CLINICAL AND LABORATORY STANDARDS INSTITUTE®

3rd Edition

EP09c

Measurement Procedure Comparison and Bias Estimation Using Patient Samples

This guideline covers the design of measurement procedure comparison experiments using patient samples and subsequent data analysis techniques used to determine the bias between two *in vitro* diagnostic measurement procedures.

A guideline for global application developed through the Clinical and Laboratory Standards Institute consensus process.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Clinical and Laboratory Standards Institute

Setting the standard for quality in medical laboratory testing around the world.

The Clinical and Laboratory Standards Institute (CLSI) is a not-for-profit membership organization that brings together the varied perspectives and expertise of the worldwide laboratory community for the advancement of a common cause: to foster excellence in laboratory medicine by developing and implementing medical laboratory standards and guidelines that help laboratories fulfill their responsibilities with efficiency, effectiveness, and global applicability.

Consensus Process

Consensus—the substantial agreement by materially affected, competent, and interested parties—is core to the development of all CLSI documents. It does not always connote unanimous agreement but does mean that the participants in the development of a consensus document have considered and resolved all relevant objections and accept the resulting agreement.

Commenting on Documents

CLSI documents undergo periodic evaluation and modification to keep pace with advances in technologies, procedures, methods, and protocols affecting the laboratory or health care.

CLSI's consensus process depends on experts who volunteer to serve as contributing authors and/or as participants in the reviewing and commenting process. At the end of each comment period, the committee that developed the document is obligated to review all comments, respond in writing to all substantive comments, and revise the draft document as appropriate.

Comments on published CLSI documents are equally essential and may be submitted by anyone, at any time, on any document. All comments are managed according to the consensus process by a committee of experts.

Appeal Process

When it is believed that an objection has not been adequately considered and responded to, the process for appeal, documented in the CLSI *Standards Development Policies and Processes,* is followed.

All comments and responses submitted on draft and published documents are retained on file at CLSI and are available upon request.

Get Involved—Volunteer!

Do you use CLSI documents in your workplace? Do you see room for improvement? Would you like to get involved in the revision process? Or maybe you see a need to develop a new document for an emerging technology? CLSI wants to hear from you. We are always looking for volunteers. By donating your time and talents to improve the standards that affect your own work, you will play an active role in improving public health across the globe.

For additional information on committee participation or to submit comments, contact CLSI.

Clinical and Laboratory Standards Institute
950 West Valley Road, Suite 2500
Wayne, PA 19087 USA
P: +1.610.688.0100
F: +1.610.688.0700
www.clsi.org
standard@clsi.org

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

EP09c, 3rd ed.
June 2018
Replaces EP09-A2-IR and EP09-A3

Measurement Procedure Comparison and Bias Estimation Using Patient Samples

Jeffrey R. Budd, PhD
A. Paul Durham, MA
Thomas E. Gwise, PhD
Douglas M. Hawkins, PhD
Mark Holland, PhD
Beimar Iriarte, MS
Anders Kallner, MD, PhD
Kristian Linnet, MD, PhD
Robert Magari, PhD
Jeffrey E. Vaks, PhD

Abstract

Clinical and Laboratory Standards Institute guideline EP09—*Measurement Procedure Comparison and Bias Estimation Using Patient Samples* is written for laboratorians and manufacturers. It describes procedures for determining the bias between two measurement procedures, and it identifies factors for consideration when designing and analyzing a measurement procedure comparison experiment using patient samples. An overview of the measurement procedure comparison experiment includes considerations for both manufacturers and laboratorians. Details on how to create difference and scatter plots for visual inspection of the data are provided. Once the data are characterized, various methods are introduced for quantifying the relationship between two measurement procedures, including bias estimates and regression techniques. The final chapter contains recommendations for manufacturers' evaluation of bias and statement format for bias claims.

Clinical and Laboratory Standards Institute (CLSI). *Measurement Procedure Comparison and Bias Estimation Using Patient Samples*. 3rd ed. CLSI guideline EP09c (ISBN 978-1-68440-006-5 [Print]; ISBN 978-1-68440-007-2 [Electronic]). Clinical and Laboratory Standards Institute, 950 West Valley Road, Suite 2500, Wayne, Pennsylvania 19087 USA, 2018.

The Clinical and Laboratory Standards Institute consensus process, which is the mechanism for moving a document through two or more levels of review by the health care community, is an ongoing process. Users should expect revised editions of any given document. Because rapid changes in technology may affect the procedures, methods, and protocols in a standard or guideline, users should replace outdated editions with the current editions of CLSI documents. Current editions are listed in the CLSI catalog and posted on our website at www.clsi.org. If you or your organization is not a member and would like to become one, or to request a copy of the catalog, contact us at: Telephone: +1.610.688.0100; Fax: +1.610.688.0700; E-Mail: customerservice@clsi.org; Website: www.clsi.org.

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Copyright ©2018 Clinical and Laboratory Standards Institute. Except as stated below, any reproduction of content from a CLSI copyrighted standard, guideline, derivative product, or other material requires express written consent from CLSI. All rights reserved. Interested parties may send permission requests to permissions@clsi.org.

CLSI hereby grants permission to each individual member or purchaser to make a single reproduction of this publication for use in its laboratory procedures manual at a single site. To request permission to use this publication in any other manner, e-mail permissions@clsi.org.

Suggested Citation*

CLSI. *Measurement Procedure Comparison and Bias Estimation Using Patient Samples*. 3rd ed. CLSI guideline EP09c. Wayne, PA: Clinical and Laboratory Standards Institute; 2018.

Previous Editions:
January 1986, April 1993, December 1995, September 2002, July 2010, August 2013

ISBN 978-1-68440-006-5 (Print)
ISBN 978-1-68440-007-2 (Electronic)
ISSN 1558-6502 (Print)
ISSN 2162-2914 (Electronic) Volume 38, Number 12

* EP09 was corrected in June 2018, and the code was revised to EP09c.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Committee Membership

Consensus Council

Dennis J. Ernst, MT(ASCP), NCPT(NCCT)
Chairholder
Center for Phlebotomy Education
USA

Mary Lou Gantzer, PhD, FACB
Vice-Chairholder
USA

J. Rex Astles, PhD, FACB, DABCC
Centers for Disease Control and Prevention
USA

Lucia M. Berte, MA, MT(ASCP)SBB, DLM, CQA(ASQ)CMQ/OE
Laboratories Made Better!
USA

Karen W. Dyer, MT(ASCP), DLM
Centers for Medicare & Medicaid Services
USA

Thomas R. Fritsche, MD, PhD, FCAP, FIDSA
Marshfield Clinic
USA

Loralie J. Langman, PhD, DABCC, FACB, F-ABFT
Mayo Clinic
USA

Ross J. Molinaro, PhD, MLS(ASCP)CM, DABCC, FACB
Siemens Healthcare Diagnostics, Inc.
USA

James R. Petisce, PhD
BD Diagnostic Systems
USA

Andrew Quintenz
Bio-Rad Laboratories, Inc.
USA

Robert Rej, PhD
New York State Department of Health – Wadsworth Center
USA

Zivana Tezak, PhD
FDA Center for Devices and Radiological Health
USA

Expert Panel on Evaluation Protocols

James H. Nichols, PhD, DABCC, FACB
Chairholder
Vanderbilt University School of Medicine
USA

Paula Ladwig, MS, MT(ASCP)
Vice-Chairholder
Mayo Clinic
USA

Valeria L. Alcon, PhD
Health Canada
Canada

Julianne Cook Botelho, PhD
Centers for Disease Control and Prevention
National Center for Environmental Health
USA

Jeffrey R. Budd, PhD
Beckman Coulter
USA

Mark D. Kellogg, PhD, MT(ASCP), DABCC, NACB
Boston Children's Hospital
USA

Marina V. Kondratovich, PhD
FDA Center for Devices and Radiological Health
USA

Robert J. McEnroe, PhD
USA

Jeffrey E. Vaks, PhD
Roche Molecular Diagnostics
USA

Document Development Committee on Method Comparison and Bias Estimation Using Patient Samples

Jeffrey R. Budd, PhD
Chairholder
Beckman Coulter
USA

A. Paul Durham, MA
USA

Thomas E. Gwise, PhD
FDA Center for Drug Evaluation and Research
USA

Anders Kallner, MD, PhD
Karolinska Hospital
Stockholm, Sweden

Kristian Linnet, MD, PhD
University of Copenhagen
Copenhagen, Denmark

Staff

Clinical and Laboratory Standards Institute
USA

Luann Ochs, MS
Project Manager

Megan L. Tertel, MA, ELS
Editorial Manager

Catherine E.M. Jenkins
Editor

Kristy L. Leirer, MS
Editor

Laura Martin
Editor

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Acknowledgment

CLSI, the Consensus Council, the Expert Panel on Evaluation Protocols, and the Document Development Committee on Method Comparison and Bias Estimation Using Patient Samples gratefully acknowledge the following individuals for their thorough review of this guideline and the resulting corrections and clarifications reflected in the corrected edition:

Jeffrey R. Budd, PhD
Chairholder
Beckman Coulter
USA

Douglas M. Hawkins, PhD
University of Minnesota Medical
Center - Fairview
USA

A. Paul Durham, MA
USA

Mark Holland, PhD
Beckman Coulter
USA

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Contents

Abstract i

Committee Membership iii

Foreword vii

Chapter 1: Introduction 1

1.1 Scope 1
1.2 Standard Precautions 1
1.3 Terminology 2

Chapter 2: Introduction to Measurement Procedure Comparison and Bias Estimation 7

2.1 Process Flow Chart 7
2.2 Introduction 8

Chapter 3: Preparing for the Measurement Procedure Comparison Study 13

3.1 Planning for Measurement Procedure Comparison Studies 13
3.2 Considerations for the Medical Laboratory 18

Chapter 4: Performing the Measurement Procedure Comparison Study 19

4.1 Measurement Procedure Familiarization Period 19
4.2 Quality Control 19
4.3 Running the Study 20
4.4 Inspection of Data During Collection 20
4.5 Documentation of Rejected Data 20

Chapter 5: Visual Data Analysis 21

5.1 Visual Data Review 21
5.2 Scatter Plots 21
5.3 Difference Plots 22
5.4 Inspect Plots for Underlying Characteristics 23

Chapter 6: Quantitative Analysis 31

6.1 Estimating Bias From Difference Plots 31
6.2 Fitting a Line to Scatter Plots (Regression Analysis) 36
6.3 Bias and Regression Parameters With Confidence Intervals 41

Chapter 7: Comparisons Within a Measurement Procedure 43

7.1 Comparing Two Conditions Within an Already Validated Measurement Procedure 43
7.2 Sample Type Comparisons 43
7.3 Other Comparisons 43

Chapter 8: Interpreting Results and Comparing to Performance Criteria 45

8.1 Interpreting the Study Results 45
8.2 Manufacturer's Statement of Bias Performance Claims 46
8.3 Medical Laboratory's Statement of Bias Performance 47

Chapter 9: Conclusion 48

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Contents (Continued)

Chapter 10: Supplemental Information 48

References 49

Appendix A. Deming Regression 51

Appendix B. Weighted Deming Regression 56

Appendix C. A Practical Example Illustrating Bias Estimation and Measurement Procedure Comparison Techniques 59

Appendix D. Example Datasets 67

Appendix E. Detecting Aberrant Results (Outliers) 71

Appendix F. Confidence Interval of a Median Estimate of Bias Between Measurement Procedures 74

Appendix G. Ordinary Linear Regression 80

Appendix H. Weighted Least Squares Regression (Weighted Ordinary Linear Regression) 83

Appendix I. Passing-Bablok Regression 90

Appendix J. Profile Weighted Deming Regression 93

Appendix K. Iterative Approaches for Estimating Confidence Intervals for Bias and Regression Parameters 96

The Quality Management System Approach 102

Related CLSI Reference Materials 103

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Foreword

Measurement procedure comparison is one of the most common techniques used by both manufacturers and medical laboratorians to estimate the bias of an *in vitro* diagnostic (IVD) measurement procedure relative to a comparator. It involves the comparison of results from patient samples from two measurement procedures intended to measure the same component (eg, measurand concentration) with the key determination being the estimate of bias between them.

A number of different scenarios exist in which measurement procedure comparison studies are indicated. For both the manufacturer and the medical laboratorian, the ideal scenario is the comparison of a candidate measurement procedure to a generally accepted standard or reference measurement procedure. In the case of a manufacturer, this involves the establishment and perhaps validation of performance claims for bias. In the case of a laboratorian, it involves introducing a measurement procedure into the laboratory, including verification of its manufacturer's claims (specifications). The scope of the experimental and data-handling procedures for these two purposes differs. In either case the assumption that the reference measurement procedure provides "true" values means that bias (systematic measurement error) is estimated.

Quite commonly, however, there is no standard or reference measurement procedure. The manufacturer instead compares a candidate measurement procedure to the most appropriate measurement procedure currently available. The laboratorian usually compares the candidate and an available procedure. Then, there might not be a "true" value and the "difference," rather than the "bias," is estimated.

Given the variety of performance characteristics of IVD measurement procedures, a single experimental design is not appropriate for all types of laboratory and manufacturer measurement procedure comparisons. Therefore, performance characteristics such as measuring interval and precision profile are taken into account in structuring an experiment for comparing two measurement procedures. Multiple worked examples are presented.

This guideline is intended to promote effective and correct data analysis and reporting using standard experimental and statistical methods.

Manufacturers of medical laboratory measurement procedures or devices should use this guideline to establish and standardize their bias performance claims. Many different forms have been used for such claims, and they have not always been sufficiently specific to allow user verification.

Overview of Changes

This guideline replaces the previous edition of the approved guideline, EP09-A2-IR, published in 2010. Several changes were made in this edition, including:

- Broader coverage of measurement procedure comparison applications
- More reasons for comparisons based on patient samples (factor comparisons [eg, sample tube types])
- Visualization/exploration of data using difference plots
- Regression descriptions including weighted options, Deming, and Passing-Bablok techniques
- Measurement of bias using difference plots
- Measurement of bias at clinical decision points
- Computation of confidence intervals for all parameters
- Outlier detection using extreme studentized deviate
- Relocation of most of the detailed mathematical descriptions to the appendixes

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

This guideline was corrected in 2018 and replaces the original third edition of the approved guideline, EP09-A3, published in 2013. Corrections were made as follows:

- Reorganizing the content to emphasize the process of performing a measurement procedure comparison
- Clearly specifying that manufacturers should use regression analysis to characterize bias
- Adding information on using precision profile information in performing Deming regressions
- Adding more information on determining confidence intervals for bias estimates at specified concentrations using regression fits
- Making corrections to the description of the Passing-Bablok regression technique
- Adding a detailed description of the bootstrap iterative technique for bias estimation
- Correcting minor miscellaneous errors in equations

NOTE: Due to the complex nature of the calculations in this guideline, it is recommended that the user have access to a computer and statistical software.

NOTE: The content of this guideline is supported by the CLSI consensus process and does not necessarily reflect the views of any single individual or organization.

Key Words

Alternative regression methods, bias, evaluation protocol, experimental design, linear regression, measurement procedure comparison, outliers, quality control, residuals

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

Measurement Procedure Comparison and Bias Estimation Using Patient Samples

Chapter 1: Introduction

This chapter includes:

- Guideline's scope and applicable exclusions
- Standard precautions information
- "Note on Terminology" that highlights particular use and/or variation in use of terms and/or definitions
- Terms and definitions used in the guideline
- Abbreviations and acronyms used in the guideline
- Symbols used in the guideline

1.1 Scope

This guideline provides recommendations for designing an experiment and selecting methods to quantify systematic measurement error (bias or difference) between measurement procedures based on comparing patient samples. It provides both difference plot and regression procedures to determine the relationship between two measurement procedures either across their measuring intervals or at selected concentrations. Intended users of this guideline are manufacturers of *in vitro* diagnostic (IVD) reagents—which includes those who create laboratory-developed tests—as well as regulatory authorities and medical laboratory personnel.

This guideline is for use with measurement procedures that provide quantitative numerical results. This guideline is not intended for use with ordinal IVD examinations, commonly referred to as qualitative procedures (see CLSI document EP12[1]). This guideline is not intended to provide information on evaluation of random error (see CLSI documents EP05[2] and EP15[3]) or to determine the total error inherent in a comparison of measurement procedures (see CLSI document EP21[4]). It is not intended to measure the variability of multiple replicates collected during the measurement of a sample, nor is it intended to measure the bias of individual measurements such as those resulting from sample interference (as covered in CLSI document EP07[5]).

1.2 Standard Precautions

Because it is often impossible to know what isolates or specimens might be infectious, all patient and laboratory specimens are treated as infectious and handled according to "standard precautions." Standard precautions are guidelines that combine the major features of "universal precautions and body substance isolation" practices. Standard precautions cover the transmission of all known infectious agents and thus are more comprehensive than universal precautions, which are intended to apply only to transmission of bloodborne pathogens. Published guidelines are available that discuss the daily operations of diagnostic medicine in humans and animals while encouraging a culture of safety in the laboratory.[6] For specific precautions for preventing the laboratory transmission of all known infectious agents from laboratory instruments and materials and for recommendations for the management of exposure to all known infectious diseases, refer to CLSI document M29.[7]

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

1.3 Terminology

1.3.1 A Note on Terminology

CLSI, as a global leader in standardization, is firmly committed to achieving global harmonization whenever possible. Harmonization is a process of recognizing, understanding, and explaining differences while taking steps to achieve worldwide uniformity. CLSI recognizes that medical conventions in the global metrological community have evolved differently in different countries and regions and that legally required use of terms, regional usage, and different consensus timelines are all important considerations in the harmonization process. CLSI recognizes its important role in these efforts, and its consensus process focuses on harmonization of terms to facilitate the global application of standards and guidelines.

Essentially, new documents must adhere to the latest international vocabulary of metrology[8] whenever an ambiguity in the interpretation or understanding of terms occurs. Recently, many definitions have become more explicit and understandable, but the international language is difficult and compact. The international vocabulary of metrology deals with general metrology and terminology that should be useful for most disciplines that measure quantities.

The understanding of a few terms has changed during the last decade as the concepts have developed. Particularly, *trueness* (measurement trueness) is defined as expressing the closeness of agreement between the average of an infinite number of replicate measurements and a reference value; and *precision* (measurement precision) is defined as closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions. Consequently, *accuracy* (measurement accuracy) is the closeness of agreement between a measured value and a true quantity value of a measurand. Thus, this concept comprises both trueness and precision, and applies to a single result. *Measuring interval* has replaced *reportable range* when referring to "a set of values of a measurand for which the error of a measuring instrument (test) is intended to lie within specified limits." An *interval* $[a;b]$ is delineated by two limits a and b ($b > a$), whereas a *range* ($\mathrm{r}[a;b]$) is expressed as the difference between b and a ($b - a$). Thus, the range of the interval $[a;b]$ is the difference ($b - a$) that is denoted by $r[a;b]$.

The term *measurand* is used when referring to the quantity intended to be measured instead of *analyte* (component represented in the name of a measurable quantity). The term *measurement procedure* replaces *analytical method* and *assay* for a set of operations, used in the performance of particular measurements according to a given method.

Verification focuses on whether specifications of a measurement procedure can be achieved, whereas *validation* verifies that the procedure is fit for purpose.

NOTE: Mandates are generally reserved for CLSI standards but are occasionally allowed in CLSI guidelines. In CLSI guidelines, use of the term "must" is either 1) based on a requirement or 2) indicative of a necessary step to ensure patient safety or proper fulfillment of a procedure. The working group evaluated use of the term "must" and deemed it appropriate.

1.3.2 Definitions

accuracy (of measurement) – closeness of agreement between a measured quantity value and a true quantity value of a measurand[8]; **NOTE 1:** The concept "measurement accuracy" is not a quantity and is not given a numerical quantity value. A measurement is said to be more accurate when it offers a smaller measurement error[8]; **NOTE 2:** The term "measurement accuracy" should not be used for "measurement trueness," and the term "measurement precision" should not be used for "measurement accuracy," which, however, is related to both these concepts[8]; **NOTE 3:** "Measurement accuracy" is sometimes understood as closeness of agreement between measured quantity values that are being attributed to the measurand.[8]

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

bias (of measurement) – difference between the expectation of the test results and an accepted reference value[9]; estimate of a systematic measurement error[8]; **NOTE 1:** Because trueness, the agreement between a measured value and a reference value, is a concept and cannot be expressed numerically, it is estimated using bias; **NOTE 2:** Bias at a specified concentration can be determined using the slope and intercept from a measurement procedure comparison regression; **NOTE 3:** In laboratory verification comparison studies using difference plots, when bias is relatively constant and in the same direction (positive or negative) within a specified measuring interval, bias can be expressed as an average over that specified interval.

correlation coefficient (r)//determination coefficient (r^2) – a measure of the linear relationship between two random variables; **NOTE 1:** It ranges from −1 to 1, ie, from perfect negative to perfect positive linear relationship; **NOTE 2:** $r = 0$ indicates no observed linear relationship.

decision point (medical decision point) – a concentration of the measurand that is used as a threshold for making a medical statement; **NOTE:** Often, decision points refer to reference limits, but other concentrations, such as from medical guidelines, are also used.

Deming regression – a method to estimate slope and intercept parameters from a measurement procedure comparison experiment with allowance for both measurement procedures to have imprecision; **NOTE:** The measurement error for each measurement procedure is accounted for in the estimation procedure.[10]

difference plot – a plot of the difference between a measured value and a reference concentration plotted on the y-axis vs the reference concentration on the x-axis; **NOTE 1:** Often, a dashed line is drawn at zero difference; **NOTE 2:** The reference concentration is often expressed as the average of the results of the measurements; **NOTE 3:** The difference may be expressed relative to the reference concentration; **NOTE 4:** For presentation of data for validation purposes, the reference concentration should be comparative measurement procedure values.

imprecision – the random dispersion of a set of replicate measurements and/or values expressed quantitatively by a statistic, such as standard deviation or coefficient of variation.

least squares regression – the method of statistically placing the location of the estimated line or curve among the data so that the sum of the squares of the distances of each data point from the line in the perpendicular direction from the x-axis (ie, parallel to the y-axis) is minimized; **NOTE:** It allows the direct algebraic computation of the coefficients and an estimate of their uncertainty.

measurand – quantity intended to be measured[8]; **NOTE 1:** The specification of a measurand requires knowledge of the kind of quantity, description of the state of the phenomenon, body, or substance carrying the quantity, including any relevant component, and the chemical entities involved[8]; **NOTE 2:** The measurement, including the measuring system and the conditions under which the measurement is carried out, might change the phenomenon, body, or substance such that the quantity being measured may differ from the measurand as defined. In this case, adequate correction is necessary[8]; **EXAMPLE 1:** S-Creatinine concentration is frequently measured using an enzyme-based technique resulting in a color reaction. Results will be biased due to different specificity of the measurement procedures and different quantities measured; **EXAMPLE 2:** The length of a steel rod in equilibrium with the ambient temperature of 23°C will be different from the length at the specified temperature of 20°C, which is the measurand. In this case, a correction is necessary[8]; **NOTE 3:** In chemistry, "analyte," or the name of a substance or compound, is a term sometimes used for "measurand." This usage is erroneous because these terms do not refer to quantities.[8]

measuring interval – set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental measurement uncertainty, under defined conditions[8]; **NOTE:** The lower limit of a measuring interval should not be confused with detection limit.[8]

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.
© Clinical and Laboratory Standards Institute. All rights reserved.

ordinary linear regression (OLR) – least squares linear regression that usually refers to nonweighted least squares regression; **NOTE:** OLR can also be described as uniformly weighted ordinary least squares regression.

outlier – the observation in a sample, so far separated in value from the remainder as to suggest that it may be from a different population, or the result of an error in measurement; **NOTE 1:** It is also defined as a number of a set of values that is inconsistent with the other numbers of the set[11]; **NOTE 2:** Statistical tests can be used to identify outliers, but visual inspection of data can serve as a sensitive indicator of when such tests are useful.

Passing-Bablok regression – nonparametric procedures to estimate slope and intercept parameters from a measurement procedure comparison experiment.[12,13]

precision (of measurement) – closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions[8]; **NOTE 1:** Measurement precision is usually expressed numerically by measures of imprecision, such as standard deviation, variance, or coefficient of variation under the specified conditions of measurement[8]; **NOTE 2:** The "specified conditions" can be, for example, repeatability conditions of measurement, intermediate precision conditions of measurement, or reproducibility conditions of measurement[8,14]; **NOTE 3:** Measurement precision is used to define measurement repeatability, intermediate measurement precision, and measurement reproducibility.[8]

replicate – a value resulting from a repeat analysis of the same specimen.

reproducibility (measurement) – measurement precision under reproducibility conditions of measurement.[8]

reproducibility condition (of measurement) – condition of measurement, out of a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects[8]; **NOTE 1:** The different measuring systems may use different measurement procedures[8]; **NOTE 2:** A specification should give the conditions changed and unchanged, to the extent practical.[8]

residual – the difference between a given data point and its predicted value; **NOTE:** As used in EP09, for evaluating a value predicted by a regression fit.

sample – one or more parts taken from a specimen[15]; **EXAMPLE:** A volume of serum taken from a larger volume of serum[15]; **NOTE 1:** A sample is prepared from the patient specimen and used to obtain information by means of a specific laboratory test; **NOTE 2:** The system from which a sample is taken might not be of the same type as that of the measurand. For example, a given blood sample may serve for measurement of pH in plasma hemoglobin concentration in erythrocytes.

scatter plot//scatter diagram – a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data; **NOTE 1:** The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis; **NOTE 2:** Also called scatter chart, scattergram, scatter diagram, or scatter graph.

trueness (of measurement) – closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value[8]; **NOTE 1:** The measure of trueness is usually expressed in terms of bias[9]; **NOTE 2:** Measurement trueness is inversely related to systematic measurement error but is not related to random measurement error[8]; **NOTE 3:** "Measurement accuracy" should not be used for "measurement trueness"[8] and vice versa.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

validation – verification, where the specified requirements are adequate for an intended use[8]; **EXAMPLE:** A measurement procedure, ordinarily used for the measurement of mass concentration of nitrogen in water, may be validated also for measurement of mass concentration of nitrogen in human serum[8]; **NOTE:** An external claims verification conducted by the manufacturer for the purpose of validating that the measurement procedure is fit for the purpose of quantitatively determining the concentration of a measurand in a medical laboratory setting is an example of a validation.

verification – provision of objective evidence that a given item fulfills specified requirements[8]; **EXAMPLE 1:** Confirmation that a given reference material as claimed is homogeneous for the quantity value and measurement procedure concerned, down to a measurement portion having a mass of 10 mg[8]; **EXAMPLE 2:** Confirmation that performance properties or legal requirements of a measuring system are achieved[8]; **EXAMPLE 3:** Confirmation that a target measurement uncertainty can be met[8]; **NOTE 1:** When applicable, measurement uncertainty should be taken into consideration[8]; **NOTE 2:** The item may be, eg, a process, measurement procedure, material, compound, or measuring system[8]; **NOTE 3:** The specified requirements may be, eg, that a manufacturer's specifications are met[8]; **NOTE 4:** Verification should not be confused with calibration. Not every verification is a validation.[8]

weighted regression – a parametric regression analysis technique that weights the influence of individual patient results based upon a predefined criterion; **NOTE 1:** Weighted regression is often applied to both least squares (*y* results weighted) and Deming (*x* and *y* results weighted) regressions; **NOTE 2:** The weighting scheme typically assigns influence to a result that is inversely related to its expected variance.

1.3.3 Abbreviations and Acronyms

CI	confidence interval
CV	coefficient of variation
IVD	*in vitro* diagnostic
OLR	ordinary linear regression
QC	quality control
SD	standard deviation
SE	standard error
WLS	weighted least squares

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

1.3.4 Symbols Used in the Text

The following symbols are used in this guideline. See Table 1 for a list of terms used to describe the two measurement procedures being compared.

X	*X* characteristic of a comparative measurement procedure
Y	*Y* characteristic of a candidate measurement procedure
N	total number of samples
r	correlation coefficient
x	observation from comparative measurement procedure
x_i	estimate of comparative measurement procedure's value for sample number *i*. This may either be a single comparative measurement procedure replicate from a sample or the average of multiple comparative measurement procedure replicates from that sample.
y	observation from candidate measurement procedure
y_i	estimate of candidate measurement procedure's value for sample number *i*. This may either be a single candidate measurement procedure replicate from a sample or the average of multiple candidate measurement procedure replicates from that sample.
d_i	difference between comparative and candidate measurement procedures for sample number *i*
$\bar{d}$	average of sample result differences between comparative and candidate measurement procedures
z_i	position on horizontal axis of a difference plot for sample number *i*
a	*y* intercept of the regression line
b	slope of the regression line
k	rank order of the samples
S_{yx}	standard deviation of residuals of regression (standard error of estimate)
X_c	selected concentration of comparative measurement procedure (eg, medical decision level)

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

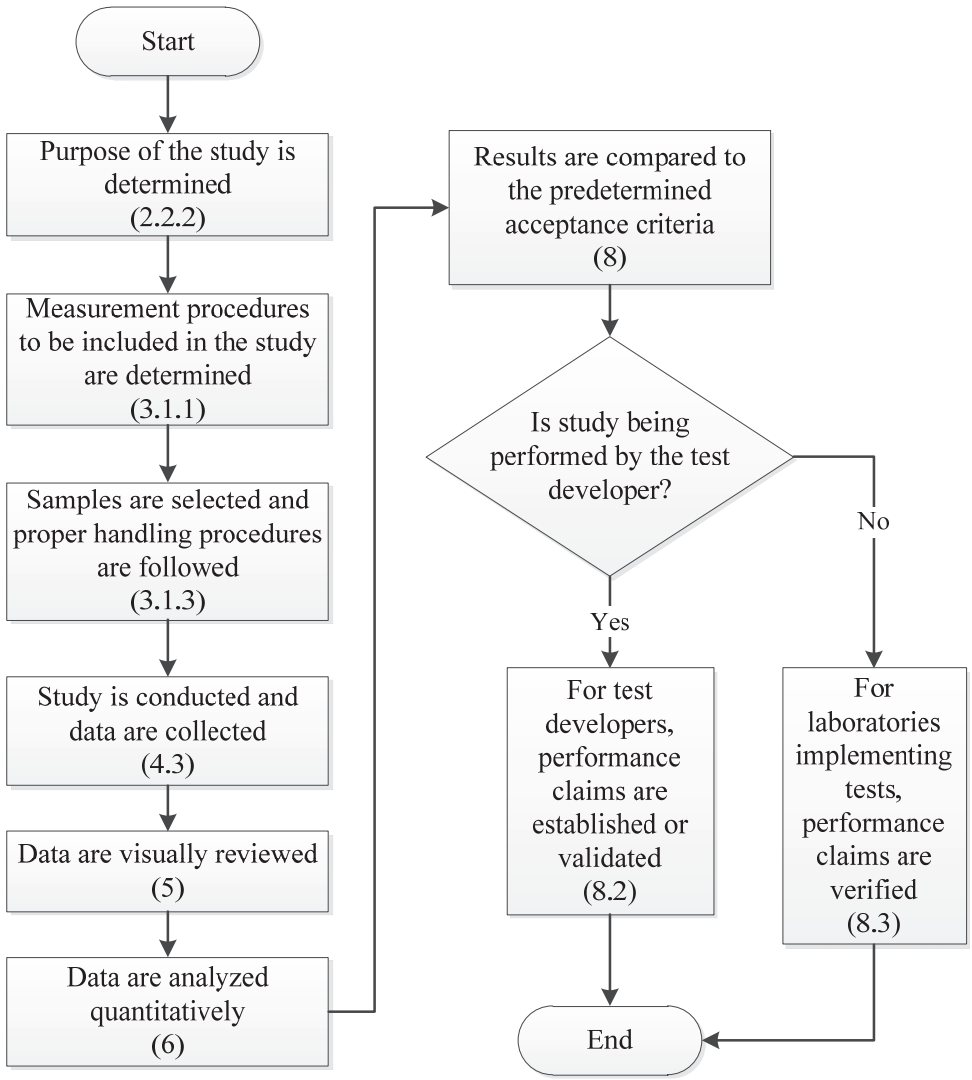
Chapter 2: Introduction to Measurement Procedure Comparison and Bias Estimation

This chapter includes:

- Process flow chart
- Introduction to measurement procedure comparison
- Overview of the study
- Purposes for performing measurement procedure comparisons

2.1 Process Flow Chart

Figure 1 shows the measurement procedure comparison process for test developers and laboratories. Applicable chapters and subchapters for each topic are included parenthetically.



[*] Five basic symbols are used in process flow charts: oval (signifies the beginning or end of a process), arrow (connects process activities), box (designates process activities), diamond (includes a question with alternative "Yes" and "No" responses), pentagon (signifies another process).

Figure 1. Process for Measurement Procedure Comparison[*]

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

2.2 Introduction

The purpose of this guideline is to establish good practices for measuring systematic differences of a measurement procedure over its measuring interval in a population of patient samples, relative to a comparative measurement procedure or reference method. Difference plots are used to visually portray the relationship between measurement procedures to evaluate whether the relationship is consistent with a constant difference or as a constant percentage difference (constant CV) over the measuring interval. The plots can be used when performing laboratory verification to determine the bias estimate from such plots through either an average or a median. Given the knowledge gained from the difference plots, users are provided with regression fit options to characterize bias in terms of slope and intercept and bias estimates at selected concentrations.

This guideline describes multiple types of studies in which measurement procedures are compared, each of which has its own experimental design parameters. These study designs dictate differences in the number of factors to incorporate, including the number of samples, and the number of replicates for each sample. The situations covered in this guideline assume a study is comparing two procedures that measure the same quantity by using measurement procedure results from study samples.

In selecting an analysis technique for a set of data, a stepwise process is described that starts with visual data inspection using difference and scatter plots. The data from difference plots can then be used to estimate the bias (or percent bias) between measurement procedures. For medical laboratorians, the data might not need additional analysis. The guideline continues, however, by describing various regression techniques and their underlying assumptions that help determine which technique should be used in each situation. Such techniques can, in many cases, provide more robust estimates of bias, so medical laboratories might wish to use them. Manufacturers use regression techniques in almost all cases. The goal throughout the guideline is to propose a set of techniques for determining the relationship between measurement procedures and to detail the strengths and weaknesses of these techniques for given situations.

A brief description of measurement procedure comparison scenarios is provided in the following subchapters. Subchapter 2.2.1 is a general overview common to all scenarios. Subchapters 2.2.2.1 and 2.2.2.2 are intended for manufacturers or research laboratories that have created a candidate measurement procedure. Subchapter 2.2.2.3 is intended for the typical medical laboratorian.

2.2.1 Overview of the Measurement Procedure Comparison Study

The purpose of a measurement procedure comparison study is first determined, along with the two measurement procedures to be used in the study. The three primary purposes for such a study are introduced in Subchapter 2.2.2. Other purposes for such studies are described in Chapter 7.

For any well-conducted study, the personnel performing the measurements must be familiar with the instrument systems used in the study. The familiarization period is described in Subchapter 4.1.

There are numerous considerations for conducting any measurement procedure comparison study. General considerations, including sample selection and handling, are covered in Subchapter 3.1.3, and those specific only for studies by medical laboratories are covered in Subchapter 3.2.

Once the data are collected for the study, they must be reviewed to determine if the goals of data collection have been met and to characterize the data for interval and distribution of measurements and other factors that can dictate what techniques are used for data analysis. Visual data inspection techniques are suggested in Chapter 5 for this purpose.

The outcome of a measurement procedure comparison study is a quantification of the relationship between two such procedures. This relationship can be expressed, when applicable, as a bias over a selected interval

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

or a bias at a selected concentration. For validation studies, the overall relationship (bias) is expressed using the estimated slope and intercept. The techniques that can be used to supply such estimates are described in Chapter 6.

Finally, Chapter 8 discusses the steps for comparing estimated bias to acceptance criteria and for stating performance claims.

Throughout this guideline, the terms in Table 1 are used to describe the measurement procedures to be compared.

Table 1. Measurement Procedure Terminology

***X* Characteristic Plotted on the Horizontal, x-axis**	***Y* Characteristic Plotted on the Vertical, y-axis**
Independent variable	Dependent variable
Comparative measurement procedure or reference measurement procedure	Candidate measurement procedure

2.2.2 Primary Purposes for Measurement Procedure Comparisons

Certain purposes for which measurement procedure comparison studies are performed impose experimental design parameters on their analysis and presentation. In particular, assay manufacturers and developers conducting such a study for regulatory registrations or approval (eg, for the purpose of establishing and validating claims for a relationship or bias relative to another quantitative measurement procedure) must perform a suitable regression analysis of the data. In this setting, difference plotting can play a supporting role but does not constitute, on its own, an acceptable way to summarize the study. In contrast, when a medical laboratory conducts a measurement procedure comparison study to fulfill regulatory requirements for "verifying accuracy [trueness]" when introducing a commercial assay, difference plotting generally suffices, though regression analysis is often desirable.

2.2.2.1 Establishing the Relationship Between Measurement Procedures by the Manufacturer

Manufacturers must establish the relationship of any candidate measurement procedure for measurand quantification with a comparative measurement procedure, ideally a reference measurement procedure. Typically, when such a reference measurement procedure is available, the desired result of the comparison is no significant bias between them. However, often a new (candidate) measurement procedure is developed as an improvement over a comparative measurement procedure (eg, an automated IVD procedure to replace a microplate procedure). In such situations the primary goal is to establish the bias between them. At least 100 patient samples with measurand values spanning the common measuring interval of the two measurement procedures are recommended for establishment of bias claims. Influential factors can be included in the experimental design such as calibration, run, day, reagent lot, calibrator lot, and instrument. The average result of multiple sample replicates may be used for both procedures to decrease the uncertainty of the bias estimate.

2.2.2.2 Validation by the Manufacturer

The second measurement procedure comparison situation a manufacturer might encounter is the validation that an IVD measurement procedure meets the claims already established for it. Such a study, more than any other type of measurement procedure performance study, can be the key to validating that the measurement procedure is fit for the purpose of quantitatively determining the concentration of the measurand in a medical laboratory setting. Therefore, such an experiment is usually performed at one or more external sites, because the goal is to show that the candidate measurement procedure has low bias in comparison to a trusted comparative measurement procedure in an in-use situation.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Generally speaking, to validate that a candidate measurement procedure is fit for purpose, it is evaluated per its instructions for use. Usually, the procedure uses only a single replicate. Statistically, this fact does not invalidate the estimation of bias using averages of multiple replicates. Provided each replicate represents equivalent information (ie, order has no influence), each represents a result from the procedure per the instructions for use, so averaging the results from multiple replicates simply improves bias estimation. Multiple replicates also permit Deming regression (see Appendixes A and B) without needing prior knowledge of measurement procedure imprecision. For the visual displays of relative bias presented in Chapters 5 and 6, replicate results from each measuring procedure (candidate, comparator) are averaged before being plotted.

However, given that the primary purpose of this type of verification study is to validate whether the candidate measurement procedure is fit for purpose, regulatory authorities might require using a set of single candidate results from individual replicates matched to average comparator results across multiple replicates, or using a set of single candidate results matched with single comparator results. In these two cases, the visual displays described in Chapters 5 and 6 can be from individual measurements rather than averaged results. Analysis techniques remain the same, except the use of Deming regression necessitates prior knowledge of measurement procedure imprecision.

At least 100 patient samples with measurand values spanning as much of the common measuring interval of both measurement procedures as feasible are recommended for such a validation, and the study should be conducted at each site over three to five days at minimum. Typically, the measurement procedure's bias claims are verified when the estimated bias is within a predetermined acceptance criterion.

2.2.2.3 Measurement Procedure Introduction to the Medical Laboratory

Medical laboratories typically perform measurement procedure comparison studies when they are introducing an IVD product into their testing menu. The candidate measurement procedure typically replaces one currently used in the laboratory. A decision to bring in a candidate measurement procedure is often based, at least in part, on the performance results provided by the manufacturer through either performance claims or postmarket study comparisons. In either case, when the comparison data to be performed by the laboratory is available from the manufacturer, the desired outcome is to verify the manufacturer-supplied bias performance. Otherwise the goal is to independently quantify the bias (difference). The bias can help determine what changes, if any, need to be made in reporting results from the candidate measurement procedure. Such changes might include changes in reference intervals (see CLSI document EP28[16]) or medical decision values to reflect the difference between measurement procedure results.

To perform the analysis methods described in this guideline, medical laboratories should attempt to measure at least 40 patient samples that span the common measuring interval of the measurement procedures. Single, nonreplicated sample measurements are typically obtained for each measurement procedure. When sample volume and time constraints permit, the average result of multiple sample replicates can be used for both procedures to decrease the uncertainty of the bias estimate.

2.2.2.4 Summary of Measurement Procedure Comparison Studies

Table 2 lists the characteristics typical of each type of study.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Table 2. Typical Study Characteristics

Type of Study	Conducted by	Number of Samples	Number of Candidate MP Replicates Used	Number of Candidate MP Systems	Recommended Analyses to Be Used to Determine Bias
MP claims establishment	Manufacturer	≥100	1 or more	1 or more	Regression
MP validation	Manufacturer	≥100	1	1 or more	Regression
MP introduction (verification)	Laboratory	≥ 40	1 or more	1	Difference plot or regression

Abbreviation: MP, measurement procedure.

There might or might not be constraints on the number of replicates for the comparative measurement procedure. Whenever multiple replicates are used, the average of a sample's replicates is typically used as the result estimate for each sample on that measurement procedure.

There are no restrictions on the number of candidate instrument systems used by a manufacturer. A laboratory usually has only one candidate instrument system that it is introducing into the laboratory. The goal is usually to determine how that system compares to the current instrument system.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

This page is intentionally left blank.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Chapter 3: Preparing for the Measurement Procedure Comparison Study

This chapter includes:

- Planning for measurement procedure comparison studies
- Considerations for the medical laboratory

3.1 Planning for Measurement Procedure Comparison Studies

This subchapter covers experimental considerations for the manufacturer. In most cases, these considerations are in common with those for the medical laboratory. The considerations unique to the medical laboratory are covered in Subchapter 3.2.

When a manufacturer or research laboratory creates an IVD measurement procedure, a comparison study is typically performed to help determine whether this candidate measurement procedure successfully quantifies the measurand. Successful quantification is determined when the bias over the measuring interval (characterized by a regression slope and intercept), or the bias measured at specified measurand concentrations, are within pre-established specifications when comparing a candidate to a comparative measurement procedure. Ideally, the comparative measurement procedure for such a comparison is an accepted standard or reference measurement procedure with low imprecision and bias. However, the comparative measurement procedure is often a commercially available measurement procedure from another manufacturer, which can have limitations with regard to imprecision, measuring interval, linearity, and specificity for the measurand. This situation can be especially true for an automated candidate measurement procedure that is being designed to replace a manual procedure. In this case, the comparative measurement procedure can have significantly higher imprecision than the candidate.

Regardless of the imprecision inherent in the comparative measurement procedure, during the research and development process, the manufacturer can perform multiple measurement procedure comparisons between the comparative and the candidate measurement procedures to create a robust calibration scheme. Such testing is beyond the scope of this guideline.

3.1.1 Comparative Measurement Procedure

The measurement procedure comparison experiment gives an estimate of the bias between two measurement procedures and estimates for bias at any specified concentration. The estimates of the concentration for a sample from the comparative measurement procedure should ideally have the following characteristics:

- Have lower uncertainty than estimates of the concentration for that same sample from the candidate measurement procedure, which can often be achieved by averaging replicates, when needed.

- Be free from known interferences, whenever possible. This should be true for both the comparative and the candidate measurement procedures.

- Use the same units as the candidate measurement procedure or have the ability to be converted to the same units.

- Be traceable to reference material or reference measurement procedures, whenever possible.

The experiment does not segregate the various sources of bias into those coming from each of the measurement procedures being compared (see CLSI document EP14[17] for information on detection of

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

matrix interference). Interference effects can contribute as much as imprecision effects to a difference between measurement procedures in any given sample. (Proper characterization of interference effects on each measurement procedure can be determined by a separate experiment; see CLSI document EP07.[5])

The goal for manufacturers for both claims establishment and validation studies is to determine the relationship between the candidate and comparative measurement procedures over the broadest interval possible. Every attempt must be made to collect samples that cover the entire measuring interval of the measurement procedure rather than just the reference interval or the medical decision points.

The measurement of measurand concentration is restricted by the analytical measuring intervals of the two measurement procedures (ie, the concentrations for which they overlap). The analytical measuring interval is the measurand concentration interval claimed by the manufacturer to provide acceptable performance. Ideally, the measuring interval of the comparative measurement procedure is at least as wide as the measuring interval of the candidate measurement procedure so that bias at the limits of the analytical measuring interval can be compared.

Sometimes a candidate measurement procedure is developed to meet an unmet medical need for an extended interval of measurand values. When a reference measurement procedure is available that covers this extended interval, it can be used as the comparative measurement procedure. Often, however, there is no reference method and the comparative measurement procedure has a restricted measuring interval. In such a case, when the comparative measurement procedure's dilution capabilities have been verified, then unmodified high concentration samples on the candidate measurement procedure can be compared to diluted samples on the comparative measurement procedure to cover the extended measuring interval.

3.1.2 Factors Affecting the Measurement Procedure Comparison

Many experimental factors affect the bias estimate from a measurement procedure comparison study. Random factors such as within-run and between-run (including between-day) variability play a role. It is generally expected that such random factors will not create a systematic shift in bias but only affect the variability of the bias estimate. Therefore, an experimental design that increases the replication over such factors (eg, number of runs, number of days, or number of replicates per run) will decrease the bias estimate uncertainty and thus the width of the confidence interval (CI) for the bias estimate.

Other factors can cause systematic shifts of the bias estimate. These can include shifts due to operator, calibrations, instruments, reagent lots, and calibrator lots. The manufacturer should have an idea of which factors provide the highest potential for such bias shifts. The ideal IVD measurement procedure uses the calibration procedure to eliminate such shifts, but not all IVD measurement procedures are ideal. When such a factor is significant, its effect can also be moderated by increasing the number of instances of that factor in the study (eg, increasing the number of calibrations).

3.1.2.1 Establishment of Relationship Study Design

The goal of a measurement procedure comparison study to establish a relationship is to determine the candidate measurement procedure's bias relative to the comparative measurement procedure. In some cases, comparisons are to a reference or other comparative measurement procedure with an expected bias to the candidate measurement procedure. The manufacturer determines customer needs and performs testing on the candidate measurement procedure during its development process. Both of these sources of information should be used to determine an expectation for bias, which can be different than zero, and the factors that should be considered for such a study.

An establishment study is an analytical study typically performed at the manufacturer's site that answers the question, "What is the relationship between the candidate and the comparative measurement procedures?" It is up to the manufacturer to determine the rigor of the study based on knowledge

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

accumulated on the candidate measurement procedure being tested. During assay development, the manufacturer determines the potential influence of various factors on candidate measurement procedure bias. Because the manufacturer later validates the claim established by this study, the manufacturer can determine for itself how the establishment study is designed.

A study, for example, could be conducted over three to five days with a relatively equivalent number and concentration distribution of samples run each day. Besides the day factor, the manufacturer may choose to include additional factors in the study design such as reagent lot, calibrator lot, calibration, instrument, and operator.

If the comparative measurement procedure is a well-controlled standard or reference measurement procedure, then none of these factors need to be considered for this measurement procedure, because each of the individual comparative results would be defined as true (within expected random error). Often the manufacturer can be hampered by the lack of information on the variability for the comparative measurement procedure and the lack of access to multiple instruments and reagent lots. In either case, each subset of comparative measurement procedure results is ideally generated proximal to the same time as possible to the same subset of results generated by the candidate measurement procedure.

Averaging multiple replicates can reduce an estimate of bias's imprecision for an establishment study. However, the available sample volume frequently limits the number of replicates. For such a study during establishment, if three or more replicates are available, the use of the median for the sample value estimate rather than the average is a reasonable alternative.

3.1.2.2 Manufacturer's Validation Study Design

The goal of a manufacturer's validation study, usually conducted at one or more medical laboratory sites, is to show that the candidate measurement procedure can meet bias specifications when run under typical medical laboratory conditions. A measurement procedure can also be shown to be acceptable when a known difference (non-zero) is confirmed. This can be seen as a validation that the measurement procedure is fit for the purpose of quantitatively determining the concentration of a measurand in a medical laboratory setting.

The study should be performed over a minimum of three to five days. Other factors similar to those mentioned for an establishment of relationship study can be considered in conducting such a validation experiment. Typically, however, complex study designs are not feasible in an experiment conducted at an external medical laboratory. However, performing multiple calibrations during the study and combining data from more than one medical laboratory may increase the robustness of bias estimates.

Both measurement procedures should be run as described in their instructions for use. Therefore, if only one replicate is used for providing a measurement, then a typical strategy is to collect only one replicate to protect against the possibility that subsequent replicates are not exchangeable with the first.[18]

Using a single replicate per instructions for use also ensures the study validates that a candidate measurement procedure is fit for purpose. Statistically, this fact does not invalidate the estimation of bias using averages of multiple replicates. Provided each replicate represents equivalent information (ie, order has no influence), each represents a result from the procedure per the instructions for use, so averaging of the results on multiple replicates simply improves bias estimation.

When multiple replicates are being collected for either procedure, then, as mentioned in Subchapter 3.1.4.1, either the average or median results may be used in the calculation of the sample value for that measurement procedure. Generally, the average is the summary of choice, unless use of the median is needed to offset the effect of skewness in the distribution of replicate values.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

The resulting estimates of bias (characterized by slope and intercept or calculated at specified concentrations) should be compared to acceptance criteria established before study initiation.

3.1.3 Study Samples

The goal for any measurement comparison study is to determine the relationship between the candidate and comparative procedures using unmodified patient samples that cover the common measuring intervals.

Patient samples should be collected and handled according to accepted good laboratory practice and manufacturers' recommendations. Any medical, demographic, or analytical (eg, hemolysis, icterus, and lipemia/turbidity indexes) information available on the patient samples should be retained.

When it is desired to perform replicate determinations or to retain enough sample for possible retesting, and the sample volume needed cannot be obtained from a single patient, then "minipools" can be made by mixing samples from multiple patients (two, when possible) with approximately the same concentration of measurand and similar disease histories, whenever possible. In cases in which specific measurand concentrations are desirable but not available for spanning the entire measuring interval, two samples with disparate concentrations of measurand may be pooled. When the samples are whole blood, the samples need to be serologically compatible to be mixed. Any such modified patient samples should be noted in the listing of results, and ideally also in any graphical representation of the data. Ideally, modified samples should comprise a small portion of the samples in the study (eg, no more than 20%).

NOTE: The process of pooling can mask sample-to-sample characteristics by averaging out unique or sample-specific influences and thus can lead to an optimistic picture of the comparability of the two measurement procedures. To minimize this effect, material from an individual patient sample should be used in no more than one minipool sample aliquot to be tested. However, in cases of rare diseases or specimens, it might be necessary to use an individual patient sample to prepare multiple aliquots at different measurand concentrations.

When the biological specimens under study are tissue samples rather than body fluid aliquots, heterogeneity is a concern. Handling of such samples is covered in CLSI document I/LA28.[19] Their characteristics are not amenable to the measurement of multiple replicates and certainly prohibit the pooling of multiple samples.

Occasionally, upon sample collection there are many more low concentration samples than ones of higher concentration. In such cases, the first option is to see whether the sample set can be supplemented with additional high concentration samples. For some measurement procedures, the incidence rate of such high samples is so low that such supplementation is not feasible. A visualization technique for such datasets is provided in Subchapter 5.4.4.

When unmodified patient samples are difficult to obtain at specific portions of the measuring interval, other options can be investigated such as dilution or depletion to obtain lower concentrations, or spiking analyte into unmodified samples to obtain high concentration samples. However, such techniques should only be used as a last resort because modified samples are likely to have commutability limitations between the two different measurement procedures. In fact, modified samples have been shown to have commutability limitations for different reagent lots for the same measurement procedure.[20] When such techniques are used, they should be labeled as such in any plots in which they are presented and analysis should be performed both with and without the modified samples.

In any study comparing candidate to comparative measurement procedures, only one of the sample types (eg, serum or plasma) recommended for both measurement procedures should be used. The sample type used in the comparison should be stated.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Samples for the measurement procedure comparison study might need to be stored as they are accumulated, especially when the two measurement procedures are in different locations. The medical laboratory should ensure that variation related to damage or deterioration due to transport or storage does not affect either measurement procedure. To the extent possible, the split samples used on the candidate and comparative measurement procedures should be treated in a similar manner with respect to storage and handling.

3.1.4 Number of Samples and Replicates

For establishment and validation studies, manufacturers should use at least 100 samples that meet the criteria stated in Subchapters 3.1.1 and 3.1.3.

When two medically relevant populations have been shown to provide different relationships between the candidate and comparative measurement procedures, 100 samples from each such population should be included in the study. Such differences may be stated in the manufacturer's product labeling or in the medical literature. As an example, two immunoassays might exhibit one relationship on samples from pregnant female patients and another relationship on samples from male patients, due perhaps to different concentrations of a cross-reacting substance in the two populations that has different effects on the two measurement procedures. In another example, two measurement procedures for parathyroid hormone might diverge markedly from one another on samples from dialysis patients even when they agree closely on samples from patients with normal kidney function.

A sufficient volume of each sample should be obtained so that the number of replicates specified for the candidate and comparative measurement procedures can be run.

For a manufacturer's establishment of bias performance, the matched sample-to-sample bias comparison necessitates that the average of each sample concentration be used in the data analysis. When multiple replicates are available, they should be averaged to estimate each sample concentration. The underlying assumption behind this averaging of results is that both replicates, from each measurement procedure, are attempting to measure the same, unchanging quantity for that sample,[21] and that an average therefore reduces the uncertainty (standard error) of the estimate for that sample. When three or more replicates are available, the use of the median rather than the average for the sample value estimate is a reasonable alternative.

When the manufacturer is conducting the study to verify that its stated claims are being met through a validation study, the measurement of a sample's concentration must be derived as it would be during the intended use of the measurement procedure. Therefore, when a candidate measurement procedure uses only one replicate to produce patient results, a typical strategy is to collect one replicate per sample for this procedure in the manufacturer's validation (see Subchapters 2.2.2.2 and 3.1.2.2). When, however, a measurement procedure (ie, a manual measurement procedure) specifies that two replicates be averaged to get a patient result (eg, enzyme-linked immunosorbent measurement procedure), the average of two replicates should be used in the validation study. For such a study, the manufacturer might or might not have more leeway to collect and average multiple replicates from the comparative measurement procedure. The use of multiple replicates for either procedure is reasonable only in cases in which sample results are demonstrated not to be dependent on replicate order.

3.1.5 Sample Sequence

During a prospective study, sample sets may be run as they are collected or they may be organized into sets (eg, that cover the measuring interval) for later daily batch testing. Within each set, all samples should be run in random sequence for both measurement procedures. When, within each set of samples, individual replicates are run independently (eg, in a manual method), the replicates should be randomized as well. For a random access, automated instrumentation, there is no need to randomize individual replicates within each set of samples.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

3.1.6 Time and Duration

For a given sample, measurement by the comparative and candidate procedures should occur within a time span consistent with measurand stability. When possible, use samples drawn the day of the analysis. When stored samples are used, they should all be stored in a manner that ensures their stability and meets the stated storage conditions for both the candidate and the comparative measurement procedures. Samples should be stored in the same manner for both procedures to avoid introducing storage condition as a variable.

3.2 Considerations for the Medical Laboratory

The purpose of a measurement procedure comparison conducted by a medical laboratory can either be the verification of bias performance claimed by the manufacturer or independent quantification of bias. Either goal can be achieved through the use of this guideline.

The considerations covered in Subchapter 3.1 for manufacturers also cover studies conducted by medical laboratories, except for the unique considerations covered below in this subchapter.

3.2.1 Comparative Measurement Procedure Selection

For the comparative measurement procedure, the medical laboratory's current measurement procedure, the measurement procedure used by the manufacturer in the labeled claims, or a recognized reference measurement procedure is used. The current measurement procedure is the most likely choice because the medical laboratory usually wants to understand the difference between a measurement procedure to be introduced and the one it has been using. The medical laboratory should be aware that any comparative method, other than a reference measurement procedure, can have some vulnerability to interfering substances and matrix effects.

3.2.2 Number of Samples and Replicates

Medical laboratories should analyze at least 40 samples that meet the criteria stated in Subchapters 3.1.1 and 3.1.3 to establish the bias (difference) between measurement procedures. More samples improve the confidence in the statistical estimates and increase the opportunity to incorporate the effects of unexpected interfering substances (individual idiosyncratic biases).

For the medical laboratory, single measurements per procedure are acceptable when deemed appropriate by the laboratory director.

For measurement procedure introduction into a medical laboratory (see Subchapter 2.2.2.3), the matched sample-to-sample bias comparison necessitates that the best estimate of each sample concentration be used. Therefore, when multiple replicates are available, they should be averaged (or the median taken) to estimate each sample concentration. The underlying assumption behind this averaging of results is that the replicates from each measurement procedure are attempting to measure the same, unchanging quantity for that sample[21] and that an average therefore reduces the uncertainty of the sample concentration estimate.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Chapter 4: Performing the Measurement Procedure Comparison Study

This chapter includes:

- Measurement procedure familiarization period
- QC
- Running the study
- Inspection of data during the collection

4.1 Measurement Procedure Familiarization Period

The operators of both the candidate and the comparative measurement procedures should be familiar with the following items:

- Operation of the instrument systems and procedural steps to perform the measurement procedures
- Maintenance procedures of these instrument systems
- Sample preparation methods
- Calibration and measurement procedure quality monitoring functions

Manufacturers' training programs, when offered, can be a part of the familiarization period. Medical laboratory personnel must set up and operate the instrument systems long enough to ensure that the operators understand all procedures and can properly operate them. Five days are recommended for the measurement procedure familiarization period. For extremely simple instrument systems, a shorter period can suffice; for complex, multichannel instrument systems, a longer period might be needed.

The operators should practice analyzing unmodified patient samples to bring to their attention all possible contingencies (eg, error flags, error correction) that might arise during routine operation of either instrument system. Data collected during this period can be used as objective evidence that the personnel are qualified to perform the measurement procedures being compared, but not as study data. The measurement procedure familiarization period is not complete until the operators can perform the measurement procedures with confidence.

The familiarization period is optional for a manufacturer who is conducting the study under its own standard procedures as long as the personnel conducting the study have been trained on these procedures and have experience running both measurement procedures.

4.2 Quality Control

4.2.1 Calibration and Procedure Control

Calibrating both the candidate and comparative measurement procedures at the start of the study is recommended to ensure that each is in conformance with all QC parameters. When necessary, the instruments should be recalibrated as indicated in the instructions for use or the laboratory's operating procedure for either measurement procedure.

4.2.2 Quality Control Procedures

Appropriate QC procedures should be followed during the experiment. Control charts should be kept, and any run that appears out of control on either measurement procedure should be repeated on that measurement procedure.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

4.3 Running the Study

When possible, the same split samples should be run on both the candidate and the comparative measurement procedures at roughly the same time following the established study design (see Chapter 3). Synchronization is less critical when the samples are stored frozen, which might be necessary when the two instrument systems running the measurement procedures are at separate locations. Nevertheless, when samples are grouped for running at specified times, those same grouped samples should ideally be run at those times on both measurement procedures (eg, group one is run on day one for each instrument system). When problems are detected with a sample result, any retesting should be done as soon as possible.

The data collection process should ensure accurate, complete, and timely accumulation of study results including notation of any study or instrument errors.

4.4 Inspection of Data During Collection

Data should be inspected during its collection because blunders of data collection, such as misalignment of sample names with results from a sample, are easier to catch during testing than after testing has been completed. At this stage, such errors with assignable cause can be corrected with no controversy. Later, it might be difficult to determine the cause of discrepant data, and therefore difficult to correct such an error. Multiple replicates of each measurement make the determination of cause much easier, whether at this stage or later upon data visualization.

Results from all replicates should be retained and presented in a data table, along with the averages, if the average is to be used as that sample's concentration estimate. When data are manually transcribed in the table, each entry should be inspected against the original instrument value to detect possible transcription errors. The completed data table should be inspected for obvious outliers. Such outliers should be investigated for potential errors caused by instrumentation, human, or procedural errors. This is an initial review that might not detect all such errors. Later, when the data are inspected visually (see Chapter 5), this table can again be reviewed to see whether individual replicate errors are causing errors in the sample concentration estimates.

Data collected during a time when an instrument system indicates that an error condition exists should be documented but not included in the final data analysis.

Any data for which the operator can document that an error was made should be recorded, but not included, in the final analysis.

4.5 Documentation of Rejected Data

Any situation for which data is rejected (replicates or samples) should be carefully documented along with any discovered causes and problems. Documentation records should be retained.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Chapter 5: Visual Data Analysis

This chapter includes:

- Visual data review considerations
- Scatter plots and difference plots
- Inspecting the plots for underlying characteristics

5.1 Visual Data Review

Having completed the collection of data for analysis, the next step is the visual review of the data. Such a review is useful to evaluate whether the desired interval is adequately covered, to screen for the presence of exceptional values or outliers, to get an initial understanding of the difference between the measurement procedures, and to determine how best to characterize the variability of these differences across the overlapping measuring interval provided by the two measurement procedures. Two robust and flexible tools for this review are the scatter plot and the difference plot.[22]

The examples in this chapter and Chapter 6 are instructional for the purpose of describing methods for visual data review and quantitative analysis. They are not intended to imply acceptance of particular aspects of study design. Study design is discussed elsewhere (see Chapters 3 and 4) for the three types of studies considered in the guideline. In particular, the examples do not have the number of samples needed for manufacturer studies, and in many cases the number of samples at higher concentrations are smaller than would be found in an optimal design. Such examples are offered because many of the difficulties that various visualization and analysis techniques try to mitigate are also mitigated by a larger, more evenly spaced sample set. The data tables for most of the examples are provided in Appendixes C and D. They are denoted as Table C1 or D1, D2, etc.

5.2 Scatter Plots

A scatter plot presents the results of a measurement procedure comparison study, with the comparative measurement procedure on the x-axis and the candidate measurement procedure on the y-axis. These plots can be visually inspected to determine the underlying variability characteristics of this relationship. Foreknowledge of what these plots might show can be gleaned from prior precision studies using CLSI document EP05.[2]

Two measurement procedure comparisons are presented in Figures 2 and 3. Figure 2 displays a consistent amount of variability (SD) across the measuring interval (from Table D1). Figure 3 displays a variability that is proportional to concentration, which can be expressed as a constant CV, for which CV = SD / concentration (from Table D2; see Appendix D for these and other tables that contain data for the figures that follow). This constant CV is seen as a cone of data converging to a point at the lower left-hand corner and opening up to the upper right.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

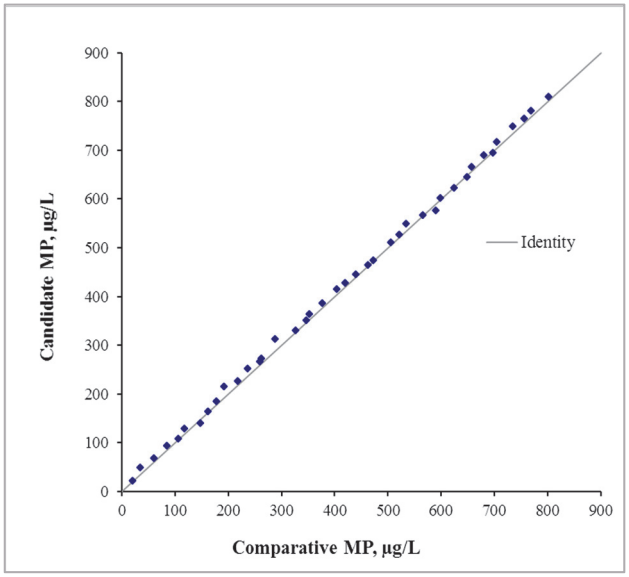


Figure 2. Scatter Plot With Constant SD (From Table D1)
Abbreviation: MP, measurement procedure.

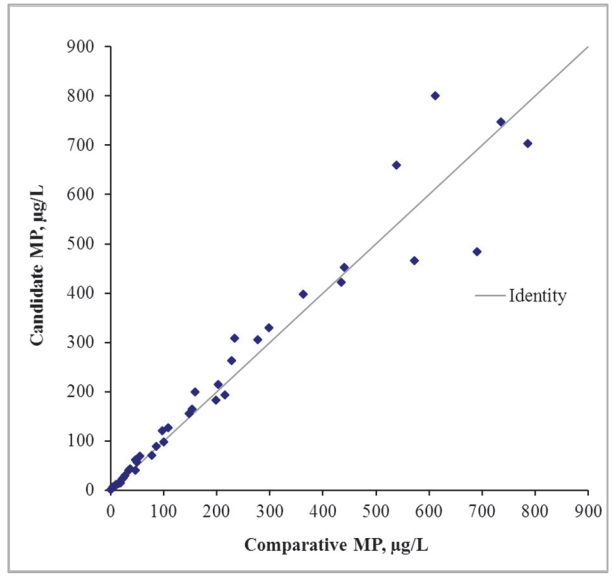


Figure 3. Scatter Plot With Constant CV (From Table D2)
Abbreviation: MP, measurement procedure.

5.3 Difference Plots

A difference plot[23] presents the results of a measurement procedure comparison study, with the measurand concentration on the horizontal axis and the difference between the candidate and the comparative measurement procedures on the vertical axis (see Figure 4). Bland-Altman[22] is an example of a difference plot. Such plots can be visually inspected to determine the underlying variability characteristics of this relationship.

The user must select from four types of difference plots based on two factors. The first factor is determined by whether the user wishes to see the comparative method as the truth against which the candidate method is compared or to see the average of the two methods as the best estimate of the true value for a sample. In the first case, the horizontal axis of the plot is the result from the comparative measurement procedure.[24] This option should be used for presenting validation data. In the second case, advocated by Bland and Altman,[25] the horizontal axis is the average of the two measurement procedures' results.

When a reference measurement procedure is the comparative measurement procedure, its results should be used on the horizontal axis. A manufacturer may wish to use the most common measurement procedure as the comparative measurement procedure. In this case, when the comparative measurement procedure is not considered a reference, the average result of the two measurement procedures (candidate and comparative) may be used on the horizontal axis for data visualization during the establishment phase of test development.

A medical laboratory may use its current measurement procedure as the comparative measurement procedure and may consider it to be a reference because the goal is to compare the known behavior of its current procedure against the unknown candidate measurement procedure. In this case, the results for the comparative measurement procedure should be used on the horizontal axis.

The second factor is whether the variability of the differences between the two measurement procedures is constant or proportional to the concentration on the horizontal axis. In the first instance, the magnitude of the difference is assumed to be essentially the same across the entire interval of concentrations (see Figure 2). In the second instance, the magnitude of the difference is assumed to be proportional to concentration (see Figure 3). Because this characteristic of the relationship might not be known beforehand, it is suggested

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

that both reporting unit and percent difference plots be created and inspected (eg, Figures 4A and 4B). The equations for creating each of the four plots are provided in Table 3.

Table 3. Formulas for Creating Difference Plots

Horizontal Axis (*z*)	Vertical Axis: Difference (*d*) Is Constant (Constant SD)	Vertical Axis: Difference (*d*) Is Proportional to Concentration (Constant CV)
Comparative measurement procedure (used for validation data)	$z_i = \text{concentration} = x_i$ $d_i = \text{difference} = y_i - x_i$ (1)	$z_i = x_i$ $d_i = \frac{(y_i - x_i)}{x_i}$ (2)
Average of the two procedures	$z_i = \frac{(x_i + y_i)}{2}$ $d_i = y_i - x_i$ (3)	$z_i = \frac{(x_i + y_i)}{2}$ $d_i = (y_i - x_i) / \frac{(x_i + y_i)}{2}$ (4)

Abbreviations: CV, coefficient of variation; SD, standard deviation.

The parameters for Table 3 are:

- x_i is the result of the comparative measurement procedure for patient sample number *i*.
- y_i is the result of the candidate measurement procedure for patient sample number *i*.
- (z_i, d_i) are the resultant coordinates on the difference plot for patient sample number *i*.

5.4 Inspect Plots for Underlying Characteristics

The optimal technique used to determine the bias between the candidate and comparative measurement procedures is highly dependent upon whether the data meet specific underlying assumptions. It should first be determined whether the variability of differences between the two measurement procedures is constant or proportional to concentration.

5.4.1 Constant Difference Variability (Constant Standard Deviation)

When the variability of the differences between the candidate and comparative measurement procedures is constant, the two plots appear as they do in Figure 4:

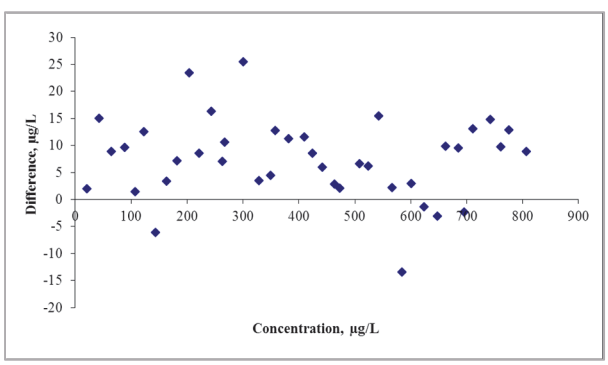


Figure 4A. Reporting Units Difference Plot

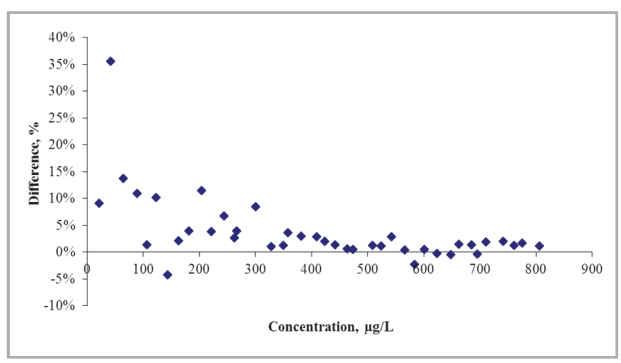


Figure 4B. Percent Difference Plot

Figure 4. Constant Difference Variability Between Measurement Procedures (From Table D1). The spread of the differences remains consistent across the range of concentration on the reporting units difference plot, but the spread of the differences increases significantly with decreasing concentration on

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

the percent difference plot. When the difference is constant across the interval of concentration, the reporting unit difference plot provides the better representation of the difference between measurement procedures.

5.4.2 Proportional Difference Variability (Constant Coefficient of Variation)

When the variability of the differences between the candidate and comparative measurement procedures is proportional to concentration, the two plots appear as they do in Figure 5:

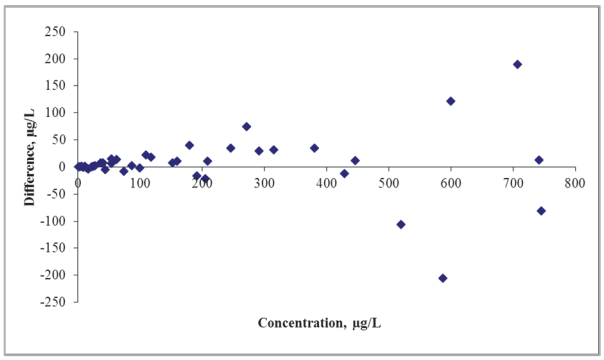


Figure 5A. Reporting Units Difference Plot

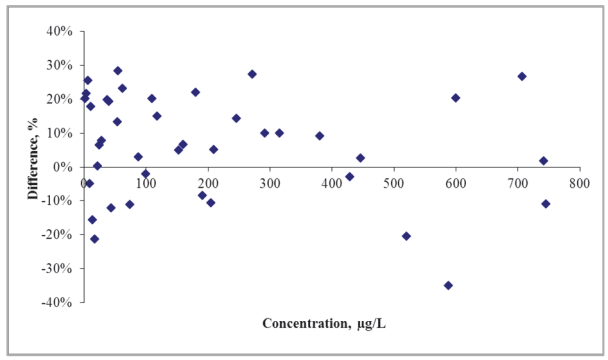


Figure 5B. Percent Difference Plot

Figure 5. Proportional Difference Variability Between Measurement Procedures (From Table D2)

In this case, the situation is reversed from Figure 4. The reporting units difference plot provides a spread of differences that gets wider as the horizontal axis concentrations increases, but the percent difference plot provides a spread that is consistent over the horizontal axis. In this case, the percent difference plot provides the better representation of the difference between measurement procedures.

5.4.3 Mixed Difference Variability (Standard Deviation and Coefficient of Variation)

Often a measurement procedure exhibits a mixture of these two characteristics with the differences being constant at low concentrations and proportional to concentration at higher concentrations. Such a measurement procedure is shown in Figure 6.

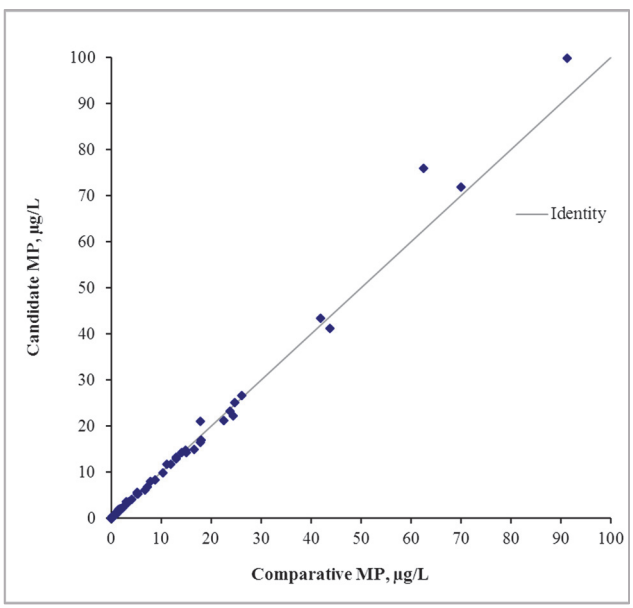


Figure 6A. Scatter Plot

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

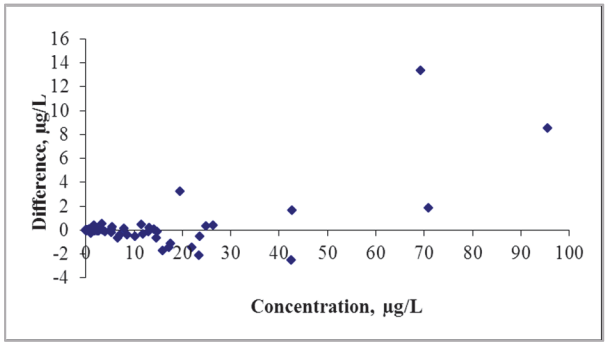


Figure 6B. Reporting Units Difference Plot

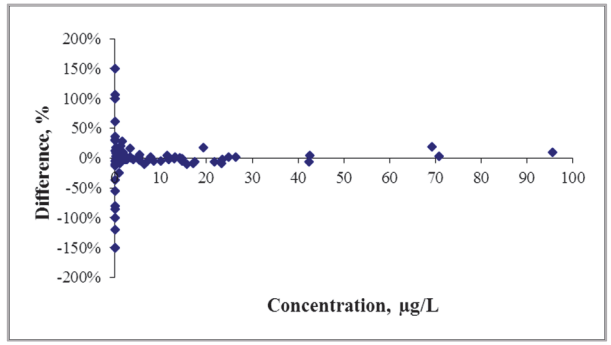


Figure 6C. Percent Difference Plot

Figure 6. Mixed Difference Variability Between Measurement Procedures (From Table C1)
Abbreviation: MP, measurement procedure.

Through inspection of the plot (or the underlying data used to generate the plot), an estimate of the concentration at which the relationship changes from a constant difference to a proportional difference can be obtained. This concentration can also be estimated by formal statistical analysis (called change point analysis), which is beyond the scope of this guideline.

The samples in this example and in Figure 7A do not evenly cover the measuring interval. More samples at higher concentrations should be collected to meet the needs of a manufacturer's study.

5.4.4 Ranked Order Difference Plot

Regardless of how diligently samples are collected, occasionally a final dataset has subintervals in which there are few data points. Most commonly these are at higher concentrations for which only patients with relatively rare disease states provide results. In such cases, usually in measurement procedures with proportional variability, the individual points are widely separated at higher concentrations (see Figure 7), making determination of variability characteristics more challenging.

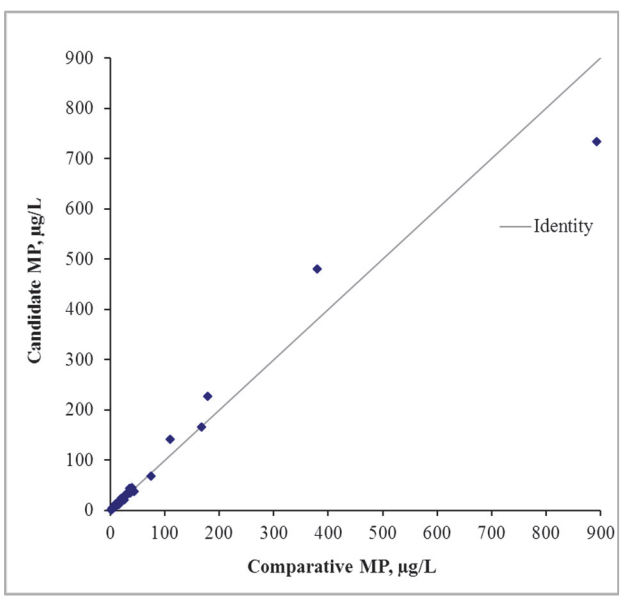


Figure 7A. Scatter Plot

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

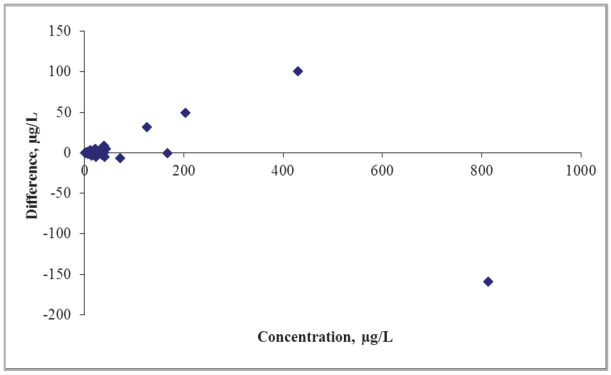


Figure 7B. Reporting Units Difference Plot

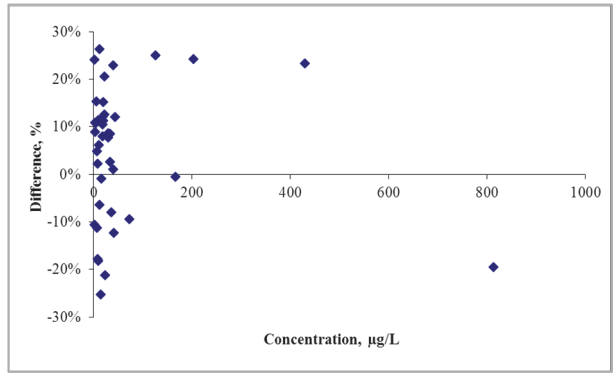


Figure 7C. Percent Difference Plot

Figure 7. Proportional Difference Variability Between Measurement Procedures (From Table D3)
Abbreviation: MP, measurement procedure.

A ranked order difference plot can help in visualizing such a dataset. The first step in creating such a plot is to rank the *N* data points by concentration from lowest to highest and assign them numbers from 1 to *N* in that order. The options for the vertical axis are the same as those shown in Table 3.

The formulas for creating ranked order difference plots are provided in Table 4.

Table 4. Formulas for Creating Ranked Order Difference Plots

Horizontal Axis (*z*)	Vertical Axis: Difference (*d*) Is Constant (Constant SD)	Vertical Axis: Difference (*d*) Is Proportional to Concentration (Constant CV)
Samples ranked by comparative measurement procedure results	$z_k = \text{Rank}\,(x_i)$ $d_k = y_k - x_k$ (5)	$z_k = \text{Rank}\,(x_i)$ $d_k = \frac{(y_k - x_k)}{x_k}$ (6)
Samples ranked by average of the two procedures	$z_k = \text{Rank}\frac{(x_i + y_i)}{2}$ $d_k = y_k - x_k$ (7)	$z_k = \text{Rank}\frac{(x_i + y_i)}{2}$ $d_k = (y_k - x_k)/(\frac{x_k + y_k}{2})$ (8)

Abbreviations: CV, coefficient of variation; SD, standard deviation.

The parameters for Table 4 are:

- *k* is the rank order of the samples (ranked by concentration and then by order of data collection in the case of ties in concentration).
- x_k is the result of the comparative measurement procedure for patient sample number *k*.
- y_k is the result of the candidate measurement procedure for patient sample number *k*.
- (z_k, d_k) are the resultant coordinates on the difference plot for patient sample number *k*.

The proportional difference plot (see Figure 7C) is shown in Figure 7D, with this optional horizontal axis.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

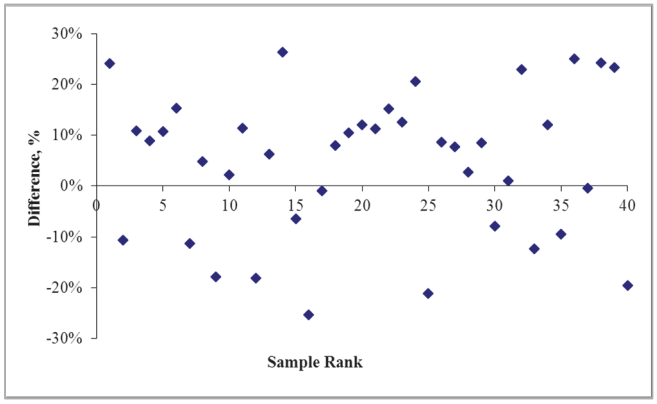


Figure 7D. Difference Plot With Ranked Order Horizontal Axis (From Table D3)

The constant proportional difference is quite apparent when viewed in this way. Such a technique is also useful for understanding the concentration at which the relationship changes from a constant difference to a proportional difference. The mixed difference example from Figure 6 is presented in Figure 8, using the ranked sample number technique.

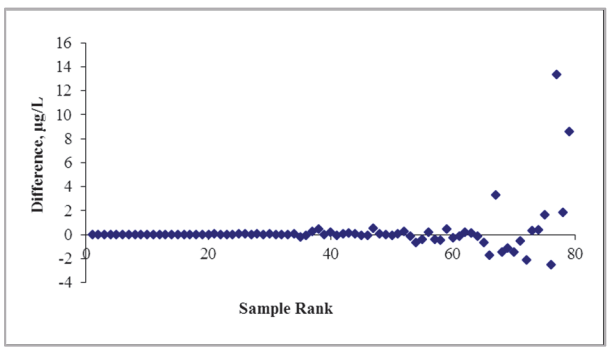


Figure 8A. Reporting Units Difference Plot

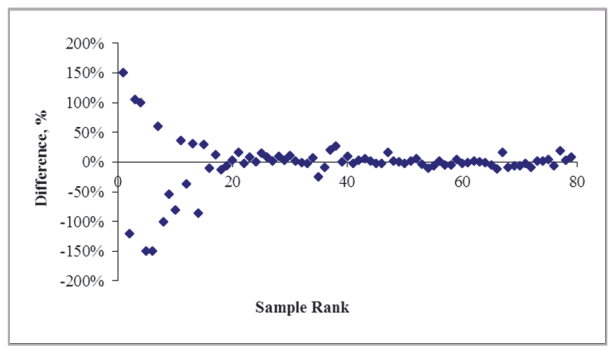


Figure 8B. Percent Difference Plot

Figure 8. Mixed Difference Variability Between Measurement Procedures (From Table C1)

The point at which the mixed difference plot changes from constant to proportional is much easier to determine with this optional horizontal axis. The goal of picking horizontal (displaying the *z*-values) and vertical (displaying the *d*-values) axes is to create a difference plot that most readily displays the relationship between the two measurement procedures so that the underlying assumptions can be visually inspected.

Inspecting Figure 8A, there appears to be constant variability from sample 1 through 35 to perhaps 55. Figure 8B displays relatively consistent proportional differences from sample 79 down to sample 40 to perhaps 30. It is important to note that portrayal of data in a ranked order plot as shown in Figure 8 enables one to better determine the region in which this change occurs than does portrayal in a more typical difference plot, as shown in Figure 6.

5.4.5 When Bias Changes With Concentration

In some cases, bias can vary across the measuring interval irrespective of the variability pattern. In Figure 9, the variability of the differences is consistent across the measuring interval, but the magnitude of the difference (bias) changes in a linear fashion.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

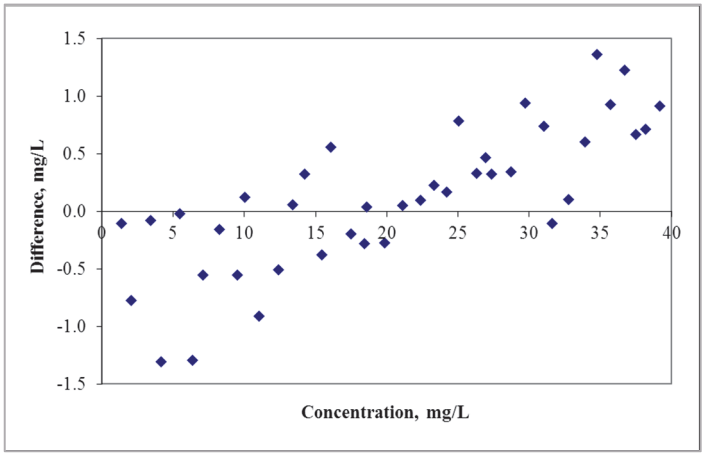


Figure 9. Bias Changes With Concentration

5.4.6 Nonlinear Relationship

Figure 10 illustrates a special case of nonconstant differences. In this dataset the variabilities of the differences are proportional to concentration; however, the magnitudes of the difference between the candidate and comparative measurement procedures change across the measuring interval in a nonlinear manner.

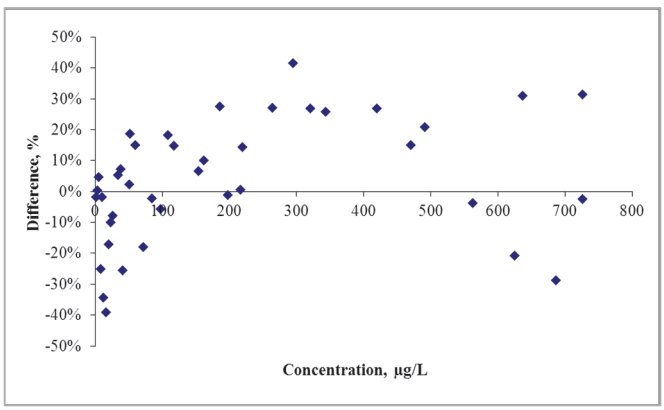


Figure 10. Nonlinear Relationship

5.4.7 Visualizing Anomalous Results

Both scatter plots and difference plots as shown in Figure 11 are useful in visualizing anomalous results.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

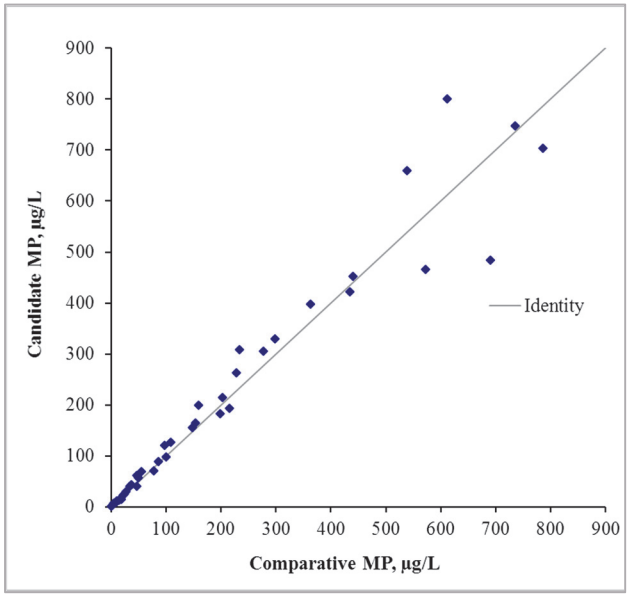


Figure 11A. Scatter Plot

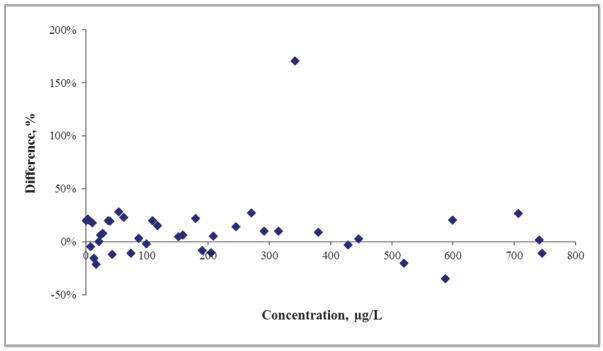


Figure 11B. Percent Difference Plot: Horizontal Axis = Average of Both Procedures

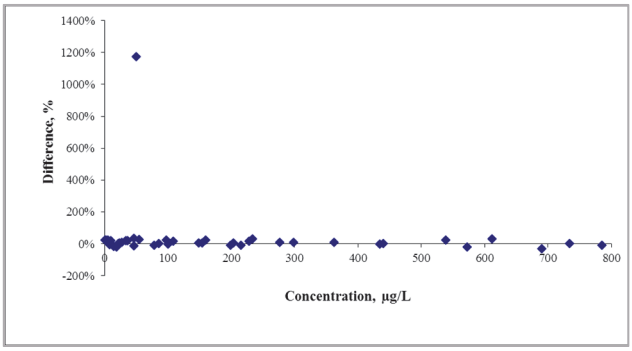


Figure 11C. Percent Difference Plot: Horizontal Axis = Comparative Procedure

Figure 11. Single, Outlying Point (From Table D4)
Abbreviation: MP, measurement procedure.

The ability of difference plots to demonstrate anomalous data using the two horizontal axis options is shown in Figure 11. Figure 11B shows each point using its distance from the identity line in the direction perpendicular to that line. This is similar to the way an outlying point is typically perceived on a scatter plot (see Figure 11A). Figure 11C also shows the distance from the identity line, but in the scatter plot's y-axis direction. For low concentration samples, such a view inflates the proportional distance of an outlying point from the other points. For visualizing anomalous results, it can be useful to view both types of difference plots.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

This page is intentionally left blank.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Chapter 6: Quantitative Analysis

This chapter includes:

- Estimating bias from difference plots
- Fitting a line to scatter plots (regression analysis)
- Determining bias and regression parameters with confidence intervals (CIs)

6.1 Estimating Bias From Difference Plots

Quantitative techniques can be applied to both difference plots and scatter plots to estimate bias.

When verifying the performance of a measurement procedure in a medical laboratory (in which $N = 40$ is sufficient), only the bias estimation from difference plots is needed. When additional analysis is needed, such as for validation studies, see Subchapter 6.2, which covers regression techniques of estimating bias. The underlying assumption of computing bias from difference plots is that some part of the relationship for either constant difference plots or proportional difference plots can have a linear constant bias, either as an absolute difference (constant SD) or as a proportional difference (constant CV). Therefore, the overall bias estimate can be used for any concentration within that interval. The use of regression analysis per Subchapter 6.2, on the other hand, provides a unique estimate of bias at any specified concentration.

6.1.1 Constant Standard Deviation

When, on review of the difference plots, the distribution of bias appears to be consistent across the measuring interval on the reporting units difference plot (see Figure 12), then an estimate of the bias between measurement procedures can be made by using the average (or median) of the individual differences between the measurement procedures. For verification purposes, this is the bias estimate for any concentration within the measuring interval.

In Figure 12, the vertical distribution is displayed to the right of the difference plot. Note that the histogram to the right shows the typical bell-shaped normal distribution.

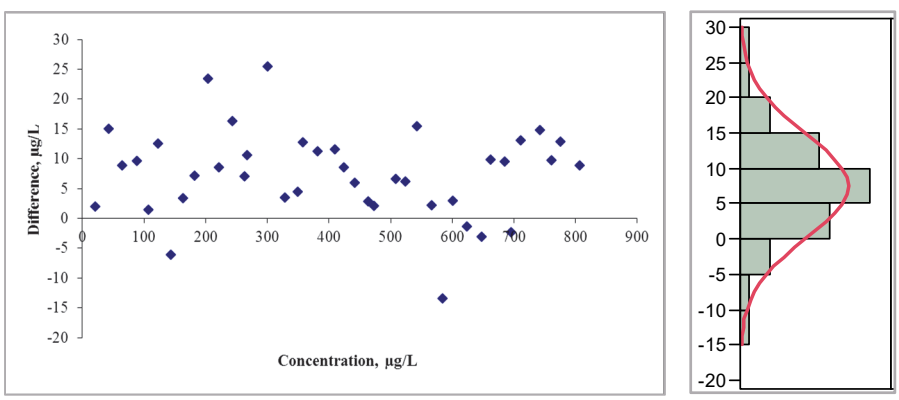


Figure 12. Reporting Units Difference Plot (From Table D1)

For a constant SD difference relationship, the differences to use are provided by the equation $d_i = y_i - x_i$. For a relationship with a nonskewed vertical distribution (as seen in Figure 12), bias is computed as the average of all such differences.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

$$\bar{d} = \sum_{i=1}^{N} d_i / N \tag{9}$$

For the distribution in Figure 12, this average result is 7.5 µg/L, which is the appropriate estimate for the entire measured interval from 20 to 800 µg/L.

Figure 13 provides a view of a constant difference relationship with an outlying point that creates a skewed vertical distribution. This outlying point is readily seen in the histogram to the right.

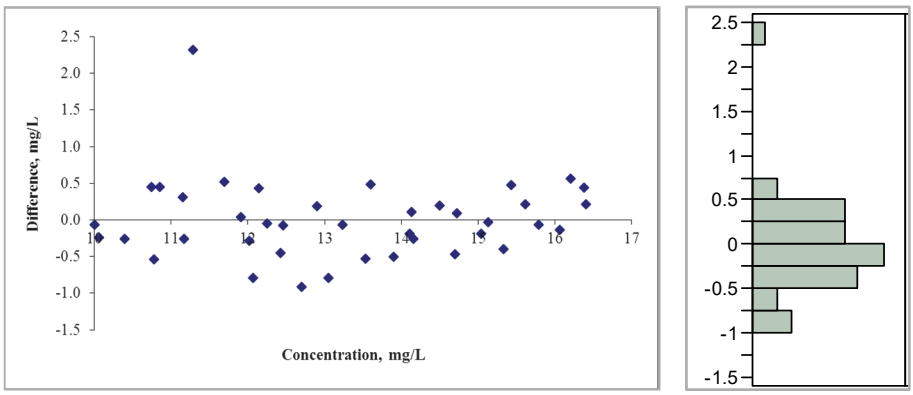


Figure 13. Reporting Units Difference Plot (From Table D5)

For a constant difference relationship with a skewed vertical distribution, compute the bias as the median of the difference values. For the distribution in Figure 13, this median result is −0.07 mg/L, which is the appropriate estimate for the entire measured interval from 10 to16 mg/L.

6.1.2 Constant Coefficient of Variation

Figure 14 includes the percent difference plot of Figure 7D, with the horizontal axis being the ranked order number of the samples. The distribution of bias appears to remain constant over all samples.

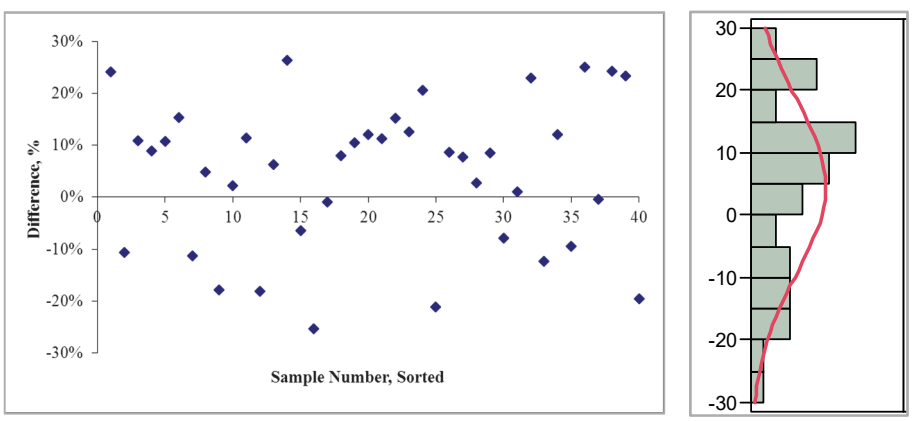


Figure 14. Percent Difference Plot (From Table D3)

Note the histogram on the right of Figure 14. Although it does not strictly follow the classical bell-shaped curve, it does not display outlying point(s) separated from the primary distribution. Therefore, for this constant CV distribution, no outlying percent differences significantly skew the results, so the average proportional difference can be used. For this distribution the average is 4.6%. In other instances when significant skewness is detected, the median can be used. Once the distribution is determined to be proportionally consistent across a range of results, then the bias is calculated only in the vertical direction,

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

and the horizontal axis is irrelevant. This means that both the average bias across the measuring interval and the bias at any specified concentration is estimated by this same calculated bias.

For the constant SD calculations above, the procedures for calculating both the horizontal and the vertical axis results are irrelevant to the calculation of bias. For constant CV calculations, however, the vertical difference axis is a ratio of difference to concentration. In this case, the user must decide between equations (2) and (4) in Table 3; in other words, dividing by the comparative measurement procedure or by the average of the candidate and comparative measurement procedures. The same difference calculations for *d* (vertical axis) are presented as equations (6) and (8) in Table 4, as seen in Figure 14.

In Figure 15, the constant CV data from Figure 11C are presented showing the difference divided by the comparative measurement procedure.

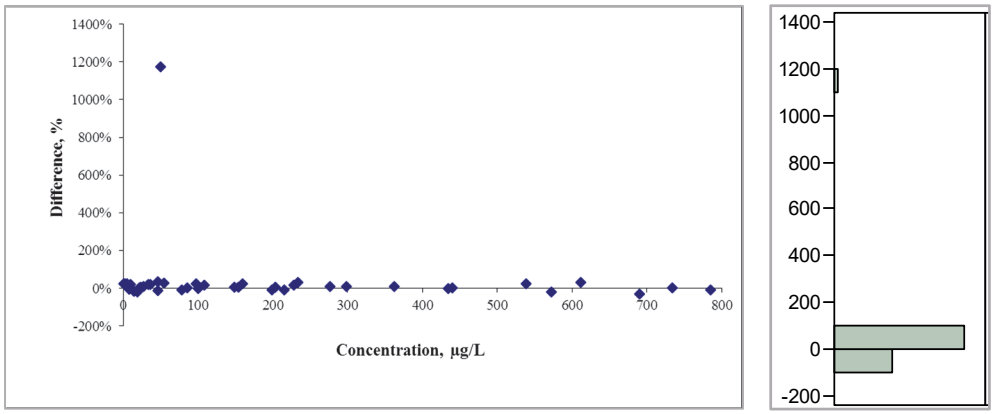


Figure 15. Percent Difference Plot (From Table D4)

In this case, the median result is 7.5% and the average result is 36.5%. Clearly, the median result is the best estimate of central tendency in this case, but the difference plot is too compressed to tell the user whether this estimate is usable over the interval of the measurements. Appendix E presents a method to determine whether a result can statistically be declared an outlier. When an outlier is identified (see Figure 15) the data can additionally be presented without this outlier, as seen in Figure 16.

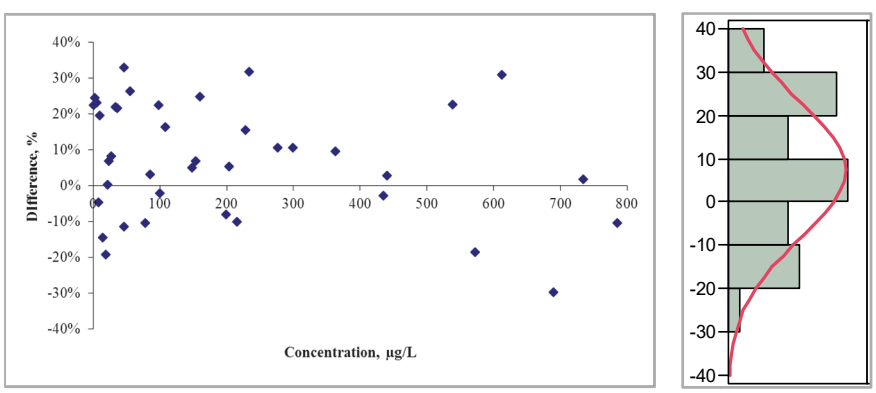


Figure 16. Percent Difference Plot Without Outlier (From Table D4)

Figure 16 shows consistent proportional differences across the measured interval of this study. Therefore, for verification purposes, a single bias estimate can reasonably be used to represent the bias within this interval from approximately 0 to 800 µg/L. This bias estimate is the median of all sample results, including the outlier, which was stated above as 7.5%. (Note that in this case, the average result after excluding the outlier is also 7.5%.)

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

6.1.3 Mixed Variability (Standard Deviation and Coefficient of Variation)

When a mixed variability model is observed, the sample rank, *k,* that best separates the constant SD portion of the dataset from the constant CV portion is determined (see Figure 8). For the low concentration portion, the estimate of bias as a constant difference relationship over the samples having rank 1 to *k* is computed. For the high concentration portion, the estimate of percent bias as a proportional difference relationship over the samples having rank $k+1$ to N is computed.

At least 20 samples in each subgroup are needed for a reasonably accurate estimate of bias for both intervals. After computing the bias (or percent bias) of each subgroup, this estimate plus the applicable concentration interval of each sample subgroup is reported. The data from Figures 6 and 8 were analyzed in this way. The lowest 40 points (from 0 to 1.8 µg/L) had an average offset of 0.20 µg/L. The highest 39 points (from 1.8 to 96 µg/L) had an average proportional offset of 0.43%. See Appendix C to see how these results were computed.

When the precision profiles of measurement procedures have been characterized, the expected behavior of a comparison study, including, for example, a constant SD at low concentrations and a constant CV at high concentrations, can be provided as acceptance criteria. An example of this type of analysis is included in Appendix H.

6.1.4 When Bias Changes With Concentration

When the bias changes over the measuring interval in a linear fashion, the dataset is inappropriate for the bias estimation techniques described in this subchapter. In these instances the user should perform a regression analysis as described in Subchapter 6.2.

To determine if such a change over the measuring interval is significant, a regression can be performed on the difference plot. When the contribution of slope is significantly different from zero with an approximate 95% level of confidence, a nonconstant difference is present, as seen in Figure 17. An ordinary linear regression (OLR) was performed to determine the line fit in Figure 17 because, in this case, the variability around the line is relatively uniform across the concentration interval of the data. When this assumption is not met, other regression techniques[26] (beyond the scope of this guideline) can be more appropriate.

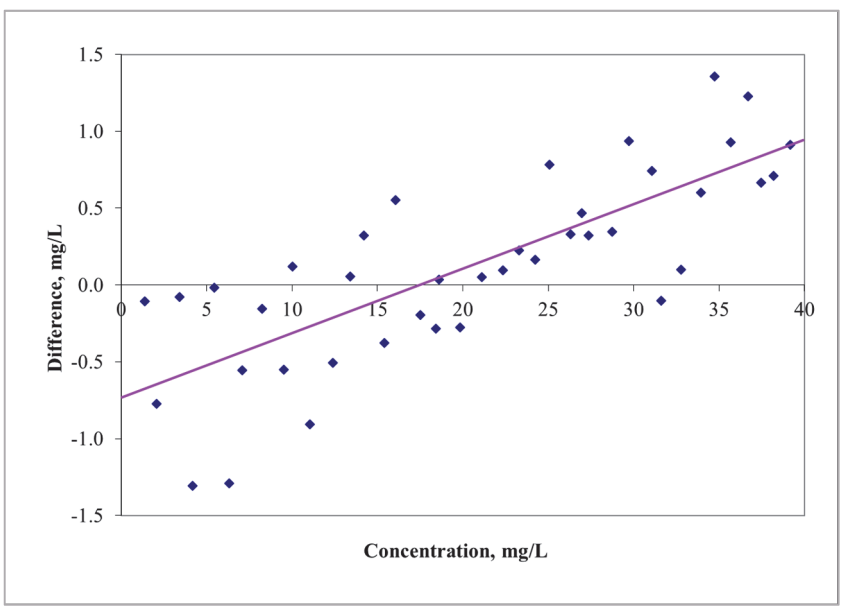


Figure 17. Regression Fit of Bias vs Concentration (Same Data as Figure 9)

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

6.1.5 Nonlinear Relationship

When the bias displays a nonlinear relationship with concentration, as seen in Figure 10, neither difference plots nor regressions over the entire measuring interval are appropriate. One option for such data is an analysis that describes fitting a line (perhaps nonlinear) to the points provided in a difference plot.[27] Such analysis is beyond the scope of this guideline.

When the goal of the verification study is to provide a bias estimate for a specific medical decision concentration x_c, the points around that concentration can be used to provide such an estimate. A difference plot of a subset of points around the medical decision concentration can be constructed. At least 20 such points should be selected by either selecting the nearest 10 points above and 10 points below the concentration or selecting an interval of concentrations around the medical decision concentration. The above selection of points should be performed based on a list of results ranked on the average of the candidate and comparative measurement procedures (ie, equations 7 or 8 in Table 4). Selecting points for computing such an estimate based on a list ranked on a single measurement procedure would improperly bias the results.[28]

6.1.6 Vertically Skewed Distribution (Aberrant Results)

When one sample or a small number of samples is causing the perceived skewness in the y-axis direction (see Figure 11), the techniques outlined in Appendix E are used to determine whether they are true outliers. If so, the possible cause of the outlying result should be investigated. When the point(s) cannot be eliminated for cause, the vertical distribution remains nonsymmetrical and the bias should be estimated using the median difference or median percent difference. When the decision is made to remove statistically identified outliers from the dataset, analyses should be performed both with and without the outliers.

6.1.7 Confidence Interval of the Bias Estimate

Once the average (or median) bias and its CI have been determined, they can also be used to evaluate bias at appropriate medical decision concentrations based on the acceptance criteria. Note that bias estimates and their CIs of individual concentrations using difference plots can be used during development and during laboratory verification, but not for validation purposes.

6.1.7.1 Bias Estimate

With a symmetrical distribution of differences (either SD or CV), the average is used as the bias estimate. The computation of the standard error (SE) is needed to determine the CI for this estimate:

$$\mathrm{SE}(\bar{d}) = \sqrt{\frac{\sum_{i=1}^{N}(d_i - \bar{d})^2}{N(N-1)}} = \frac{SD}{\sqrt{N}}, \tag{10}$$

in which d_i is the difference between the candidate and comparative measurement procedures for each sample i and $\bar{d}$ is the average of all such differences (see equation 9).

Assuming the differences follow a normal (gaussian) distribution, the CI is computed by multiplying the SE by the factor derived from the confidence desired (typically 95%) and the sample size by using Student t-distribution and adding and subtracting the result from the average estimate. For a sample size N of 20 ($N-1=19$ degrees of freedom), this 95% factor is 2.093, for $N=40$ it is 2.023, and for $N=100$ it is 1.984. In the first case, the 95% CI would be from (average −2.093SE) to (average +2.093SE).

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

6.1.7.2 Median Bias Estimate

Although the CI for an average estimate of bias can be computed by an equation, computing the CI around the estimate of median bias calls for a nonparametric method. The interval it produces is close to but might not be the standard 95%, because it depends upon the number of points used in the estimate. As the number of points decreases, the CI becomes less precise. See Appendix F for examples of computing the CI of a median bias for 100 points.

6.2 Fitting a Line to Scatter Plots (Regression Analysis)

Regression, as an analysis technique, is applicable in a wider range of situations than difference plots. For some, it also provides a more intuitive comparison of each sample point between the two measurement procedures. As with difference plots, regression analysis techniques have underlying assumptions that must be met. Difference plots provide many of the answers with respect to such assumptions. Other assumptions related to imprecision can be checked against the results of CLSI document EP05[2] precision studies. Therefore, good first steps in deciding which regression technique to use is reviewing prior precision studies and creating difference plots, as described above, to characterize the distribution of differences.

As mentioned above, studies conducted to introduce a candidate measurement procedure by medical laboratories might not need analyses beyond difference plot analysis. However, when some difference plot assumptions are not met, the best choice is to continue on to regression analysis. For manufacturers conducting establishment or validation studies, regression analysis is necessary.

The initial goal of regression analysis in a measurement procedure comparison study is to fit a straight line through the data presented, generating a scatter plot with the comparative measurement procedure on the x-axis and the candidate measurement procedure on the y-axis. The default assumption is trying to demonstrate exact concordance (identity) between both measurement procedures. In other words, when the comparative measurement procedure provides a result of 1 for a sample (x_1), the candidate measurement procedure also provides a result of 1 (y_1), and when the comparative measurement procedure provides a result of 100 for another sample (x_2), the candidate measurement procedure also provides a result of 100 (y_2).

It is easy to draw a line, defined as $y=a+bx$, through two such perfect results, in which x is the comparative measurement procedure result, y is the candidate measurement procedure result, a is the intercept of the line to the y-axis, and b is the slope of the line. In this case, the slope would be:

$$(y_1 - y_2) \,/\, (x_1 - x_2) = (1 - 100)/(1 - 100) = 1.0 \tag{11}$$

The intercept would then be:

$$\mathrm{a} = y_2 - bx_2 = 100 - 1.0 \bullet 100 = 0.0 \tag{12}$$

Therefore, the equation for a measurement procedure comparison between two perfectly concordant measurement procedures has an intercept of zero and a slope of one.

In some cases, it is known beforehand that there will not be such a perfect relationship. One example is a candidate measurement procedure that is trying to more closely match an international standard. The comparative measurement procedure in this example was introduced before the standard was developed. It is thought that the known positive proportional difference is 20%. In this instance, the expected result of a measurement procedure comparison study is an intercept of zero and a slope of 1.2.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

6.2.1 Constant Standard Deviation

Two measurement procedure comparisons are presented below. The first (see Figure 18A), using the data from Figure 2, confirms that there is a consistent amount of variability (SD) across the measuring interval. The amount of scatter around this relationship is small, especially when the entire range of measurements is considered. The same scatter (SD), and relationship of scatter to concentration, is presented in Figure 18B, but the interval over which measurements are made is far more restricted.

The correlation coefficient (r) is often squared to provide the coefficient of determination (r^2). This is the fraction of variance in y explained by a least squares regression line fitted through the data. The theory and calculations for computing r and for fitting a least squares regression line through such data is presented in Appendix G. The computed r^2 for these two examples is 0.999 for Figure 18A and 0.961 for Figure 18B.

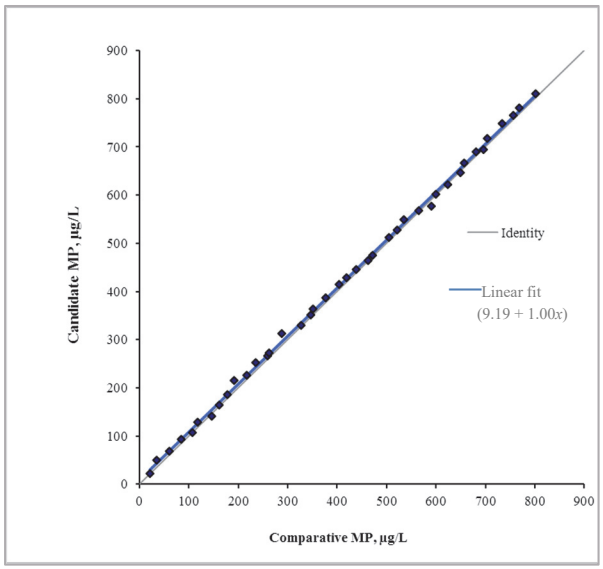


Figure 18A. Ordinary Linear Regression Fit to Highly Correlated Data

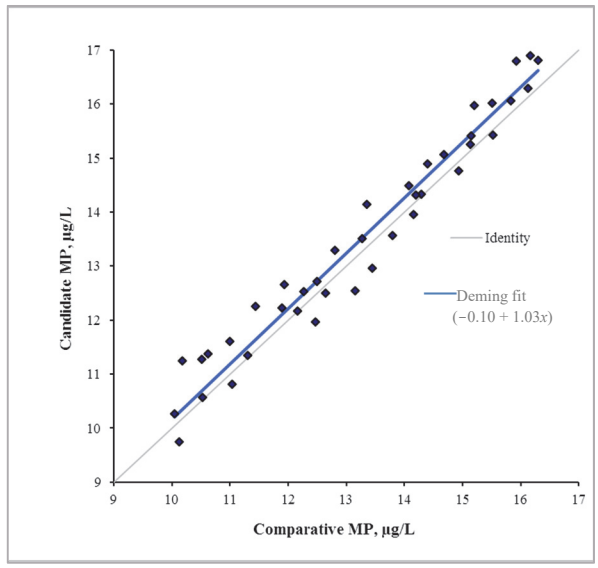


Figure 18B. Deming Fit to Less Correlated Data

Figure 18. Regression Fit to Data With Constant SD (Figure 18A Data From Table D1; Figure 18B Data From Table D6)
Abbreviation: MP, measurement procedure.

In relationships such as those in Figure 18 in which there is a constant SD, an even distribution of points over the measured interval, and an r^2 greater than 0.95, OLR (see Appendix G) provides results consistent with other techniques. However, if any of these assumptions are broken, then other techniques, mentioned below, should be used.

In cases of constant difference variability (constant SD), the estimate of r^2 should be used only as an indicator of the strength of a regression fit and as a rule-of-thumb determination of whether an OLR might be acceptable. It should not be used as a criterion to accept or reject the candidate measurement procedure. For multiple replicate situations in which replicate averages are used for each measurement, the correlation coefficient, although an accurate representation of the amount of between-sample variability explained by the regression fit, overstates the amount of total variability (within and between sample) explained by a regression fit to the individual replicate results. Finally, the correlation coefficient should not be used as an indication of linearity. Such an evaluation can be performed by following the procedures in CLSI document EP06.[29]

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

The argument for using OLR for instances similar to those in Figure 18 is basically that it is adequate. A better case can be made for using a constant SD Deming[10] regression for such instances. OLR attempts to minimize the differences between the points and the fitted line as measured in the vertical (*y*) direction. This technique assumes that only the candidate measurement procedure has inherent imprecision. This is never true; even a comparative measurement procedure comprised of samples made gravimetrically from standard material still has imprecision associated with mass determinations. Deming regression as shown in Figure 18B (see Appendix A) allows the imprecision of both measurement procedures to be taken into account. When the imprecision estimates of both measurement procedures are known, such as from precision studies conducted using CLSI document EP05,[2] this knowledge can be used to determine the fit.[30]

The consistency of a regression fit using a Deming regression is easily shown by switching the x-axis and the y-axis and redoing the regression. Using Deming regression, the relationship in most cases is more consistent. Using OLR, the two results are usually inconsistent.

In conclusion, for cases in which there is constant difference variability (constant SD) across the measuring interval of the two measurement procedures, constant SD Deming is recommended as the default regression technique.

6.2.2 Constant Coefficient of Variation

As seen in Chapter 5, measurement procedure comparisons with constant difference variability should not use the same difference plot techniques as those with proportional difference variability. This is also true for regression techniques.

Datasets exhibiting proportional difference variability do not meet the underlying assumptions for either OLR or constant SD Deming regression. Instead of a constant SD, such datasets exhibit a constant CV. Much like Figure 3, the scatter plot displays the points opening up like a trumpet with the distribution narrow on the lower left and wider on the upper right (see also Figure 19).

Weighted least squares (WLS) regression can take such a distribution into account. The specific weighting known as constant CV least squares regression gives each point a weight inversely proportional to the square of the concentration on the x-axis. Thus, points farther to the right on the scatter plot have less influence on the regression line fit than do points on the left because the points to the right are expected to be more scattered. A discussion of this regression technique is provided in Appendix H.

WLS regression has some of the same faults as OLR. First, the assumption of zero imprecision for the comparative measurement procedure is false. Second, the results are even less likely to be consistent if the axes are switched.

The constant CV Deming regression can solve both of these issues. The specific weighting in constant CV Deming regression works in a similar fashion as WLS. Points to the upper right have less influence on the regression fit than do points to the lower left. As in the previous discussion of Deming regression, the assumption is that both measurement procedures have some inherent variability, in this case expressed as CV. A regression line fit to the data in Figure 3 using constant CV Deming is shown in Figure 19. A discussion of constant CV Deming regression is provided in Appendix B.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

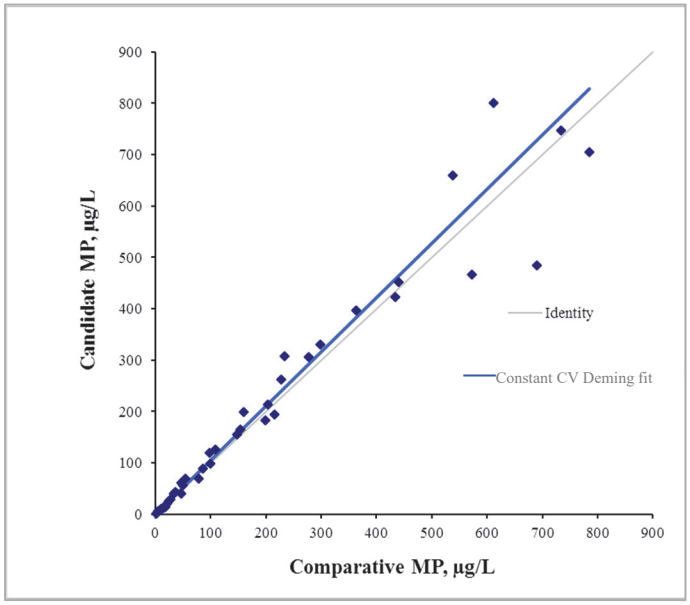


Figure 19. Constant CV Deming Regression Fit to Data With Constant CV (From Table D2)
Abbreviations: CV, coefficient of variation; MP, measurement procedure.

Neither WLS nor constant CV Deming regressions are possible when points with a concentration of zero for either of the two measurement procedures are compared, because weighting of points based on the calculation of CV = SD / concentration is not feasible. However, because the measuring interval is defined as being above the limit of detection or limit of quantitation (see CLSI document EP17[31]) and within the interval of linearity, zero concentration points should not be part of a measurement procedure comparison dataset. In cases of data exhibiting constant CV over the measuring interval, the recommended default regression technique should be constant CV Deming regression.

All examples in this guideline assume a single estimate of concentration for each measurement procedure for each sample. When this estimate is the average of multiple replicates, imprecision data are lost that could be used to determine the relative imprecision of the two measurement procedures. The resultant imprecision ratio between the two measurement procedures can be automatically computed by some Deming regression software packages. This ratio, often referred to as lambda, is an input into any Deming regression. When multiple replicates or such software are not available, previously generated imprecision results from studies based on CLSI document EP05[2] can be used to estimate this ratio. When such studies are not available or similar measurement procedures are being compared, the best default estimate for this ratio is 1.0. Without any knowledge of the ratio, for some purposes, the ratio can be varied to assess the sensitivity of the Deming regression to its value. These considerations hold true for both constant SD Deming regression (see Appendix A) and weighted (constant CV) Deming regression (see Appendix B).

Some users advocate the transformation of data before plotting constant CV datasets. Both logarithmic and power functions have been used for such transformations. The advantage of such techniques is that the scatter around the line fit can be made to resemble an evenly spaced, constant SD dataset. Such views of the data can help determine whether the spread of points at high concentrations is due solely to a constant CV relationship or to potential outlying points. After transformation, an OLR or constant SD Deming regression fit can usually be performed. The resultant line equation can be used to estimate bias at any concentration after the reverse transformation. However, the slope of the transformed data line fit cannot be used for comparison to bias acceptance criteria, because such criteria are stated in reference to untransformed results.

Finally, for constant CV datasets, many advocate the use of Passing-Bablok regression (see Subchapter 6.2.3). This regression method, as with any nonparametric technique, needs a higher sample size than a

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

parametric technique to reproducibly provide the same results. However, given the suggested sample size for manufacturers (100) and even medical laboratories (40) it is a viable technique for constant CV datasets.

6.2.3 Mixed Variability (Standard Deviation and Coefficient of Variation)

It is assumed that measurement procedure comparison data have already been explored using difference plots before the initiation of regression analysis. The data in Figure 6B showed, at low concentrations, that the variability of the differences was constant. In Figure 6C the same data showed, at higher concentrations, that there was proportional difference. In such a case, the data display neither a constant SD nor a constant CV over the entire measuring interval. When the data show no significant offset at low concentrations using difference plots (ie, distribution of differences overlap the zero bias line), the default recommendation of a constant CV Deming regression is a reasonable option. In the case shown in Figure 6, however, the influence of the highest concentration points causes an inflated estimate of proportional difference (see Figure 20A). The influence of each such point would have been less pronounced had more points been collected at higher concentrations, and might have reduced the need to look beyond the constant CV Deming regression technique.

Passing-Bablok regression is a nonparametric technique that, when fitting a line through the data, makes no assumptions about the distribution of the data points. It essentially draws a median line through the data (ie, there is always close to an equal number of points on either side of a Passing-Bablok fit). Because it makes no distribution assumptions, it is an appropriate technique to use for datasets, such as those from Figure 6 that break the assumptions made in other techniques. A description of Passing-Bablok regression can be found in Appendix I. The result of a Passing-Bablok regression on the data from Figure 6 is shown in Figure 20B. Again, if this had been a manufacturer's dataset using 100 or more samples, additional samples at higher concentrations should have been collected to more evenly cover the measuring interval.

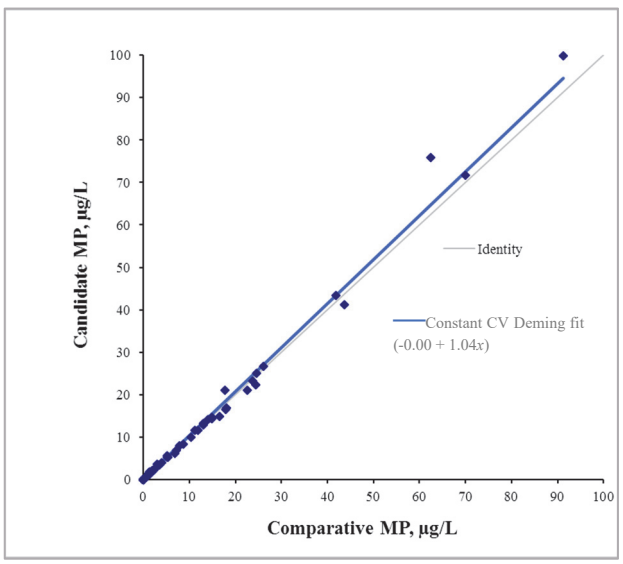


Figure 20A. Constant CV Deming Fit

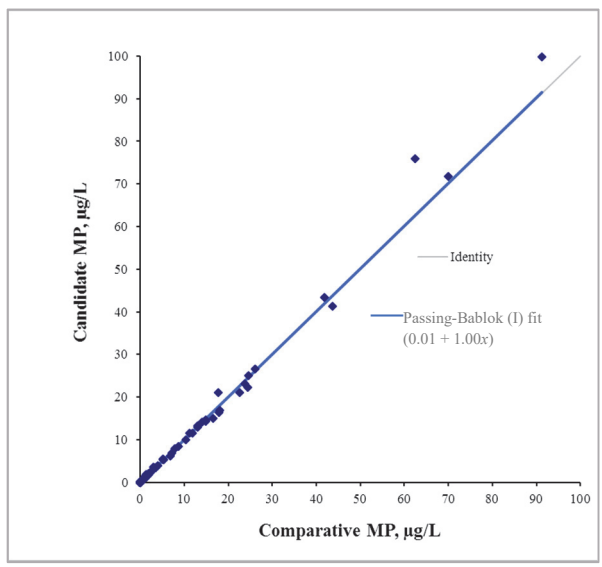


Figure 20B. Passing-Bablok Fit

Figure 20. Regression Fits With Mixed Variability Data (From Table C1)
Abbreviation: MP, measurement procedure.

In the earlier difference plot analysis described in Subchapter 6.1.3, the data previously seen in Figure 6 had an average offset at low concentrations of about 0.20 µg/L and an average proportional offset at high concentrations of 0.43%. The Deming regression fit from Figure 20A gives an intercept of 0.00 µg/L and a higher proportional offset of 4.0%. The Passing-Bablok regression fit from Figure 20B more closely matches the difference plot estimates with an intercept of 0.01 µg/L and a proportional offset of 0.3%. See Appendix C for a complete set of results for these data.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

As illustrated in this example, constant CV Deming regression is more resistant to the influence of a few high concentration samples than unweighted regression techniques, but it does not totally eliminate their influence. In such cases, if the relationship's linearity can be assumed, Passing-Bablok regression can provide a more robust alternative.

In theory, when the precision profiles of both measurement procedures are known or can be approximated from available data, a better alternative would be a version of Deming regression making direct use of this measurement procedure- and concentration-dependent precision information.[32] See Appendix B for a discussion of this approach.

6.2.4 Aberrant Results

Aberrant results create a distribution that is nonsymmetrical, and possibly skewed. Such distributions do not meet the underlying assumptions of either OLR or Deming regression. Therefore, in such situations, a Passing-Bablok regression should be performed.

6.2.5 Nonlinear Distribution

A nonlinear distribution does not meet the underlying assumptions of any of the regression techniques because they all assume linearity and that the two measurement procedures measure the same quantity. In instances in which these assumptions are not met, the difference plot techniques described in Subchapter 6.1.5 should be used.

6.3 Bias and Regression Parameters With Confidence Intervals

All regression analyses mentioned above provide an estimate of the relationship between the candidate and comparative measurement procedures by fitting a line through the data with the equation $y=a+bx$ in which a is the intercept and b is the slope.

The regression equation provided by any of these methods can be used to estimate the bias between the candidate measurement procedure (Y) and the comparative measurement procedure (X) along the vertical axis (ie, the y-axis) at any value within the interval of measured comparative values.

When an OLR or weighted OLR is performed, the CI (typically 95%) of this estimate can be computed directly. See Appendixes G and H for these computation descriptions. The equation for a CI of a bias estimate from an unweighted Deming regression is provided in Appendix A.

For all other regression techniques, the CI of any bias estimate cannot be directly computed through an equation. For these techniques, combining the CI of the slope and the CI of the intercept does not directly compute the CI of the bias estimate in the vertical (y-axis) direction at a specified comparative value. Instead, an iterative technique can be used to create a set of data from the N regression points. For each such set created, a regression line can be fit and the bias estimate can be made in the vertical (y-axis) direction. Using these determinations, the SE of the bias estimate can be computed. Similarly, the SE of the slope and intercept can also be determined. These SE estimates can, in turn, be used to determine CI estimates.

One common method for performing this iteration is the jackknife technique, in which each sample is withdrawn in turn from the dataset to create N sample sets of $N-1$ samples each. This technique, which can be applied to unweighted or weighted Deming regression, is described in detail in Appendix K1. A second method is the bootstrap technique in which any number of sample sets are created by sampling randomly with replacement from the original set of samples.[33] The bootstrap technique must be used for Passing-Bablok regression. This technique, which can be used with any regression model, is described in detail in Appendix K2.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

This page is intentionally left blank.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Chapter 7: Comparisons Within a Measurement Procedure

This chapter includes:

- Comparing two conditions within an already validated measurement procedure
- Sample type comparisons
- Other comparisons

7.1 Comparing Two Conditions Within an Already Validated Measurement Procedure

Manufacturers or laboratories might wish to compare two conditions within an already validated or released measurement procedure. The same analysis methods mentioned earlier in this guideline are applicable. Such studies can be performed to estimate bias across sample tube types, raw material lots, reagent lots, calibrator lots, or other factors. The scope of the experimental and data-handling procedures for this purpose is smaller than that for claims establishment or validation studies.

For these types of studies, the assumption is that measurement procedure performance has been established and validated, including bias, imprecision, and linearity across its measuring interval. There is no need to ensure that the full measuring interval is covered by such a study; the only need is to provide a reasonable interval of concentration measurements covering the medical decision points and both diseased and nondiseased areas of the measuring interval. Because of these considerations, a sample size of 40 is adequate.

In such studies the data are obtained using the same measurement procedure under two conditions; consequently, the comparison and candidate data have very similar performance characteristics except for the condition being examined.

7.2 Sample Type Comparisons

The relationship between measurements from different tube types or sample types, as mentioned in CLSI document I/LA21,[34] is often of interest to manufacturers. Because only a single factor of a measurement procedure is to be characterized, there is no need to sample other factors such as instrument, day, or calibration. Such studies can be conducted on a single instrument, on a single reagent lot, on a single day.

For such studies, aliquots of both sample types collected from a patient can typically be run on the same instrument, within the same timeframe, on the same lot of reagent material, and on the same calibration. When multiple replicates are run for each sample type, the average of the replicates should be used as the individual measurement for each aliquot.

7.3 Other Comparisons

Other study examples include comparisons typically run by a medical laboratory, such as between lots of reagent material or between instruments of the same or a different manufacturer (ie, production and backup instrument). Unmodified patient samples are easier to collect for such studies than for sample type studies, because only a single sample is needed per patient. A reagent lot comparison should be run on the same instrument with both lots run within a short timeframe. An instrument comparison should also be done within a short timeframe, keeping all factors as consistent as possible. Other bias estimation techniques with smaller sample sizes may be used in such instances such as those specified in CLSI document EP31.[35]

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.
© Clinical and Laboratory Standards Institute. All rights reserved.

This page is intentionally left blank.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Chapter 8: Interpreting Results and Comparing to Performance Criteria

This chapter includes:

- Interpreting the study results
- Manufacturer's statement of bias performance claims
- Medical laboratory's statement of bias performance

8.1 Interpreting the Study Results

The difference between a comparative measurement procedure and a candidate replacement measurement procedure is of interest over their entire common measuring interval (usually expressed as slope) or at one or more medically significant concentrations. The CI of the measured bias (average bias over the measuring interval or bias at a specified concentration; see Subchapter 6.3) is compared with the defined acceptable bias limits. Each manufacturer or medical laboratory should develop its own criteria (in consultation with medical staff or the technical literature). These criteria should be predefined before the measurement procedure comparison study, especially when the study will be evaluated by a regulatory authority. These criteria should include the decision rules regarding the possible outcomes of the bias estimate and its 95% CI, as shown in Figure 21. In this figure, each solid dot is an example bias estimate and the vertical line is its 95% CI.

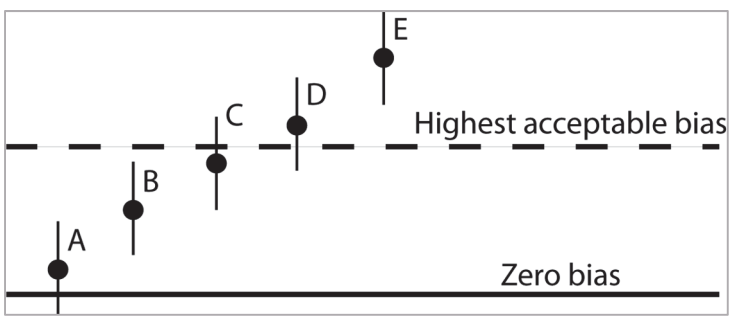


Figure 21. Possible Bias Estimate Outcomes

Outcome A is an instance in which the 95% CI of the measured bias includes zero. The outcome statement can therefore be that no significant bias was seen between the candidate and comparative measurement procedures.

In outcome B, the predefined acceptable bias limit includes the 95% CI for estimated bias. Therefore, the bias of the candidate measurement procedure to the comparative measurement procedure meets the bias acceptance criteria with 95% confidence and is acceptable for the defined application. Please note that this statement is also valid for outcome A.

When the estimated bias is within the predefined acceptable bias limit but the CI is not (outcome C), it cannot be said that the bias acceptance criteria was met with 95% confidence. Because the estimated bias is less than the highest acceptable bias limit, it can be concluded that the bias is acceptable. However, an alternate conclusion that an inappropriate percent of sample results will have bias outside the limit is also possible.

When the estimated bias is outside of the predefined acceptable bias limit but the CI is not (outcome D), the data do not show a bias outside of the bias acceptance criteria with 95% confidence. Because the CI includes the highest acceptable bias limit, it still can be concluded that the bias is acceptable. However, there is less confidence in this conclusion than for outcome C.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

When the estimated bias and its CI are outside of the predefined acceptable bias limit (outcome E), the performance of the candidate measurement procedure is not acceptable for the defined application.

Instead of concluding a candidate measurement procedure is not acceptable, the above set of outcomes could instead drive a medical laboratory to adjust reference intervals using the results of the comparison study per CLSI document EP28.[16] This is not the case for widely accepted medical decision concentrations. Such cutoff points are established through extensive medical studies or medical experience and therefore typically cannot be changed.

When a manufacturer has provided comparison data for the candidate measurement procedure, the stated bias can replace the acceptance criteria in the analyses above. When the manufacturer's claim for bias is included in the 95% CI (outcomes C or D), the medical laboratory can conclude that the candidate measurement procedure meets the bias claims of the manufacturer.

8.2 Manufacturer's Statement of Bias Performance Claims

It is expected that manufacturers provide claims based on regression analysis in which *X* is the comparative measurement procedure and *Y* is the candidate measurement procedure. The following items should be included in a manufacturer's claim for measurement procedure comparison bias:

- The total number of samples (points) used in the measurement procedure comparison. Each sample provides only one point to the comparison.
  - When samples are excluded from the analysis, the number of such samples must be stated along with the reason for their exclusion.

- The interval of collected data (the highest and lowest value of *x* included in the regression).

- The comparative method, and its calibration traceability if known, used in the measurement procedure comparison.

- Whether individual determinations were used in the comparison or averages of replicate determinations and, if averages were used, how many repetitions within each average. This should be noted for both *X* and *Y*.

- The number of days, instruments, reagent lots, calibrator lots, and calibration cycles used to collect the data on *Y*.

- The slope and intercept of the fitted linear regression line (by any method), along with their CIs.

- The bias calculated from the regression line at stated medical decision points (either at generally recognized decision points or at the extremes of the reference interval) along with the CI of each bias estimate.

- A scatter plot of the observed data, using identical scales and intervals for the *x* and *y* axes, with *all* data indicated. The scatter plot should include the fitted regression line with its 95% CI, when possible, and the line of identity ($y=x$). For display purposes, manufacturers might need to provide an additional scatter plot of one candidate replicate (x) versus one comparative replicate (Y) if multiple replicates were used.

- The method used to fit the linear regression line (eg, OLR, weighted regression, Deming, Passing-Bablok).

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

- In cases in which least squares regression is used, the following parameters should be provided:
  - The SD of residuals from regression s_{yx} (defined in Appendix G).
  - The correlation coefficient (r) or determination coefficient (r^2).

8.3 Medical Laboratory's Statement of Bias Performance

The laboratory might wish to provide a statement of bias determined through either regression analysis or difference plot analysis. For regression analysis, the same, or a subset of the information listed above for manufacturers can be provided. For difference plot analysis, the details on the measurement procedure comparison study can be the same. Instead of slope and intercept, however, the laboratory should describe the interval over which a constant SD was found and what that bias was in measurement units, and the interval over which a constant CV was found and what that bias was in percent difference. The bias at any medical decision point that falls within an interval will be the bias seen within that interval.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Chapter 9: Conclusion

Techniques for comparing IVD measurement procedures and quantifying differences between them are described in this guideline. It starts with preparing for and performing studies for collecting the needed data. Examples of difference plots and scatterplots are provided as ways to visualize this comparison data. Measurements of bias based on difference plots can be used by laboratories for verification studies but regression analyses based on scatter plots are necessary for quantification of differences by IVD manufacturers. Such regression analyses are extensively covered including which techniques to use depending upon data characteristics revealed by data visualization. The fit of the selected regression model can then be used to determine bias at result values corresponding to medical decision thresholds. Various situations are described for using these techniques within a measurement procedure. Finally, once differences have been determined, suggestions are made for stating the outcome of the measurement procedure comparison.

Chapter 10: Supplemental Information

This chapter includes:

- References
- Appendixes
- The Quality Management System Approach
- Related CLSI Reference Materials

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

References

1 CLSI. *User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline—Second Edition.* CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

2 CLSI. *Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline—Third Edition.* CLSI document EP05-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

3 CLSI. *User Verification of Precision and Estimation of Bias; Approved Guideline—Third Edition.* CLSI document EP15-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

4 CLSI. *Evaluation of Total Analytical Error for Quantitative Medical Laboratory Measurement Procedures.* 2nd ed. CLSI document EP21. Wayne, PA: Clinical and Laboratory Standards Institute; 2016.

5 CLSI. *Interference Testing in Clinical Chemistry.* 3rd ed. CLSI guideline EP07. Wayne, PA: Clinical and Laboratory Standards Institute; 2018.

6 Miller JM, Astles R, Baszler T, et al.; Biosafety Blue Ribbon Panel, Centers for Disease Control and Prevention (CDC). Guidelines for safe work practices in human and animal medical diagnostic laboratories. *MMWR Suppl.* 2012;61(1):1-102.

7 CLSI. *Protection of Laboratory Workers From Occupationally Acquired Infections; Approved Guideline—Fourth Edition.* CLSI document M29-A4. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

8 Bureau International des Poids et Mesures (BIPM). *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms* (VIM, 3rd edition, JCGM 200:2012). http://www.bipm.org/en/publications/guides/vim.html. Accessed March 28, 2018.

9 ISO. *Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions.* ISO 5725-1. Geneva, Switzerland: International Organization for Standardization; 1994.

10 Mandel J. *The Statistical Analysis of Experimental Data.* Mineola, New York: Dover Publications; 1984:282-292.

11 WHO. Expert Committee on Biological Standardization. *Glossary of Terms for Biological Substances Used for Texts of the Requirements.* WHO unpublished document BS/95.1793. Geneva, Switzerland: World Health Organization; 1991.

12 Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. *J Clin Chem Clin Biochem.* 1983;21(11):709-720.

13 Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in clinical chemistry, Part II. *J Clin Chem Clin Biochem.* 1984;22(6):431-445.

14 ISO. *Accuracy (trueness and precision) of measurement methods and results – Part 3: Intermediate measures of the precision of a standard measurement method.* ISO 5725-3. Geneva, Switzerland: International Organization for Standardization; 1994.

15 ISO. *Medical laboratories – Requirements for quality and competence.* ISO 15189. Geneva, Switzerland: International Organization for Standardization; 2012.

16 CLSI. *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition.* CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute; 2010.

17 CLSI. *Evaluation of Commutability of Processed Samples; Approved Guideline—Third Edition.* CLSI document EP14-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

18 Carroll RJ, Ruppert D. The use and misuse of orthogonal regression in linear errors-in-variables models. *Am Stat.* 1996; 50(1):1-6.

19 CLSI. *Quality Assurance for Design Control and Implementation of Immunohistochemistry Assays; Approved Guideline—Second Edition.* CLSI document I/LA28-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2011.

20 Miller WG, Erek A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. *Clin Chem.* 2011;57(1):76-83.

21 Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat.* 2007;17(4):571-582.

22 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1(8476):307-310.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

[23] Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8(2):135-160.

[24] Krouwer JS. Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med.* 2008;27(5):778-780.

[25] Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet.* 1995;346(8982):1085-1087.

[26] Carstensen B. *Comparing Clinical Measurement Methods: A Practical Guide.* Chichester, United Kingdom: John Wiley & Sons Ltd; 2010.

[27] Hawkins DM. Diagnostics for conformity of paired quantitative measurements. *Stat Med.* 2002;21(13):1913-1935.

[28] Keller T, Butz H, Lein M, et al. Discordance analysis characteristics as a new method to compare the diagnostic accuracy of tests: example of complexed versus total prostate-specific antigen. *Clin Chem.* 2005;51(3):532-539.

[29] CLSI. *Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline.* CLSI document EP06-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2003.

[30] Linnet K. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clin Chem.* 1998;44(5):1024-1031.

[31] CLSI. *Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline—Second Edition.* CLSI document EP17-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.

[32] Martin RF. General Deming regression for estimating systematic bias and its confidence interval in method comparison studies. *Clin Chem.* 2000;46(1):100-104.

[33] Davison AC, Hinkley DV. *Bootstrap Methods and their Application.* Cambridge, UK: Cambridge University Press; 1997.

[34] CLSI. *Clinical Evaluation of Immunoassays; Approved Guideline—Second Edition.* CLSI document I/LA21-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

[35] CLSI. *Verification of Comparability of Patient Results Within One Health Care System; Approved Guideline (Interim Revision).* CLSI document EP31-A-IR. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix A. Deming Regression

Abbreviation for Appendix A

SD standard deviation

With measurement errors in both measurement procedures being compared, the linear model describing the relationship between the measurement procedure results *X* and *Y* can be expressed as:

$$Y = a + b(X + \varepsilon_X) + \varepsilon_Y, \tag{A1}$$

in which:

a, b	=	intercept and slope of the linear model, and
ε_X, ε_Y	=	random errors in the *X* and *Y* measurement procedures.

Equation (A1) parameters (a, b) can be estimated with data using regular Deming regression under the following assumptions: the random errors ε_X, ε_Y are independent (across the measurement procedures, specimens, and replicates) and normally distributed with zero averages and constant, measurand-level–independent SDs, $\sigma(\varepsilon_X)$, $\sigma(\varepsilon_Y)$.

The SDs, $\sigma(\varepsilon_X)$, $\sigma(\varepsilon_Y)$, of the random errors are practically constant for measurement procedures with small analytical measuring intervals of the measurand, such as electrolytes. In other cases, the SDs of random measurement errors are often approximately proportional to the measurand level over a large proportion of the measuring interval. In such cases, constant coefficient of variation Deming analysis is more appropriate, as described in Appendix B.

The information on the SDs of the random errors of measurements' approximate constancy or proportionality to the measurand level is often available from the manufacturers' specifications. When such information is not available, it can be obtained by calculating the SDs of the replicated results of measurements for the tested samples, plotting those vs respective replicate averages, and visually examining the graph. The SDs, $\hat{\sigma}(\varepsilon_{X_i})$, $\hat{\sigma}(\varepsilon_{Y_i})$, of the replicate measurements are calculated for each sample using the following equations (assuming the same number of replicates, *R*, for each measurement procedure, *X* and *Y*, and each of the *N* samples tested):

$$\hat{\sigma}(\varepsilon_{X_i}) = \sqrt{\frac{1}{R-1}\sum_{j=1}^{R}(X_{ij} - \bar{X}_i)^2}, \text{ and} \tag{A2}$$

$$\hat{\sigma}(\varepsilon_{Y_i}) = \sqrt{\frac{1}{R-1}\sum_{k=1}^{R}(Y_{ik} - \bar{Y}_i)^2}, \tag{A3}$$

in which:

i	=	sample number; $i = 1, 2, \ldots, N$, and
j, k	=	replicate number; $j, k = 1, 2, \ldots, R$.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.
©Clinical and Laboratory Standards Institute. All rights reserved.

Appendix A. (Continued)

In the calculations below, $\bar{X}_i$,$\bar{Y}_i$ are the replicate averages when the SDs of the random measurement errors are approximately constant, and they are the averages of the logarithms of replicate measurement results when the SDs are approximately proportional to the measurand level. The replicate averages for these measurement results obtained with the *i*th sample ($i = 1, 2, \ldots, N$) are calculated as:

$$\bar{X}_i = \frac{1}{R}\sum_{j=1}^{R} X_{ij} \tag{A4}$$

$$\bar{Y}_i = \frac{1}{R}\sum_{k=1}^{R} Y_{ik} \tag{A5}$$

Equation (A1) can be rewritten for the replicate averages as:

$$\bar{Y}_i = a + b(\bar{X}_i + \varepsilon_X) + \varepsilon_Y, \tag{A6}$$

in which $\varepsilon_{\bar{X}}$, $\varepsilon_{\bar{Y}}$ are random errors of the replicate averages.

Regular Deming regression provides unbiased minimum variance estimates of the equation A6 parameters (a, b)[1] with modified notation; equation for b assumes positive $\hat{\sigma}_{\bar{X}\bar{Y}}$, which is the case with medical laboratory measurement procedures:

$$b = \frac{\hat{\sigma}^2_{\bar{Y}} - \hat{\lambda}\hat{\sigma}^2_{\bar{X}} + \sqrt{\left(\hat{\sigma}^2_{\bar{Y}} - \hat{\lambda}\hat{\sigma}^2_{\bar{X}}\right)^2 + 4\hat{\lambda}\hat{\sigma}^2_{\bar{X}\bar{Y}}}}{2\hat{\sigma}_{\bar{X}\bar{Y}}} \tag{A7}$$

$$a = \bar{\bar{Y}} - b\bar{\bar{X}} \tag{A8}$$

The parameters used in equations (A7) and (A8) are calculated using formulas (A9) to (A14).

$$\hat{\sigma}^2_{\bar{X}} = \frac{1}{N}\sum_{i=1}^{N}(\bar{X}_i - \bar{\bar{X}})^2 \tag{A9}$$

$$\hat{\sigma}^2_{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2 \tag{A10}$$

$$\hat{\sigma}_{\bar{X}\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}(\bar{X}_i - \bar{\bar{X}})(\bar{Y}_i - \bar{\bar{Y}}) \tag{A11}$$

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix A. (Continued)

$$\bar{\bar{X}} = \frac{1}{N}\sum_{i=1}^{N}\bar{X}_i \tag{A12}$$

$$\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i \tag{A13}$$

$$\hat{\lambda} = \frac{\hat{\sigma}^2(\varepsilon_{\bar{Y}})}{\hat{\sigma}^2(\varepsilon_{\bar{X}})} \tag{A14}$$

in which:

N	=	number of samples used for fitting model (A3).
$\bar{\bar{X}}, \bar{\bar{Y}}$	=	averages across measurement results obtained with X and Y measurement procedures with samples (grand averages).
$\hat{\sigma}^2_{\bar{X}}, \hat{\sigma}^2_{\bar{Y}}, \hat{\sigma}_{\bar{X}\bar{Y}}$	=	average squares and average cross-product of the deviations of the replicate averages of results of measurement obtained with the X and Y measurement procedures from the respective grand averages.
$\hat{\lambda}$	=	ratio of the variances of random errors of the two measurement procedures (within-run or repeatability when data are collected in a single run).

The constant, measurand-level–independent, random error variance estimates, $\hat{\sigma}^2(\varepsilon_X)$, $\hat{\sigma}^2(\varepsilon_Y)$, are calculated as follows[2] (the equations are modified for the same numbers of replicates, R, for both measurement procedures, X and Y, and each of N specimens):

$$\hat{\sigma}^2(\varepsilon_X) = \frac{1}{N(R-1)}\sum_{i=1}^{N}\sum_{j=1}^{R}(X_{ij} - \bar{X}_i)^2 \tag{A15}$$

$$\hat{\sigma}^2(\varepsilon_Y) = \frac{1}{N(R-1)}\sum_{i=1}^{N}\sum_{k=1}^{R}(Y_{ik} - \bar{Y}_i)^2 \tag{A16}$$

Each of the above variances has $N(R-1)$ degrees of freedom.

The variances of the averages of R replicate results of measurement are R times smaller than the variances of the individual results given in equations (A15) and (A16):

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix A. (Continued)

$$\hat{\sigma}^2(\varepsilon_{\bar{X}}) = \frac{1}{NR(R-1)}\sum_{i=1}^{N}\sum_{j=1}^{R}(X_{ij} - \bar{X}_i)^2 \tag{A17}$$

$$\hat{\sigma}^2(\varepsilon_{\bar{Y}}) = \frac{1}{NR(R-1)}\sum_{i=1}^{N}\sum_{k=1}^{R}(Y_{ik} - \bar{Y}_i)^2 \tag{A18}$$

The equations for the estimates of the variances of the intercept, σ_a^2, and slope, σ_b^2, and their covariance, σ_{ab}, in Deming regression (large sample size approximation) are as follows (modified from Miller)[1]:

$$\hat{\sigma}_a^2 = \frac{1}{N}\left[\hat{\sigma}_{\bar{Y}}^2 - 2b\hat{\sigma}_{\bar{X}\bar{Y}} + b^2\hat{\sigma}_{\bar{X}}^2 + \frac{\bar{\bar{X}}^2 b^2}{\sigma_{\bar{X}\bar{Y}}^2}(\hat{\sigma}_{\bar{X}}^2\hat{\sigma}_{\bar{Y}}^2 - \hat{\sigma}_{\bar{X}\bar{Y}}^2)\right] \tag{A19}$$

$$\hat{\sigma}_b^2 = \frac{b^2}{N\hat{\sigma}_{\bar{X}\bar{Y}}^2}(\hat{\sigma}_{\bar{X}}^2\hat{\sigma}_{\bar{Y}}^2 - \hat{\sigma}_{\bar{X}\bar{Y}}^2) \tag{A20}$$

$$\hat{\sigma}_{ab} = -\frac{\bar{\bar{X}}b^2}{N\hat{\sigma}_{\bar{X}\bar{Y}}^2}(\hat{\sigma}_{\bar{X}}^2\hat{\sigma}_{\bar{Y}}^2 - \hat{\sigma}_{\bar{X}\bar{Y}}^2) \tag{A21}$$

Assuming the Y_is follow a normal (gaussian) distribution, the above variances have $N-2$ degrees of freedom, and the $100(1-\gamma)\%$ confidence intervals for the slope and intercept are:

$$a \pm t(N-2, 1-\tfrac{\gamma}{2})\hat{\sigma}_a, \text{ and} \tag{A22}$$

$$b \pm t(N-2, 1-\tfrac{\gamma}{2})\hat{\sigma}_b, \tag{A23}$$

in which σ_a, σ_b = SDs of the intercept and slope estimates found as square roots of the respective variances in equations (A19) and (A20).

$t\left(N-2, 1-\frac{\gamma}{2}\right) = 100(1-\gamma)$ percentile of the *t*-distribution with $N-2$ degrees of freedom.

The estimates of the intercept and slope are correlated. Using the variances and the covariance of the estimates allows for obtaining the joint elliptical confidence region for these parameters. Description of the method of obtaining the joint confidence region is beyond the scope of this guideline.

The predicted bias (B_c) at a given medical decision level X_c is:

$$\hat{B}_c = a + (b-1)X_c \tag{A24}$$

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix A. (Continued)

The standard error for the bias can be calculated from the variances of intercept and slope, and their covariance (equations A19, A20, A21) as follows:

$$\hat{\sigma}_{\text{Bias}} = \sqrt{\hat{\sigma}_a^2 + X_c^2\hat{\sigma}_b^2 + 2X_c\hat{\sigma}_{ab}} \tag{A25}$$

The use of the above formulas for calculating $\hat{\sigma}_a$ and $\hat{\sigma}_b$ are not appropriate when the large sample approximation and other conditions mentioned in Miller are not satisfied.[1] The jackknife approach provided in Appendix K1 can be implemented under less restrictive conditions and is recommended in general situations. The bootstrap, repeatedly collecting *N* samples with replacement from the original samples, also provides a similarly less restrictive methodology to compute the standard errors.[3]

References for Appendix A

1 Miller RG Jr. *Beyond ANOVA, Basics of Applied Statistics.* New York, NY: Wiley; 1986:220-230.

2 Kendall M, Stuart A. *The Advanced Theory of Statistics, Volume 2: Inference and Relationship.* 4th ed. London, England: Griffin; 1979:406-407.

3 Davison AC, Hinkley DV. *Bootstrap Methods and their Application.* Cambridge, UK: Cambridge University Press; 1997.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix B. Weighted Deming Regression

Abbreviations for Appendix B

CV coefficient of variation
SD standard deviation

For measurement procedures with extremely wide measuring intervals, the analytical SD is seldom constant. Rather, a proportional relationship might apply. In a situation in which proportional analytical errors for the measurement procedures are compared, the optimal approach is a weighted form of Deming regression analysis that takes into account the relationship between random error and measurand concentration. For a given sample measured by two analytical measurement procedures, *X* and *Y*:

$$x_i = X_{\text{Target}_i} + \varepsilon_{X_i} \tag{B1}$$

$$y_i = Y_{\text{Target}_i} + \varepsilon_{Y_i} \tag{B2}$$

$$Y_{\text{Target}_i} = \alpha + \beta X_{\text{Target}_i} \tag{B3}$$

x_i and y_i are the measured values, X_{Target_i} and Y_{Target_i} are the corresponding target values, ε_{X_i} and ε_{Y_i} are the random analytical error terms of the measurement procedures *X* and *Y*, α is the regression intercept, and β is the regression slope. The analytical SDs are assumed to be proportional to the target values (CV):

$$\sigma_X = \text{CV}_X X_{\text{Target}} \text{ and } \sigma_Y = \text{CV}_Y Y_{\text{Target}} \tag{B4}$$

Given a proportional relationship for the random errors, a weighted procedure assigns larger weights to measurements in the low range; the low-range measurements are more precise than measurements at higher concentrations that are subject to larger random errors. More specifically, distances from (x_i, y_i) to the line are inversely weighted according to the squared analytical SDs (variances) at a given concentration that express the random error. The regression line is then estimated so that the sum of squared weighted differences is minimized. The regression procedure is most conveniently performed using dedicated software. The principle of the computations is outlined below. Weighted averages, weighted sums of squares, and a weighted cross product are computed:

$$\bar{X}_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad \bar{Y}_w = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i} \tag{B5, B6}$$

$$u_w = \sum_{i=1}^{N} w_i (x_i - \bar{X}_w)^2 \tag{B7}$$

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix B. (Continued)

$$q_w = \sum_{i=1}^{N} w_i (y_i - \bar{Y}_w)^2 \tag{B8}$$

$$p_w = \sum_{i=1}^{N} w_i (x_i - \bar{X}_w)(y_i - \bar{Y}_w) \tag{B9}$$

The slope and intercept are estimated as[1,2]:

$$b = \frac{(\lambda q_w - u_w) + \sqrt{(u_w - \lambda q_w)^2 + 4\lambda p^{2w}}}{2\lambda p_w} \tag{B10}$$

$$a_0 = \bar{Y}_w - b\bar{X}_w \tag{B11}$$

The weights are obtained by an iterative approach:

$$z_i = \frac{\widehat{\mathrm{X}}_{\mathrm{Target}_i} + \lambda \widehat{\mathrm{Y}}_{\mathrm{Target}_i}}{1 + \lambda}, \text{and} \tag{B12}$$

$$w_i = 1/[h(z_i)]^2, \tag{B13}$$

in which z_i estimates the true concentration of target *i*, and *h* is the fitted variance model (see Appendix J) of the variance. In Linnet's original proposal of constant CV, $h(z_i) = z_i$,[1]

$$W_i = \frac{1}{\left(\frac{\widehat{\mathrm{X}}_{\mathrm{Target}_i} + \lambda \widehat{\mathrm{Y}}_{\mathrm{Target}_i}}{1 + \lambda}\right)^2}, \tag{B14}$$

but other variance models can be used when appropriate.[1,2] It is here presumed that the ratio λ between the squared SDs (variances) for the random error components is constant throughout the measuring interval.

$$\lambda = {\sigma_x^2}\Big/{\sigma_y^2} = {\mathrm{var}(\varepsilon_X)}\Big/{\mathrm{var}(\varepsilon_Y)} \tag{B15}$$

λ can be based on the analytical CVs obtained from quality control results, for example. Otherwise, λ can, as default, be assigned the value 1. Without any knowledge of the ratio, for some purposes, the ratio can be varied to assess the sensitivity of the Deming regression to its value.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix B. (Continued)

Bias at medical decision level(s) is calculated based on the estimates of slope and intercept. The jackknife approach provided in Appendix K1 can be used to calculate standard errors of regression parameters and SE of bias.

References for Appendix B

[1] Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. *Stat Med.* 1990;9(12):1463-1473.

[2] Linnet K. Evaluation of regression procedures for methods comparison studies. *Clin Chem.* 1993;39(3):424-432.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix C. A Practical Example Illustrating Bias Estimation and Measurement Procedure Comparison Techniques

Abbreviations for Appendix C

%Diff	percent difference
CI	confidence interval
CV	coefficient of variation
Diff	difference
MP	measurement procedure
OLR	ordinary linear regression
SD	standard deviation
WLS	weighted least squares

A manufacturer decides to perform a comparison to determine whether one lot of reagent is equivalent to another lot for a measurement procedure (MP) with a measuring interval from near zero to 100 µg/L, for which equivalence is concluded when the bias is within ± 6% or 0.06 µg/L, whichever is greater. In this example, the same measurement procedure is used in both instances so data from the two lots are labeled MP X and MP Y. After accumulating the recommended 40 samples from an external site, it is discovered that all the samples are within a 0 to 10-µg/L measuring interval. It is decided to have the site accumulate additional samples in order to cover the entire measuring interval, resulting in a total of 79 samples. Two replicates are collected using each lot for each sample, and the average of each replicate pair is computed. The difference and the percent difference between lots are computed. The average over both lot results is used as the divisor for computing percent difference, because both are known to have similar imprecision performance and neither can be considered a reference lot. The dataset in Table C1, used previously for Figure 6 in Subchapter 5.4.3 of this guideline and Figure 20 in Subchapter 6.2.3 of this guideline, has been ranked by this average concentration (average).

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix C. (Continued)

Table C1. Mixed Variability Example Dataset

MP X	MP Y	Average	Order	Diff	%Diff		MP X	MP Y	Average	Order	Diff	%Diff
0.004	0.001	0.003	1	−0.003	−100.0%		1.773	1.945	1.859	41	0.172	9.3%
0.001	0.007	0.004	2	0.006	150.0%		1.917	1.991	1.954	42	0.074	3.8%
0.007	0.001	0.004	3	−0.006	−150.0%		1.978	2.084	2.031	43	0.106	5.2%
0.007	0.001	0.004	4	−0.006	−150.0%		2.315	2.373	2.344	44	0.058	2.5%
0.004	0.012	0.008	5	0.008	100.0%		2.371	2.329	2.350	45	−0.042	−1.8%
0.012	0.004	0.008	6	−0.008	−100.0%		2.681	2.623	2.652	46	−0.058	−2.2%
0.004	0.013	0.009	7	0.009	105.9%		3.034	3.580	3.307	47	0.546	16.5%
0.014	0.006	0.010	8	−0.008	−80.0%		3.287	3.348	3.318	48	0.061	1.8%
0.014	0.008	0.011	9	−0.006	−54.5%		3.469	3.472	3.471	49	0.003	0.1%
0.008	0.015	0.012	10	0.007	60.9%		4.063	3.979	4.021	50	−0.084	−2.1%
0.030	0.012	0.021	11	−0.018	−85.7%		5.186	5.264	5.225	51	0.078	1.5%
0.018	0.026	0.022	12	0.008	36.4%		5.404	5.244	5.324	52	−0.160	−3.0%
0.026	0.018	0.022	13	−0.008	−36.4%		5.243	5.529	5.386	53	0.286	5.3%
0.030	0.041	0.036	14	0.011	31.0%		6.811	6.149	6.480	54	−0.662	−10.2%
0.040	0.036	0.038	15	−0.004	−10.5%		7.215	6.815	7.015	55	−0.400	−5.7%
0.037	0.050	0.044	16	0.013	29.9%		7.792	7.961	7.877	56	0.169	2.1%
0.045	0.051	0.048	17	0.006	12.5%		8.719	8.348	8.534	57	−0.371	−4.3%
0.051	0.045	0.048	18	−0.006	−12.5%		10.365	9.885	10.125	58	−0.480	−4.7%
0.150	0.142	0.146	19	−0.008	−5.5%		11.154	11.608	11.381	59	0.454	4.0%
0.173	0.179	0.176	20	0.006	3.4%		11.878	11.588	11.733	60	−0.290	−2.5%
0.194	0.230	0.212	21	0.036	17.0%		13.001	12.864	12.933	61	−0.137	−1.1%
0.224	0.220	0.222	22	−0.004	−1.8%		13.041	13.246	13.144	62	0.205	1.6%
0.244	0.264	0.254	23	0.020	7.9%		14.037	14.152	14.095	63	0.115	0.8%
0.338	0.340	0.339	24	0.002	0.6%		14.942	14.272	14.607	64	−0.670	−4.6%
0.645	0.653	0.649	25	0.008	1.2%		14.838	14.692	14.765	65	−0.146	−1.0%
0.607	0.703	0.655	26	0.096	14.7%		16.637	14.921	15.779	66	−1.716	−10.9%
0.641	0.697	0.669	27	0.056	8.4%		17.873	16.436	17.155	67	−1.437	−8.4%
0.666	0.739	0.703	28	0.073	10.4%		18.031	16.918	17.475	68	−1.113	−6.4%
0.744	0.768	0.756	29	0.024	3.2%		17.757	21.047	19.402	69	3.290	17.0%
0.766	0.861	0.814	30	0.095	11.7%		22.538	21.096	21.817	70	−1.442	−6.6%

Licensed to: George Otero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/14/2018
Downloaded on 9/14/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Licensed to: George Odero
This document is protected by copyright. Downloaded from 9/4/2018.

Appendix C. (Continued)

Table C1. (Continued)

MP X	MP Y	Average	Order	Diff	%Diff		MP X	MP Y	Average	Order	Diff	%Diff
0.884	0.863	0.874	31	−0.021	−2.4%		24.358	22.259	23.309	71	−2.099	−9.0%
0.871	0.883	0.877	32	0.012	1.4%		23.720	23.210	23.465	72	−0.510	−2.2%
0.880	0.877	0.879	33	−0.003	−0.3%		24.655	24.996	24.826	73	0.341	1.4%
0.893	0.955	0.924	34	0.062	−6.7%		26.155	26.577	26.366	74	0.422	1.6%
1.038	0.811	0.925	35	−0.227	−24.6%		43.709	41.220	42.465	75	−2.489	−5.9%
1.090	1.000	1.045	36	−0.090	−8.6%		41.801	43.464	42.633	76	1.663	3.9%
1.200	1.479	1.340	37	0.279	20.8%		62.516	75.876	69.196	77	13.360	19.3%
1.389	1.833	1.611	38	0.444	27.6%		69.923	71.797	70.860	78	1.874	2.6%
1.774	1.729	1.752	39	−0.045	−2.6%		91.235	99.802	95.519	79	8.567	9.0%
1.767	1.772	1.770	40	0.005	0.3%							

Abbreviations: %Diff, percent difference; Diff, difference; MP, measurement procedure.

The data are plotted in a scatter plot (Figure C(a)) and in constant difference (Figure C(b)) and proportional difference (Figure C(c)) plots. Note that the concentration distribution of samples in this example is still not evenly spaced over the measuring interval. To expand this into a typical 100-sample manufacturer's study, additional higher concentration samples should be collected.

Appendix C. (Continued)

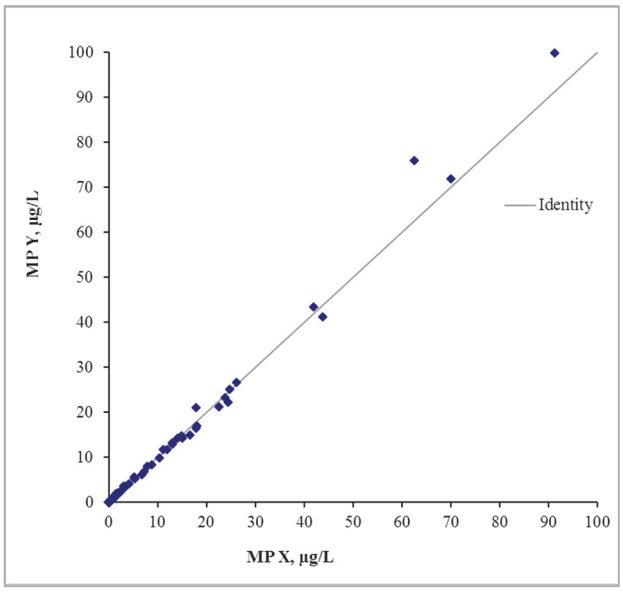


Figure C(a). Lot Comparison (see Figure 6A, Subchapter 5.4.3 of this guideline)
Abbreviations: MP X, measurement procedure X; MP Y, measurement procedure Y.

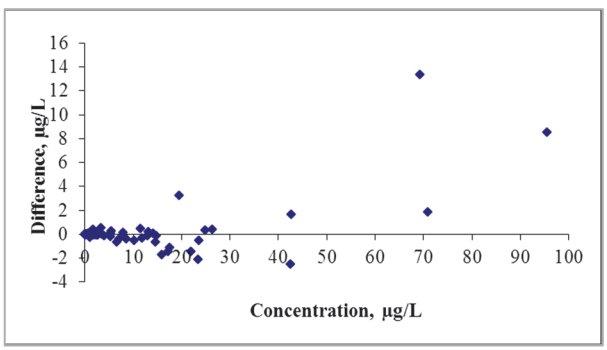


Figure C(b). Constant Difference Plot (see Figure 6B, Subchapter 5.4.3 of this guideline)

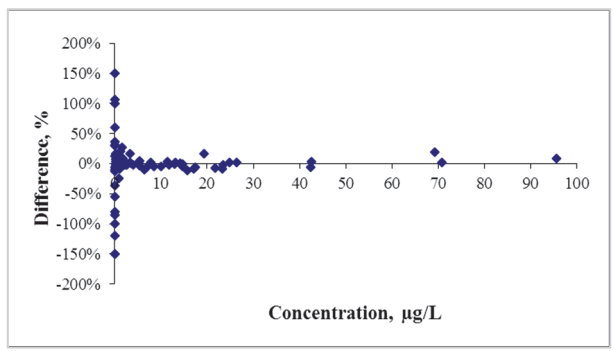


Figure C(c). Proportional Difference Plot (see Figure 6C, Subchapter 5.4.3 of this guideline)

The plots visually indicate that the dataset has mixed variability with constant SD at lower concentrations and constant CV at higher concentrations. Because it is difficult to determine the concentration at which the transition between these two occurs, the difference plots are replotted with rank order as the horizontal axis.

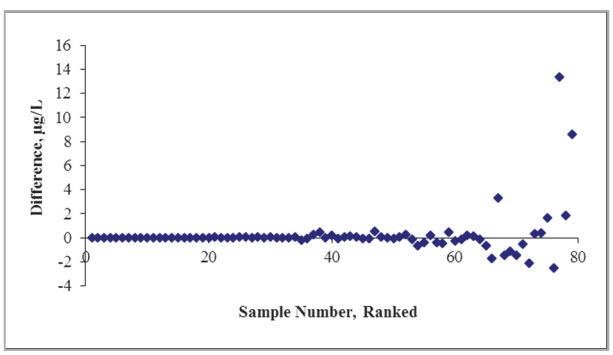


Figure C(d). Constant Difference Plot (see Figure 8A, Subchapter 5.4.4 of this guideline)

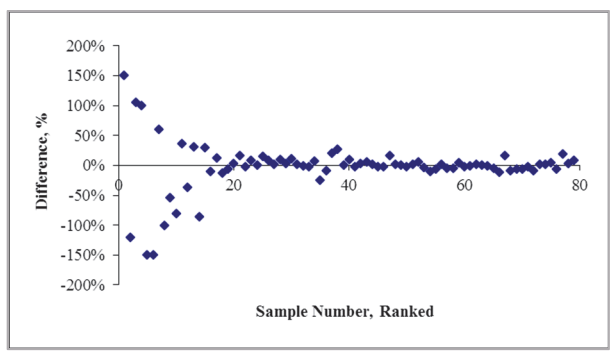


Figure C(e). Proportional Difference Plot (see Figure 8B, Subchapter 5.4.4 of this guideline)

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix C. (Continued)

Inspecting Figure C(d), there appears to be a constant variability from sample 1 though at least sample 35. Figure C(e) displays relatively consistent proportional differences from sample 79 down to at least sample 40. Given the option of picking any sample number from 35 to 40 as the point of change of the relationship from constant to proportional, it is decided to divide the data into two equal-sized sets of results, from sample numbers 1 through 40 and from 41 to 79. The concentration at which the relationship changes from constant to proportional SD can be estimated by formal statistical analysis (called change point analysis), but is beyond the scope of this guideline.

The average difference of the low concentration dataset is 0.020 µg/L with a 95% confidence interval (CI) of −0.010 to 0.051 µg/L and covers the interval from 0 to 1.8 µg/L. The 95% CI of estimated constant bias is covered by the prespecified acceptance criterion (± 0.06 µg/L). It can therefore be concluded (with 95% confidence) that the criterion for lot equivalence was met at lower concentrations.

The average difference of the high concentration dataset is 0.43% with a 95% CI of −1.83% to 2.69% and covers the interval from 1.8 to 100 µg/L. The 95% CI of estimated proportional bias is covered by the prespecified acceptance criterion (± 6%). It can therefore be concluded (with 95% confidence) that the criterion for lot equivalence was met at higher concentrations.

This analysis would have been adequate for a characterization, but, for illustrative purposes, all the regression models introduced are used below to analyze the data. A review of the examples in Subchapter 6.2 of this guideline demonstrates that a constant CV Deming regression would be a reasonable choice. However, given the small number of influential, high concentration samples, a Passing-Bablok regression would be the best choice. Technical discussions on the suitability of various regression techniques are referenced.[1-4]

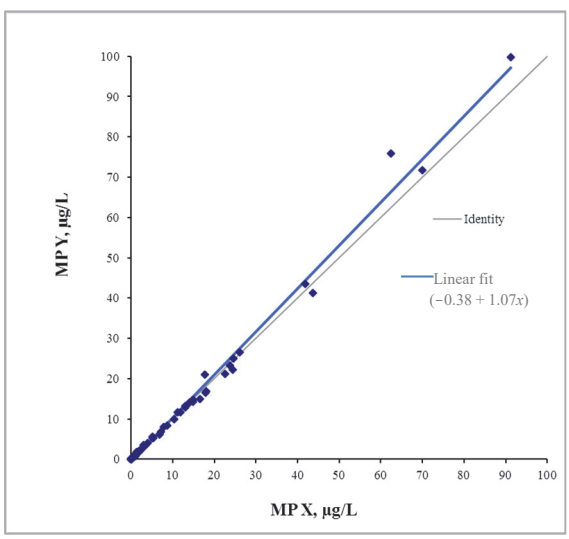


Figure C(f). Scatter Plot With Ordinary Linear Regression (OLR) Fit
Abbreviations: MP X, measurement procedure X; MP Y, measurement procedure Y.

The OLR fit, seen in Figure C(f), demonstrates two difficulties with this technique for this dataset. First, the three highest concentration data points are so influential that the line is forced through the middle of them, regardless of the other 76 points, resulting in an estimate of positive bias (slope = 1.07). Second, the resulting pivot of line causes a low intercept, which misses most of the near-zero results (intercept = −0.36 µg/L).

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix C. (Continued)

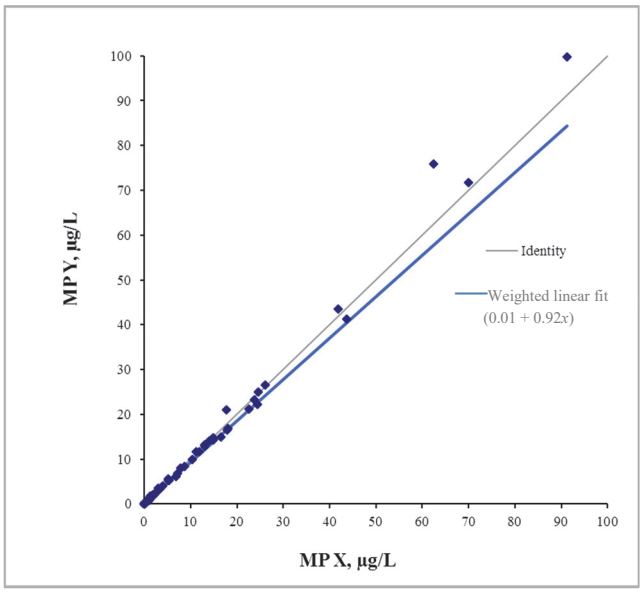


Figure C(g). Scatter Plot With Weighted Least Squares (WLS) Fit
Abbreviations: MP X, measurement procedure X; MP Y, measurement procedure Y.

The weighted least squares (WLS) regression, seen in Figure C(g), solves the problem at the low end of the measuring interval (intercept = 0.01 µg/L), but the heightened influence of these low results causes the fitted line to miss all of the data points above 30 µg/L, resulting in a negative bias estimate (slope = 0.92).

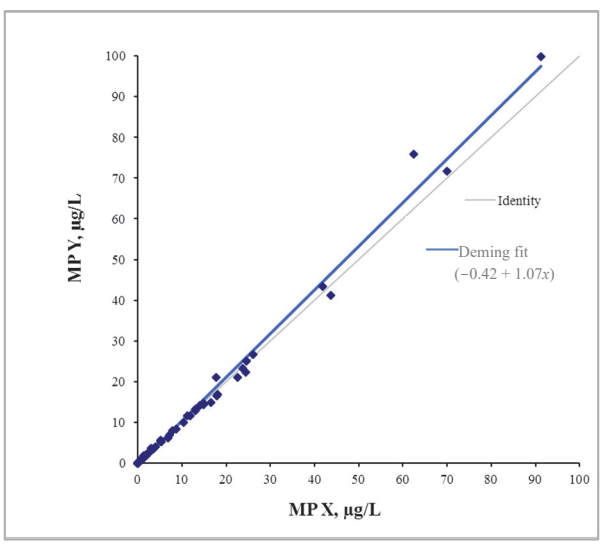


Figure C(h). Scatter Plot With Deming Fit
Abbreviations: MP X, measurement procedure X; MP Y, measurement procedure Y.

The Deming fit (nonweighted), seen in Figure C(h), has the same difficulties with datasets displaying proportional variability as the OLR fit. The high concentration points are very influential (slope = 1.07), which also causes a poor fit through low concentration points (intercept = −0.42 µg/L).

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix C. (Continued)

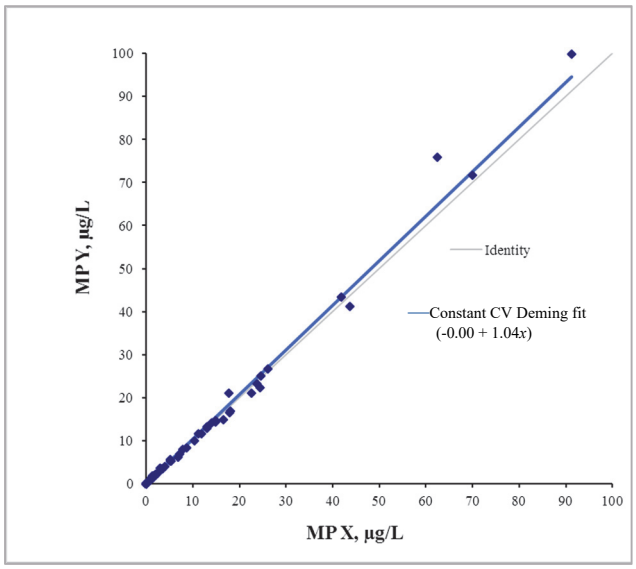


Figure C(i). Scatter Plot With Constant CV Deming Fit (see Figure 20A, Subchapter 6.2.3 of this guideline)
Abbreviations: MP X, measurement procedure X; MP Y, measurement procedure Y.

The constant CV Deming fit, seen in Figure C(i), much like the WLS fit, ensures that the line is drawn through the lowest concentration points (intercept = 0.00 µg/L). However, the constant CV Deming fit is not as heavily influenced by the few high concentration points (slope = 1.04).

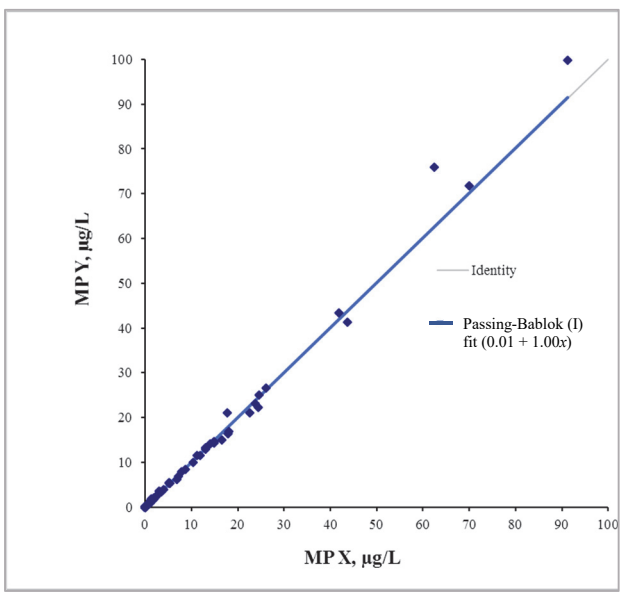


Figure C(j). Passing-Bablok Algorithm I Fit (see Figure 20B, Subchapter 6.2.3 of this guideline)
Abbreviations: MP X, measurement procedure X; MP Y, measurement procedure Y.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix C. (Continued)

The Passing-Bablok algorithm (see Passing and Bablok Part I reference from Appendix I), seen in Figure C(j), provides the estimates of slope (1.00, 95% CI, 0.98 to 1.02 µg/L) and intercept (0.01 µg/L, 95% CI, −0.01 to 0.01 µg/L). These parameters are not unduly influenced by either the low or the high concentration points. The slope 95% CI of 0.98 to 1.02 is equivalent to a 95% CI of proportional bias of −2% to +2%.

This CI is covered by the prespecified acceptance criterion of ± 6%. It can therefore be concluded (with greater than 95% confidence) that the criterion for lot equivalence has been met.

This equation could be used to estimate the bias between the two measurement procedures at any concentration of interest. Assuming a medical decision point of 5 µg/L for MP X and the unrounded estimates for slope and intercept, the resultant estimate for MP Y would be 0.0055 + 1.0028 • 5 = 5.019 µg/L. Expressed as a proportional bias, this is (5.019 − 5.000) / average (5.000, 5.019) = 0.37%.

Although the jackknife technique described in Appendix K1 is very amenable to providing the 95% CI of the bias estimate parametric regression methods such as OLR or Deming, using this technique for Passing-Bablok gives a result of 0.32% to 0.39%. Such an unrealistically small interval can be obtained for this nonparametric method because the exclusion of an individual point has little effect on the outcome. For Passing-Bablok, a more realistic result is obtained using the bootstrap technique, mentioned in Subchapter 6.3 of this guideline and more fully described in Appendix K2. When 79 samples were selected with replacement from the sample population for 1000 individual regressions, the 95% distribution interval of bias estimates was found to be −2.02% to +1.94%. This interval is covered by the prespecified acceptance criterion of ± 6%. It can therefore be concluded (with greater than 95% confidence) that the criterion for lot equivalence has been met at this medical decision point. The bootstrap technique can also be used with Deming regression to obtain CIs of regression parameters and bias estimates.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix D. Example Datasets

Abbreviations for Appendix D

CV coefficient of variation
MP measurement procedure
SD standard deviation

Tables D1 to D6 represent data from several figures within EP09. The appropriate Appendix D table is referenced parenthetically in each figure legend.

Table D1. Constant SD Example Dataset 1

Sample	Comparative MP, µg/L	Candidate MP, µg/L		Sample	Comparative MP, µg/L	Candidate MP, µg/L
1	20.379	22.331		21	419.455	428.058
2	34.751	49.751		22	439.121	445.067
3	60.277	69.165		23	462.059	464.932
4	83.777	93.426		24	472.317	474.433
5	106.269	107.723		25	504.759	511.372
6	116.743	129.249		26	520.683	526.851
7	146.795	140.653		27	534.388	549.849
8	161.256	164.652		28	564.996	567.208
9	178.083	185.256		29	590.481	577.021
10	191.946	215.370		30	599.080	602.060
11	217.536	226.060		31	623.936	622.545
12	235.636	251.961		32	649.103	646.031
13	259.064	266.115		33	657.008	666.813
14	261.709	272.312		34	680.382	689.887
15	287.760	313.284		35	696.740	694.387
16	326.337	329.828		36	704.163	717.294
17	347.114	351.629		37	734.406	749.166
18	351.462	364.261		38	755.933	765.697
19	375.992	387.253		39	768.454	781.376
20	403.530	415.137		40	801.763	810.653

Abbreviations: MP, measurement procedure; SD, standard deviation.

Table D2. Constant CV Example Dataset 1

Sample	Comparative MP, µg/L	Candidate MP, µg/L		Sample	Comparative MP, µg/L	Candidate MP, µg/L
1	0.881	1.078		21	108.408	126.106
2	2.872	3.575		22	148.179	155.735
3	4.975	6.431		23	153.802	164.340
4	8.351	7.961		24	159.521	199.267
5	9.753	11.666		25	198.730	182.707
6	14.450	12.362		26	203.215	213.992
7	18.552	14.992		27	215.483	193.963
8	21.520	21.581		28	227.755	263.157
9	23.481	25.084		29	233.649	308.018
10	26.710	28.909		30	298.821	330.309
11	33.259	40.577		31	276.827	306.107

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix D. (Continued)

Table D2. (Continued)

Sample	Comparative MP, µg/L	Candidate MP, µg/L		Sample	Comparative MP, µg/L	Candidate MP, µg/L
12	45.907	40.702		32	362.759	397.674
13	36.180	43.981		33	439.989	452.057
14	49.853	57.015		34	434.477	422.399
15	46.330	61.638		35	572.399	466.581
16	54.798	69.268		36	689.940	484.590
17	78.002	69.833		37	538.316	660.415
18	85.464	88.143		38	612.061	801.227
19	100.030	98.011		39	734.508	747.958
20	97.877	119.859		40	785.566	704.414

Abbreviations: CV, coefficient of variation; MP, measurement procedure.

Table D3. Constant CV Example Dataset 2

Sample	Comparative MP, µg/L	Candidate MP, µg/L		Sample	Comparative MP, µg/L	Candidate MP, µg/L
1	0.998	1.271		21	17.682	19.790
2	1.851	1.664		22	17.976	20.930
3	2.564	2.858		23	20.654	23.438
4	2.786	3.046		24	19.805	24.340
5	3.508	3.905		25	25.907	20.946
6	4.613	5.378		26	27.394	29.873
7	6.254	5.584		27	27.928	30.175
8	6.456	6.774		28	32.294	33.176
9	7.772	6.497		29	31.654	34.473
10	7.427	7.593		30	37.321	34.472
11	8.013	8.978		31	38.867	39.258
12	9.964	8.305		32	34.828	43.839
13	10.498	11.172		33	43.114	38.103
14	10.104	13.164		34	40.274	45.428
15	12.476	11.694		35	75.226	68.413
16	15.932	12.349		36	109.740	141.146
17	14.889	14.752		37	166.803	166.030
18	16.681	18.071		38	178.471	227.842
19	16.600	18.433		39	379.574	479.814
20	16.921	19.093		40	893.271	734.152

Abbreviations: CV, coefficient of variation; MP, measurement procedure.

Table D4. Constant CV Example Dataset With Outlier

Sample	Comparative MP, µg/L	Candidate MP, µg/L		Sample	Comparative MP, µg/L	Candidate MP, µg/L
1	0.881	1.078		21	108.408	126.106
2	2.872	3.575		22	148.179	155.735
3	4.975	6.125		23	153.802	164.340
4	8.351	7.961		24	159.521	199.267
5	9.753	11.666		25	198.730	182.707
6	14.450	12.362		26	203.215	213.992

.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix D. (Continued)

Table D4. (Continued)

Sample	Comparative MP, µg/L	Candidate MP, µg/L		Sample	Comparative MP, µg/L	Candidate MP, µg/L
7	18.552	14.992		27	215.483	193.963
8	21.520	21.581		28	227.755	263.157
9	23.481	25.084		29	233.649	308.018
10	26.710	28.909		30	298.821	330.309
11	33.259	40.577		31	276.827	306.107
12	45.907	40.702		32	362.759	397.674
13	36.180	43.981		33	439.989	452.057
14	49.853	635.000		34	434.477	422.399
15	46.330	61.638		35	572.399	466.581
16	54.798	69.268		36	689.940	484.590
17	78.002	69.833		37	538.316	660.415
18	85.464	88.143		38	612.061	801.227
19	100.030	98.011		39	734.508	747.958
20	97.877	119.859		40	785.566	704.414

Abbreviations: CV, coefficient of variation; MP, measurement procedure.

Table D5. Constant SD Example Dataset With Outlier

Sample	Comparative MP, g/L	Candidate MP, mg/L		Sample	Comparative MP, mg/L	Candidate MP, mg/L
1	10.041	9.973		21	13.270	13.203
2	10.184	9.943		22	13.794	13.264
3	10.121	12.442		23	13.358	13.840
4	10.520	10.263		24	14.287	14.026
5	11.042	10.506		25	14.203	14.015
6	11.298	11.040		26	14.150	13.651
7	10.514	10.968		27	14.080	14.187
8	10.995	11.305		28	14.680	14.767
9	10.626	11.079		29	14.404	14.597
10	11.434	11.956		30	14.932	14.464
11	11.890	11.925		31	15.146	15.119
12	12.161	11.875		32	15.132	14.948
13	12.274	12.228		33	15.525	15.128
14	11.927	12.356		34	15.196	15.671
15	12.469	11.674		35	15.508	15.722
16	12.647	12.200		36	15.824	15.758
17	12.499	12.422		37	16.130	15.991
18	13.154	12.239		38	15.925	16.492
19	13.449	12.656		39	16.161	16.600
20	12.804	12.996		40	16.300	16.511

Abbreviations: MP, measurement procedure; SD, standard deviation.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix D. (Continued)

Table D6. Constant SD Example Dataset 2

Sample	Comparative MP, mg/L	Candidate MP, mg/L		Sample	Comparative MP, mg/L	Candidate MP, mg/L
1	10.041	10.273		21	13.270	13.503
2	10.184	11.243		22	13.794	13.564
3	10.121	9.742		23	13.358	14.140
4	10.520	10.563		24	14.287	14.326
5	11.042	10.806		25	14.203	14.315
6	11.298	11.340		26	14.150	13.951
7	10.514	11.268		27	14.080	14.487
8	10.995	11.605		28	14.680	15.067
9	10.626	11.379		29	14.404	14.897
10	11.434	12.256		30	14.932	14.764
11	11.890	12.225		31	15.146	15.419
12	12.161	12.175		32	15.132	15.248
13	12.274	12.528		33	15.525	15.428
14	11.927	12.656		34	15.196	15.971
15	12.469	11.974		35	15.508	16.022
16	12.647	12.500		36	15.824	16.058
17	12.499	12.722		37	16.130	16.291
18	13.154	12.539		38	15.925	16.792
19	13.449	12.956		39	16.161	16.900
20	12.804	13.296		40	16.300	16.811

Abbreviations: MP, measurement procedure; SD, standard deviation.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018

© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix E. Detecting Aberrant Results (Outliers)

Abbreviations for Appendix E

CV coefficient of variation
ESD extreme studentized deviate
SD standard deviation

The primary goal of detecting aberrant results (outliers) is to enable troubleshooting. Outliers must be found before their underlying cause can be investigated. In an ideal situation, outliers are detected during data collection as suggested in Subchapter 4.4 of this guideline when their cause can more likely be identified.

This guideline provides techniques to obtain robust estimates of bias for both difference plots and regression analyses. For difference plots, using the median rather than the average to estimate bias over an interval of measurements reduces or eliminates the outlier's undue influence on the result. In a similar manner, use of a Passing-Bablok regression provides similar robust regression estimates in the presence of outliers.

The outlier detection process is simplified by using difference plots. The techniques in Subchapter 5.4 of this guideline are followed to characterize whether the data are from a constant SD or constant CV relationship. In those cases in which there are mixed variability relationships, the data can be split into two sets: concentrations in which there is a constant SD, and concentrations in which there is a constant CV. Within an identified dataset, the differences are expressed as either differences or percent differences. There should be at least 20 samples in any such set. To align with Subchapter 5.4 of this guideline, d_j represents a result from the distribution seen in a difference plot.

The detection of aberrant results reduces to a detection of an outlier from within a single distribution. The generalized extreme studentized deviate (ESD) technique, which assumes that the distribution of the vast majority of data points is normal (gaussian), can be used when the number of outliers is unknown, and becomes more robust as the number of samples increases. To perform this technique:

1. Set the significance level (α), which will be used to detect outliers. Typical values are 0.05 or 0.01.

2. Determine whether there are potential outliers from a graphical or other review of the dataset. No more than 5% of sample results can be flagged as outliers. Set the upper bound on number of potential outliers (h) at this 5% level, rounding down to a whole number. (For 44 samples, h equals 2. For 112 samples, h equals 5.)

3. For each dataset, determine whether one or more suspect results can be statistically deemed outliers based upon the generalized ESD test.[1-3] When the results are not determined to be outliers, they should be retained in the dataset.
   a. Compute the average $(\bar{d})$ and SD, including the suspected outliers.

   b. Find the maximum observed deviate from the average deviate scaled in terms of SD (for $j = 1, 2, \ldots N$):

$$ESD_1 = \frac{\max\left(\left|d_j - \bar{d}\right|\right)}{SD} \tag{E1}$$

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix E. (Continued)

Repeat this calculation to obtain the ESD_i for all potential outliers for $i = 1, 2, \ldots h$. Each subsequent calculation of ESD_i is performed after removing the previously identified potential outlier from the dataset. Thus at each iteration the number of results is reduced by one, then $\bar{d}$, *SD*, and the (ESD_i) are computed again (ie, to look for outlier 2, the number of samples remaining is $j = 1, 2, \ldots N-1$).

c. Corresponding to the number of test statistics (h), compute the following h critical values:

$$\lambda_i = \frac{t_{v,p}(N-i)}{\sqrt{(N-i+1)(v+t_{v,p}^2)}}, \tag{E2}$$

in which N is the initial number of samples in the dataset, and $i = 1, 2, \ldots h$,

$$v = N - i - 1, \tag{E3}$$

$$p = 1 - \frac{\alpha}{2(N-i+1)}, \tag{E4}$$

and $t_{v,p}$ is the one-sided $100p$ percentage point from Student t-distribution with v degrees of freedom and probability $= p$.

d. The number of outliers is determined by finding the largest i such that $ESD_i > \lambda_i$. If ESD_h and ESD_{h+1} are equal (a tie), then neither one should be seen as an outlier.

When an outlier is detected, part of the investigation should be separating out the individual replicates for that sample to determine whether the aberrant result is due to a single replicate. This can be a key finding that can help determine the cause of the aberrant result.

The dataset from Appendix F (Table F2) is an example of an instance in which an obvious outlier would warrant the use of the generalized ESD test. This dataset provides the following results:

Setting $\alpha = 0.01$, with $N = 100$ and then $h = 5$. Table E1 lists each subsequent iteration.

Table E1. Example Results

Parameter	$i=1$	$i=2$	$i=3$	$i=4$	$i=5=h$
Average ($\bar{d}$)	0.01%	0.57%	0.35%	0.15%	−0.03%
SD	9.15%	7.25%	6.94%	6.69%	6.45%
ESD_i	6.09	3.01	2.78	2.75	2.14
λ_i	3.75	3.75	3.75	3.74	3.74
Bias	−55.70% $j=3$	22.41% $j=75$	19.64% $j=29$	18.57% $j=44$	13.78% $j=26$

Abbreviations: *ESD,* extreme studentized deviate; *SD*, standard deviation.
Symbols: j, the row in Table F2 in Appendix F in which each bias was obtained; λ_i, critical value.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix E. (Continued)

In this series of calculations the only case in which $ESD_i > \lambda_i$ is the first iteration. This is therefore the only identified outlier. When, due to results of this analysis, one or more points are removed from the dataset, the analysis should be performed with and without the identified outliers.

References for Appendix E

1 Rosner B. Percentage points for a generalized ESD many-outlier procedure. *Technometrics.* 1983;25(2):165-172.

2 NIST. Section 1.3.5.17: Detection of outliers. In: *NIST/SEMATECH e-Handbook of Statistical Methods.* http://www.itl.nist.gov/div898/handbook/. Accessed March 28, 2018.

3 Hypothesis testing: two-sample inference. In: Rosner B. *Fundamentals of Biostatistics.* 8th ed. Boston, MA: Cengage Learning; 2016:279-298.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix F. Confidence Interval of a Median Estimate of Bias Between Measurement Procedures

Abbreviation for Appendix F

CI confidence interval

In instances in which the distribution of results does not follow a normal (gaussian) distribution, the median is a more robust estimator of central tendency than the mean. This appendix provides two procedures to compute the median and its confidence interval (CI). The first procedure is based on a normal approximation to the binomial, and the second procedure is based on the Wilcoxon distribution-free signed rank test.

Experiment

Table F1 shows the experimental layout of a measurement procedure comparison study, in which y_i is the result with the candidate measurement procedure and x_i is the result with the comparative measurement procedure.

Table F1. Experimental Layout for a Measurement Procedure Comparison Study

Sample	x_i	y_i
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
…	…	…
N	x_N	x_N

Assumptions:

- Let $d_i = y_i - x_i$, for $i = 1,\ldots N$. The differences $d_1,\ldots d_N$ are mutually independent.
- Each d_i comes from a continuous population, not necessarily the same, that is symmetrical about a common median θ. Each d_i can be expressed in units or as a percentage of the sample concentration.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Licensed to: George Odero
This document is protected by copyright. Downloaded on 9/4/2018.

Appendix F. (Continued)

Example:

Table F2. Median Bias and CI Example Data

Sample	x	y	$(y-x)/x$		Sample	x	y	$(y-x)/x$
1	0.52	0.49	−5.77%		51	27.13	24.42	−9.99%
2	0.99	0.97	−2.02%		52	25.72	27.72	7.78%
3	1.49	0.66	−55.70%		53	24.45	24.17	−1.15%
4	1.99	1.87	−6.03%		54	26.39	27.65	4.77%
5	2.52	2.56	1.59%		55	28.76	28.12	−2.23%
6	3.20	2.91	−9.06%		56	26.37	29.74	12.78%
7	3.37	3.53	4.75%		57	27.63	28.28	2.35%
8	3.95	3.96	0.25%		58	28.78	27.89	−3.09%
9	4.85	4.51	−7.01%		59	27.74	29.19	5.23%
10	4.96	5.46	10.08%		60	32.59	29.34	−9.97%
11	5.33	5.85	9.76%		61	31.48	32.00	1.65%
12	6.00	5.98	−0.33%		62	30.31	31.99	5.54%
13	7.02	6.62	−5.70%		63	32.30	30.50	−5.57%
14	7.15	6.92	−3.22%		64	33.11	32.73	−1.15%
15	7.76	7.47	−3.74%		65	31.33	32.76	4.56%
16	7.71	7.56	−1.95%		66	31.45	31.62	0.54%
17	8.59	8.08	−5.94%		67	31.37	32.35	3.12%
18	9.87	8.94	−9.42%		68	33.54	33.46	−0.24%
19	8.75	9.34	6.74%		69	32.74	32.71	−0.09%
20	10.48	9.81	−6.39%		70	33.21	36.07	8.61%
21	10.16	9.78	−3.74%		71	32.57	32.26	−0.95%
22	11.17	10.91	−2.33%		72	35.85	37.90	5.72%
23	11.83	11.13	−5.92%		73	37.04	34.18	−7.72%
24	11.79	12.17	3.22%		74	35.23	36.43	3.41%
25	12.29	12.63	2.77%		75	34.54	42.28	22.41%
26	11.39	12.96	13.78%		76	35.45	35.52	0.20%
27	13.67	14.93	9.22%		77	39.35	37.29	−5.24%

Appendix F. (Continued)

Table F2. (Continued)

Sample	x	y	$(y-x)/x$		Sample	x	y	$(y-x)/x$
28	12.93	12.91	−0.15%		78	40.13	39.27	−2.14%
29	12.83	15.35	19.64%		79	37.98	41.93	10.40%
30	16.78	14.71	−12.34%		80	41.87	42.29	1.00%
31	14.72	15.92	8.15%		81	41.14	42.46	3.21%
32	16.53	15.32	−7.32%		82	43.39	37.68	−13.16%
33	17.17	17.04	−0.76%		83	38.93	39.71	2.00%
34	16.82	15.68	−6.78%		84	43.28	42.52	−1.76%
35	18.39	18.17	−1.20%		85	42.48	46.82	10.22%
36	17.68	17.38	−1.70%		86	42.55	44.16	3.78%
37	19.30	18.82	−2.49%		87	45.17	39.29	−13.02%
38	19.53	20.98	7.42%		88	44.18	45.78	3.62%
39	19.77	18.42	−6.83%		89	45.12	41.71	−7.56%
40	20.48	18.77	−8.35%		90	40.93	46.03	12.46%
41	21.08	20.34	−3.51%		91	48.80	46.89	−3.91%
42	21.31	21.37	0.28%		92	49.47	43.86	−11.34%
43	21.64	21.21	−1.99%		93	45.21	47.88	5.91%
44	20.52	24.33	18.57%		94	48.44	46.26	−4.50%
45	24.30	23.68	−2.55%		95	45.07	43.64	−3.17%
46	21.30	21.72	1.97%		96	43.72	47.45	8.53%
47	24.13	23.59	−2.24%		97	49.74	52.83	6.21%
48	23.99	24.19	0.83%		98	47.59	54.06	13.60%
49	22.19	24.19	9.01%		99	48.61	54.09	11.27%
50	23.83	23.75	−0.34%		100	53.08	49.53	−6.69%

Licensed to: George Otero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/14/2018
© 2014 Clinical and Laboratory Standards Institute. All rights reserved.

Appendix F. (Continued)

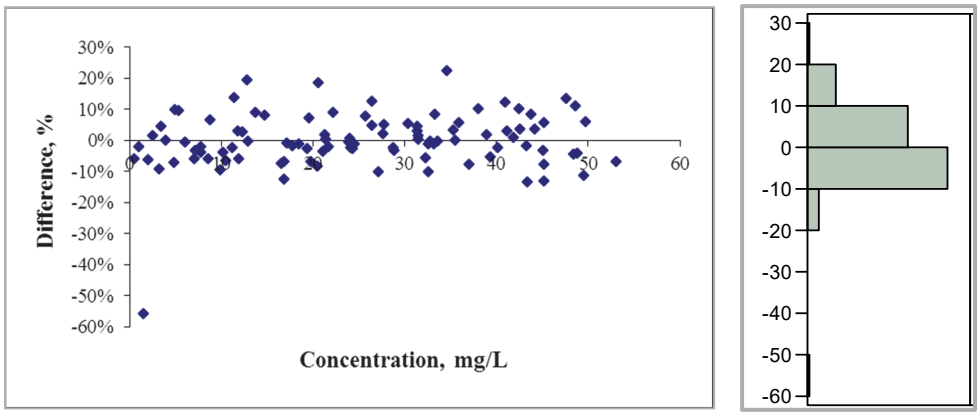


Figure F1. Proportional Difference Plot

F1 Procedure Based on the Normal Approximation to the Binomial

The point estimator is the simple median of the differences as used in this guideline.

$$\hat{\theta} = median(d_i, i = 1, \ldots, N) \tag{F1}$$

For $N = 100$, after ranking these differences in order, the median difference is the average of the differences at positions 50 and 51 (ie, position 50.5).

From the normal approximation to the binomial, an approximate $1-\alpha$ CI for the median is obtained by counting off $1/2N^{1/2}\eta_\alpha$, in which η_α is the upper $1/2\alpha$ significance point of a standard normal variate. For example, for $N = 100$, $\alpha = 0.05$, the approximation is $1/2N^{1/2}\eta_\alpha = 5(1.96) = 9.8$. Rounding out 50.5 ± 9.8, the 96.4% CI is provided by positions 40 and 61. The CI at this percentage can be provided or an estimate of the 95% CI can be obtained through interpolation. The resultant estimates from the data in Table F2 are thus:

$N = 100$
Median = −0.335%
96.4% CI = −2.020% to 1.590% (positions 40 and 61)
95% CI = −1.998% to 1.152% (interpolated)
94.3% CI = −1.990% to 1.000% (positions 41 and 60)

F2 Procedure Based on the Wilcoxon Distribution-Free Signed Rank Test

The Wilcoxon Distribution-Free Signed Rank Test procedure might give somewhat different estimates, but because more differences are included in the calculations, the CI is more robust and therefore perhaps more appropriate for smaller sample sizes (eg, 40 samples).

F2.1 Hodges-Lehmann Point Estimator of θ, $\hat{\theta}$

The Hodges-Lehmann point estimator[1] is given by:

$$\hat{\theta} = median\left(\frac{d_i + d_j}{2}, i \le j = 1, \ldots, N\right) \tag{F2}$$

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix F. (Continued)

Let $W^{(1)} \leq \ldots \leq W^{(M)}$ denote the ordered values $\frac{d_i+d_j}{2}$, or Walsh average.

From the number of pairs of differences $M=N(N+1)/2$, it follows that:

when M is odd $\rightarrow k = \frac{M-1}{2}$, then $\hat{\theta} = W^{(k+1)}$; (F3)

when M is even $\rightarrow k = \frac{M}{2}$, then $\hat{\theta} = \frac{W^{(k)}+W^{(k+1)}}{2}$. (F4)

F2.2 Tukey Two-Sided Confidence Interval for $\boldsymbol{\theta}$, $(\boldsymbol{\theta_L}, \boldsymbol{\theta_U})$:

$$\theta_L = W^{(C_\alpha)} \tag{F5}$$

$$\theta_U = W^{(t_{\alpha/2})} \tag{F6}$$

$$C_\alpha = \frac{N(N+1)}{2} + 1 - t_{\alpha/2} \tag{F7}$$

The position $t_{\alpha/2}$ for various values of n is tabulated in probability tables associated with the Wilcoxon Signed Rank Test statistic distribution[2] and is defined as the value under the null distribution, of the Wilcoxon Signed Rank T statistic, such that $P(T \geq t) = \alpha/2$. That is, when the probability of a value greater than or equal to $T = \alpha/2, t_{\alpha/2} = T$.

$W^{(1)} \leq \ldots \leq W^{(M)}$ are the ordered values of the $\frac{d_i+d_j}{2}$ averages, $1 \leq i \leq j \leq N$, used in computing $\hat{\theta}$. That is, θ_L is the $\frac{d_i+d_j}{2}$ average that occupies the position C_α in the list of M ordered $\frac{d_i+d_j}{2}$s. And θ_U is the $\frac{d_i+d_j}{2}$ average that occupies the position $t_{\alpha/2}$ in the ordered list.

The calculation of $W^{(1-M)}$s is the cumbersome part that needs software for trueness. For example, in an experiment of $N=9$, the number of $W^{(1-M)}$s, M, would be $M = N(N+1)/2 = 45$; for an experiment involving $N=50$, $M=1275$.

For the experiment of $N=50$, the upper bound of the $100(1-\alpha)\%$ CI ($\alpha = 0.0495$) corresponds to $W^{(t_{0.0495/2})}$, which would be located in position $t_{0.0495/2} = 841$, and the lower bound corresponds to $W^{(C_\alpha)}$, which would be located in $C_{0.0495} = 435$.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix F. (Continued)

The following estimates, using the data in Table F2 and the plot in Figure F1, demonstrate this computational technique.

$N = 100$
Median = −0.05%
95% CI = −1.41% to 1.61%

References for Appendix F

1 Hollander M, Wolfe DA. *Nonparametric Statistical Methods.* 2nd ed. New York, NY: John Wiley & Sons, Inc.; 1999.

2 Harter HL, Owen DB, eds. *Selected Tables in Mathematical Statistics, Volume 1.* Providence, RI: American Mathematical Society; 1973.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix G. Ordinary Linear Regression

Abbreviations for Appendix G

OLR ordinary linear regression
SD standard deviation

G1 Correlation

The results of an ordinary linear regression (OLR) analysis are valid only when certain assumptions about the data are true. One of these assumptions is that the *X* variable is known without error. In the medical laboratory, this is not true because every measurement has intrinsic error. However, when the range of measured concentrations is sufficiently wide, the effect of this error on the regression estimates can be considered negligibly small. The correlation coefficient, *r,* can be used as a rough guide to assess the adequacy of the *X* range in overcoming this problem. The formula for *r* is as follows:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \quad \text{(G1)}$$

in which x_i is the best estimate of the concentration of measurand in sample "*i*" from the comparative measurement procedure (average over all replicates from that sample), y_i is a similar estimate using the candidate measurement procedure (average over all replicates for that sample), and:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{(G2)}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} \quad \text{(G3)}$$

Using the underlying assumptions of OLR, a practical rule has been that the range of *X* can be considered adequate if $r \geq 0.975$ (or, equivalently, if $r^2 \geq 0.95$). Under these assumptions, an *r* that satisfies this rule indicates that the error in *X* is adequately compensated by the range of data, and OLR can be used to estimate the slope and intercept. When the data do not fit the assumptions of OLR, this practice is not valid.

NOTE: This procedure assesses the *range* of the data; it does not measure the *distribution* of the data within the measuring interval. One must still obtain an even distribution of data throughout the measuring interval.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix G. (Continued)

G2 Regression Fit

For the set of paired observations (x_i, y_i) the slope (b) and the y-intercept (a) are calculated according to the following formulas:

The regression parameters are:

$$b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{G4}$$

$$a = \bar{y} - b\bar{x} \tag{G5}$$

Thus, the computed line is described by the following equation:

$$\hat{y}_i = a + bx_i \tag{G6}$$

G3 Residuals

The difference, measured in the Y direction, between a given data point and the regression line is called the *residual* for that point. The SD from regression S_{yx} is the SD of these residuals and is thus a measure of the "scatter" of the points around the regression line. The residual for a point (x_i, y_i) can be calculated using the following formula:

$$\text{Residual}_i = y_i - \hat{Y}_i = y_i - (a + bx_i) \tag{G7}$$

The SD from regression is given by performing the following calculations:

$$s_{yx} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2}} \tag{G8}$$

The estimate of the predicted bias ($\hat{B}_c$) at a given medical decision level, X_c, is given by:

$$\hat{B}_c = a + (b-1)X_c \tag{G9}$$

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix G. (Continued)

The 95% confidence interval for $\hat{B}_{\mathrm{c}}$ (the true bias at X_{c}) is given by:

$$[\hat{B}_{\mathrm{c,low}}, \hat{B}_{\mathrm{c,high}}] = \hat{B}_{\mathrm{c}} \pm t(N-2, 0.975) s_{yx} \sqrt{\frac{1}{N} + \frac{(X_{\mathrm{c}} - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}} \ , \qquad \text{(G10)}$$

in which t is the Student t-distribution.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix H. Weighted Least Squares Regression (Weighted Ordinary Linear Regression)

Abbreviations for Appendix H

CI	confidence interval
OLR	ordinary linear regression
SD	standard deviation
SE	standard error
WLS	weight least squares

Constant SD assumption is generally unrealistic for most of the medical laboratory measurement procedure comparisons. The implementation of ordinary linear regression (OLR) in these cases might not be appropriate. The presence of unequal SD is evidenced by the inspection of the difference plots, as explained in Subchapter 5.3 of this guideline. The inspection of residual plots after fitting an OLR can also indicate the presence of unequal SD.

This approach is called "weighted" because it introduces weights that are inversely related to the square of the SD at a particular concentration as:

$$w_i = \frac{1}{\sigma^{2_i}}, \tag{H1}$$

in which σ_i is the SD at that particular concentration.

The SD is often assumed to be proportional to the concentration of the candidate measurement procedure. When such an assumption is made, the weights can be directly determined from the square of this concentration. Alternatively, if a precision profile of the candidate measurement procedure is available, then the SD of each sample and, thus, the weight for each sample can be computed from this profile.

However, SD might or might not be proportional to the concentration, and precision profiles might or might not be available. Repeated measurements on each sample can provide estimates of repeatability at each concentration, but the number of replicates should be large enough to obtain realistic estimates of σ_i.

When information about σ_i is not known, w_i needs to be estimated from the data. The approach described by Neter et al.[1] is presented below.

First, the slope (*b*) and intercept (*a*) are calculated using OLR. Then, residuals are calculated as:

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i) \tag{H2}$$

The absolute value of the residuals is an estimate of SD, as $\sigma_i = |e_i|$. SD function is calculated by regressing σ_i to X_i using OLR. Assuming a proportional relationship between SD and concentration, the linear equation is expressed as:

$$\hat{\sigma}_i = a_\sigma + b_\sigma X_i, \tag{H3}$$

in which a_σ and b_σ are intercept and slope. Use the fitted values from the SD function to estimate the weight as:

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix H. (Continued)

$$\hat{w}_i = \frac{1}{\hat{\sigma}^{2i}} \tag{H4}$$

The weighted average of the reference and candidate measurement procedures is calculated as:

$$\bar{X}_w = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i}, \qquad \bar{Y}_w = \frac{\sum_{i=1}^{N} y_i w_i}{\sum_{i=1}^{N} w_i} \tag{H5, H6}$$

The slope and intercept are calculated as:

$$b_w = \frac{\sum_{i=1}^{N} w_i x_i y_i - \frac{\sum_{i=1}^{N} w_i x_i \sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}}{\sum_{i=1}^{N} w_i x_i^2 - \frac{(\sum_{i=1}^{N} w_i x_i)^2}{\sum_{i=1}^{N} w_i}} \tag{H7}$$

$$a_w = \bar{Y}_w - b_w \bar{X}_w \tag{H8}$$

This process is repeated several times by using the residuals from the weighted least squares (WLS) fit to re-estimate the SD function, and then obtain revised weights. Usually, after three to four iterations the differences between consecutive estimates of slope and intercept are negligible.

The denominator in the formula of the weighted slope is the weighted sum of squares of X:

$$SSXw = \sum_{i=1}^{N} w_i x_i^2 - \frac{(\sum_{i=1}^{N} w_i x_i)^2}{\sum_{i=1}^{N} w_i} \tag{H9}$$

The residuals are calculated using the weighted slope and intercept as:

$$e_i = Y_i - (a_w + b_w x_i) \tag{H10}$$

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix H. (Continued)

The SD of the regression is:

$$s_{yx} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2 w_i}{N-2}} \tag{H11}$$

The standard errors (SEs) of slope and intercept are:

$$\hat{\sigma}_b = \frac{S_{yx}}{SSXw} \tag{H12}$$

$$\hat{\sigma}_a = s_{yx}\sqrt{\frac{1}{\sum_{i=1}^{N} w_i} + \frac{\bar{X}_w^2}{SSXw}} \tag{H13}$$

The $100\,(1-\gamma)\%$ confidence intervals (CIs) for the slope and intercept are:

$$a_w \pm t(N-2, 1-\gamma/2)\hat{\sigma}_a, \text{ and} \tag{H14}$$

$$b_w \pm t(N-2, 1-\gamma/2)\hat{\sigma}_b, \tag{H15}$$

in which $t(N-2, 1-\gamma/2)$ is the $100(1-\gamma)$ percentile of the *t*-distribution with $N-2$ degrees of freedom. Bias at medical decision level X_c is calculated as:

$$\hat{B}_c = a_w + (b_w - 1)X_c \tag{H16}$$

SE of bias is:

$$\hat{\sigma}_{\text{Bias}} = s_{yx} = \sqrt{\frac{1}{\sum_{i=1}^{N} w_i} + \frac{(X_c - \bar{X}_w)^2}{SSXw}} \tag{H17}$$

Assuming the Y_is follow a normal (gaussian) distribution, the $100(1-\gamma)\%$ CI for bias is:

$$\hat{B}_c \pm t(N-2, 1-\gamma/2)\hat{\sigma}_{\text{Bias}} \tag{H18}$$

Data analysis for comparing platelets on two analyzers (candidate vs comparative) is shown for illustration purposes. The shape of the scatter of the difference plot shown in Figure H1 indicates that SD is not constant throughout the measuring interval. The data for this example can be found in Table H2.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix H. (Continued)

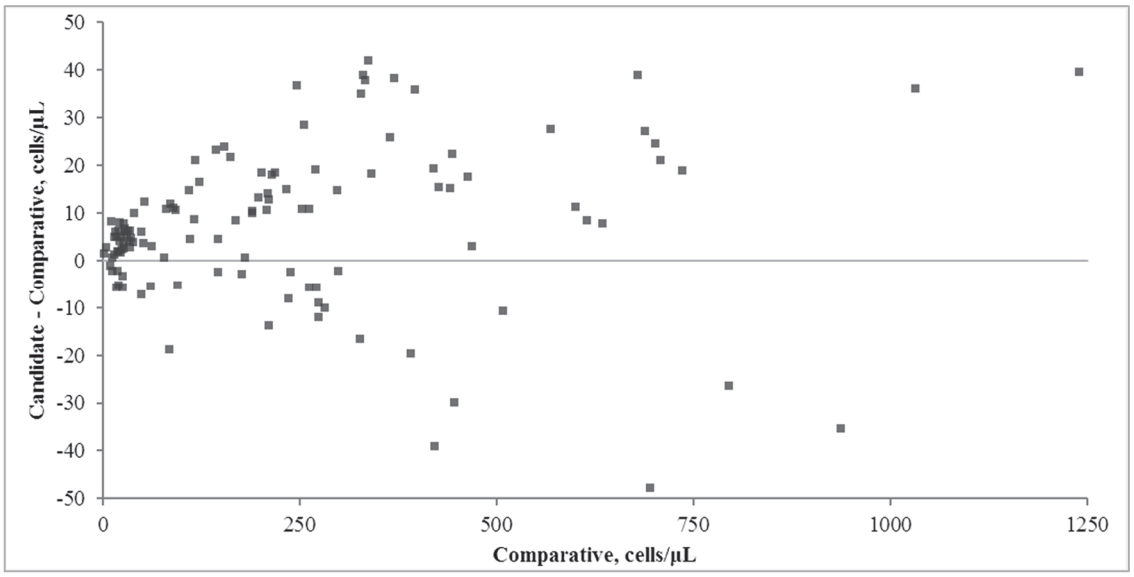


Figure H1. Difference Plots Comparing Platelet Results From Two Analyzers

The estimates of the regression parameter using the WLS approach are shown in Table H1.

Table H1. The Estimates of Regression

	Estimate	SE	Lower	Upper
Intercept	3.013	1.073	0.889	5.138
Slope	1.021	0.007	1.007	1.035
Regression SE	1.222			

Abbreviation: SE, standard error.

It is expected that bias for the candidate measurement procedure is within ± 10 cells/µL or 5% of the values of the comparative measurement procedure throughout the measuring interval. The estimated bias, 95% confidence limits, and specifications are graphically shown in Figure H2. Bias and concentration are shown on the original scale as cells per microliter.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix H. (Continued)

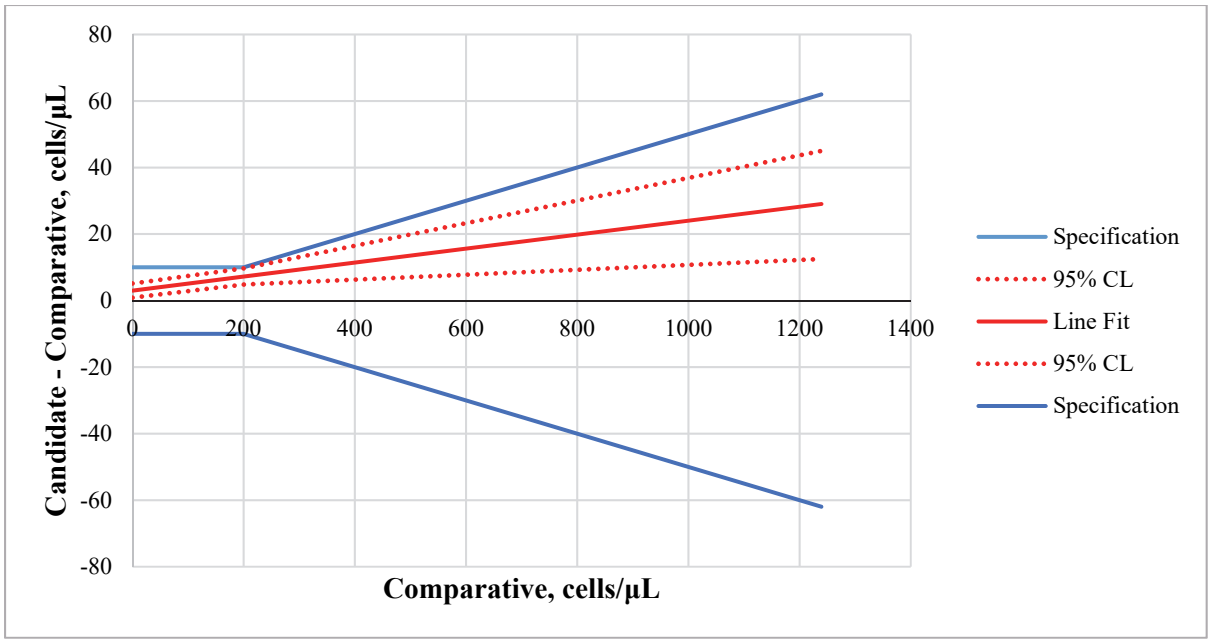


Figure H2. Estimated Bias, Confidence Limits, and Specifications
Abbreviation: CL, confidence limit.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix H. (Continued)

Table H2. Data Comparing Platelet Results From Two Analyzers (A = Comparative, B = Candidate) in cells/µL

Sample	A	B		Sample	A	B		Sample	A	B
1	1.5	3.0		41	60.2	54.8		81	270.1	289.2
2	4.0	6.9		42	61.5	64.6		82	271.3	265.7
3	9.2	8.0		43	78.0	78.6		83	273.5	264.5
4	10.2	18.5		44	80.6	91.4		84	274.2	262.2
5	11.2	9.0		45	84.4	65.7		85	281.1	271.1
6	12.4	13.0		46	85.3	97.2		86	297.0	311.7
7	14.8	19.7		47	89.0	100.0		87	298.7	296.5
8	14.8	16.0		48	92.6	103.2		88	326.7	310.2
9	15.9	21.9		49	94.9	89.6		89	327.1	362.1
10	16.4	10.8		50	108.6	123.4		90	329.6	368.5
11	17.6	22.6		51	110.4	115.0		91	332.8	370.6
12	18.1	15.9		52	115.6	124.4		92	337.4	379.5
13	18.1	20.0		53	116.9	138.1		93	340.1	358.3
14	19.2	14.0		54	122.7	139.2		94	364.8	390.6
15	19.6	25.9		55	143.6	166.8		95	370.1	408.4
16	19.9	21.8		56	146.1	143.7		96	390.6	371.0
17	20.4	24.5		57	146.2	150.8		97	395.7	431.7
18	21.2	29.2		58	154.5	178.5		98	419.3	438.7
19	22.0	27.0		59	161.7	183.4		99	421.3	382.3
20	22.2	24.0		60	167.7	176.1		100	426.3	441.8
21	23.4	25.8		61	176.6	173.7		101	440.4	455.6
22	25.2	22.0		62	179.7	180.4		102	443.4	465.8
23	25.5	19.7		63	188.9	198.9		103	446.2	416.4
24	25.6	33.4		64	189.0	199.4		104	462.7	480.3
25	26.3	30.0		65	197.9	211.1		105	467.7	470.7
26	26.4	28.9		66	201.7	220.1		106	507.4	496.7
27	27.5	34.3		67	207.7	218.3		107	568.3	595.9
28	28.2	34.3		68	209.2	223.4		108	599.6	611.0
29	30.3	35.8		69	210.5	196.8		109	613.8	622.3
30	31.4	37.8		70	210.9	223.8		110	633.5	641.3
31	32.9	37.1		71	214.1	232.2		111	678.6	717.5
32	33.9	40.3		72	218.6	237.1		112	687.6	714.9
33	34.3	37.1		73	232.9	247.9		113	695.1	647.3
34	35.3	40.0		74	235.0	227.0		114	701.0	725.6
35	38.4	42.2		75	237.8	235.3		115	708.3	729.5
36	39.2	49.3		76	246.1	283.0		116	735.6	754.5
37	48.2	41.0		77	252.6	263.5		117	794.8	768.5
38	49.0	55.0		78	254.9	283.5		118	937.0	901.6
39	51.3	55.0		79	261.4	272.3		119	1031.9	1068.0
40	52.2	64.6		80	262.4	256.6		120	1239.3	1279.0

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix H. (Continued)

Reference for Appendix H

[1] Neter J, Kutner MH, Wasserman W, Nachtsheim CJ. *Applied Linear Statistical Models.* 4th ed. Chicago, IL: McGraw-Hill/Irwin; 1996.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix I. Passing-Bablok Regression

The method comparison procedure of Passing and Bablok allows one to describe the linear relationship ($Y = \alpha + \beta X$) between two quantitative measurement procedures. No assumption of normal (gaussian) distribution is needed. The slope of the Passing-Bablok regression line is the adjusted median of all possible slopes of the lines connecting data point pairs. The intercept is the median of intercepts that can be subsequently calculated from the data point pairs using the Passing-Bablok slope just described.

I1 Assumptions

Two measuring systems yielding quantitative values are to be compared. A total of N individual specimens is measured. Let i represent one of the N specimens.

The random variables X and Y represent outcomes of two measurement systems. The random variables can be expressed as the sums of the expected values of their respective distributions and associated measurement errors.

$$x_i = x_i^* + \xi_i \tag{I1}$$

$$y_i = y_i^* + \eta_i \tag{I2}$$

in which x_i^*, y_i^* are the expected values of X and Y, respectively, and ξ_i, η_i are realizations of random error terms. Both come from the same type of distribution. Their variances need not be constant within the sample range, but they should remain proportional.

The following structural relationship can be modeled:

$$y_i^* = \alpha + \beta x_i^* \tag{I3}$$

I2 Estimating Slope and Intercept

Given n measurement pairs (x_i, y_i), all pairs of measurements i, j with $1 \leq i < j \leq n$ are generated. For each pair, the slope S_{ij} of the line connecting the two points is calculated: $S_{ij} = (y_i - y_j)/(x_i - x_j)$. In any pair for which $x_i = x_j$, if $y_i < y_j$ then $S_{ij} = -\infty$, and if $y_i > y_j$, then $S_{ij} = +\infty$. If $x_i = x_j$ and $y_i = y_j$ then S_{ij} is undefined and should be ignored. N is written for the number of S_{ij} remaining after any that are undefined are removed. When there are no undefined S_{ij}, N is $n(n-1)/2$.

The median of the set of possible slopes is a biased estimator of β. Passing and Bablok propose adjusting the median by K, in which K is the number of values of S_{ij} when $S_{ij} < -1$. Ranking the slopes in ascending order $(S_{(\text{rank order})})$ the unbiased estimator of β (b) is given by:

$$b = S_{\left(\frac{N+1}{2}+K\right)} \text{ if } N \text{ is odd and } b = \frac{1}{2}\left(S_{\left(\frac{N}{2}+K\right)} + S_{\left(\frac{N}{2}+1+K\right)}\right) \text{ if } N \text{ is even.} \tag{I4}$$

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix I. (Continued)

The intercept is defined as:

$$a = \text{median}\{y_i - bx_i\} \tag{I5}$$

I3 Confidence Bounds

To calculate confidence bounds for β at the γ level, let $w_{\gamma/2}$ be the $(1-\gamma/2)$ quantile of the standard normal distribution. And let:

$$C_\gamma = w_{\frac{\gamma}{2}}\sqrt{\frac{n(n-1)(2n+5)}{18}} \tag{I6}$$

and

$M_1 = \frac{N-C_\gamma}{2}$ and $M_2 = N - M_1 + 1$, in which M_1 is rounded to the nearest integer. (I7, I8)

Then, a confidence interval at the γ level for β is:

$$S_{(M_1+K)} \leq \beta \leq S_{(M_2+K)} \tag{I9}$$

To create confidence bounds for the intercept, let b_L and b_U represent the lower and upper confidence limits for β, and a_L and a_U represent the lower and upper confidence limits for α, then:

$$a_L = \text{median}\{y_i - b_U x_i\} \tag{I10}$$
$$a_U = \text{median}\{y_i - b_L x_i\}. \tag{I11}$$

The bootstrap technique is also a robust method to estimate the confidence intervals for slope, intercept, and bias estimate at a medical decision point (see Appendix K2).

I4 Discussion

Passing and Bablok note that because the values of S_{ij} are not independent, the median of the $S_{ij}s$ is a biased estimator of β.[1] The authors attempt to account for this bias by using the adjustment K as described above. Different implementations of the Passing-Bablok can get slightly different results. In a follow-up article, the authors discuss the performance of their procedure compared to several others and sample size needs.[2] In another article, Bablok et al.[3] describe modifications of the original procedure discussed above. This amended procedure should not be used for measurement procedure comparison and bias estimations. For additional discussion of the performance of the method, see the original papers.[1,3,4]

References for Appendix I

1 Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. *J Clin Chem Clin Biochem.* 1983;21(11):709-720.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545. Downloaded on 9/4/2018.

Appendix I. (Continued)

[2] Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in clinical chemistry, Part II. *J Clin Chem Clin Biochem.* 1984;22(6):431-445.

[3] Bablok W, Passing H, Bender R, Schneider B. A general regression procedure for method transformation. Application of linear regression procedures for method comparison studies in clinical chemistry, Part III. *J Clin Chem Clin Biochem.* 1988;26(11):783-790.

[4] Linnet K. Evaluation of regression procedures for methods comparison studies. *Clin Chem.* 1993;39(3):424-432.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix J. Profile Weighted Deming Regression

Abbreviations for Appendix J

CV coefficient of variation
R-L Rocke-Lorenzato
SD standard deviation

Most measurement procedures have neither a constant, concentration-independent SD nor a constant CV that hold throughout its measuring interval. When neither a constant nor a proportional SD can be assumed for one or both measurement procedures being compared, there are two options:

- A regression method like the nonparametric Passing-Bablok discussed in Appendix I can be used, making no assumptions about the measurement procedures' random variability

- A parametric regression method can be adopted that draws, for its weights, on a parametric model of imprecision as a function of concentration, determined for each of the measurement procedures across their measuring intervals, or at least across the spectrum of values exhibited in the comparison study

Considerable work relevant to the second option has been published in the medical chemistry literature: work on precision profiles (also known as variance functions) and weighted Deming-like regression methods based on them.[1-12] Unfortunately, this has yet to translate into software applications widely adopted for precision profiling and regression analysis in the medical laboratory community.

Validation studies or the equivalent—means and within-laboratory ("total") SDs for samples in precision studies similar to those in CLSI document EP05,[13] supplemented by the lower limit of quantitation and possibly limit of detection statistics—serve as the basis for constructing a precision profile. Measurement procedure comparison studies involving at least two measurements per measurement procedure per sample provide an alternative, but less desirable, basis.

In general, either the two-parameter Rocke-Lorenzato (R-L) model, Equation J1, or one of the three-parameter Sadler models, eg, Equation J2, should suffice for generating a variance function that does justice to a measurement procedure's concentration-dependent imprecision. (These are fitted to the data by maximum likelihood rather than nonlinear least squares techniques.) The Sadler models (more complex, hence more flexible) were originally developed for immunoassays. The simpler R-L model is likely to suffice for many chemistry assays and for assays with relatively compact measuring intervals.

R-L: $$\text{Variance} = \text{SD}^2 = a + bX^2 \quad \text{(J1)}$$

Sadler: $$\text{Variance} = \text{SD}^2 = (a + bX)^c \quad \text{(J2)}$$

Other models have been proposed and can be suitable. However, models like the "power" function, Equation J3, which imply that imprecision shrinks to nothing at the limit as concentration approaches zero should be avoided, because the implied behavior is unrealistic.

Power: $$\text{CV} = a \bullet X^b \quad \text{(J3)}$$

Intuitively, the R-L model and its variants yield profiles reflecting a gradual change from constant, non-zero SD at low concentrations to proportional SD, ie, constant CV, at high concentrations.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix J. (Continued)

The resulting variance model can be used as shown in Appendix B to provide weights for a Deming regression. Another use is illustrated by Figure 6 in which one panel is appropriate for constant SD and the other for constant CV. The scaled differences are given by:

$$D_i = \frac{(Y_i - X_i)}{s_i}, \tag{J4}$$

in which S_i is the variance-modeled SD of data pair *i* can be plotted against X_i. When the variance model is correct, this plot should have constant width. For example, fitting the R-L model to the dataset of Figure 6 gives the variance model:

$$\mathrm{SD}^2 = 0.0082^2 + (0.0873\mathrm{X})^2 \tag{J5}$$

Using this model to scale the differences gives Figure J1 in which measurement procedure X (MPX) is shown on log scale so that the left end of the range can be seen. The model seems to have captured the variance fairly well as evidenced by the reasonably constant height of the plot.

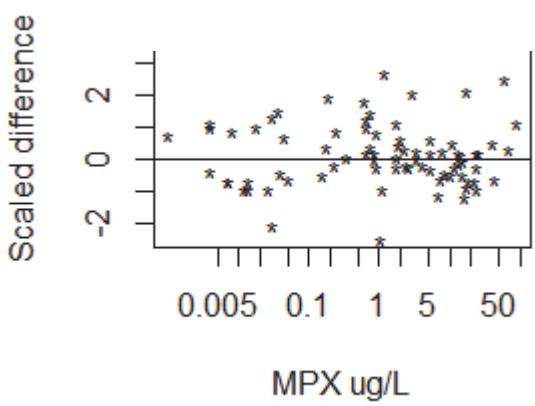


Figure J1. Scaled Difference Plot
Abbreviation: MPX, measurement procedure X.

References for Appendix J

1 Ekins RP. The precision profile: its use in assay design, assessment and quality control. In: Hunter WM, Corrie JET, eds. *Immunoassays for Clinical Chemistry.* 2nd ed. Edinburgh: Churchill Livingstone, 1983:75-105.

2 Munson PJ, Rodbard D. Statistical comparison of two assays methods: weighted linear regression with errors in X and Y. In: Wilson DW, Gaskell SJ, Kemp K, eds. *Quality Control in Clinical Endocrinology: Proceedings of the Eighth Tenovus Workshop, Cardiff, May 1979.* Cardiff: Alpha Omega Publishing Ltd; 1981:89-98.

3 Sadler WA, Smith MH. A reliable method of estimating the variance function in immunoassay. *Comp Stat Data Anal.* 1985;3:227-239.

4 Rocke DM, Lorenzato S. A two-component model for measurement error in analytical chemistry. *Technometrics.* 1995;(2)37:176-184.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix J. (Continued)

[5] Martin RF. General Deming regression for estimating systematic bias and its confidence interval in method-comparison studies. *Clin Chem.* 2000;46(1):100-104.

[6] Sadler WA. Error models for immunoassays. *Ann Clin Biochem.* 2008;45(Pt 5):481-485.

[7] Sadler WA. Joint parameter confidence regions improve the power of parametric regression in method-comparison studies. *Accred Qual Assur.* 2010;15(10):547-554.

[8] Thompson M. Precision in chemical analysis: a critical survey of uses and abuses. *Anal Methods.* 2012;4:1598-1611.

[9] Sadler WA. Australasian Association of Clinical Biochemists (AACB). Regression Accuracy and Power Simulations (Version 4.0). https://www.aacb.asn.au/professionaldevelopment/useful-tools/regression-accuracy-and-power-simulations-version-40. Accessed March 28, 2018.

[10] Hawkins DM. A general variance model in methods comparison. *J Chemometrics.* 2013;27(11):414-419.

[11] Hawkins DM, Weckwerth C. Errors in variables regression with value-censored data. *J Chemometrics.* 2016;30(6):332-335.

[12] Sadler WA. Australasian Association of Clinical Biochemists (AACB). Variance Function Program (Version 14.0). https://www.aacb.asn.au/professionaldevelopment/useful-tools/variance-function-program-version-140. Accessed March 28, 2018.

[13] CLSI. *Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline—Third Edition.* CLSI document EP05-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord-351545. Downloaded on 9/4/2018.

© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix K. Iterative Approaches for Estimating Confidence Intervals for Bias and Regression Parameters

Abbreviations for Appendix K

CI confidence interval
SE standard error

K1 Jackknife Technique

Candidate and comparative measurement procedures are compared on the same N samples. This is referred to as the full set. Deming slope (b), intercept (a), and bias are calculated at medical decision points (B_c) as described in Appendixes A and B. The jackknife technique should not be used for Passing-Bablok regression as described in Appendix I.

Subset 1 is created by removing the first sample only. This subset has $N-1$ samples. Subset 2 is created by removing the second sample only. This process is continued for each ith sample until N subsets with $N-1$ samples are created.

For each subset, the slope (b_i), intercept (a_i), and bias are calculated at medical decision points (B_{ci}) using the formulas previously described.

For each of the parameters above, the deviations of the subsets from the full set and the average deviation are calculated as:

Deviations:

$$\delta_{bi} = Nb - (N-1)b_i \quad \text{(K1)}$$
$$\delta_{ai} = Na - (N-1)a_i \quad \text{(K2)}$$
$$\delta_{Bi} = NB_c - (N-1)B_{ci} \quad \text{(K3)}$$

Average deviation:

$$\bar{\delta}_b = \frac{\sum_{i=1}^{N} \delta_{bi}}{N} \quad \text{(K4)}$$

$$\bar{\delta}_a = \frac{\sum_{i=1}^{N} \delta_{ai}}{N} \quad \text{(K5)}$$

$$\bar{\delta}_{Bc} = \frac{\sum_{i=1}^{N} \delta_{Bci}}{N} \quad \text{(K6)}$$

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix K. (Continued)

Standard errors (SEs) are calculated as:

SE of slope:

$$\hat{\sigma}_b = \sqrt{\frac{\sum_{i=1}^{N}(\delta_{bi} - \bar{\delta}_b)^2}{N(N-1)}} \tag{K7}$$

SE of intercept:

$$\hat{\sigma}_a = \sqrt{\frac{\sum_{i=1}^{N}(\delta_{ai} - \bar{\delta}_a)^2}{N(N-1)}} \tag{K8}$$

SE of bias at medical decision level:

$$\hat{\sigma}_{Bc} = \sqrt{\frac{\sum_{i=1}^{N}(\delta_{Bci} - \bar{\delta}_{Bc})^2}{N(N-1)}} \tag{K9}$$

Alternatively, the SE of the bias at a medical decision level can be computed directly after performing the jackknife iterative calculations on the slope and intercept parameters only by using the covariance formula below.

$$\hat{\sigma}_{ab} = \frac{\sum_{i=1}^{N}(\delta_{ai} - \bar{\delta}_a)(\delta_{bi} - \bar{\delta}_b)}{N(N-1)} \tag{K10}$$

Alternate SE of bias at medical decision level:

$$\hat{\sigma}_{B_c} = \sqrt{\hat{\sigma}_a{}^2 + X_c^2\hat{\sigma}_b{}^2 + 2X_c\hat{\sigma}_{ab}} \tag{K11}$$

For each of these estimates (slope, intercept, bias) the confidence interval (CI) of the estimate is given by:

$$CI(Estimate) = Estimate \pm t(N-2, \alpha)\hat{\sigma}_{Estimate}, \tag{K12}$$

in which *Estimate* is slope (a), intercept (b), or bias (B_c); N is the sample size; α is the significance level (typically 95%); and t is the Student t-distribution critical value.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix K. (Continued)

K2 Bootstrap Technique

Two basic approaches to obtaining bootstrap CIs for regression parameters can be used: percentile method and the standard bootstrap CI approximation.[1] For the percentile CI, the 2.5th and 97.5th percentiles of the bootstrap sample distribution for the parameter of interest are taken as the lower and upper limits of the 95% CI. For the standard bootstrap CI, the empirical estimates of SEs are obtained from bootstrap samples and are subsequently used to create CIs based on the appropriate *t*-distribution critical value. Two bootstrap algorithms are presented because the percentile method provides a nonparametric CI estimate, and the standard bootstrap provides a convenient method to compute CIs for the estimate of bias across many concentration values after completion of the bootstrapping algorithm. Bootstrap techniques can be applied for any regression methodology, including Deming and Passing-Bablok.

NOTE 1: The basic bootstrapping approaches outlined below can result in conservative CIs estimates for the regression intercept and bias estimates near the analytical measuring interval lower limit. More advanced bootstrap algorithms may be used to improve the statistical performance of the CI limits.

NOTE 2: Bootstrapping is a stochastic technique, so users should not expect to observe identical results from one run of the algorithm to the next with the same dataset. However, when greater precision is desired for the CI limits, the number of bootstrap samples (B) can be increased until results reported to the desired degree of rounding are stable from one run of the bootstrap algorithm to the next.

The procedure for implementing bootstrap CIs is:

1. Draw a large number (typically at least B = 1000 for percentile intervals or at least B = 200 for *t* intervals) of random bootstrap samples with replacement from the original set of (*x*,*y*) pairs of observations, eg, (*x*,*y*) • (1), (*x*,*y*) • (2), …, (*x*,*y*) • (B). Here (*x*,*y*) • (*j*) represents the *j*th bootstrap resampling of *N* (x_i, y_i) pairs drawn with replacement from the original *N* observations.

2. For each bootstrap sample (*x*,*y*) • (*j*), estimate the regression parameter(s) of interest. For example, compute the regression intercept and slope.

   $\hat{a}(j)$
   $\hat{b}(j)$

   For the percentile method, also compute the estimate of bias at specified medical decision point concentration value (X_c) for each bootstrap sample, if desired.

   $$\hat{B}_c(j) = \hat{a}(j) + \left(\hat{b}(j) - 1\right)X_c \qquad \text{(K13)}$$

3. For standard bootstrap only, calculate the mean, sample standard deviation, and covariance to compute the SE for each parameter of interest.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Appendix K. (Continued)

Bootstrap parameter estimate means:

$$\bar{b} = \frac{\sum_{j=1}^{B} \hat{b}(j)}{B} \tag{K14}$$

$$\bar{a} = \frac{\sum_{j=1}^{B} \hat{a}(j)}{B} \tag{K15}$$

Bootstrap parameter estimate SEs:

SE of slope:

$$\hat{\sigma}_b = \sqrt{\frac{\sum_{j=1}^{B} [\hat{b}(j) - \bar{b}]^2}{B-1}} \tag{K16}$$

SE of intercept:

$$\hat{\sigma}_a = \sqrt{\frac{\sum_{j=1}^{B} [\hat{a}(j) - \bar{a}]^2}{B-1}} \tag{K17}$$

Covariance between intercept and slope:

$$\hat{\sigma}_{ab} = \frac{\sum_{j=1}^{B} [\hat{a}(j) - \bar{a}][\hat{b}(j) - \bar{b}]}{B-1} \tag{K18}$$

SE of bias at medical decision level:

$$\hat{\sigma}_{B_c} = \sqrt{\hat{\sigma}_a{}^2 + X_c^2 \hat{\sigma}_b{}^2 + 2X_c \hat{\sigma}_{ab}} \tag{K19}$$

4. Compute the CI limits. For the percentile method, calculate the 2.5th and 97.5th percentile of the observed distribution for each parameter of interest across bootstrap samples $j = 1,\ldots,B$ to represent the lower and upper limits of the 95% CI.

   For the standard bootstrap, use equation K20 to obtain 95% confidence intervals for each estimate (slope, intercept, bias).

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Appendix K. (Continued)

$$CI(Estimate) = Estimate \pm t(N-2,\alpha)\hat{\sigma}_{Estimate} \quad \text{(K20)}$$

In equation K20, the *t*-distribution is used to provide a conservative interval and the degrees of freedom ($N-2$) are the number of data points minus the number of parameters being estimated by the regression fit (slope and intercept).

Reference for Appendix K

[1] Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci.* 1986;1(1):54-75.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

This page is intentionally left blank.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

The Quality Management System Approach

Clinical and Laboratory Standards Institute (CLSI) subscribes to a quality management system (QMS) approach in the development of standards and guidelines that facilitates project management, defines a document structure using a template, and provides a process to identify needed documents. The QMS approach applies a core set of "quality system essentials" (QSEs), basic to any organization, to all operations in any health care service's path of workflow (ie, operational aspects that define how a particular product or service is provided). The QSEs provide the framework for delivery of any type of product or service, serving as a manager's guide. The QSEs are:

Organization	Personnel	Process Management	Nonconforming Event Management
Customer Focus	Purchasing and Inventory	Documents and Records	Assessments
Facilities and Safety	Equipment	Information Management	Continual Improvement

EP09 covers the QSE indicated by an "X." For a description of the other documents listed in the grid, please refer to the Related CLSI Reference Materials section.

Organization	Customer Focus	Facilities and Safety	Personnel	Purchasing and Inventory	Equipment	Process Management	Documents and Records	Information Management	Nonconforming Event Management	Assessments	Continual Improvement
		M29				X EP05 EP06 EP07 EP12 EP14 EP15 EP17 EP21 EP28 EP31 I/LA21 I/LA28					

Path of Workflow

A path of workflow is the description of the necessary processes to deliver the particular product or service that the organization or entity provides. A laboratory path of workflow consists of the sequential processes: preexamination, examination, and postexamination and their respective sequential subprocesses. All laboratories follow these processes to deliver their services, namely quality laboratory information.

EP09 does not cover any of the medical laboratory path of workflow processes. For a description of the documents listed in the grid, please refer to the Related CLSI Reference Materials section.

Preexamination				Examination			Postexamination	
Examination ordering	Sample collection	Sample transport	Sample receipt and processing	Examination	Results review and follow-up	Interpretation	Results reporting and archiving	Sample management
EP31		EP31	I/LA28	EP07 I/LA28	EP31 I/LA21	EP14	I/LA28	

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

Related CLSI Reference Materials*

EP05 **Evaluation of Precision of Quantitative Measurement Procedures. 3rd ed., 2014.** This document provides guidance for evaluating the precision performance of quantitative measurement procedures. It is intended for manufacturers of quantitative measurement procedures and for laboratories that develop or modify such procedures.

EP06 **Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach. 1st ed., 2003.** This document provides guidance for characterizing the linearity of a method during a method evaluation; for checking linearity as part of routine quality assurance; and for determining and stating a manufacturer's claim for linear range.

EP07 **Interference Testing in Clinical Chemistry. 3rd ed., 2018.** This guideline provides background information, guidance, and experimental procedures for investigating, identifying, and characterizing the effects of interferents on clinical chemistry test results.

EP12 **User Protocol for Evaluation of Qualitative Test Performance. 2nd ed., 2008.** This document provides a consistent approach for protocol design and data analysis when evaluating qualitative diagnostic tests. Guidance is provided for both precision and method-comparison studies.

EP14 **Evaluation of Commutability of Processed Samples. 3rd ed., 2014.** This document provides guidance for evaluating the commutability of processed samples by determining if they behave differently than unprocessed patient samples when two quantitative measurement procedures are compared.

EP15 **User Verification of Precision and Estimation of Bias. 3rd ed., 2014.** This document describes the estimation of imprecision and of bias for clinical laboratory quantitative measurement procedures using a protocol that can be completed within as few as five days.

EP17 **Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures. 2nd ed., 2012.** This document provides guidance for evaluation and documentation of the detection capability of clinical laboratory measurement procedures (ie, limits of blank, detection, and quantitation), for verification of manufacturers' detection capability claims, and for the proper use and interpretation of different detection capability estimates.

EP21 **Evaluation of Total Analytical Error for Quantitative Medical Laboratory Measurement Procedures. 2nd ed., 2016.** This guideline provides manufacturers and end users with an understanding of concepts related to total analytical error (TAE) for quantitative measurement procedures. An experimental protocol and data analysis method are provided to estimate TAE based upon a comparison of methods experiment with patient specimens, and to assess it relative to a pre-established goal for clinical acceptability.

EP28 **Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory. 3rd ed., 2010.** This document contains guidelines for determining reference values and reference intervals for quantitative clinical laboratory tests.

EP31 **Verification of Comparability of Patient Results Within One Health Care System. 1st ed., 2012.** This document provides guidance on how to verify comparability of quantitative laboratory results for individual patients within a health care system.

I/LA21 **Clinical Evaluation of Immunoassays. 2nd ed., 2008.** This document addresses the need for clinical evaluation of new immunoassays and new applications of existing assays, as well as multiple assay formats and their uses. As a guide to designing and executing a clinical evaluation, this document will aid developers of "in-house" assays for institutional use, developers of assays used for monitoring pharmacologic effects of new drugs or biologics, and clinical and regulatory personnel responsible for commercializing products.

I/LA28 **Quality Assurance for Design Control and Implementation of Immunohistochemistry Assays. 2nd ed., 2011.** This document provides guidelines for the development of validated diagnostic, prognostic, and predictive immunohistochemical assays.

* CLSI documents are continually reviewed and revised through the CLSI consensus process; therefore, readers should refer to the most current editions.

Licensed to: George Odero
This document is protected by Copyright. CLSI order #Ord:351545, Downloaded on 9/4/2018.

Related CLSI Reference Materials (Continued)

M29 **Protection of Laboratory Workers From Occupationally Acquired Infections. 4th ed., 2014.** Based on US regulations, this document provides guidance on the risk of transmission of infectious agents by aerosols, droplets, blood, and body substances in a laboratory setting; specific precautions for preventing the laboratory transmission of microbial infection from laboratory instruments and materials; and recommendations for the management of exposure to infectious agents.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018
© Clinical and Laboratory Standards Institute. All rights reserved.

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

Explore the Latest Offerings From CLSI!

As we continue to set the global standard for quality in laboratory testing, we are adding products and programs to bring even more value to our members and customers.

By becoming a CLSI member, your laboratory will join 1,600+ other influential organizations all working together to further CLSI's efforts to improve health care outcomes. You can play an active role in raising global laboratory testing standards—in your laboratory, and around the world.

Find out which membership option is best for you at **www.clsi.org/membership.**

Find what your laboratory needs to succeed! CLSI U provides convenient, cost-effective continuing education and training resources to help you advance your professional development. We have a variety of easy-to-use, online educational resources that make *e*Learning stress-free and convenient for you and your staff.

See our current educational offerings at **www.clsi.org/education.**

When laboratory testing quality is critical, standards are needed and there is no time to waste. eCLIPSE™ Ultimate Access, our cloud-based online portal of the complete library of CLSI standards, makes it easy to quickly find the CLSI resources you need.

Learn more and purchase eCLIPSE at **clsi.org/eCLIPSE.**

For more information, visit www.clsi.org today.

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

950 West Valley Road, Suite 2500, Wayne, PA 19087 USA

P: +1.610.688.0100 **Toll Free (US):** 877.447.1888 **F:** +1.610.688.0700

E: customerservice@clsi.org **www.clsi.org**

PRINT ISBN 978-1-68440-006-5

ELECTRONIC ISBN 978-1-68440-007-2

Licensed to: George Odero
This document is protected by copyright. CLSI order #Ord-351545, Downloaded on 9/4/2018.