EP5-A
Vol. 19 No. 2
Replaces EP5-T2
February 1999 Vol. 12 No. 4

Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline

This document provides guidance for designing an experiment to evaluate the precision performance of clinical chemistry devices; recommendations on comparing the resulting precision estimates with manufacturer's precision performance claims and determining when such comparisons are valid; as well as manufacturer's guidelines for establishing claims.

NCCLS...

Serving the World's Medical Science Community Through Voluntary Consensus

NCCLS is an international, interdisciplinary, nonprofit, standards-developing and educational organization that promotes the development and use of voluntary consensus standards and guidelines within the healthcare community. It is recognized wordwide for the application of its unique consensus process in the development of standards and guidelines for patient testing and related healthcare issues. NCCLS is based on the principle that consensus is an effective and cost-effective way to improve patient testing and healthcare services.

In addition to developing and promoting the use of voluntary consensus standards and guidelines, NCCLS provides an open and unbiased forum to address critical issues affecting the quality of patient testing and health care.

PUBLICATIONS

An NCCLS document is published as a standard, guideline, or committee report.

Standard A document developed through the consensus process that clearly identifies specific, essential requirements for materials, methods, or practices for use in an unmodified form. A standard may, in addition, contain discretionary elements, which are clearly identified.

Guideline A document developed through the consensus process describing criteria for a general operating practice, procedure, or material for voluntary use. A guideline may be used as written or modified by the user to fit specific needs.

Report A document that has not been subjected to consensus review and is released by the Board of Directors.

CONSENSUS PROCESS

The NCCLS voluntary consensus process is a protocol establishing formal criteria for:

- The authorization of a project
- The development and open review of documents
- The revision of documents in response to comments by users
- The acceptance of a document as a consensus standard or guideline.

Most NCCLS documents are subject to two levels of consensus–"proposed" and "approved." Depending on the need for field evaluation or data collection, documents may also be made available for review at an intermediate (i.e., "tentative") consensus level.

Proposed An NCCLS consensus document undergoes the first stage of review by the healthcare community as a proposed standard or guideline. The document should receive a wide and thorough technical review, including an overall review of its scope, approach, and utility, and a line-by-line review of its technical and editorial content.

Tentative A tentative standard or guideline is made available for review and comment only when a recommended method has a well-defined need for a field evaluation or when a recommended protocol requires that specific data be collected. It should be reviewed to ensure its utility.

Approved An approved standard or guideline has achieved consensus within the healthcare community. It should be reviewed to assess the utility of the final document, to ensure attainment of consensus (i.e., that comments on earlier versions have been satisfactorily addressed), and to identify the need for additional consensus documents.

NCCLS standards and guidelines represent a consensus opinion on good practices and reflect the substantial agreement by materially affected, competent, and interested parties obtained by following NCCLS's established consensus procedures. Provisions in NCCLS standards and guidelines may be more or less stringent than applicable regulations. Consequently, conformance to this voluntary consensus document does not relieve the user of responsibility for compliance with applicable regulations.

COMMENTS

The comments of users are essential to the consensus process. Anyone may submit a comment, and all comments are addressed, according to the consensus process, by the NCCLS committee that wrote the document. All comments, including those that result in a change to the document when published at the next consensus level and those that do not result in a change, are responded to by the committee in an appendix to the document. Readers are strongly encouraged to comment in any form and at any time on any NCCLS document. Address comments to the NCCLS Executive Offices, 940 West Valley Road, Suite 1400, Wayne, PA 19087, USA.

VOLUNTEER PARTICIPATION

Healthcare professionals in all specialities are urged to volunteer for participation in NCCLS projects. Please contact the NCCLS Executive Offices for additional information on committee participation.

Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline

Abstract

Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline (NCCLS Document EP5-A) provides guidance and procedures for evaluating the precision of in vitro diagnostic devices and includes recommendations for manufacturers in evaluating their devices and methods when establishing performance claims. Included are guidelines for the duration, procedures, materials, data summaries, and interpretation techniques that are adaptable for the widest possible range of analytes and device complexity. Experiments and analysis procedures presented in this document are applicable to a wide variety of methods and instrumentation. A balance is created in the document between complexity of design and formulae, and simplicity of operation. Definitions for "between day," "between run," "within run," and "total," when applied to precision, are provided.

(NCCLS. *Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline.* NCCLS document EP5-A [ISBN 1-56238-368-X]. NCCLS, 940 West Valley Road, Suite 1400, Wayne, PA 19087-1898 USA, 1999.)

EP5-A
ISBN 1-56238-368-X
February 1999 ISSN 0273-3099

Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline

Volume 19 Number 2

John W. Kennedy, Chairholder
R. Neill Carey, Ph.D.
Richard B. Coolen, Ph.D.
Carl C. Garber, Ph.D.
Henry T. Lee, Jr.
Jacob B. Levine, M.B.A.
Iris M. Osberg

Suggested Citation

NCCLS. *Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline*. NCCLS document EP5-A (ISBN 1-56238-368-X). NCCLS, 940 West Valley Road, Suite 1400, Wayne, PA 19087-1898, USA, 1999.

Proposed Guideline
December 1981

Tentative Guideline
December 1982

Tentative Guideline—Second Edition
March 1992

Approved Guideline
February 1999

ISBN 1-56238-368-X
ISSN 0273-3099

Committee Membership

Area Committee on Evaluation Protocols

Carl C. Garber, Ph.D. **Chairholder**	**Quest Diagnostics, Incorporated** **Teterboro, New Jersey**
Jan S. Krouwer, Ph.D. **Vice Chairholder**	**Chiron Diagnostics Corporation** **Medfield, Massachusettes**

Subcommittee on Evaluation of Precision

John W. Kennedy **Chairholder**	**Medstat Consultants** **Palo Alto, California**
R. Neill Carey, Ph.D.	Peninsula Regional Medical Center Salisbury, Maryland
Richard B. Coolen, Ph.D.	Ortho-Clinical Diagnostics Rochester, New York
Carl C. Garber, Ph.D.	Quest Diagnostics, Incorporated Teterboro, New Jersey
Henry T. Lee, Jr.	FDA Center for Devices and Radiological Health Rockville, Maryland
Jacob B. Levine, M.B.A.	Bayer Corporation Tarrytown, New York
Iris M. Osberg	Children's Hospital Denver, Colorado

Advisors

Stanley Bauer, M.D.	Beth Israel Medical Center New York, New York
James O. Westgard, Ph.D.	University of Wisconsin Madison, Wisconsin
Donald M. Powers, Ph.D. *Board Liaison*	Ortho-Clinical Diagnostics Rochester, New York
Thomas R. King, M(ASCP) *Staff Liaison*	NCCLS Wayne, Pennsylvania
Patrice E. Polgar *Editor*	NCCLS Wayne, Pennsylvania

ACTIVE MEMBERSHIP (as of 1 JANUARY 1999)

Sustaining Members

Abbott Laboratories
American Association for Clinical Chemistry
Bayer Corporation
Beckman Coulter, Inc.
Becton Dickinson and Company
bioMérieux Vitek, Inc.
College of American Pathologists
Dade Behring Inc.
Ortho-Clinical Diagnostics, Inc.
Pfizer Inc
Roche Diagnostics, Inc.

Professional Members

American Academy of Family Physicians
American Association of Bioanalysts
American Association of Blood Banks
American Association for Clinical Chemistry
American Association for Respiratory Care
American Chemical Society
American Medical Technologists
American Public Health Association
American Society for Clinical Laboratory Science
American Society of Hematology
American Society for Microbiology
American Society of Parasitologists, Inc.
American Type Culture Collection, Inc.
Asociacion Espanola Primera de Socorros
Asociacion Mexicana de Bioquimica Clinica A.C.
Assn. of Public Health Laboratories
Assoc. Micro. Clinici Italiani-A.M.C.L.I.
Australasian Association of Clinical Biochemists
British Society for Antimicrobial Chemotherapy
Canadian Society for Medical Laboratory Science—Société Canadienne de Science de Laboratoire Médical
Canadian Society of Clinical Chemists
Clinical Laboratory Management Association
College of American Pathologists
College of Medical Laboratory Technologists of Ontario
College of Physicians and Surgeons of Saskatchewan
Commission on Office Laboratory Accreditation
Institut für Stand. und Dok. im Med. Lab. (INSTAND)
International Council for Standardization in Haematology
International Federation of Clinical Chemistry
International Society for Analytical Cytology
Italian Society of Clinical Biochemistry
Japan Society of Clinical Chemistry
Japanese Committee for Clinical Laboratory Standards
Joint Commission on Accreditation of Healthcare Organizations
National Academy of Clinical Biochemistry
National Society for Histotechnology, Inc.
Ontario Medical Association Laboratory Proficiency Testing Program
Ordre professionnel des technologistes médicaux du Québec
RCPA Quality Assurance Programs PTY Limited
Sociedade Brasileira de Analises Clinicas
Sociedade Brasileira de Patologia Clinica
Sociedad Espanola de Quimica Clinica
VKCN (The Netherlands)

Government Members

Armed Forces Institute of Pathology
Association of Public Health Laboratory Directors
BC Centre for Disease Control
Centers for Disease Control and Prevention
Chinese Committee for Clinical Laboratory Standards
Commonwealth of Pennsylvania Bureau of Laboratories
Department of Veterans Affairs
Deutsches Institut für Normung (DIN)
FDA Center for Devices and Radiological Health
FDA Division of Anti-Infective Drug Products
Federacion Bioquimica de la Provincia (Argentina)
Health Care Financing Administration
Instituto Scientifico HS. Raffaele (Italy)
Iowa State Hygienic Laboratory
Manitoba Health
Massachusetts Department of Public Health Laboratories
Michigan Department of Public Health
National Association of Testing Authorities - Australia
National Institute of Standards and Technology
National Institutes of Health
Ohio Department of Health
Oklahoma State Department of Health
Ontario Ministry of Health
Saskatchewan Health-Provincial Laboratory
South African Institute for Medical Research
Swedish Institute for Infectious Disease Control

Industry Members

AB Biodisk
Abbott Laboratories
AccuMed International, Inc.
Accumetrics, Inc.
Amersham Pharmacia Biotech
Ammirati Regulatory Consulting
Asséssor

Avecor Cardiovascular, Inc.
Avocet Medical, Inc.
Bayer Corporation - Elkhart, IN
Bayer Corporation - Middletown, VA
Bayer Corporation - Tarrytown, NY
Bayer Corporation - West Haven, CT
Bayer-Sankyo Co., Ltd.
Beckman Coulter, Inc.
Beckman Coulter, Inc. - Palo Alto, CA
Beckman Instruments (Japan) Ltd.
Becton Dickinson and Company
Becton Dickinson Consumer Products
Becton Dickinson Immunocytometry Systems
Becton Dickinson Italia S.P.A.
Becton Dickinson Microbiology Systems
Becton Dickinson VACUTAINER Systems
bioMérieux Vitek, Inc.
Biometrology Consultants
Bio-Rad Laboratories, Inc.
Biotest AG
Bristol-Myers Squibb Company
Canadian Reference Laboratory Ltd.
CASCO●NERL Diagnostics
Checkpoint Development Inc.
Chiron Diagnostics Corporation - International Operations
Chiron Diagnostics Corporation - Reagent Systems
Clinical Lab Engineering
COBE Laboratories, Inc.
Combact Diagnostic Systems Ltd.
Control Lab (Brazil)
Cosmetic Ingredient Review
Cubist Pharmaceuticals
Cytometrics, Inc.
Dade Behring Inc. - Deerfield, IL
Dade Behring Inc. - Glasgow, DE
Dade Behring Inc. - Marburg, Germany
Dade Behring Inc. - Miami, FL
Dade Behring Inc. - Sacramento, CA
Dade Behring Inc. - San Jose, CA
DAKO A/S
Diagnostic Products Corporation
DiaSorin
Eiken Chemical Company, Ltd.
Enterprise Analysis Corporation
Fort Dodge Animal Health
Fujisawa Pharmaceutical Co. Ltd.
Gen-Probe
Glaxo-Wellcome, Inc.
Greiner Meditech, Inc.
Health Systems Concepts, Inc.
Helena Laboratories
Hoechst Marion Roussel, Inc.
Hybritech, Incorporated
Hycor Biomedical Inc.
I-STAT Corporation
Instrumentation Laboratory
Integ, Inc.
International Technidyne Corporation
Kendall Sherwood-Davis & Geck
Labtest Sistemas Diagnosticos Ltda.
LifeScan, Inc. (a Johnson & Johnson Company)
LifeSign, LLC
Lilly Research Laboratories
Medical Device Consultants, Inc.
Medical Laboratory Automation Inc.
MediSense, Inc.
Merck & Company, Inc.
Neometrics Inc.
Nichols Institute Diagnostics (Div. of Quest Diagnostics, Inc.)
Nissui Pharmaceutical Co., Ltd.
Nippon Becton Dickinson Co., Ltd.
Norfolk Associates, Inc.
North American Biologicals, Inc.
OBC Associates
Olympus Corporation
Optical Sensors, Inc.
Organon Teknika Corporation
Ortho-Clinical Diagnostics, Inc. (England)
Ortho-Clinical Diagnostics, Inc. (Raritan, NJ)
Ortho-Clinical Diagnostics, Inc. (Rochester, NY)
Oxoid Inc.
Oxoid LTD (U.K.)
Pfizer Inc
Pharmacia & Upjohn
Procter & Gamble Pharmaceuticals, Inc.
The Product Development Group
Radiometer America, Inc.
Radiometer Medical A/S
David G. Rhoads Associates, Inc.
Rhône-Poulenc Rorer
Roche Diagnostics GmbH
Roche Diagnostics, Inc.
Roche Diagnostic Systems (Div. Hoffmann-La Roche Inc.)
Roche Laboratories (Div. Hoffmann-La Roche Inc.)
The R.W. Johnson Pharmaceutical Research Institute
Sarstedt, Inc.
Schering Corporation
Schleicher & Schuell, Inc.
Second Opinion
SenDx Medical, Inc.
Showa Yakuhin Kako Company, Ltd.
SmithKline Beecham Corporation
SmithKline Beecham (NZ) Ltd.
SmithKline Beecham, S.A.
Streck Laboratories, Inc.
Sysmex Corporation (Japan)
Sysmex Corporation (Long Grove, IL)
Vetoquinol S.A.
Vysis, Inc.
Wallac Oy
Warner-Lambert Company
Wyeth-Ayerst
Xyletech Systems, Inc.
YD Consultant
Zeneca

Trade Associations

Association of Medical Diagnostic Manufacturers
Health Industry Manufacturers Association
Japan Association Clinical Reagents Ind. (Tokyo, Japan)
Medical Industry Association of Australia

Associate Active Members

20th Medical Group (Shaw AFB, SC)
67th CSH Wuerzburg, GE (NY)
121st General Hosptial (CA)
Acadiana Medical Laboratories, LTD (LA)
Advocate Laboratories (IL)
The Aga Khan University Medical Center (Pakistan)
Alabama Reference Laboratory
Allegheny General Hospital (PA)
Allegheny University of the Health Sciences (PA)
Allina Laboratories (MN)
Alton Ochsner Medical Foundation (LA)

Anzac House (Australia)
Associated Regional & University Pathologists (UT)
Baptist St. Anthony's Health Network (TX)
Baystate Medical Center (MA)
Brazileiro De Promocao (Brazil)
Bristol Regional Medical Center (TN)
Brookdale Hospital Medical Center (NY)
Brooke Army Medical Center (TX)
Brooks Air Force Base (TX)
Broward General Medical Center (FL)
Calgary Laboratory Services (Calgary, AB, Canada)
Cardinal Glennon Children's Hospital (MO)
Central Kansas Medical Center
Champlain Valley Physicians Hospital (NY)
Children's Hospital (LA)
Children's Hospital Medical Center (Akron, OH)
Clendo Lab (Puerto Rico)
Colorado Mental Health Institute at Pueblo
Columbia Tulsa Regional Medical Center (OK)
Commonwealth of Kentucky
Community Medical Center (NJ)
CompuNet Clinical Laboratories (OH)
Consolidated Laboratory Services (CA)
Covance CLS (IN)
Danville Regional Medical Center (VA)
Detroit Health Department (MI)
Duke University Medical Center (NC)
Duzen Laboratories (Turkey)
E.A. Conway Medical Center (LA)
East Texas Medical Center
Elmhurst Memorial Hospital (IL)
Emory University Hospital (GA)
Fairview-University Medical Center (MN)
Florida Hospital Alta Monte
Florida Hospital East Orlando
Foothills Hospital (Calgary, AB, Canada)
Fox Chase Cancer Center (PA)
Fresenius Medical Care/Life Chem (NJ)
Grady Memorial Hospital (GA)
Guthrie Clinic Laboratories (PA)
Hacettepe Medical Center (Turkey)
Harris Methodist Fort Worth (TX)
Harris Methodist Northwest (TX)
Hartford Hospital (CT)
Health Alliance Laboratory (OH)
Health Network Lab (PA)
Health Sciences Centre (Winnipeg, MB, Canada)
Hoag Memorial Hospital Presbyterian (CA)
Holmes Regional Medical Center (FL)
Holzer Medical Center (OH)
Hopital de Chicoutimi (Chicoutimi, PQ, Canada)
Hopital Saint Pierre (Belgium)
Hunter Area Pathology Service (Australia)
International Health Management Associates, Inc. (IL)
Intermountain Health Care Laboratory Services (UT)
Jacobi Medical Center (NY)
John Randolph Hospital (VA)
Johns Hopkins Medical Institutions (MD)
Kaiser Permanente (CA)
Kenora-Rainy River Regional Laboratory Program (Dryden, Ontario, Canada)
Klinicni Center (Slovenia)
La Rabida Children's Hospital (IL)
LabCorp (NC)
Laboratoire de Santé Publique du Quebec (Canada)
Lancaster General Hospital (PA)
Langley Air Force Base (VA)
Loma Linda University Medical Center (CA)
Los Angeles County and USC Medical Center (CA)
Louisiana State University Medical Center
Lutheran Hospital (WI)
Main Line Clinical Laboratories, Inc. (PA)
Massachusetts General Hospital
MDS Metro Laboratory Services (Burnaby, BC, Canada)
MDS-Sciex (Concord, ON, Canada)
Med-Chem Laboratories Ltd. (Scarborough, ON, Canada)
Medical Center Hospital (TX)
Memorial Medical Center (LA)
Memorial Medical Center (IL)
Mercy Health System (PA)
Mercy Hospital (NC)
Methodist Hospital (TX)
Methodist Hospital Indiana
Methodist Hospitals of Memphis (TN)
Milton S. Hershey Medical Center (PA)
Mississippi Baptist Medical Center
Monte Tabor-Centro Italo-Brazileiro De Promocao (Brazil)
Montefiore Medical Center (NY)
Montreal Children's Hospital (Canada)
Mount Sinai Hospital (NY)
Mount Sinai Hospital (Toronto, Ontario, Canada)
National Genetics Institute (CA)
Naval Hospital Cherry Point (NC)
Nebraska Health System
New Britain General Hospital (CT)
New England Medical Center Hospital (MA)
The New York Blood Center
New York State Department of Health
New York University Medical Center
NorDx (ME)
North Carolina Laboratory of Public Health
North Coast Clinical Laboratory, Inc. (OH)
North Shore University Hospital (NY)
Northwestern Memorial Hospital (IL)
Ohio State University Hospitals
Olin E. Teague Medical Center (TX)
Our Lady of Lourdes Hospital (NJ)
Our Lady of the Resurrection Medical Center (IL)
Permanente Medical Group (CA)
Providence Health System (OR)
Providence Medical Center (WA)
Queen Elizabeth Hospital (Prince Edward Island, Canada)
Queensland Health Pathology Services (Australia)
Quest Diagnostics (PA)
Quest Diagnostics Incorporated (NJ)
Quintiles Laboratories, Ltd. (GA)
Regions Hospital
Research Medical Center (MO)
Riyadh Armed Forces Hospital (Saudi Arabia)

Saint Mary's Regional Medical Center (NV)
St. Alexius Medical Center (ND)
St. Anthony Hospital (CO)
St. Boniface General Hospital (Winnipeg, Canada)
St. Francis Medical Center (CA)
St. John Hospital and Medical Center (MI)
St. John Regional Hospital (St. John, NB, Canada)
St. Joseph Hospital (NE)
St. Joseph's Hospital - Marshfield Clinic (WI)
St. Luke's Hospital (PA)
St. Luke's Regional Medical Center (IA)
St. Luke's-Roosevelt Hospital Center (NY)
St. Mary Hospital (NJ)
St. Mary Medical Center (CA)
St. Mary of the Plains Hospital (TX)
St. Vincent's Hospital (Australia)
San Francisco General Hospital (CA)
Seoul Nat'l University Hospital (Korea)
Shanghai Center for the Clinical Laboratory (China)
Shands Healthcare (FL)
SmithKline Beecham Clinical Laboratories (GA)
South Bend Medical Foundation (IN)
South Western Area Pathology Service (Austraila)
Speciality Laboratories, Inc. (CA)
Stanford Health Services (CA)
Stormont-Vail Regional Medical Center (KS)
Sun Health-Boswell Hospital (AZ)
Sunrise Hospital and Medical Center (NV)
Sutter Health (CA)
Timmins & District Hospital (Timmons, ON, Canada)
The Toledo Hospital (OH)
Tri-City Medical Center (CA)
Tripler Army Medical Center (HI)
Trumbull Memorial Hospital (OH)
Tulane Medical Center Hospital & Clinic (LA)
Twin Lake Regional Medical Center
UCSF Medical Center (CA)
UNC Hospitals (NC)
Unilab Clinical Laboratories (CA)
United Clinical Laboratories (IA)
University of Alabama - Birmingham Hospital
University of Alberta Hospitals (Canada)
University of Chicago Hospitals (IL)
University Hospital (IN)
University Hospital (Gent) (Belgium)
University Hospital (London, Ontario, Canada)
University Hospital of Cleveland (OH)
The University Hospitals (OK)
University of Medicine & Dentistry, NJ University Hospital
University of Michigan
University of the Ryukyus (Japan)
University of Texas Medical School at Houston
University of Virginia Medical Center
University of Washington
UPMC Bedford Memorial (PA)
USAF Medical Center (OH)
UZ-KUL Medical Center (Belgium)
VA (Albuquerque) Medical Center (NM)
VA (Dayton) Medical Center (OH)
VA (Denver) Medical Center (CO)
VA (Indianapolis) Medical Center (IN)
VA (Kansas City) Medical Center (MO)
VA Outpatient Clinic (OH)
VA (Tuskegee) Medical Center (AL)
Viridae Clinical Sciences, Inc. (Vancouver, BC, Canada)
ViroLogic, Inc. (CA)
ViroMed Laboratories, Inc. (MN)
Waikato Hospital (New Zealand)
Walter Reed Army Institute of Research (MD)
Warde Medical Laboratory (MI)
Warren Hospital (NJ)
Washoe Medical Center (NV)
Watson Clinic (FL)
William Beaumont Hospital (MI)
Williamsburg Community Hospital (VA)
Wilford Hall Medical Center (TX)
Wilson Memorial Hospital (NY)
Winchester Hospital (MA)
Winn Army Community Hospital (GA)
Yonsei University College of Medicine (Korea)
York Hospital (PA)
Zale Lipshy University Hospital (TX)

Contents

Foreword

Current clinical chemistry literature contains numerous examples of product evaluations. Many of these use the basic concepts that are included in this guideline. While more complex and customized experimental designs have been used for both published studies and regulatory purposes in special cases, there still appears to be a strong need in the clinical community for the basic approaches to quantitative precision assessment to be described along with their rationales.

In order to address this need, the subcommittee has drawn on the experience of users, representatives of industry, statisticians, chemists, laboratory personnel, and medical personnel for developing this guideline. The extremely wide variety of in vitro diagnostic devices currently available made it apparent to us that a single experimental design would not be appropriate for all devices. Therefore, we have constructed this guideline to give primarily conceptual guidance on the duration, procedures, materials, data summaries, and interpretation techniques that would be adaptable for the widest possible range of analytes and device complexity. We have tried to illustrate each step of the evaluation with an example of a typical experimental design.

At each step in developing this protocol, we chose carefully among the many recommendations for duration, inclusion of quality control, and methods of determining the components of precision. We have tried to create a balance in the document between complexity of design and formulae, and simplicity of operation. We have included an appendix (Appendix C) that provides guidelines for modifying the design and calculations when appropriate.

The earlier editions of EP5 were widely reviewed by the clinical laboratory testing community and generated a variety of remarks. The subcommittee thanks all for their recommendations. We carefully reviewed each comment and made changes in the document where appropriate. Not all viewpoints could be accommodated, however. Comments on EP5-T2 and subcommittee responses are included at the end of the this document. Review and comment on this edition is encouraged.

Key Words

Evaluation protocol, experimental design, medical devices, outlier, precision, quality control.

Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline

1 Introduction

1.1 Purpose

This document provides guidelines to design an experiment to evaluate the precision performance characteristics of clinical chemistry devices. In many cases, these techniques can be used in areas other than chemistry.

When using a modification of the device, a user needs to verify that essential performance characteristics of the device have not changed. Comparison to original claimed precision performance may not be valid. Examples of typical modifications are the use of reagents, specimen sources, calibrating or control materials or operating procedures that are different from those stated in the manufacturer's labeling (instructions for use).

1.2 Overview of the General Precision Evaluation Experiment

1.2.1 General Guidelines

Proper evaluation of an analytical device requires:

- Sufficient time to become familiar with the mechanics of operation and maintenance of the device according to the manufacturer's instructions;
- Sufficient time to become familiar with the steps of the evaluation protocol;
- Maintenance of the device in proper quality control during the entire period of the evaluation;
- Sufficient data and an appropriate experiment that is of sufficient duration to generate adequate samples. (Data and experiment duration are critical in that precision estimates should have a sufficient number of degrees of freedom. This should properly reflect long-term performance of the device during a period of routine workload in the laboratory or for the customers); and
- Statistically valid data analysis procedures.

How "sufficient data" is defined will depend on the ultimate use for the data and how well the precision of the device is determined.

1.2.2 Device Familiarization Period

The first step is to become familiar with the device, and with all aspects of set-up, operation, maintenance, and other factors in its routine laboratory use. Users can do this after or concurrently with the training period suggested by the manufacturer.

1.2.3 Protocol Familiarization Period

The first five operating days of the precision evaluation experiment should be used to become familiar with the experimental protocol itself. While practicing the experiment, any serious problems with the device should be detected, data that may be usable at the end of the experiment collected, and preliminary acceptability tests for precision and for other performance characteristics that are not addressed in these guidelines, such as linearity, drift, etc. can be performed. A quality control program must be maintained to ensure that the results represent true performance.

1.2.4 Precision Evaluation Experiment

Once familiarity with the device is achieved, the precision evaluation experiment can be started. A minimum of 20 operating days is recommended for the precision evaluation experiment. Because day-to-day precision may be a large component of the total precision in analytic performance, performance must be evaluated long enough to ensure that total precision is adequately estimated.

During each of the testing days, two separate runs (when “runs” are an important component of the target device operating procedure) with two test samples at each of at least two levels of analyte concentration should be analyzed. In addition to the test samples, at least one quality control sample in each run should be analyzed. Use of the laboratory’s routine quality control procedures and materials (if appropriate) on the

device during the evaluation is recommended. If “runs” do not constitute an aspect of the device under consideration, then the four samples at each level should be analyzed throughout the day.

1.2.5 Completing the Precision Experiment

After this protocol familiarization period, the experiment should be continued for 15 more days. At the end of each five operating days, the control limits on a set of quality control charts should be recalculated and all data checked for acceptability. If this process identifies outliers, every attempt should be made to determine the cause of the problem. Data may not be rejected without valid justification, since this will lead to understatement of precision performance. When the precision experiment is completed, the appropriate statistical calculations on the data are performed. If the evaluator determines that a “learning curve” affected data during the protocol familiarization period, this data may be excluded and replaced with an equal amount of data collected at the end of the originally scheduled evaluation period.

1.2.6 Comparison to Other Precision Evaluations

Other methods still sometimes used for evaluating precision consist of a single run of 20 observations (for within-run) and still sometimes single observation (or just a few) at a given concentration each day for 10 or 20 days for total imprecision (usually incorrectly calculated and erroneously labeled day-to-day precision). This method has serious drawbacks, and is specifically not recommended in this protocol.

When a single run is used to estimate within-run imprecision, there is a significant risk that the operating conditions in effect at the time of that single run may not reflect usual operating parameters, thus adversely affecting the estimate. Furthermore, there is no way to determine how representative of expected performance that single run may be. For this reason, this document recommends that within-run precision be estimated by "pooling" the within-run performance over many runs, thus insuring a more robust and representative estimate that should extrapolate to future performance under a variety of routine conditions.

For total imprecision, while the procedure recommended herein requires some cumbersome calculations, it will be independent of the number of days and runs within a day used to estimate it (which traditional methods are not); it correctly combines the effect of within-run, between-run, and between-day components of precision (which will vary in relative size from method to method), and avoids the errors in using incorrect terms (such as "day-to-day") for total imprecision.

Lack of a standard for how to adjust the correct statistical estimate of total precision for the number of observations per run and per day have, in the past, led to a multitude of misleading performance estimates by both manufacturers and laboratorians.

1.3 Statistical Power of Precision Estimates

1.3.1 Precision and Confidence

When designing an evaluation experiment, it must be decided beforehand how well the "true" precision of the device is to be determined. Each time a certain precision protocol is run, an estimate of the "true" precision of the device is obtained. When this same protocol is rerun in the same laboratory with a device that is in control, a different estimate of the precision will result even though the "true" precision is the same.

These estimates of precision might be expected to scatter around the "true" precision, and the estimates obtained from more observations to cluster more closely around the "true" precision. In a sense, more "confidence" in an estimate is based on a larger number of observations. The more "confidence" in an estimate, the more "statistical power" to detect performance that is different from the claim.

1.3.2 Statistical Comparison with Manufacturer

With this precision evaluation experiment, estimates can be compared for within-run and total precision with those from the manufacturer. The statistical power of such a

comparison can be calculated, that is, how much the estimates statistically differ from claimed performance based on the number of degrees of freedom of these estimates.

This extremely important concept can be used to illustrate that an estimate of within-run precision based on 100 degrees of freedom can detect relatively small deviations from claimed performance. Likewise, an estimate of within-run precision based on only, for example, 10 degrees of freedom will detect only major departures from claimed performance and thus, a test based on such an estimate has low statistical power.

If the estimate has 40 degrees of freedom, there is a greater statistical power and the estimate can detect smaller, though still clinically important, departures from claimed performance. This is an important aspect in the design of any evaluation experiment.

1.4 Standard Precautions

Because it is often impossible to know which might be infectious, all patient blood specimens are to be treated with standard precautions. For specific precautions for preventing the laboratory transmission of bloodborne infection from laboratory instruments and materials; and recommendations for the management of bloodborne exposure, refer to NCCLS document M29—*Protection of Laboratory Workers from Instrument Biohazards and Infectious Disease andTransmitted by Blood, Body Fluids, and Tissue*.

2 Device Familiarization Period

2.1 Purpose

The operation, maintenance procedures, methods of sample preparation, and calibration and monitoring functions required must be learned. Many manufacturers of clinical chemistry devices provide operator training. The device should be set up and operated in the individual laboratory long enough to understand all of the procedures involved to avoid problems during the actual evaluation of its performance. Analyzing actual sample material, including pools, controls, leftover serum (if appropriate), or any other test materials appropriate for the device should be practiced.

All possible contingencies (such as error flags, error correction, calibration, etc.) that might arise during routine operation should be carefully monitored. Data should not be collected during this period. The device familiarization period is not complete until the user can demonstrate that he/she can operate the device properly.

2.2 Duration

A five-day familiarization period is adequate for most devices. A shorter or longer period may be appropriate, depending on the complexity of the device and the skill level of the operator.

3 Protocol Familiarization Period (Users and Manufacturers)

3.1 Purpose

An evaluation experiment often involves steps not ordinarily encountered during routine laboratory conditions. To keep these unfamiliar steps from adversely affecting the results of the evaluation experiment, the experiment should be practiced for some time before starting the protocol. Use of this period will ensure understanding of the protocol. The experiment should be run as described in the next section using the regular test materials and quality control materials in the laboratory.

3.2 Duration

This protocol familiarization should be continued until data is obtained without operational difficulty for a minimum of five operating days. This period can be extended as necessary for complex devices.

3.3 Use of Data

The data collected without operational difficulty during those five or more days should be incorporated into the estimation of precision along with data collected during subsequent operation of the protocol, if in the opinion of the evaluator this data is consistent with subsequent data. All data should be subjected to quality control acceptability checks as described below.

3.4 Quality Control Procedures

When estimating precision, it should be assumed that the device is operating in a stable condition while collecting the data. To justify this assumption, the performance of the device with quality control samples should be monitored. During the protocol familiarization period, routine quality control procedures on the device should be used. The trial control limits should be calculated after completing this phase of data collection. If these trial control limits do not reasonably agree with the manufacturer's performance claims for the device, the manufacturer should be contacted before the experiment is continued.

3.5 Additional Evaluations

While practicing the experiment, other features of the device can be checked. Linearity, recovery, or any other feature not discussed in these guidelines can be tested. These tests should be used to see if there are any serious problems with the device. If there are any problems, the manufacturer should be contacted to determine the cause of the problem. The decision of whether the device is acceptable should not be made solely on the basis of these limited, preliminary tests.

3.6 Preliminary Precision Test

At or near the end of the protocol familiarization period, an initial test of within-run precision should be conducted. Twenty aliquots of an appropriate test material (or a complete "batch" if less than 20) should be assayed in sequence. The standard deviation and coefficient of variation of the results should be calculated. If a considerable discrepancy from expected results is found, the manufacturer should be contacted and no further testing should be conducted until the problem is solved. It should be emphasized that this single run test *is not* sufficient to judge the acceptability of the device. It can only identify problems that should be solved before continuing the evaluation. This data is used only for this one-time verification.

4 Precision Evaluation Experiment

4.1 Components of Precision

The objective of the precision evaluation experiment is to estimate the total precision of the device. Intuitively, total precision is the variability of the device when used over an indefinitely long period. To some degree, several sources of variability contribute to this long-term precision. Generally, it is sufficient to design the experiment so that all these sources will influence the total precision estimate without trying to determine the relative size of each source or component. Terms used to describe the time-related components of total precision include:

- Within-run
- Between-run
- Within-day
- Day-to-day (also called between-day).

Of these, the within-run precision component and the total precision are generally of most interest. The experiment described in this section was designed to provide estimates of the total precision and within-run precision of the device during operation in the laboratory. It was not attempted to incorporate in this experiment separate estimates of other possibly significant sources of variability such as calibrator or reagent lot differences or technologist/operator differences, but suggested that manufacturers include such factors. Included, but not estimated individually, are other factors that influence precision (e.g., sample preparation, test material stability, carryover, and drift; refer to NCCLS document EP10—*Preliminary Evaluation of Quantitative Clinical Laboratory Methods*).

4.2 Reagents and Calibration Materials

A single lot of reagents and calibration materials may be used for the entire protocol, but interpretation (and explicit labeling, when appropriate) of results must include this fact, and results may underestimate true long-term total imprecision. Introducing several lots of these materials will increase the observed variability, and although the experiment does not allow for separately estimating the effects of these factors as described herein, may better

represent the real precision performance of the device.

4.3 Test Materials

4.3.1 Matrix

The test materials should be selected to simulate the characteristics of the appropriate clinical specimens. Stable frozen pools are preferred when appropriate and possible. When necessary, stable, commercially avai-lable, protein-based materials may be used.

4.3.2 Concentrations

Test materials should be chosen carefully by considering several criteria. Two concen-trations are recommended, although more may be used.

Concentrations that span a significant portion of the analytic (reportable) range of the device whenever possible should be selected. If more than two concentrations are available, additional concentrations as close as possible to the "medical decision levels" used in the laboratory should be chosen. To compare evaluation results to published performance claims, concentrations that correspond to the levels in those claims should be chosen.

4.4 Number of Runs and Days

4.4.1 General Guidelines

The experiment and calculations described in this document are one example of an evaluation design. This experiment and its calculations are an example of a balanced design (a fully nested Model II ANOVA), which is appropriate for most clinical chemistry systems and devices. Other designs may be more appropriate for specific systems, but the required calculations and statistical interpre-tations will be different.

The precision evaluation experiment requires a sufficient amount of data so that the estimates of precision properly reflect the true precision parameters of the device. A minimum of 20 acceptable operating days is generally neces-sary to achieve this result, except in situations where this is known not to be a factor. During the first five days of the experiment, the user should become familiar with the protocol as described in Section 3.0.

A short-run method has a run duration of less than 2 hours, while a long-run method (such as RIA) has a considerably longer "run," generally done once per shift. For long-run methods, the one run per day procedure in Appendix C should be used; for short-run procedures, the evaluation samples may be tested anywhere in the run. For the purposes of the analysis of variance, an evaluation run is a discrete time period of data collection designed to enable the estimation of variability (or drift) within a day. For some devices, such as random-access, discrete, or unitary devices, the concept of a “run” may not be appropriate. In this case, samples should be run randomly throughout a working shift to simulate the actual operation of the device.

4.4.2 Specific Procedures

See sections 1.2.2 (Device Familiarization Period) and 1.2.3 (Protocol Familiarization Period) for initial steps in the evaluation process.

The following steps shall be taken within each day:

(1) Analyze two runs or batches.

(2) If a run must be rejected because of quality control procedures or operating difficulties, conduct an additional run *after* an investi-gation is conducted to identify and correct the cause of the problem.

(3) Within each run or batch, analyze two aliquots of test material for each concen-tration used.

(4) Include in each run the quality control samples ordinarily used to judge the acceptability of the run or day.

(5) Change the order of analysis of test materials and quality control samples for each run or day.

(6) To simulate actual operation, include at least ten patient samples in each run whenever possible.

(7) Separate the runs performed each day by a minimum of two hours.

4.5 Recording the Data

Appendix A contains examples of data recording sheets to summarize data. This type of summary is valuable in the statistical analysis described below. If the number of runs, days, or observations is changed, a similar sheet should be created, the resulting data transcribed onto it, and the necessary calculations adjusted accordingly.

4.6 Quality Control Procedures

4.6.1 General Guidelines

Normal quality control procedures should be conducted during the precision evaluation experiment. At least one quality control sample at an appropriate concentration in each run should be included. If two or more concentrations for quality control are ordinarily used, this method should be continued throughout the evaluation experiment.

4.6.2 Statistical Quality Control Charts

Preliminary statistical quality control charts should be set up for the device at the end of the protocol familiarization period (i.e., the first five acceptable days of the precision data collection period). The following procedure should be followed:

(1) Calculate the center lines, warning limits, and out-of-control limits from these initial data according to usual practices.

(2) Plot all subsequent quality control data on the charts.

(3) If at any point an out-of-control condition is detected, determine the cause, eliminate the offending point, and then repeat the run. It is suggested since there is low statistical power with these preliminary estimates, that ±3 S.D.s be used as indications for investigation, and ±4 S.D.s be used for rejection.

(4) After each of the five days of data collection, recalculate the center lines and control limits of each chart from all acceptable data collected thus far

(5) If the previously acceptable results are now unacceptable, continue the precision experiment to obtain the proper number of days

(6) Maintain a record of the number of rejected runs.

4.7 Detection of Outliers

A detection criterion for outliers must be defined to use during the precision evaluation experiment. The detection criterion is needed to be certain that operational problems will not unduly distort the resulting data and precision estimates.

Assuming appropriate quality control procedures will be used during the experiment, a fairly weak (low power) test is suggested to detect gross outliers in the data. The outlier test is derived from the data collected during the preliminary precision test. Data collected during each run of the precision evaluation experiment are in pairs (duplicates). The following test should be used:

(1) If the absolute value of the difference between the replicates exceeds 5.5 times the standard deviation determined in the preliminary precision test (see Section 3.6), the pair should be rejected.

(2) If such an outlier is found, the cause of the problem should be investigated, and the run repeated for that analyte. The value 5.5 is derived from the upper 99.9% value of the Studentized range. **NOTE:** This test should be used when the concentration of the preliminary test material is reasonably close to the concentration of the evaluation test material.

The evaluator may wish to schedule additional days of evaluation at the outset of the investigation, to allow for potential run rejections, if needed. If more that 5% of the runs need to be rejected and no assignable cause can be found, then the investigator should consider the possibility that the device is not sufficiently stable to allow a valid variability assessment.

4.8 Statistical Computations for Precision

After collecting the data and transcribing them onto an appropriate recording sheet, the calculations described in this section should be performed. A sample completed recording sheet can be found in Appendix B, along with the associated calculations. Separate calculations should be performed for each concentration, and all data checked against the outlier criteria described in Section 4.7.

4.8.1 Within-Run Precision Estimate

The estimate of within-run precision is derived from the following formula:

$$S_{wr} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{2}(X_{ij_1} - X_{ij_2})^2}{4I}} \qquad (1)$$

where:

I = total number of days (generally 20)
j = run number within day (1 or 2)
X_{ij_1} = result for replicate 1, run j on day i
X_{ij_2} = result for replicate 2, run j on day i.

Two results are needed on each of two runs for every day to use the above formula. If only one run is available on a given day, that run should not be used with this formula. See Appendix C for formulas to use when there is only one run on each day. As long as there are no more than 10% of the evaluation days with missing runs (i.e., only one run) in the two-run-per-day experiment, the resulting statistical calculations will be valid.

4.8.2 Total Precision Estimates

Several quantities are required to determine total precision estimates. The calculations below will be needed:

$$A = \sqrt{\frac{\sum_{i=1}^{I}(\bar{X}_{i_1\cdot} - \bar{X}_{i_2\cdot})^2}{2I}} \qquad (2)$$

where:

I = number of days (with two runs)
$\bar{X}_{i_1\cdot}$ = average result run 1, day i (average of the two replicates)
$\bar{X}_{i_2\cdot}$ = average result run 2, day i (average of the two replicates).

A is calculated by squaring the difference between the first run analysis average and the second run analysis average for each day, summing up these quantities for all days, dividing by 2I and taking the square root. This calculation data from a day with only one run should not be included.

The second quantity is:

$$B = \sqrt{\frac{\sum_{i=1}^{I}(\bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot\cdot\cdot})^2}{I - 1}} \qquad (3)$$

where:

I = number of days
$\bar{X}_{i\cdot\cdot}$ = average of all results day i
$\bar{X}_{\cdot\cdot\cdot}$ = average of all results.

This is the standard deviation of the daily means, as identified in Data Sheet #3 in Appendix A.

The following is calculated as:

$S^2_{dd} = B^2 - A^2/2$
$S^2_{rr} = A^2 - S^2_{wr}/2$

(set to 0 if negative).

Setting the (possibly) negative variance components to zero follows a widely used convention in statistics. If these calculations are performed with a software package, adherence to the above convention should be assured.

The estimate of total precision is then calculated with the following standard deviation formula:

$$S_T = \sqrt{S^2_{dd} + S^2_{rr} + S^2_{wr}} \qquad (4)$$

A different result will be obtained from this formula for S_T from that obtained by calculating the standard deviation of all data observed (without regard to day or run). The above formula is the correct way to estimate the total

precision SD because it properly weights the between-day, between-run, and within-run components. The coefficient of variation corresponding to this estimate of total precision standard deviation should be calculated by dividing S_T by the analyte concentration of the test material and multiplying by 100. The result should be expressed as a percentage.

NOTE: Use these quantities A and B to estimate within-day and day-to-day components of precision. (See Appendix C.)

4.9 Comparison with Manufacturer's Claims or Other Performance Criteria

The precision estimates obtained in the previous section should be compared to performance claims for the precision of the device. The chi-square (X^2) statistic as described below should be used. To use this method, the performance claim is expressed as a point estimate (i.e., a standard deviation). The within-run and total precision estimates should be compared separately.

4.9.1 Within-Run Precision Comparison

The performance claim standard deviation σ_{wr} should be denoted. The chi-square test uses the square of both the user's and manufacturer's estimates of within-run precision. The number of "degrees of freedom" associated with S_{wr}^2 (the user's estimated within-run variance) must be known to perform this test. In the experiment described in this protocol, S_{wr}^2 will have as many degrees of freedom as there were data pairs (replicates within runs) used to calculate it. Thus, this will be equal to the number of runs during the experiment, which will be denoted R below. The test involves calculating the following:

$$X^2 = \frac{S_{wr}^2 \cdot R}{\sigma_{wr}^2} \qquad (5)$$

where:

S_{wr}^2 = square of the user's estimated within-run standard deviation

σ_{wr}^2 = square of the manufacturer's claim of within-run standard deviation

R = the total number of runs (degrees of freedom for S_{wr}^2).

The calculated X^2 should be compared with a statistical table of X^2 values, using the upper 95% critical value with R degrees of freedom (see Table 1). If the calculated value is less than this table value, then the estimate is not significantly different from the claimed value, and this part of the precision claim is accepted.

NOTE: The estimate may be larger than the manufacturer's claim, and still *not* be significantly different.

4.9.2 Comparison of Total Precision Estimate

A chi-square test similar to that described above should be used to compare the estimate of total precision to that claimed by the manufacturer, or to that required by the medical application at the user's institution. Unlike the within-run estimate, however, computing the exact number of degrees of freedom for S_T involves a complicated calculation. Because of the structure of the protocol, the user cannot assume that all observations are independent, a necessary assumption before the customary estimate for degrees of freedom (total number observations - 1 can be used). The formula below for T degrees of freedom for S_T takes into account this lack of independence.

Let:

$ME = S_{wr}^2$ (mean square for within-run)
$MR = 2A^2$ (mean square for runs)
$MD = 4B^2$ (mean square for days)

where:

S_{wr}, A and B are defined in Section 4.7(2). Then,

$$T = \frac{I(2ME + MR + MD)^2}{2ME^2 + MR^2 + \frac{I}{I-1} MD^2} \qquad (6)$$

The nearest integer to this calculated value should be used as the appropriate degrees of freedom for S_T.

Using this value, the appropriate statistic is as follows:

$$X^2 = \frac{S_T^2 \cdot T}{\sigma_T^2} \qquad (7)$$

where:

S_T^2 = square of user's estimate of total standard deviation

σ_T^2 = square of manufacturer's claim of total standard deviation, or medically required standard deviation

T = degrees of freedom for S_T.

If the calculated X^2 is less than the critical upper 95% X^2 value, (from Table 1), the precision performance as indicated by the total precision estimate is acceptable.

If the calculated X^2 is greater than the critical upper 95% X^2 value, the precision performance is not within the claimed limits, or is not acceptable for the defined medical application.

The user's estimate can be larger than the SD claimed by the manufacturer and still be acceptable. Since the user experiment is based on a limited number of observations, there are expected sampling errors of the calculated S_{wr} and S_t around the true values. The larger the user experiment, the closer the estimates will be to the true value. The chi-square test is used to determine if your estimates are *significantly* larger than those provided by the manufacturer.

Including a list of analyte specific "acceptable" standard deviations is not within the scope of this document. It is suggested that the medical staff of the user's institution be consulted or the technical literature examined to develop an appropriate numerical definition or standard for acceptable standard deviation of each analyte.

5 Use of These Guidelines by Manufacturers to Establish Precision Performance

5.1 Factors to be Considered

The experiment described in this document can be used by manufacturers to establish precision performance claims for within run, total standard deviations (point estimates), and coefficients of variation. However, the goal of the manufacturer should be to establish these point estimates with sufficient rigor so that they will be valid over a wide variety of operating environments that individual users may encounter in the routine use of the method, device, or instrument.

The manufacturer may choose to employ a single reagent lot, calibration cycle, device, and operator for a minimum of 20 days to estimate total precision. This approach minimizes the effects of factors which increase long term imprecision, and increases the risk that individual users may not be able to achieve similar results in their laboratories. This risk may be reduced by incorporating multiple devices, operators, reagent lots, calibrator lots and calibration cycles (if appropriate), which will generally increase the precision standard deviation. Including additional sources of variation should better reflect the range of results that will be experienced by customers.

If these experiments are used by manufacturers to establish precision claims, the resulting labeling *must* include a statement as to the number of days, runs, devices, operators, calibration cycles, calibrator lots, and assay reagent lots that were included in the evaluation.

5.2 Incorporating Multiple Factors

Two approaches are available for incorporating the effects of multiple factors in the data. The first method is to perform the basic two runs-per-day experiment described in this document, but using multiple reagent lots, calibration cycles, operators, and instruments over the course of the 20 or more days of the evaluation experiment. The data may be analyzed and summarized according to the formulas provided, but will now reflect the influence on precision performance of those factors which were incorporated into the design of the experiment. The estimates will then better reflect the range of precision likely to be experienced in users' laboratories.

The second method permits the use of multiple instruments providing more than two runs per day. In this situation, the general nested analysis of variance should be used to determine the components of variance that pertain to each individual instrument, incorporating multiple reagent and calibrator lots, calibration cycles, and operators. The estimates for each instrument may then be "pooled" to create the precision performance

claims, or results may be presented individually. This pooling may be done ONLY when *each* instrument is evaluated with multiple operators, reagent lots and calibrator lots, *not* when there is only one such factor level per device. This method will reflect variations in precision performance between different instruments without incorporating the actual instrument-to-instrument component which would not be applicable to single-instrument users. The exact calculations for this general procedure are beyond the scope of this document but may be found in standard references on the analysis of variance.

It must be noted that regulatory agencies may require identification, evaluation and estimation of variance components beyond those employed in a user precision evaluation when this guideline is employed by manufacturers. In cases where additional components need evaluation, it is suggested that a statistician be consulted for appropriate experimental designs.

5.3 Format for Statement of Claims

Labeled claims for precision performance must include the following information, except where noted as optional:

- Concentrations at which claim is made.
- Point estimates (single value parameter estimate) of within-run precision standard deviation.
- Within-run percent coefficient of variation (optional).
- Point estimate of total precision standard deviation.
- Total precision percent coefficient of variation (optional).
- Confidence intervals on within-run and total standard deviation (optional).
- Actual number of days involved in experiment, and number of sites.
- Actual total number of runs (if applicable).
- Total number of observations (optional).
- Number of instruments/devices used in the evaluation, and how results were pooled.
- Number of reagent lots.
- Number of calibration cycles and calibration lots.
- The terms "run-to-run," "between-run," "day-to-day," and "between-day" should not be used anywhere in precision performance claims statements because of the ambiguity involved in their interpretation and calculation.
- A manufacturer may elect to include a table of expected maximum observed standard deviation (tolerance limit) for the within-run and total precision S.D.s, indexed by degrees of freedom (df). This will provide the user of the method with a benchmark to indicate that small verification experi-ments may result in calculated estimates slightly higher than the published SD point estimate and still demonstrate statisticaly equivalent precision (for within and total). Table 2 may be used for the multipliers of the claimed SDs to create a table that may look like this:

Within-run SD published: 10.5 @ 240 mg/dL

df for User Experiment	Acceptable SD Maximum
10	14.2
20	13.1
30	12.6
40	12.4
100	11.7

The purpose of such a table in a labeling claim would be to simply illustrate the sometimes-confusing fact that a user verification estimate can sometimes be higher than the published SD point estimate, and still verify the claim.

Table 1. Critical Values of Chi-Square

DF of User Variance Estimate	95% Critical Value	99% Critical Value
5	11.1	15.1
6	12.6	16.8
7	14.1	18.5
8	15.5	20.1
9	16.9	21.7
10	18.3	23.2
11	19.7	24.7
12	21.0	26.2
13	22.4	27.7
14	23.7	29.1
15	25.0	30.6
16	26.3	32.0
17	27.6	33.4
18	28.9	34.8
19	30.1	36.2
20	31.4	37.6
25	37.7	44.3
30	43.8	50.9
35	49.8	57.3
40	55.8	63.7
50	67.5	76.2
60	79.0	88.4
70	90.5	100.4
75	96.2	106.4
79	100.7	111.1
80	101.9	112.3
90	113.1	124.1
100	124.3	135.6

Table 2. Tolerance Factors for User SD Estimates.

df for User SD Estimates	Upper 95% Tolerance Limit for 95% of User Estimates*
10	1.35
20	1.25
30	1.20
40	1.18
50	1.16
60	1.15
70	1.14
80	1.13
90	1.12
100	1.11

* Multiply point estimate from manufacturer experiment by this factor to obtain the upper tolerance limit.

Table 3. Symbols Used in Text

S_{wr}	estimate of within-run standard deviation
I	total number of days (generally 20)
J	number of runs within day (generally 2)
X_{ijk}	result for run j on day i (result of replicate k on run j on day i; generally k = 1 or 2)
$\bar{X}_{i_{1\cdot}}$	average result of the replicates for run 1, day i
$\bar{X}_{i\cdot\cdot}$	average of all results day i
$\bar{X}_{\cdots}$	average of all results
A	standard deviation of the run means
B	standard deviation of the daily means
S_{dd}	estimate of between day standard deviation
S_{rr}	estimate of between run standard deviation
S_T	estimate of total precision standard deviation
σ_{wr}	performance claim within-run standard deviation
R	total number of runs (degrees of freedom for S^2_{wr})
T	degrees of freedom for S_T
ME	mean square for within-run (error)
MR	mean square for runs
MD	mean square for days
σ_T	manufacturer's claim of total standard deviation or medically required standard deviation

References

1. CAP. Proceedings of the 1976 Aspen Conference, Analytical goals in clinical chemistry. Skokie, Ilinois: College of American Pathologists; 1977.

2. Box GEP, Hunter WG, Hunter JS. *Statistics for Experimenters. Study of Variation.* New York: John Wiley and Sons; 1978.

3. Bauer S, Kennedy JW. Applied statistics for the clinical laboratory. *J. Clin. Lab. Auto.* I-XI (a series of 11 individual articles). 1982; 1983; 1984.

Appendix A. Sample Data Recording Sheets

Operator_____

DATA SHEET #1. Precision Evaluation Experiment

Concentration _____ **Reagent Source/Lot**_____

Analyte_____ **Calibrator Source/Lot**_____

Device_____

		Run 1		Run 2				
		Result	Result	Result	Result	Mean	Mean	Daily
Day #	Date	1	2	1	2	Run 1	Run 2	Mean

DATA SHEET #2. Precision Evaluation Experiment

Analyte/Concentration_____ **Device**_____

	Run 1	*Run 2*	$\left(\text{Mean}_{\text{Run 1}} - \text{Mean}_{\text{Run 2}}\right)^2$
Day #	*(Rep1 — Rep2)²*	*(Rep1 — Rep2)²*	
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
Sums	*(1)*	*(2)*	*(3)*

DATA SHEET #3.

Using results labeled (1), (2), and (3) from Data Sheet #2.

Section

4.7.1 $S_{wr} = \sqrt{\frac{(1) + (2)}{4I}} =$ _____

where I = number of days

4.7.2 $A = \sqrt{\frac{(3)}{2I}} =$ _____

from Data Sheet #1

4.7.2 B = Standard deviation of "Daily Means" = _____

4.7.2 $S_T = \sqrt{\frac{2B^2 + A^2 + S_{wr}^2}{2}} =$ _____

Calculation of T (degrees of freedom for total standard deviation estimate)

$ME = S_{wr}^2 =$ _____ $MR = 2A^2 =$ _____ $MD = 4B^2 =$ _____

$$T = \frac{I(2ME + MR + MD)^2}{2ME^2 + MR^2 + \frac{I}{I - 1} MD^2}$$

= _____

= _____(rounded to nearest integer)

DATA SHEET #4. Precision Evaluation Experiment Comparison to Claims

1. **Within-Run Precision** User Concentration Level = _____

User SD _____ Claim Concentration Level = _____

User Variance (SD^2) _____ Degrees of Freedom (R) _____

Performance Claim SD _____

Variance (SD^2) _____

(A) (User Variance ÷ Claim Variance) • R = _____

(B) Critical Chi-square (from Table 1) _____

Claim Rejected (A > B) _____

Claim Accepted (A ≤ B) _____

1. **Total Precision** User Concentration Level = _____

User SD _____ Claim Concentration Level = _____

User Variance (SD^2) _____ Degrees of Freedom (T) _____

Performance Claim SD _____

Variance (SD^2) _____

(A) (User Variance ÷ Claim Variance) • T = _____

(B) Critical Chi-square (from Table 1) _____

Claim Rejected (A > B) _____

Claim Accepted (A ≤ B) _____

Note: "SD" means standard deviation, and refers to the S_{wr} (calculated), S_T (calculated), or manufacturer claim.

Appendix B. Example of Completed Sample Data Recording Sheets

Operator_____

DATA SHEET #1. Precision Evaluation Experiment

Concentration High
Analyte Glucose
Device XYZ

Reagent Source/Lot AA—Lot 87011
Calibrator Source/Lot AA—Lot 87011

		Run 1		Run 2				
Day #	Date	Result 1	Result 2	Result 1	Result 2	Mean Run 1	Mean Run 2	Daily Mean
1	7/8	242	246	245	246	244	245.5	244.75
2	7/9	243	242	238	238	242.5	238	240.25
3	7/10	247	239	241	240	243	240.5	241.75
4	7/11	249	241	250	245	245	247.5	246.25
5	7/14	246	242	243	240	244	241.5	242.75
6	7/15	244	245	251	247	244.5	249	246.75
7	7/16	241	246	245	247	243.5	246	244.75
8	7/17	245	245	243	245	245	244	244.5
9	7/18	243	239	244	245	241	244.5	242.75
10	7/21	244	246	247	239	245	243	244
11	7/22	252	251	247	241	251.5	244	247.75
12	7/23	249	248	251	246	248.5	248.5	248.5
13	7/24	242	240	251	245	241	248	244.5
14	7/25	246	249	248	240	247.5	244	245.75
15	7/28	247	248	245	246	247.5	245.5	246.5
16	7/29	240	238	239	242	239	240.5	239.75
17	7/30	241	244	245	248	242.5	246.5	244.5
18	7/31	244	244	237	242	244	239.5	241.75
19	8/1	241	239	247	245	240	246	243
20	8/4	247	240	245	242	243.5	243.5	243.5

DATA SHEET #2. Precision Evaluation Experiment

Analyte/Concentration Glucose/High Device XYZ

Day #	Run 1 $(Rep1 - Rep2)^2$	Run 2 $(Rep1 - Rep2)^2$	$\left(\text{Mean Run 1} - \text{Mean Run 2}\right)^2$
1	16	1	2.25
2	1	0	20.25
3	64	1	6.25
4	64	25	6.25
5	16	9	6.25
6	1	16	20.25
7	25	4	6.25
8	0	4	1.00
9	16	1	12.25
10	4	64	4.00
11	1	36	56.25
12	1	25	0
13	4	36	49.00
14	9	64	12.25
15	1	1	4.00
16	4	9	2.25
17	9	9	16.00
18	0	25	20.25
19	4	4	36.00
20	49	9	0.00
Sums	(1) 289	(2) 343	(3) 281.00

DATA SHEET #3.

Using results labeled (1), (2), and (3) from first calculation sheet (B1):

$$S_{wr} = \sqrt{\frac{(1) + (2)}{4I}} = \underline{\ 2.81\ } \qquad \text{(B1)}$$

where I = total number of days

$$A = \sqrt{\frac{(3)}{2I}} = \underline{\ 2.65\ }$$

B = Standard deviation of "Daily Means" = 2.34

$$S_T = \sqrt{\frac{2B^2 + A^2 + S_{wr}^2}{2}} = \underline{\ 3.60\ }$$

Calculation of T (degrees of freedom for total standard deviation estimate): (B2)

ME = S_{wr}^2 = 7.90 MR = $2A^2$ = 14.0450 MD = $4B^2$ = 21.9024 (B2)

$$T = \frac{I\,(2ME + MR + MD)^2}{2ME^2 + MR^2 + \frac{I}{I-1}MD^2}$$

= 64.76 (estimates will vary slightly due to rounding error)

= 65 (rounded to nearest integer)

DATA SHEET #4. Precision Evaluation Experiment Comparison to Claims

1. **Within-Run Precision** User Concentration Level = Gluc/High

User SD 2.81 Claim Concentration Level = 240 mg/dL

User Variance (SD^2) 7.90 Degrees of Freedom (R) 40

Performance Claim SD 2.5

Variance (SD^2) 6.25

(A) (User Variance ÷ Claim Variance) • R = 50.56

(B) Critical Chi-square (from Table 1) 55.8

Claim Rejected (A > B) _____

Claim Accepted (A ≤ B) √

1. **Total Precision** User Concentration Level = Gluc/High

User SD 3.60 Claim Concentration Level = 240 mg/dL

User Variance (SD^2) 12.96 Degrees of Freedom (T) 65

Performance Claim SD 3.4

Variance (SD^2) 11.56

(A) (User Variance ÷ Claim Variance) • T = 72.65 (note: answers may vary slightly due to rounding of intermediate results)

(B) Critical Chi-square (from Table 1) 84.8

Claim Rejected (A > B) _____

Claim Accepted (A ≤ B) √

Note: "SD" means standard deviation, and refers to the S_{wr} (calculated), S_T (calculated), or manufacturer claim.

Appendix C. Additional Statistical Considerations

C1 Modifications for One Run Per Day

For some devices, only one run per day may be needed. Correct and useful estimates of within-run and total precision standard deviations for the device can still be obtained. However, the separation of total precision into between-day and between-run, within day components is not possible. The estimate of the within-run precision standard deviation should be calculated from the following formula:

$$S_{wr} = \sqrt{\frac{\sum_{i=1}^{I} (X_{i1} - X_{i2})^2}{2I}} \qquad \text{(C1)}$$

where:

I = total number of days (generally 20)
X_{i1} = result for replicate 1 on day i
X_{i2} = result for replicate 2 on day i.

The procedures and specifics of the general protocol as described in the main document should be followed except for running only one run per day instead of two. **Note**: There are only half as many degrees of freedom in this estimate as there are with two runs per day.

C1.1 Increasing Degrees of Freedom

Two methods may be used to modify the protocol to increase the number of degrees of freedom for the within-run precision estimate.

C1.1.1 Increase Length of Experiment

The number of days in the experiment may be increased, continuing to run only two aliquots of precision test material per run. The formula above may still be used for calculations. A minimum of 30 days is recommended.

C1.1.2 Increase Number of Aliquots

More than two aliquots of material within each run for the 20 days may be analyzed. If this method is used, the within-run standard deviation should be calculated from the following formula:

$$S_{wr} = \sqrt{\frac{\sum_{i=1}^{I} \sum_{j=1}^{N} (X_{ij} - \bar{X}_{i\cdot})^2}{I \cdot (N - 1)}} \qquad \text{(C2)}$$

where:

I = total number of days
N = number of replicate analyses per run
X_{ij} = result on replicate j in run on day i
$\bar{X}_{i\cdot}$ = average (mean) of all replicates on day i.

The number of degrees of freedom in this estimate is then I times the number of replicates per run minus 1 $[I \cdot (N - 1)]$ Each run must contain the *same* number of replicates for this formula to be appropriate. **Note:** Factor 2 does not appear in the denominator, as now this formula uses the sum of squared deviations from the run mean, as opposed to the convenient shortcut of duplicate observation differences used in previous formulas (appropriate for only two observations).

C1.2 Total Precision Standard Deviation

The total precision standard deviation estimates should be calculated with the following formulas. With only one run per day, the procedure differs somewhat from the formula described in the main protocol.

Calculate:

$$B = \sqrt{\frac{\sum_{i=1}^{I} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2}{I - 1}} \qquad \text{(C3)}$$

where:

I = number of days
$\bar{X}_{i\cdot}$ = average replicates on day i
$\bar{X}_{\cdot\cdot}$ = average of all results over all days.

C1.2.1 Standard Error

B is the standard deviation of the daily means (generally called the *standard error* of the daily means). When only one run per day is performed, the estimate combines the between-day and between-run components of precision. This formula should be used regardless of the number of days or the number of replicates.

C1.2.2 Total Standard Deviation

The estimate of the total precision standard deviation from the quantity B calculated above, and the within-run standard deviation estimate S_{wr}, is as follows:

$$S_T = \sqrt{B^2 + \frac{N - 1}{N} S_{wr}^2} \qquad \text{(C4)}$$

where:

N = number of replicates per run
B = standard deviation of daily means
S_{wr}^2 = within-run variance estimate (standard deviation squared).

This formula can be used regardless of which method is used to increase the number of observations for within-run precision (additional days or additional replicates per run).

C1.2.3 Satterthwaite's Equation

Use Satterthwaite's equation to calculate the proper number of degrees of freedom for S_T (the number of degrees of freedom for S_T is denoted T in Section 4.8.2). This is the only way to obtain the proper value for use in the chi-square test of claims described in Section 4.8.2. Use the following procedure:

$ME = S_{wr}^2$ (mean square for within run)
$MD = N \bullet B^2$ (mean square for both runs and days)

Then, calculate T as:

$$T = \frac{((N - 1)ME + MD)^2}{\frac{(N - 1)ME^2}{I} + \frac{MD^2}{I - 1}} \quad \text{(C5)}$$

Use the nearest integer to this quantity as the appropriate degrees of freedom for S_T.

C2 Other Estimates Available and Derivation of Formulas

Terminology that describes day-to-day or within-day precision has created confusion. Often "day-to-day" is erroneously used to mean total precision over a long period of time. Additional confusion results because the parameters of the components of precision are independent of the type of experiment, while the calculations for these estimates differ greatly depending on the number of observations per run, runs per day, and number of days.

C2.1 Between-Day Precision

Statistically, day-to-day (more appropriately called between-day) precision is the (adjusted) standard deviation of the daily means, after removing the effects of within-run and between-run, within-day variability, on the daily averages. Think of it as an estimate of the variability of daily averages that you would expect if you could perform an infinite number of observations each day. If you conduct a single run each day, you can demonstrate that the variance of the daily averages has the following *expected value*:

$$\text{Var}(\bar{X}) = \frac{\sum_{i=1}^{I} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2}{I - 1} = S_D^2 \quad \text{(C6)}$$

expected Value: $E(S_D^2) = \sigma_{dd}^2 + \frac{\sigma_{wr}^2}{N}$

where:

$\bar{X}_{i\cdot}$ = average result on day I

$\bar{X}_{\cdot\cdot}$ = average of all results on all days

I = total number of days

σ_{dd}^2 = true (adjusted) between-day variance

σ_{wr}^2 = true within-run variance

N = number of replicates per run.

As the number of replicates per run increases, the closer the estimate will be to the true parameter (i.e., the within-run precision will have a lesser influence on the estimate). The quantity called B cannot be used on the protocol to estimate the between-day precision. An adjustment must be made to this quantity for it to be useful. The adjustment/estimation procedure depends on the number of runs per day and the number of observations per run, but *not* on the number of days used in the protocol (except in the proper calculation of the original estimates).

C2.2 Two Runs Per Day

For two runs per day and two observations per run, as described in the main protocol, the quantities A and B from Section 4.7 (2) should be used to derive the following additional estimates:

Between-day standard deviation:

$$S_{dd} = \sqrt{B^2 - \frac{A^2}{2}} \qquad \text{(C7)}$$

The quantity S_{dd} is the estimate of the "true" adjusted between-day standard deviation, σ_{dd}.

Between-run, within-day standard deviation:

$$S_{rr} = \sqrt{A^2 - \frac{S_{wr}^2}{2}} \qquad \text{(C8)}$$

If working with a new device, it may be useful to calculate these estimates for a better picture of the factors influencing the observed precision.

C2.3 Single Run Per Day

For a single run per day and two or more observations per run as described in this appendix, the procedure is somewhat different. The between-day and between-run components of precision cannot be separated. The quantity called B in this case measures the sum of these two components. The only thing to do is to remove the effect of within-run variability from the estimate by calculating the following (N is the number of replicates/run):

$$S_{dd} = \sqrt{B^2 - \frac{S_{wr}^2}{N}} \qquad \text{(C9)}$$

Note: The interpretation of the quantity S_{dd} is now the sum of the between-day and between-run within-day effects.

Also, in some instances, the quantity under the radical above may be negative, which can occur if the between-day true component value is small. If this occurs, then the estimate S_{dd} should be set to 0 (zero). This same caution should be applied to the estimates calculated above for two runs per day.

DATA CALCULATION SHEET #1. Precision Evaluation Experiment

One Run per Day

Analyte/Concentration: Glucose/High **Device: XYZ**

Day #	Run Variance (S_{wr}^2)	Run Mean
1	8.0	244
2	0.5	242.5
3	32.0	243
4	32.0	245
5	8.0	244
6	0.5	244.5
7	12.5	243.5
8	0.0	245
9	8.0	241
10	2.0	245
11	0.5	251.5
12	0.5	248.5
13	2.0	241
14	4.5	247.5
15	0.5	247.5
16	2.0	239
17	4.5	242.5
18	0.0	244
19	2.0	240
20	24.5	243.5

Grand Mean ($\bar{X}_{..}$) = 244.13

Total WR Variance (S_{wr}^2) = 7.225

Variance of Daily Means (B^2) = 8.88

Total WR SD (S_{wr}) = 2.69

SD of Daily Means (B) = 2.98

Data used are from the first run of the main protocol example.

DATA CALCULATION SHEET #2. Precision Evaluation Experiment

One Run per Day

Analyte/Concentration Glucose/High Device XYX

Calculation of Total Precision Standard Deviation:

S_{wr}^2 (from Sheet #1) : 7.225

B^2 (from Sheet #1): 8.88

N (# replicates per run): 2

(C10)

$$S_T = \sqrt{B^2 + \frac{N - 1}{N} S_{wr}^2} = \sqrt{8.88 + \frac{1}{2}(7.225)}$$

$$= 3.53$$

Calculation of proper number of degrees of freedom for S_T:

(C11)

$$I = \text{\# of days} = 20$$

$$ME = S_{wr}^2 = 7.225 \quad MD = N \cdot B^2 = 17.76$$

$$T = \frac{((N - 1)\ ME + MD)^2}{\frac{(N - 1)ME^2}{I} + \frac{MD^2}{I - 1}}$$

$$= \frac{(7.225 + 17.76)^2}{\frac{(7.225)^2}{20} + \frac{(17.76)^2}{19}}$$

$$= \frac{624.25}{19.21}$$

$$= 32.49$$

Use 32 for T.

Summary of Comments and Subcommittee Responses

EP5-T2: *Evaluation of Precision Performance of Clinical Chemistry Devices—Second Edition; Tentative Guideline*

General

1. If the user is modifying the method by using alternative reagents, whose precision claims should the results be compared to? If modifying the method removes the ability to compare to any manufacturer this again calls for the need for analyte specific SD's that are acceptable. There are no guidelines in this area.

- **The experiment described in this document estimates the precision performance parameters of a method. The results should always be compared to the laboratory's own internal criteria, which may or may not coincide with the manufacturer's claims, and which should be based on clinical utility defined locally based on population disease prevalence and demographics. A "home brew" method should be judged against these internal criteria. The issue of analyte specific acceptability criteria is not within the scope of this guideline.**

2. It is unclear if this protocol should be used in the initial evaluation period or if periodic checks should be performed. The guideline should specify this.

- **This document may be used for either purpose, although its original intent was for application during initial evaluations and establishment of precision performance.**

3. Overall this document presents a cumbersome and complicated protocol to evaluate precision. It would be very costly and time consuming for the technologists and it is unclear what true benefits are gained by this procedure. On page 2 the guideline discusses that previous precision data were misleading in performance estimates. Providing an example to illustrate the need to conform to this guideline would be useful.

- **The subcommittee believes that the analysis of 4 samples per day over a 20-day period represents essentially the minimum amount of data necessary for a good estimation of precision performance, and that this places the minimum burden in cost and labor consistent with good laboratory practice for many types of analytical methods. The protocol was designed to be approximately equivalent to standard quality control methods in terms of number of samples. In the past, clinical evaluations often were performed only for a few days, at which point a decision as to acceptability was made. This did not permit an assessment of the effects of "days" on precision, and subsequent performance of the method was found to be unacceptable, contrary to the decision made on the initial evaluation. In addition, in many cases the computation of the variability was simplified to the point that it was incorrect statistically, and resulted in incorrect decisions. While the subcommittee is aware that the description of the experiment, and the required data calculations and descriptions may appear daunting, the actual conduction of the experiment is well within the bounds of reasonableness for all but the smallest of laboratories and all but the most complex and expensive of methods.**

4. It is imperative that manufacturers are performing precision estimates in accordance with these guidelines, otherwise it would be misleading to compare statistics.

- **The subcommittee believes manufacturers should: evaluate precision according to a statistically sound experimental design; calculate the summary statistics properly according to standard definitions and theory; provide the user with complete information as to the factors included and excluded in the evaluation experiments within the labeling of the method; and use a consistent definition of terms in reporting results. Any sound experimental design may be used, one of which is described in EP5-A.**

The experiment described herein is intended to be the simplest protocol incorporating time effects which exist in all laboratories. Consistent use of defined terms and summary statistics will also permit comparison of claims from manufacturer to manufacturer. For this reason, EP5-A recommends the elimination of terms such as "Day-to-day" and "Between-day" from manufacturer labeling because these terms are ambiguous and often misapplied.

5. The tentative guideline is a valid protocol to determine precision. However, its routine use in the clinical laboratory may not be necessary, and less robust methods might be substituted that do not establish but only validate precision.

- **The use of the term "robust" in this comment seems to refer to statistical power, describing the degree of departure from a claim that is detectable by an evaluation experiment of a given number of samples and days. If an experiment uses fewer samples and days, the departure from the claim can be larger and remain undetected. Conversely, increasing the number of replicates and days increases the liklihood that a difference of a given size will be detected, and that smaller differences can be detected.**

  The subcommittee agrees that a smaller experiment might be appropriate for validating situations where the manufacturer's claim on precision performance holds, but would be inadequate to detect when the method or device exhibits higher than claimed imprecision. The term "validation" is appropriate when all is acceptable, but when there is a problem, statistical power is the measure of the ability to detect differences. The subcommittee believes the EP5-A experimental protocol incorporates the minimum acceptable power likely to be desired by the laboratory evaluator.

6. We recommend referencing EP10 as a quicker but less robust method for the clinical laboratory to validate the manufacturer's claim. Many committee members felt that EP5 is more suitable for manufacturer's evaluation of precision. However, I have always used a similar protocol to EP5 to determine precision, but I have not tested the usefulness of EP10 compared to the more robust EP5.

- **EP10 uses fewer days and replicates, but is not intended for validating manufacturer's claims. See the response to comment 5. EP10 is indeed quicker and less demanding, but its intention is different, in that it provides a preliminary indication of major differences in precision performance that need to be immediately addressed. EP5 is intended as a more complete estimation of precision performance.**

7. Throughout the document (as well as others, particularly EP7-P on Interference Testing), both "precision" and "imprecision" are used interchangeably. If I understand your usage of these terms, "imprecision" is what is measured and "precision" relates to the process. These should be clearly identified and distinguished one from the other at the beginning of each document.

- **The use of the terms "precision" and "imprecision" are interchangeable and primarily semantic and contextual, not numerical. There have been many attempts to standardize when to use each, but the result is invariably awkward English. The direct statistical analogs to both are the (unambiguous) standard deviation (and CV%), so it would seem that little confusion results from using these terms interchangeably. Readers are referred to NCCLS document NRSCL8—*Terminology and Definitions for Use in NCCLS Documents* for standardized terminology.**

8. The terms "center line,""warning limits," and "out-of-control limits" are, at best, only understood by those who have worked in a clinical setting. I have spoken with a number of clinical technicians who were not sure what these terms are, and those of us from other backgrounds are not familiar with them at all. It seems you are leaving the functional definition to the reader. We, and likely other manufacturers, are implementing your guidelines as best we can. In the interest of standardization, it might be useful for your panel to arrive at some agreement regarding their definitions and applications.

- **Earlier editions of EP5 included an extensive discussion of the use of these routine quality control/quality assurance terms from control charting. These introductions greatly expanded the size of the document, and diverted attention and focus from the actual purpose of the evaluation experiment. Rather than attempt to explain the entire field in this document, it was agreed that the reader should be referred to standard laboratory quality control papers and texts. The subcommittee did not believe it could provide an adequate background and introduction to this area. Also, there are several slightly different definitions and calculation options for these parameters and, rather than provide the reader with a roadmap to these differences, the subcommittee felt it beyond the scope of this document. Laboratory QC procedures were adapted from long-established procedures used in industry. They are covered in most clinical chemistry textbooks. (Also refer to NCCLS document C24—Internal Quality Control Testing: Principles and Definitions.)**

9. I agree with everything else in the document regarding number of replicates, runs per day, and the 20 day minimum. Also the rigorous adherence to a balanced design, despite comments to the contrary, and the way the committee resolved the issue of negative variance components. I especially appreciate item 5.3 (13) (page 12), which recommends that manufacturers assist their customers in interpreting the results of a precision experiment. This would be much simpler if the performance claim were an upper value from an uncertainty interval. For example, if the performance claim were the upper limit of a 95%/90% tolerance interval, the user would only have to calculate precision and compare it with the claim, not calculate confidence intervals or perform a Chi-Square test. This would be much more oriented to the needs of the laboratory.

- **Earlier versions of EP5 incorporated the described approach. What resulted was a variety of limits that unfortunately caused some confusion, since there is a multiplicity of options as to the likelihoods and coverages that can be combined into a claim of this type. Most importantly, it was discerned that such a tolerance limit depends on knowing exactly the size of the users' confirmation experiment. It does not apply if there are more or fewer days or replicates used by the laboratorian. Rather than create such confusion, we concluded that if the experiment provides the point estimate only, then this would be independent of the size of both the establishment and verification experiments, and the point estimate could be validated by <u>any</u> future experiment when the user estimate Ns are properly accounted for in the statistics. We also include an option for a manufacturer to create a simple table for the likely results for a user's validation results, based on the size of the user experiment. This is based on the concept of a tolerance interval, but the tolerance interval is <u>not</u> an allowable option for the claim itself.**

10. I appreciate receiving this standard. I have followed EP5 for a long time; its focus had progressively narrowed to include only the user validation of performance claims and establishing performance capabilities at a single installation. Now it has expanded again to recognize the way EP5 was being used - by manufacturers to establish performance claims for a device or instrument. I guess this is a welcome addition, since there was no such guideline for manufacturers and so there was no real comparability of performance claims for different instruments.

- **The subcommittee appreciates the comment. The expressed purpose of the document is to introduce some consistency in the use of statistical terminology. It is hoped that in conjunction with NCCLS document EP11, *Uniformity of Claims for In Vitro Diagnostic Tests*, intermethodology comparisons will be possible for the user.**

11. Although the committee tried (and did a good job) to keep the statistics at a simple level without compromising the analyses, it is not obvious the committee had clear in their mind the statistical model and assumptions they used. Thus, it seems the model should be stated in the document. I would prefer the model be given in the text, but would consider it an improvement if it were provided in an appendix. The third choice would be in one of the responses by the committee.

- **The subcommittee agreed it was not necessary to include the simplistic model involved in this experimental design because it is intended for the nonstatistical user who does not require a customized experimental design. The protocol employs the usual simple two-factor fully nested components of variance model, with replicates nested within runs, which are in turn fully nested within days.**

$$(x = \mu + \alpha_i + \beta_j + \epsilon_{ijk},\ \alpha_i \sim n(o,\sigma_d^2),\ \beta_j \sim n(o,\sigma_R^2),\ \epsilon \sim n(o,\sigma_e^2))$$

12. This document recommends running pooled samples at medical decision levels for 20 days (2 replicates per 2 runs; n = 80). For some analytes, it is not possible to obtain pooled samples which are stable for 20 days. For example, whole blood is used for some assays, such as hemoglobin, glycated hemoglobin, etc. In the case of some enzymes (LDH), material stability is very limited in any storage condition. What would you recommend to estimate precision with a similar statistical power for these analytes?

    Is it acceptable to run more runs per day to achieve a precision estimate (total n = 80)? How many replicates may be included in a run without diminishing the statistical value of the estimate? For example:

    Batch analyzer can run 10 samples. A pool is stable for 4 days.

Experiment 1	Experiment 2	Experiment 3
10 replicates	5 replicates	2 replicates
2 runs	4 runs	10 runs
4 days	4 days	4 days.

- **The experiment described in the comment can be used to estimate imprecision in the case of analytes or test materials of limited stability. It would make calculations easier to have the same number of runs and replicates for each of the experiments. There are many designs appropriate for these situations (e.g., balanced incomplete blocks, balanced complete blocks) which can be found in standard statistical textbooks on experimental design. These designs are beyond the intended scope of this document, since the calculations are more complex and may require special statistical software to compute the proper components of variance.**

Section 1.1

13. In the second paragraph I recommend including "specimen" as an example of a typical user modification. Precision for urine specimens may be different than for serum specimens, for example. The sentence could read: "Examples of typical modifications are the use of specimens, reagents..."

- **The term "specimen sources" has been added to the list.**

Section 1.2.1

14. Proper evaluation also requires having pre-established criteria for acceptable performance, and statistically valid data analysis procedures. I recommend that these be added to the list of requirements.

- **The primary purpose of this guideline is to define a procedure and corresponding statistical applications to estimate precision. We agree, the user should have pre-established requirements for precision so that objectivity can be maintained in making a decision on whether the precision that is estimated is acceptable for the user's specific application(s). These requirements may vary from user to user. The process of defining these criteria is beyond the scope of this particular**

guideline. However, we have provided a process whereby the user may compare the estimated precision to the manufacturer's claims.

15. Requirements (1) and (2) could be combined into one statement: "Sufficient time to become familiar with the mechanics of operation and maintenance of the device, and with the steps of the evaluation protocol."

- **The subcommittee believes that separating the points adds emphasis to the differences between the two objectives.**

16. Requirement (4) could be better worded; it is actually two requirements: 1) appropriate experimental design to evaluate long-term variables, and 2) collecting sufficient data for a reliable estimate. I suggest: "An appropriate experimental design so that precision estimates reflect long-term performance of the device during routine use in the laboratory." and "Sufficient data to provide a reliable estimate of total system variability."

- **The wording of this section has been modified to express these concepts.**

Section 1.2.4

17. Monitoring 2 separate runs with a minimum of two test samples at 2 concentration levels generally defines the quality control protocol used in most laboratories. I would question why the data collection on QC could not be used rather than going to the expense of adding samples in addition to QC.

- **Data collection identical to QC can indeed be used to estimate imprecision. That is the way this experiment is designed. The evaluator should be examining the data from every run, and comparing the results to expected values in order to confirm that the system/device is stable. The reason for the recommendation is that this stability assessment is complicated if only a single type of test material is employed (i.e., just using QC materials).**

Section 1.2.6

18. This section states, "single run to estimate...[entails] significant risk." Does this mean the probability is high of a poor estimate or the consequences if it should occur is severe? If the former, then what are "usual operating parameters" and how can they be identified before doing the experiment? In both cases, what does that say about one run of: a) patient sample; b) proficiency test; c) quality control?

- **The statement in question is not intended to assess the consequences to decision-making of an errant estimate. It is merely a statement of the likelihood of an unrepresentative run being mistaken for typical performance because of (1) the high variability of the estimate and (2) the lack of power inherent in ignoring the possibility that within-run imprecision estimates will differ from run to run. Thus the EP5 protocol suggests that the traditional "single-run-of-20" method of estimating within-run imprecision may be better replaced by the multiple run pooled estimate. It is not necessary to know beforehand what the "usual operating parameters" are, nor identify them before doing the experiment, but merely realize that a single run may not give you the correct picture of what they are. The problem with single-run estimates of within-run imprecision is not extrapolated to the behavior of the method on unknowns, PTs or QC samples. This is estimated by the total imprecision which includes run factors, and is the entire point of EP5-A.**

19. Another issue is the use of the word estimate. There appears to be a confounding between estimate and estimator. The latter is a rule for the former. Unless it is a degenerate rule (one that has only one value), statistical estimators are NOT values but equations or formulas. Thus, an estimator cannot be correct or incorrect for estimating except in one sense. If one is referring to a particular estimator, e.g., by name , then one can say that a particular equation is not the one with that name. That is, one can ask, "Is this the correct equation for that estimator?" The

question, "Is this the correct estimator for estimating?" is nonsense. Estimates, on the other hand, are correct or incorrect in the sense that they either match what they are estimates of or they do not. Unfortunately, when we estimate we usually don't know the correct value (if we did, why estimate) so, in these cases we can't know whether an estimate is correct.

The reason for raising this issue is that this confounding could explain the problem discussed in Section 1.2.6. When using statistical estimators to estimate parameters, it should be recognized that all statistical analyses are based on statistical models that consist of an equation of the random variables and an assumption of the probability distribution of those random variables. No model is cited in this document nor are they often cited in the clinical chemistry literature when discussing the various types of precision estimators. Perhaps this has lead to the confusion - and leads to the next type of comment.

- **The distinction made by the commentor between the two terms is not universally accepted. The term "correctness" is evaluated in statistics by a host of measures about an estimator, (e.g., bias, efficiency, robustness, admissibility). The subcommittee believes that adding a distinction between these terms would confuse most readers. As for the model used for this experimental design, it is the simplest of nested random effects two-factor models. This model was described in earlier versions of EP5 and subsequently deleted to make the guideline more palatable to the general reader.**

Section 1.3.2

20. This section states, "detect small deviations from claimed performance." Should this be relatively small deviations? If precision is poor, even 100 df can have less power than 10 df with good precision. This is especially critical given that the next paragraph mentions "clinically important departures."

- **This change has been incorporated in EP5-A.**

Section 2.1

21. In this section, and Sections 4.1 and 4.3.1, the experiment can be no better than the homogeneity and stability of the test material. The diffuse and vague references to test material should be consolidated and succinct statement provided, i.e.: "Aliquots of test material (human derived control material or patient sera) should be pooled, mixed to homogeneity, aliquoted, and stored frozen at a temperature known to stabilize the analyte. Each frozen aliquot should be thawed, remixed, and stored prior to analysis by a strict protocol known to maintain analyte homogeneity and stability.

- **It is agreed that the test material defines the limits with which an experiment can be expected to estimate the method imprecision. However, the more specific the definition of the test material and its appropriate handling procedures, the less universal the description will be. Not all materials can be frozen nor do they need to be so for the purposes of this experiment. Not all materials need to be of human origin.**

22. This section states, "until operator can confidently operate the device." Should this have an operational definition, i.e., what is meant by "confidently"? Operational definitions could be useful at other places: page 4, section 3.4, "reasonably agree"; page 5, section 3.6, "considerable discrepancy."

- **The wording has been changed to "competently." As for the others, the subcommittee believes that these (admittedly subjective) assessments are best left to the judgment of competent laboratory supervisors.**

Section 3.4

23. The requirement is stability of error conditions (baseline or non-baseline) during the experiment. This important point should be stated succinctly. The experiment may be performed because the device is suspected of improper operation.

- **The subcommittee has incorporated the phrase "operating in a stable condition" into EP5-A.**

Section 3.7

24. Would "Detection of Outliers" be better placed in the next section since it is to be applied to the precision evaluation experiment and not the protocol familiarization period?

- **The subcommittee has moved Section 3.7 to 4.7 in EP5-A.**

Section 4.2

25. The guideline indicates single reagent lot and calibration. How could you arrive at total precision unless these factors are taken into account? This is particularly true for the calibration component which can be a major contributor in precision statistics. Also lot to lot variations would be important for those tests that are not calibrated such as enzymes.

- **The document has been modified in several places to indicate the limitations on interpretation of evaluations which employ only a single reagent lot or calibration cycle, as well as limited numbers of devices and operators. This will hopefully permit use of these guidelines for widely diverse purposes by both laboratorians and manufacturers. Labeling of results mandates disclosure of each such factors included in the evaluations.**

Section 4.6.2

26. In item (1), given that there is debate over how to estimate SD's and they form the basis for QC limits and procedure, shouldn't this be more specific?

- **Not all laboratory QC procedures use standard deviations, and in fact the origins of quality control did not use them at all (ranges were used). Many modern QC systems use more robust estimates of spread such as adaptive filtering, Bayesian procedures, CUSUM's, etc. The subcommittee does not feel the references to QC should attempt to override established procedures at the study site, nor restrict the types of QC that should be used.**

27. Item (3) states, "determine cause [of out-of-control] and repeat the run." Should this say "determine cause, eliminate and [then] repeat the run"?

- **Item (3) now states, "...determine the cause, eliminate the offending point, and then repeat the run."**

28. Item 5 states "if the previously acceptable results are now unacceptable, continue the experiment...", suggesting that data be discarded. What data should be discarded? The individual datum? The replicates of the concentration level (my choice)? The entire run containing the unacceptable data (re: item #3, p. 7)?

- **The laboratorian is free to choose whichever option reflects the QC practice in the individual laboratory. Any of the options suggested here are acceptable, although eliminating only the problem analyte pair at the offending concentration has been added to the recommendation.**

29. Item 61 suggests to "maintain record of the number of rejected runs." I think this is an excellent idea if the data are used to improve. Given that nothing is done with this information, why include it? Or, could something be added to reflect the value of this information?

- **It is assumed that if a manufacturer is using this document for a regulatory submission, the regulatory agency would require such documentation. If a user is performing an acceptability check on a new method, this information will be useful in discussions with the manufacturer should apparent problems arise as a result of the evaluation. Improvement is not an issue herein, since this is not a quality control situation where the data have informational value for process control. This statement is included in EP5-A for these eventualities.**

30. Statistical quality control charts are discussed on page 7 of the document. The protocol suggests that these be created after collecting data for five days and re-evaluated every five days thereafter. If is not clear, however, just what is discarded when an out-of-control condition is observed. Most often, several concentration levels of analyte are included with each run. Are you, indeed, recommending all data from the run be discarded or only the offending replicates?

- **This has been clarified in this version. Only the data from the offending analyte and level needs to be removed. Only the offending replicates should be discarded in an evaluation experiment. Unlike a quality control situation, there is no power during this initial evaluation to determine a link between these results and the run integrity, nor is there any need to preserve patient safety by discarding the whole run. Each level is unto itself in this experiment (unless operational assignable causes are apparent to reject the entire run), and the "replace" rules work just like outlier rules.**

31. I recognize that the choice of "every five days" is somewhat arbitrary. If one stays with the five-day period and data are discarded during the reevaluation, it seems the protocol would add another re-evaluation before an additional five days have elapsed. In other words, if one day's data are discarded after fifteen days, it seems the data should be re-evaluated again the following day, the "new" fifteenth day.

- **At this time, the subcommittee believes it is unnecessary to specify how this is done. Either version is acceptable.**

Section 4.7.2 (Section 4.8.2 in EP5-A)

32. "B" in equation 3 could be explained. Is this the estimate of the day to day precision?

- **In equation 3, the "B" is the uncorrected between-day component that has not yet been adjusted for the fact that it contains contributions from days, runs, and within-run sources. In the past, this equation has been incorrectly used as an estimate of what was also erroneously labeled "day-to-day" resulting in invalid statistics. The subcommittee believed it best not to define what "B" is, but simply to use it in the calculations. There is a reference, immediately below this equation to Appendix A that describes how to calculate the correct components.**

33. This comment refers to the use of Satterthwaite's approximation to the degrees of freedom of an additive model of mean squares. Eq. 4, page 10, section 4.7.2 is

$$S_T^2 = S_{dd}^2 + S_{rr}^2 + S_{wr}^2$$

This implies a linear model of the form described above except with no interaction between the Day and Run factors. That is, the linear model could be $x_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$. This raises the question: why is there no interaction factor? Is it assumed to be zero (0)? If so, the data do not

support this assumption. Look at the Table 3 (in final question in this section). It is the interaction that is statistically significant.

- **The commentor is referred to the responses to similar questions below. Interaction terms are generally not included in nested Model II ANOVAs. Such a term in an alternative model does not indicate that the factor has any physical significance. The variation ascribed to "interaction" in this comment is included in the total imprecision estimates and the runs and days factor components. See also the statistical treatment suggested at the end of this section as an alternative calculating method which does consider possible interaction effects.**

Section 4.8.2 (Section 4.9.2 in EP5-A)

34. It states that the "user cannot assume that all observations are independent." Why not? Which ones are not independent of each other? In the unstated model of this document, if the Day and Run factors are additive (Eq. 4), don't they have to be independent of each other? If not, shouldn't there be a covariance term in the model? These questions are not answerable until the model on which the EP5-T2 analyses is based is known. On the same page, it states "[Eq. 4] is correct way to estimate...because it properly weights the...components." Given that the interaction term is missing, is this a valid inference?

- **Interaction terms are usually assumed to not be physically possible in a fully nested model, and are almost never included in the variance components (ANOVA TYPE II) model used for this document. Interaction really has no meaning in almost all situations where the random effects nested model is used, although if someone can point out a reasonable physical model that includes interactions, the subcommittee will consider modifying the model.**

  The commentor appears to be confusing the lack of independence of data points with independence and additivity of the factors themselves. The subcommittee strongly believes the simple model has been found to apply in every situation the subcommittee is aware of in the application of this type of experiment. Observations in a nested model with a significant day effect are clearly not independent, since all observations within a day will be highly correlated. The same is true of runs. The statements in the document refer to this lack of independence in the observations, which also generates the need for Satterthwaite's formula.

35. The guideline should include a list of analyte specific acceptable SDs. I feel this guideline drops the ball at this point after putting the laboratory through hoops to arrive at this information.

- **We appreciate the importance of this question, but we believe it is beyond the scope of our immediate purpose to define a protocol with statistical applications appropriate to the estimation of precision. In reality, the definition of "acceptable SDs" is a somewhat individual one, depending on specific laboratory situation and applications for a given test. The answers to this question have been fairly controversial and have been the subject of many conferences and proceedings, such as the Aspen Conference sponsered by the CAP in 1976, and the more recent conference on "Accuracy and Precision Goals in Clinical Chemistry" sponsored by the AACC in 1992. Other definitions on acceptable performance are the recommendations of the National Cholesterol Education Program (CV 〈3% and Bias 〈3%) and the requirements for proficiency testing defined by HCFA in response to CLIA '88 regulations, where maximum error (regardless of whether it is due to bias or imprecision) is defined for about 125 tests. One must then define how such a maximum error can be related to maximum SD or maximum CV. Such relationships between maximum error defined for acceptable proficiency testing and maximum SD or CV have been discussed by Westgard and Burnett (*Clin Chem.* 1990; 36: 1629-1632) and by Ehrmeyer, Laessig, et. al (*Clin Chem*. 1990; 36: 1736-1740).**

36. In equation 6, why are the degrees of freedom approximated by Satterthwaite's equation when they can be determined exactly? The reason that it "cannot be assumed that all observations are independent," doesn't make sense. Which observations may not be independent? If they are

the within-run replicates, then S_{wr} also requires an approximation for the degrees of freedom. If they are the run-to-run or day-to-day, why these but not the replicates within a run? And, how is that possible?

- **Satterthwaite's formula is a complicated but necessary part of analyzing the data from a nested components of variance (or mixed) model ANOVA experimental design. In short, it is designed to estimate the correct number of degrees of freedom based on the ratio of the size of the day-to-day component (or runs component) to that of the within-run component. If there is a large day effect, this sharply reduces the actual number of degrees of freedom the data contributes to the estimate of total imprecision. In the extreme case, if all data within a day are identical, but different on different days, then you have only (number of days – 1) degrees of freedom for total variance. Increasing the number of observations within a day contributes nothing to the estimate. It would then be incorrect to assume that you have (number of data points – 1) degrees of freedom. It is a case where you cannot determine in advance how many degrees of freedom you have, but must determine it from the data itself. For many analytic methods (such as RIA, micro-titer plate ELISAs) that perform calibration within each run, these day-to-day effects are replaced by run-to-run effects, and are often the major contributor to overall variation. In such cases, Satterthwaite's formula is absolutely necessary. On the other hand, the evaluation of unitary devices generally does not need Satterthwaite's correction, since days and runs do not physically affect results. The within-run estimate cannot be corrected and is not necessary; it is a pure estimate unconfounded with any other component. As far as the independence of the observations, see the response above. When days or runs are significant, it is obvious that the data are not independent.**

Section 5.1/5.2

37. I believe that the requirements for establishing claims and for verifying claims are vastly different. It is not logical to recommend the exact same process to establish performance claims at the factory and to verify those claims in the lab. Performance claims must be established with a great deal more rigor than a test of those claims. This places an unnecessary burden on the laboratory, both to accurately establish performance capabilities and to compare these with the manufacturer claims. Perhaps EP5 should revert to user evaluation only, and reserve performance claims for the Subcommittee on Uniformity of Claims.

    This concern centers on Sections 5.1 and 5.2. These sections allow manufacturers to base their performance claims on the results from a single experiment. ("The manufacturer may choose to employ a single reagent lot, calibration cycle, device, and operator for a minimum of 20 days to estimate total precision.") The preceding and ensuing comments suggest that the Subcommittee recognized that this could lead to artificially low estimates, but opted to allow the single experiment anyway. The statement that this "...increases the risk that individual laboratories may not be able to achieve similar results in their laboratories."...is correct, but not likely to apply because of the wide statistical limits of the Chi-Square test.

    One small point is that if the protocols to establish claims and to verify claims are the same, it is not appropriate to use the Chi-Square test to compare the two estimates. This test assumes that the performance claim is a constant (absent any statistical error), which cannot be true if it were established with the same protocol as the test variance. The Chi Square test is appropriate for this standard only if the performance claims have much lower variance than the lab estimate; otherwise an F test is appropriate.

    The sentences allowing a single device, calibration, and reagent lot should be omitted entirely from 5.1, for reasons stated above and for the reasons mentioned in the document. It should be mandatory that manufacturer claims be based on an adequate sampling of instruments, reagent lots, and calibration cycles to assure that these important sources of variability have been accurately estimated. The first paragraph of Section 5.2 should be omitted, and the second paragraph expanded. I would prefer to see tolerance limits used for performance claims; however,

the stated recommendation to pool results from different instruments would be an acceptable alternative.

- **This comment represents a series of possible alternatives to the evaluation experiment, many of which existed in previous versions of this document and which were converted into the present form by the consensus process. As in responses to previous comments in this section, the allowance of a single reagent lot, etc. is to permit the separation of: intrinsic method performance under ideal circumstances (which represents best possible) allowing only method factors to influence the precision results, from method application expected performance long-term, adding in reagent lot, instruments/kits, calibrators, calibrations, and many other factors involved in the routine use of the method.**

  The minimum experiment can produce only the lower bound on imprecision performance, which if unacceptable, is easily grounds for rejecting the method. The subcommittee agrees that manufacturers should evaluate many such factors, and convey to the user the scope of the performance testing that went into the claims in the labeling. It is not the purpose of this document to provide this guidance. The subcommittee does not agree that chi-square testing is invalid in this particular case — when a claim is made, the interpretation is that it is a parameter, not an estimate, and the manufacturer must pay the penalty if there are too few degrees of freedom. The way to ensure this is to use chi-square instead of F. It is the philosophical point of the force of a claim, not the strict theoretical form that dictates the conservative approach chosen for this guideline. It also greatly simplifies the testing procedure to use an imputed infinite number of denominator degrees of freedom.

Section 5.3

38. The number of instruments/devices (9) and calibration cycles and lots (11) should be stated and not listed as optional.

- **The subcommittee agrees and they are now specified, which is consistent with the recommendations made in EP11.**

Table 3

39. On page 17, "j" is run number index in text, so perhaps an upper case "J" should be used for "number of runs within a day."

- **This change has been incorporated in EP5-A.**

Appendix A

40. Data sheet No. 1 on page 20 only has room for 16 runs! This is obviously a formatting error. Can it be corrected before the next reprinting?

- **This change has been incorporated in EP5-A.**

Appendix B

41. Upon reviewing the guideline I noticed an error in your data. I was attempting to recalculate your number from the original data listed on page 26 and the numbers did not come out. On further review of the table, I noticed that the problem was on Run 1 on Day 9. The two results from that day are 234 and 239 but the mean is recorded as 241. It appears that the 234 should actually be 243 and that it was transcribed wrong on the table. If 243 is used, then all numbers come out.

- **This change has been incorporated in EP5-A.**

42. It would appear that the mean of Run 2, day 10 should be 243.

- **This change has been incorporated in EP5-A.**

43. I believe there is an error on Data Sheet #1. The mean of run 1, day 9 should be 235.5, which would make the daily mean 240.5. These connections, in turn, effects the calculations on Data Sheet #2 and #3.

- **This has been corrected by the change in raw data (see Comment 41).**

44. On Data sheet 4, perhaps a note about the interpolation of the x^2 critical value for S_T; if 99% x^2 values are not used, why include them on p. 15?

- **At this time, the subcommittee believes any effective discussion of interpolation would only confuse the reader. It was decided to leave this discussion out — hopefully users have access to more extensive tables, or may be able to look up interpolation rules in standard math/ statistical textbooks. For the chi-squared critical values table, some users may prefer the 99% level, particularly when evaluating many analytes from the same experiment.**

Appendix C

45. In C1.1.2, "degrees of freedom...1(Ix(N-1)" is missing a matching parenthesis.

- **This has been corrected in EP5-A.**

46. In section C1.2.2, why is it labeled "Within Run SD" when total precision is being calculated and is the verb "is" missing?

- **This has been corrected in EP5-A.**

47. In section C1.2.3, "only way to obtain" seems extreme given that Satterthwaite's equation is an approximation based on certain assumptions — other approximations are derivable from those assumptions and other assumptions.

- **The qualifier is added for emphasis to underscore the invalidity of the simpler traditional estimates for number of degrees of freedom. Given the anticipated level of statistical experience of the readership of this document, we concluded that it would not be advantageous to go into the more complex models for what was designed as a simple general experiment. More sophisticated users, particularly manufacturers, are encouraged to develop more elaborate models and evaluations that assess more factors than are covered in this document.**

48. **Please note:** A detailed discussion was received that compared the study design of EP5 (ANOVA Type II) which does not include an interaction factor with an ANOVA model that does incorporate an interaction component (interaction between runs and days). The complexity of this discussion goes beyond the functionality of this document and hence was not printed. However, those interested in the complete detailed discussion may contact the NCCLSExecutive Offices for a copy of EP5-T2 comment number 48.

- **Both study designs have merit. The fact that EP5-A does not include an interaction component in the study design does not preclude manufacturers or others developing new assays from using such a study design with interactions, especially in their research and development phases. Interactions should be identified early, and eliminated or minimized as the test system design is being developed. The primary intent of EP5 is to serve as a tool for the user to estimate precision, not during the design phase, but at the end of the whole development process, when system design has been fixed. Of course, the manufacturer has interests in using EP5 (not in design) but to use EP5 to develop claims against which the user can make comparisons. Consequently, it is**

the recommendation of the subcommittee that models with interaction should be used most effectively as a research and development tool, while the model without interaction allows the separation of variability into those components that relate to the practical day to day laboratory operation.

49. The following comment was received from an NCCLS Industry Member during the 60-day delegate vote:

In Section 1.1, the wording of the second paragraph indicates that this procedure (the EP5-A protocol) should be employed when a modification occurs to a medical device that affects performance. The authors then specifically cite the use of control products not recommended by the manufacturer as an example of such modification. It is our feeling that the wording in Section 1.1 implies to the reader that they must use the control products provided by the manufacturer or submit to the complexities of this (EP5-A) study. It is our opinion that labs will opt not to do the EP5-A study in a situation that involves only use of control products not recommended by the manufacturer. Essentially, this wording favors one class of products (those made by the manufacturer) and potentially restricts competition and could affect the business of third party suppliers of control products.

- **The above wording in Section 1.1 has been retained as originally stated. EP5-A permits the use of any material that the laboratory wishes (see also Section 4.3.1) to estimate performance. The laboratory is free to make comparisons to either internal medical requirements for precision or to manufacturer's claims (see Equation 7 in EP5-A). The issue raised is only valid when a manufacturer has specified that their own control product must be used a part of the device "system." Since most manufacturers do not lock their customers into a specific quality control material, most customers are free to use any quality control material of their choosing. Those manufacturers that do specify a control product probably do so with good reason, in which case a substitution should be validated by the laboratory.**

Related NCCLS Publications*

EP6-P **Evaluation of the Linearity of Quantitative Analytical Methods; Proposed Guideline (1986).** Method for evaluating whether an instrument or quantitative analytical method meets the manufacturer's linearity claim; guidelines for manufacturers for stating a claim of an assay's linear range.

EP7-P **Interference Testing in Clinical Chemistry; Proposed Guideline (1986).** Background information and procedures for characterizing the effects of interfering substances on test results.

EP9-A **Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline (1995).** Procedures for determining the relative bias between two clinical chemistry methods or devices; design of a method comparison experiment using split patient samples, and analysis of the data.

EP10-A **Preliminary Evaluation of Quantitative Clinical Laboratory Methods; Approved Guideline (1998).** Experimental design and data analysis for preliminary evaluation of the performance of an analytical method or device.

EP11-P **Uniformity of Claims for In Vitro Diagnostic Tests; Proposed Guideline (1996).** Guidelines to promote consistency in the content and interpretation of maximum performance claims for in vitro diagnostic testing systems.

M29-A **Protection of Laboratory Workers from Instrument Biohazards and Infectious Disease Transmitted by Blood, Body Fluids, and Tissue; Approved Guideline (1997).** Guidance on the risk of transmission of hepatitis viruses and human immunodeficiency viruses in any laboratory setting; specific precautions for preventing transmission of blood-borne infection from laboratory instruments and materials; and recommendations for the management of blood-borne exposure.

*Proposed- and tentative-level documents are being advanced through the NCCLS consensus process; therefore, readers should refer to the most recent editions.

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NCCLS ▼ 940 West Valley Road ▼ Suite 1400 ▼ Wayne, PA 19087 ▼ USA ▼ PHONE 610.688.0100

FAX 610.688.0700 ▼ E-Mail:exoffice@nccls.org.

ISBN 1-56238-368-X