Ritchie Voting Member			
Ron Romano Voting Member			
Clifford Salvesen Voting Member			
Amol Savant Voting Member			
Andrew Stevens Voting Member			
Haiying Tang Voting Member			
Prasad Tumati Voting Member			
Dan Worcester Voting Member			

ASTM SEQUENCE VI

Name	Email/Phone	Company	Attend
VOTES			
MOTION:			
Adrian Alfonso Voting Member			
Jason Bowden Voting Member			
Kevin Brodwater Voting Member			
Aleise Gauer Voting Member			
Rich Grundza Voting Member			
Jeff Hsu Voting Member			
Teri Kowalski Voting Member			
Dan Lanctot Voting Member			
Ben Maddock Voting Member			
Brienne Pentz Voting Member			
Andy Ritchie Voting Member			
Ron Romano Voting Member			
Clifford Salvesen Voting Member			
Amol Savant Voting Member			
Andrew Stevens Voting Member			
Haiying Tang Voting Member			
Prasad Tumati Voting Member			
Dan Worcester Voting Member			

Sequence VI Surveillance Panel Call Meeting Agenda

April 17, 2019 @ 2:00-3:30 EST

Webex Meeting Details Below Agenda

1. Roll Call (start 2:05 EST)

1.1. SP Membership changes and additions

2. Old Business

2.1	Update on Breakin hours wording: VIE vs VIF - Wording is not consistent and potentially confers different meaning. References via a note from Rich Grundza in Appendix A1	Rich Grundza
-----	--	--------------

3. New Business

3.1	Discussion about standards for introducing a new fuel supplier	Panel
-----	--	-------

4. Next Meeting

4.1. SP Meeting: TBD

5. Meeting Adjourned

Appendix

A1

Note from Rich Grundza regarding breakin wording:

There are some inconsistencies in the wording of the requirements for additional breakin on failing references.

Specifically, section 10.1.1.8 of test method D 8114 reads "A laboratory must run a minimum of an additional 50 h of break-in, following an unacceptable reference test, in order for to be considered a new engine for calibration purposes."

Section 10.1.1.8 of D8226 states "For a new engine to be considered for calibration purposes, a lab must run a minimum of an additional 50 h of break-in, following an unacceptable reference test.", which is not consistent.

I believe the wording in D8114 is correct, save a grammatical error i.e it should read "A laboratory must run a minimum of an additional 50 h of break-in, following an unacceptable reference test, in order to be considered a new engine for calibration purposes."

I can address the grammatical error in D8114 considering this change as editorial. We could discuss this change to D8226 further as given the Itms requirements for the VIF, this may never be pursued. As a minimum, two tests are necessary to calibrate a new stand engine and running the additional break in and calling it a new engine would not allow sufficient engine hours left to conduct non-reference oil tests.

Equivalency Testing of Two Fuels (A D.O.E method)

Since this program involves two fuels (Current fuel and the candidate fuel A&B), three types of oil (Reference oil and two candidate oils, X, Y and Z) and four labs (LA, LB, LC, and LD). The design will be a 2 x 3 x 4 factorial (i.e., all 3 factors crossed) with factor F (fuel) with 2 levels (A and B), factor O (oil) with 3 levels (X, Y and Z), and factor L (labs) with 4 levels (LA, LB, LC, LD). The response (dependent) variable is the fuel economy (FE) of the two fuels. This design allows for the estimation of main effects and interaction effects. With 24 (2 times 3 times 4) treatment group combinations and one replicate per combination, it is possible to estimate all the main effects and two-way interactions. The three-way interaction provides an estimate of the experimental error (assuming there is no three-way interaction). To avoid assuming there is no three-way interaction, the 24 combinations could be replicated, e.g., two replicates per combination would result in 48 runs, three replicates would result in 72 runs and so forth. The larger the number of replicates the more sensitive the comparison (i.e., smaller differences between fuels could be detected). An optimal number of replicates can be established if the (approximate) experimental error for the observation is known by determining the expected confidence interval width.

The analysis of the resulting data would be modeled using a linear model¹. Because all three explanatory variables F, O and L are categorical, this linear model is often referred to as an Analysis of Variance or ANOVA. Because there are three crossed factors, this would be referred to as a three-way ANOVA. The comparison of the two fuels is straight forward if there are no (or very small) interaction effects. If there are interaction effects, then the interpretation is more complicated depending on the nature of the interactions.² The usual way to provide estimates of effects is to provide the point estimates along with the corresponding 95% confidence intervals. Models with and without various interaction effects would be compared using the Akaike Information Criterion (AIC) and the model with the smallest AIC would be selected as the most representative.

If the goal is to establish that the fuels are equivalent, then conventional statistical hypothesis testing cannot be used.³ Instead, equivalence hypothesis testing is required (Munk, 2004; Wellak, 2010). The mathematics behind equivalence testing is more complicated than conventional hypothesis testing, but the application of equivalence testing is, in practice, quite simple. It requires, however, that the experimenters must establish the range for a difference that is considered negligible. If the confidence interval for the difference falls entirely within the interval that describes equivalence, then the hypothesis of equivalence is supported. (On the other hand, if the confidence interval falls entirely outside the negligible interval then one of the fuels would be shown to be superior and the other inferior. If the confidence interval is only

1 The “linear” in “linear model” means that the statistical estimates for the effects are *linear functions* of the observed data values.

2 ANOVA main effects estimates are ignored if there are non-negligible interaction effects. In this case, the main effects are arbitrary. When there are interaction effects, one approach is to analyze the experiment as a one-way ANOVA.

3 Conventional hypothesis testing either rejects the null hypothesis (e.g., of no effect or difference) or fails to reject the null hypothesis. Failure to reject does not mean the null hypothesis is accepted. This means that conventional hypothesis testing cannot be used to establish equivalency.

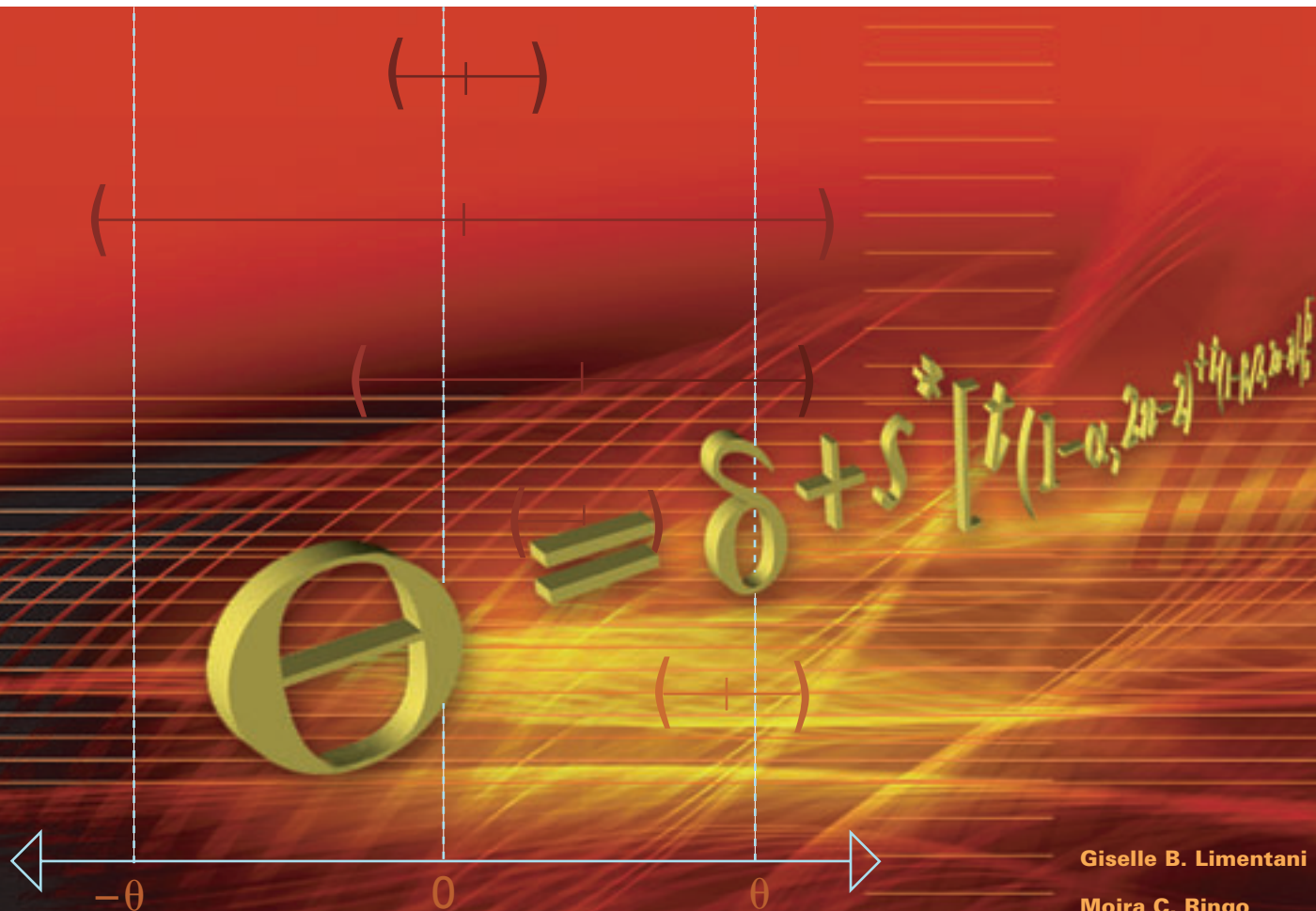
partly inside the negligible interval, then the results are indeterminant.) Setting the limits for a negligible difference requires expert judgement from test engineers.

Bibliography

ac053390m.pdf. (n.d.). Retrieved from <https://pubs.acs.org/doi/pdf/10.1021/ac053390m>.). This web page illustrates the concept of equivalence testing.

Munk, A. (2004). *Testing Statistical Hypotheses of Equivalence*. Taylor & Francis.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC.



Giselle B. Limentani

Moira C. Ringo

Feng Ye

Mandy L. Bergquist

Ellen O. McSorley

GlaxoSmithKline

BENOND

the t-Test:

Statistical Equivalence Testing

One of the most common questions considered by analytical chemists is whether replicate measurements are the same or significantly different from each other. The determination of significantly different results can be used to argue that a phenomenon is novel or to justify a claim of a significant improvement in a technique, process, or product. Science and technology are also driven by determinations of sameness, such as equivalence, control, or ruggedness.

Given the variability inherent to most instrument systems, the question of whether a measurable difference is “real” can be difficult to answer. In some cases, intuition, experience, and knowledge of the practical context

of the data can be used to inspect or “eyeball” the data to assess whether a true difference exists. For example, most of us would agree that a difference of 100% in a measurement that typically exhibits a precision of 1% is a real difference, and we would also agree that a difference of 0.01% is not significant for the same measurement. But what about less clear-cut cases in which the difference between two sets of data is similar to the precision? How much difference is “too much”? Not only does simple inspection fail in these circumstances, but a subjective process can be biased, is difficult to justify, and, most importantly, can lead to the wrong conclusion.

Statistical equivalence testing can be used to make better technical and business decisions from analytical data.

Statistical hypothesis testing offers a rigorous, objective approach to distinguishing truly significant differences in measurements from noise. Although many tests exist that are suitable for different situations, the statistical test that is most familiar is the simple-to-perform two-sample t -test (t is a probability distribution that is closely related to the standard normal distribution). However, the t -test has several limitations and may not be the most appropriate technique when the objective is to show equivalence between two data sets. We will demonstrate that the two one-sided t -test (TOST) is a better option in many cases. The two-sample t -test and TOST are distinct approaches for assessing a difference or equivalence in data; which test is used can have a significant impact on the outcome of the comparison as well as on the scientific and business decisions made as a result.

The basics

Most analytical scientists are familiar with the mean and standard deviation and how these measurements are used to make simple data comparisons. However, the calculated mean and standard deviation values, \bar{y} and s respectively, are merely estimates of the real mean and standard deviation values for a population of possible measurements. For example, for a sample data set of n independent measurements, \bar{y} and s can be calculated. These values are estimates of the mean and standard deviation of the entire population of measurements from which the sample was taken. A more informative description of the population mean is the range of probable “true” values, or confidence interval, for the mean, which is

$$\bar{y} \pm t_{(1-\alpha/2, n-1)} \cdot \frac{s}{\sqrt{n}} \quad (1)$$

for a two-sided $100(1-\alpha)\%$ confidence interval.

Note that the width of the confidence interval increases as s increases and n decreases. In other words, a data set for which s is large (noisy data) or n is small (few measurements) results in a wider confidence interval. (A narrow interval is more desirable.) The width of the confidence interval is also determined by the Student's t -value, known as t , for a given n and a given significance level α , which is related to the probability that the confidence interval includes the true mean. For most analytical applications, α is typically set at 0.05, producing a 95% confidence interval. Student's t -values may be obtained from published tables or by using the $TINV(\alpha, n-1)$ function in Excel. (Note that for some versions of Excel, it is necessary to use 2α in place of α for this function.)

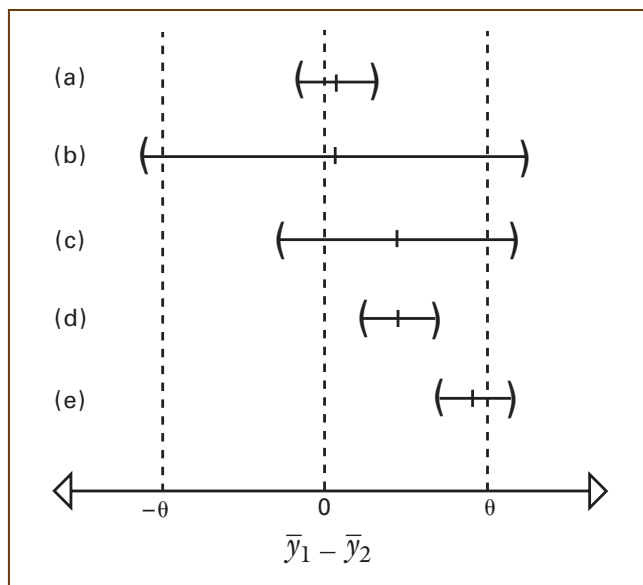


FIGURE 1. Comparison of two-sample t -test and TOST in terms of confidence intervals.

The conclusions for each scenario with a t -test and TOST, respectively, would be (a) equal and equivalent, (b, c) equal but not equivalent, (d) not equal but equivalent, (e) not equal and not equivalent.

significantly from the measurement reproducibility, which is a broader estimate of precision based on measurements by multiple analysts in multiple laboratories. In one study of several methods from a compendium for pharmaceuticals, the mean analytical repeatability was 1.5% RSD, while reproducibility was estimated at nearly double that amount (2). In the absence of a considerable amount of experience with the measurement under a range of conditions, it may be necessary to use a statistical procedure for estimating the measurement precision.

To conduct a statistical analysis, the analyst must consider the experimental objective and decide upon a null hypothesis for the test. An appropriate statistical test must then be chosen to prove that the null hypothesis is false. If sufficient evidence does not exist to prove falseness, the test defaults to the conclusion that the null hypothesis is correct but does not actually prove that it is correct. It is therefore critical to choose a null hypothesis that is the reverse of the statement the analyst wishes to prove. For example, if the analyst wishes to prove that the means from two groups of data are not equal, he or she should choose a null hypothesis in which the means are equal and then perform a test to demonstrate that this hypothesis is false. In addition, the analyst must determine how much risk of error is acceptable. In general, when a statistical test is conducted, two types of potential error can occur. The probability of a type 1 error, or α , represents the risk of rejecting the null hypothesis when it is true. A type 2 error, of probability β , occurs when the experimenter fails to reject the null hypothesis when it is false. Typical α values are 5–10%, and typical β values are 5–20%.

After the analyst has determined the appropriate error risk, it is usually necessary to estimate the sample size required for the data comparison. The procedure for estimating the necessary n depends on several factors. If s is large (poor measurement precision) and α and β are small (low risk of error is desired), the necessary n can be prohibitively large. In these cases, it may be necessary to refine the measurement system or revise the study design before proceeding. On the other hand, a small s can result

The mean is not the only parameter that is estimated with error. The estimated s , or measurement precision, can vary with the number of measurements that are made, the sample that is measured, and the manner in which the data are collected. With a 95% confidence limit and the assumption that the data follow a normal distribution, the true standard deviation for $n = 10$ can be as high as $1.8\times$ the measured s , and this error increases to $6.3\times$ the measured s for $n = 3$ (1). Moreover, the measurement repeatability (among measurements taken by a single analyst in a single laboratory) can differ

in a calculated required n of 2 or 3. Although these values may result from statistically valid calculations, the accuracy of the experimentally determined \bar{y} and s values is often very poor with small n values, and the test may result in an incorrect conclusion. It may be useful to examine risks of errors that are achievable with different n values and precisions so that the analyst can make the most informed compromise between risk and laboratory efficiency. Once valid data have been obtained, the experimenter has several choices on how to proceed with the data evaluation, depending on the research objective.

Two-sample t -test

The two-sample t -test allows comparison of the mean values of two data sets by the calculation of the test statistic

$$T = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (2)$$

in which \bar{y}_1 and \bar{y}_2 are the mean values from groups 1 and 2, s_p is an estimate of the pooled s of the measurements, and n_1 and n_2 are the number of observations for each group. The s_p of replicate sets of measurements is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (3)$$

in which s_1 and s_2 are the estimated s values for each set of measurements. The absolute value of the calculated T -value is then compared with the critical t -value for the selected significance level α (obtained from statistics tables or by using TINV in Excel). If the absolute value of the calculated T -value is greater than or equal to the critical t -value, then the data sets are declared statistically different. This test can also be performed by constructing a $100(1 - \alpha)\%$ confidence interval for the difference between two means using

$$(\bar{y}_1 - \bar{y}_2) \pm t_{(1 - \alpha/2, n_1 + n_2 - 2)} \cdot \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (4)$$

and determining whether the resulting confidence interval contains 0. If the confidence interval does not contain 0, then the means are declared not equal.

The null hypothesis of the two-sample t -test is that the mean values of the two data sets are equal; this places the burden on the analyst to prove that the mean values are in fact different. Although it is an appropriate test for proving that two data sets are different, problems arise when the two-sample t -test is used to show equivalence. First, the traditional two-sample t -test can reward the analyst for having poor precision and/or a small n . Equation 2 indicates that an increase in s_p or a decrease in n results in a smaller calculated T -value, which makes it more difficult to declare that the mean values are not equal. In the absence of substantial evidence to conclude that the mean values are different, the analyst can mistakenly default to the hypothesis that they are equal. Another problem associated with the use of the two-sample t -test is that it may lead the analyst to conclude that a statistically significant difference exists between the mean val-

ues when the magnitude of the difference is of no practical importance. This is a particular problem when the precision of the measurement is very good; a post hoc explanation of statistical significance may be required when the difference is of no practical importance. Therefore, the two-sample t -test is not well suited for showing the equivalence of mean values from two groups.

Equivalence test

An alternative to the two-sample t -test is TOST, designed specifically for bioequivalence testing of pharmaceutical products (3–6). It has recently been expanded into broader applications in pharmaceutical science (1, 7–10), process engineering (11, 12), psychology (13), medicine (14), chemistry (15), and environmental science (16). TOST begins with a null hypothesis that the two mean values are not equivalent, then attempts to demonstrate that they are equivalent within a practical, preset limit; this is conceptually opposite to the two-sample t -test procedure. Unlike the two-sample t -test, TOST appropriately penalizes poor precision and/or small n values and places the burden on the analyst to prove that the data sets are equivalent.

The design of an equivalence test can be challenging because the analyst must define an acceptance criterion on the basis of prior knowledge of the measurement as well as its intended application. This acceptance criterion θ is the limit outside which the difference in mean values should be considered practically and statistically significant. The analyst then constructs a $100(1 - 2\alpha)\%$ confidence interval for the difference between the two mean values and compares it with θ . If the confidence interval is completely contained within the interval $[-\theta, \theta]$, the mean values of the two data sets are equivalent. The use of θ establishes a priori what level of difference is acceptable.

Figure 1 displays five data comparisons and illustrates the different outcomes that arise from using the traditional two-sample t -test and TOST. The center of each confidence interval is the difference between the observed \bar{y} values. The width of the interval, which depends on the measurement precision, represents the range of plausible true differences in mean values between the data sets. If these intervals were created with the traditional two-sample t -test, for Figures 1a–c the analyst would conclude that there is no difference between the mean values because the confidence interval includes a difference of 0. The confidence intervals in Figures 1d and 1e do not include a difference of 0; therefore, the mean values would be declared different.

By contrast, if these intervals were created with TOST, the mean values of the two data sets would be declared equivalent only for Figures 1a and 1d because these confidence intervals are completely contained within the range from $-\theta$ to θ . The mean values for Figure 1d are declared equivalent even though the

confidence interval does not include 0, because the bias represented by the difference in means is small and within the interval $[-\theta, \theta]$. The confidence intervals in Figures 1b and 1c are too wide for the mean values of the data sets to be declared equivalent.

Making it easy

What is an acceptable difference between the mean values of two data sets? Choosing an appropriate θ can be a challenge. It must be greater than s/\sqrt{n} , lest the test fail simply because of imprecision rather than because of a true difference. However, θ must also be less than any specifications or standards that the testing is designed to challenge, or the test becomes too easy and will not adequately discriminate. Although some statistical software packages include TOST (17), our discussion provides a step-by-step process for performing equivalence testing with a preset θ with Excel or other commonly available computational software packages.

The first parameter that must be specified before an analyst performs statistical testing is δ , the absolute value of the true difference between the groups' mean values; δ is a hypothetical value such that if the absolute value of the observed difference is no more than δ , there is a strong probability of concluding that the two data sets represent equivalent results. An acceptable level of bias can be considered and included in δ ; however, the choice of a nonzero δ can decrease the ability of the test to distinguish a small but important difference between data sets. In addition, in the absence of extensive data, no objective basis exists for choosing a nonzero δ . Therefore, the most conservative approach is to assign a value of 0 to δ .

The second step is determining the n value needed for the test. Because the required n for TOST is related to θ and to other parameters discussed earlier, it may be helpful at this early stage to assume a range of potential θ values as a fraction of the specification or standard that the test is designed to challenge or as a multiple of s ; θ can be refined later. Then, with this series of potential θ values, along with values for α , β , δ , and s , the relationship between these variables may be solved iteratively to yield an appropriate n . Although many approaches exist for determining n (1, 10, 11, 18, 19), Excel and the following equation (20) can be used to easily calculate a simplified approximation of the required n for each group:

$$n = \frac{2s^2(z_\alpha + z_\beta)^2}{(\theta - \delta)^2} + 1 \quad (5)$$

in which the z values are the percentiles of the standard normal distribution, which are available in statistics tables or from the NORMSINV(α) function in Excel.

Table 1. θ for various n and upper limit of method precision s^* .

($\alpha = \beta = 0.05$, $\delta = 0$)

s^*	θ for $n = 5$	θ for $n = 10$	θ for $n = 12$	θ for $n = 30$
0.5	1.3	0.9	0.8	0.5
1.0	2.6	1.7	1.5	0.9
1.5	4.0	2.6	2.3	1.4
2.0	5.3	3.4	3.1	1.9
2.5	6.6	4.3	3.9	2.4
3.0	7.9	5.1	4.6	2.8

It may be helpful to compile a table of n and θ values for various combinations of α , β , δ , and s to assess the trade-offs between n and acceptable levels of risk that the test will lead to the wrong conclusion. Table 1 shows the acceptance criteria θ that are achievable for different combinations of s and n , with $\delta = 0$ and type 1 and 2 error rates of 5% ($\alpha = \beta = 0.05$). As expected, a smaller n increases the acceptance criterion within which the data would be considered equivalent, essentially making the test easier to pass but less scientifically defensible. As with any form of statistical testing, scientific and practical judgment is the key to proper implementation. Consider a situation in which θ is initially set at 0.9%; $n = 10$ would be adequate for s values $\leq 0.5\%$ because θ is $\leq 0.9\%$. If s is 1.0%, $n \geq 30$ is necessary to use $\theta = 0.9\%$.

After the appropriate sample size n is chosen, the third step is to take the first set of replicate measurements and estimate s . For better estimates of measurement precision, some approaches to equivalence testing have included multiple analysts and multiple days in data comparison studies (1). However, because of constraints on sample size, time, or resources, it is fairly common to use independent, replicate measurements from a single analyst or a single laboratory to estimate the precision of a measurement. Although this approach can lead to an underestimate of precision, it is a pragmatic compromise between appropriate statistical application and real-world constraints on resources.

To ensure that s adequately represents the true measurement precision, it is recommended that an upper confidence limit (e.g., the upper limit from a one-sided 80% confidence interval) be used as an estimate of measurement precision. The upper $100(1 - \gamma)\%$ confidence limit s^* for s may be calculated as

$$s^* = s \sqrt{\frac{n-1}{\chi^2_{(\gamma, n-1)}}} \quad (6)$$

in which $\chi^2_{(\gamma, n-1)}$ is the (100γ) th percentile of a distribution with $n - 1$ degrees of freedom (21, 22). The χ^2 is available in statistics reference tables (22) or from the CHIINV($1 - \gamma$, $n - 1$) function in Excel. This accommodation for error in the s estimate essentially enables a second lab or method to produce data with less precision, provided that the precision of the data is still equivalent to that from the original lab or method. Next, with

Table 2. Statistical significance vs practical relevance in the transfer of a dissolution method.

(θ calculated on the basis of the development laboratory s of 1.9%, with $\alpha = \beta = 0.05$, $\delta = 0$, $n = 12$)

Tablet number	% Label strength dissolved	
	Development lab	Manufacturing QC lab
1	90.8	86.2
2	88.0	87.4
3	90.5	88.2
4	90.0	89.7
5	91.0	87.3
6	86.0	87.6
7	88.3	88.0
8	89.3	86.5
9	88.9	89.6
10	91.1	89.1
11	86.2	86.1
12	91.3	86.2
\bar{y}	89.3	87.7
s	1.9	1.3
% RSD	2.1	1.5
TOST: 90% confidence interval for difference ($\theta = 3.7\%$)		
		(0.5, 2.7)
Two-sample t -test p-value		
		0.02

the desired α , β , n (for each group), δ , and s^* values, θ can be calculated by

$$\Theta = \delta + s^* [t_{(1-\alpha, 2n-2)} + t_{(1-\beta/2, 2n-2)}] \sqrt{\frac{2}{n}} \quad (7)$$

For example, suppose a researcher wishes to set an acceptance criterion for an assay that compares the content of a sample measured at two different laboratories, and the data are expressed as a percentage of the labeled content. If $\alpha = \beta = 0.05$, $\delta = 0$, and a predetermined $n = 10$ sample preparations exist for each site, statistically appropriate values for the acceptance criterion can be calculated for different levels of precision (Table 1). In all cases, the calculated θ value should be critically examined to determine whether it has practical relevance for the test that is performed (i.e., is not too small) yet is scientifically defensible for the intended application (i.e., is not too large).

After an appropriate acceptance criterion is chosen, the second set of measurements is taken and a slightly modified version of Equation 4 is used to calculate the confidence interval for the difference in mean values. Note that it is necessary to use α instead of $\alpha/2$ for the equivalence test.

$$(\bar{y}_1 - \bar{y}_2) \pm t_{(1-\alpha, n_1+n_2-2)} \cdot \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (8)$$

Alternatively, the Excel Analysis ToolPak can be used to simplify this procedure (12). Finally, this confidence interval is compared with the θ determined in the previous step. If the confidence interval for the difference in mean values is completely contained within $[-\theta, \theta]$, the mean values are considered equivalent. If the confidence interval contains some values outside $[-\theta, \theta]$, the test

has not provided sufficient evidence that the mean values are equivalent.

It is possible to declare the mean values equivalent if the confidence interval for the difference in mean values does not include 0, provided that the interval does not include any values outside $[-\theta, \theta]$. A confidence interval that does not include 0 suggests bias in the measurements, which may need to be examined further; however, the TOST conclusion that the mean values are equivalent says that the bias is less than the acceptable θ . Note that for Equation 8 to be valid, the s values of the two data sets should be similar. An appropriate variance test, such as Levene's (23), should be used to further evaluate significant differences in measurement precision.

Practical differences vs statistical significance

One of the more common criticisms of the traditional t -test is that it cannot distinguish between statistically significant and "scientifically relevant" differences. To evaluate how equivalence testing works in practice, a series of data comparison studies known as method transfers were conducted. In each of these studies, the analysis of a pharmaceutical product by a second laboratory was compared with the analysis by the original laboratory to assess whether the second laboratory applied the method in an equivalent manner.

Table 2 shows the data for the dissolution analysis of a tablet product from the development laboratory and the manufacturing QC laboratory. From data from the development lab, s was estimated at 1.9%. With the 80% upper confidence limit for this s , $\delta = 0$, and $\alpha = \beta = 0.05$, θ is calculated with Equation 7 to be 3.7% dissolved. Then, with the data in Table 2 and Equation 8,

Table 3. Comparison of the effect of poor precision on the outcome of TOST vs a two-sample t -test.

(θ calculated on the basis of the development laboratory s^* from previous experiments of 1.5%, with $\alpha = \beta = 0.05$, $\delta = 0$, $n = 6$)

Tablet number	% Label strength dissolved	
	Development lab	Contract lab
1	82	74
2	92	70
3	78	84
4	85	76
5	77	90
6	79	77
\bar{y}	82	79
s	5.6	7.3
% RSD	6.9	9.2
Difference between \bar{y} values		
		3
TOST: 90% confidence interval for difference ($\theta = 3.5\%$)		
		(-3.1, 10.5)
Two-sample t -test p-value		
		0.35

a 90% confidence interval for the difference between the laboratory mean values is calculated to be 0.5–2.7%. Because the difference between the means is less than the required 3.7%, the null hypothesis that the mean values are not equivalent is disproved, and the laboratory methods are declared equivalent. Thus, the method is successfully transferred to the manufacturing QC laboratory.

When a traditional two-sample *t*-test with $\alpha = 0.05$ is used to compare the data in Table 2, sufficient evidence exists (p-value = 0.02) for the analyst to conclude that the laboratory means are not equal. In this case, the p-value is the probability of observing a *T*-value that is more extreme than the observed *T*-value when the means are equal. The conclusion reached with the traditional two-sample *t*-test is problematic, because the difference in \bar{y} values (1.6%) is a scientifically acceptable difference for this test. This example highlights a key advantage of TOST over a two-sample *t*-test for showing equivalence—TOST allows small, scientifically irrelevant differences to exist without leading to the conclusion that the laboratory means are not equivalent.

Consequences of poor precision

Table 3 is an example of a tablet dissolution method that was transferred from a development laboratory to a contract laboratory during the early stages of product development. In this study, $n = 6$ for each laboratory because of limited sample availability. With an initial s^* estimate of 1.5% from previous analyses, θ was calculated at 3.5%, which would generally be considered acceptable for a method of this type. Notice that the actual s from each laboratory was much larger than the initial estimate of 1.5%. This was determined to be the result of poor sample homogeneity caused by degradation during storage. When the data are compared via an equivalence test with $\theta = 3.5\%$, sufficient evidence does not exist to declare the laboratories' methods equivalent. Moreover, when the upper limit of the experimentally observed s of 5.6% is used to calculate the minimum achievable θ for the test, $\theta = 19\%$, which is considerably larger than would be acceptable for a dissolution test. Even without performing the second set of measurements, the analyst can see that the method precision and study design are insufficient to meet the objective, and it is appropriate to conclude that the method transfer is a failure.

If the laboratory mean values had been compared with a two-sample *t*-test ($\alpha = 0.05$), the p-value of 0.35 indicates that the data would not have provided sufficient evidence to conclude that the laboratories were different. In other words, the laboratory mean values would have been declared equal and the method transfer would have been deemed a success. In this case, the traditional two-sample *t*-test would not have rejected the hypothesis that the data sets are equal, because s was too large relative to the difference between the \bar{y} values. This example highlights another

key advantage of TOST over a two-sample *t*-test—TOST appropriately penalizes the analyst if the observed variance is too large.

We thank Bob Barraza, Sue Long, Chris Kidd, Gerald Monger, John Allen, and Charles Goss for helpful suggestions.

Statistical hypothesis testing offers a rigorous, objective approach to distinguishing truly significant differences in measurements from noise.

Giselle B. Limentani is a director in product development for GlaxoSmithKline in Research Triangle Park, N.C. She has more than 20 years of experience developing new drugs in the pharmaceutical industry. She recently presented a paper on analytical method transfer at the Pharmaceutical and Biomedical Analysis 2004 conference. Moira C. Ringo is an investigator in product development at GlaxoSmithKline. Feng Ye is a senior manager in quality engineering for Amgen in Thousand Oaks, Calif. Mandy L. Bergquist is a principal statistician in research and development at GlaxoSmithKline. Ellen O. McSorley is an independent statistical consultant in Cary, N.C. Address correspondence to Limentani at 5 Moore Dr., Research Triangle Park, NC 27709 or Giselle.B.Limentani@gsk.com.

References

- (1) Kringle, R.; et al. *Drug Inf. J.* **2001**, *35*, 1271–1288.
- (2) Horwitz, W. *J. AOAC Int.* **1977**, *60*, 1355–1363.
- (3) Westlake, W. J. *Biometrics* **1976**, *32*, 741–744.
- (4) Schuirmann, D. J. *Biometrics* **1981**, *37*, 617.
- (5) Westlake, W. J. *Biometrics* **1981**, *37*, 589–594.
- (6) Schuirmann, D. J. *J. Pharmacokinet. Biop.* **1987**, *15*, 657–680.
- (7) Hartmann, C.; et al. *Anal. Chem.* **1995**, *67*, 4491–4499.
- (8) *Average, Population, and Individual Approaches to Establishing Bioequivalence (Draft Guidance)*; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research: Washington, DC, Aug 1999.
- (9) *Statistical Approaches to Establishing Bioequivalence*; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research: Washington, DC, Feb 2001.
- (10) Tubert-Bitter, P.; et al. *J. Clin. Epidemiol.* **2000**, *53*, 1268–1274.
- (11) Stein, J.; Doganaksoy, N. *Qual. Engin.* **1999–2000**, *12*, 105–110.
- (12) Richter, S. J.; Richter, C. *Qual. Engin.* **2002**, *14*, 375–380.
- (13) Rogers, J. L.; Howard, K. I.; Vessey, J. T. *Psychol. Bull.* **1993**, *113*, 553–565.
- (14) Munk, A.; Hwang, J. T.; Brown, L. D. *Biometr. J.* **2000**, *42*, 531–552.
- (15) Roy, T. J. *Math. Chem.* **1997**, *21*, 103–109.
- (16) McBride, G. *Aust. NZ. J. Stat.* **1998**, *41*, 19–29.
- (17) Luzar-Stiffler, V.; Stiffler, C. *J. Comput. Inf. Tech.* **2002**, *3*, 233–239.
- (18) Chow, S.-C.; Shao, J.; Wang, H. *Stat. Med.* **2003**, *22*, 55–68.
- (19) Zhang, P. *J. Biopharm. Stat.* **2003**, *13*, 529–538.
- (20) Bristol, D. R. *Commun. Statist. Theory Methods* **1993**, *22*, 1953–1961.
- (21) Seely, R. J.; Munyakazi, L.; Haury, J. *Biopharm.* **2001**, *10*, 28–34.
- (22) Chow, S. C.; Liu, J. P. *Statistical Design and Analysis in Pharmaceutical Science*; Marcel Dekker: New York, 1995.
- (23) Milliken, G. A.; Johnson, D. E. *Analysis of Messy Data, Volume 1: Designed Experiments*; Chapman & Hall: New York, 1992; pp 19–22.