NIST SPECIAL PUBLICATION 260-125

U.S. DEPARTMENT OF COMMERCE/Technology Administration
National Institute of Standards and Technology

Standard Reference Materials:

Statistical Aspects of the Certification of Chemical Batch SRMs

Susannah B. Schiller

The National Institute of Standards and Technology was established in 1988 by Congress to "assist industry in the development of technology . . . needed to improve product quality, to modernize manufacturing processes, to ensure product reliability . . . and to facilitate rapid commercialization . . . of products based on new scientific discoveries."

NIST, originally founded as the National Bureau of Standards in 1901, works to strengthen U.S. industry's competitiveness; advance science and engineering; and improve public health, safety, and the environment. One of the agency's basic functions is to develop, maintain, and retain custody of the national standards of measurement, and provide the means and methods for comparing standards used in science, engineering, manufacturing, commerce, industry, and education with the standards adopted or recognized by the Federal Government.

As an agency of the U.S. Commerce Department's Technology Administration, NIST conducts basic and applied research in the physical sciences and engineering, and develops measurement techniques, test methods, standards, and related services. The Institute does generic and precompetitive work on new and advanced technologies. NIST's research facilities are located at Gaithersburg, MD 20899, and at Boulder, CO 80303. For more information contact the Public Inquiries Desk, 301-975-3058.

NIST Special Publication 260-125

Standard Reference Materials:

Statistical Aspects of the Certification of Chemical Batch SRMs

Susannah B. Schiller

Statistical Engineering Division
Computing and Applied Mathematics Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899-0001

U.S. DEPARTMENT OF COMMERCE, Michael Kantor, Secretary
TECHNOLOGY ADMINISTRATION, Mary L. Good, Under Secretary for Technology
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, Arati Prabhakar, Director

Issued July 1996

National Institute of Standards and Technology Special Publication 260–125
Natl. Inst. Stand. Technol. Spec. Publ. 260–125, 48 pages (July 1996)
CODEN: NSPUE2

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1996

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402–9325

Preface

Standard Reference Materials (SRMs) as defined by the National Institute of Standards and Technology (NIST) are well-characterized materials, produced in quantity and certified for one or more physical or chemical properties. They are used to assure the accuracy and compatibility of measurements throughout the Nation. SRMs are widely used as primary standards in many diverse fields in science, industry, and technology, both within the United States and throughout the world. They are also used extensively in the fields of environmental and clinical analysis. In many applications, traceability of quality control and measurement processes to the national measurement system is carried out through the mechanism and use of SRMs. For many of the Nation's scientists and technologists, it is therefore of more than passing interest to know the details of the measurements made at NIST in arriving at the certified values of the SRMs produced. The NIST Special Publication 260 Series is a series of papers reserved for this purpose.

The 260 Series is dedicated to the dissemination of information on different phases of the preparation, measurement, certification, and use of NIST SRMs. In general, much more detail will be found in these papers than is generally allowed, or desirable, in scientific journal articles. This enables the user to assess the validity and accuracy of the measurement processes employed, to judge the statistical analysis, and to learn details of techniques and methods utilized for work entailing greatest care and accuracy. These papers also should provide sufficient additional information so SRMs can be utilized in new applications in diverse fields not foreseen at the time the SRM was originally issued.

Inquiries concerning the technical content of this paper should be directed to the author(s). Other questions concerned with the availability, delivery, price, and so forth, will receive prompt attention from:

Standard Reference Materials Program
Bldg. 202, Rm. 204
National Institute of Standards and Technology
Gaithersburg, MD 20899
Telephone: (301) 975-6776
FAX: (301) 948-3730

Thomas E. Gills, Chief
Standard Reference Materials Program

OTHER NIST PUBLICATIONS IN THIS SERIES

Trahey, N.M., ed., NIST Standard Reference Materials Catalog 1995-96, NIST Spec. Publ. 260 (1995 Ed.). PB95-232518/AS

Michaelis, R.E., and Wyman, L.L., Standard Reference Materials: Preparation of White Cast Iron Spectrochemical Standards, NBS Misc. Publ. 260-1 (June 1964). COM74-11061**

Michaelis, R.E., Wyman, L.L., and Flitsch, R., Standard Reference Materials: Preparation of NBS Copper-Base Spectrochemical Standards, NBS Misc. Publ. 260-2 (October 1964). COM74-11063**

Michaelis, R.E., Yakowitz, H., and Moore, G.A., Standard Reference Materials: Metallographic Characterization of an NBS Spectrometric Low-Alloy Steel Standard, NBS Misc. Publ. 260-3 (October 1964). COM74-11060**

Hague, J.L., Mears, T.W., and Michaelis, R.E., Standard Reference Materials: Sources of Information, Publ. 260-4 (February 1965). COM74-11059**

Alvarez, R., and Flitsch, R., Standard Reference Materials: Accuracy of Solution X-Ray Spectrometric Analysis of Copper-Base Alloys, NBS Misc. Publ. 260-5 (February 1965). PB168068**

Shultz, J.I., Standard Reference Materials: Methods for the Chemical Analysis of White Cast Iron Standards, NBS Misc. Publ. 260-6 (July 1965). COM74-11068**

Bell, R.K., Standard Reference Materials: Methods for the Chemical Analysis of NBS Copper-Base Spectrochemical Standards, NBS Misc. Publ. 260-7 (October 1965). COM74-11067**

Richmond, M.S., Standard Reference Materials: Analysis of Uranium Concentrates at the National Bureau of Standards, NBS Misc. Publ. 260-8 (December 1965). COM74-11066**

Anspach, S.C., Cavallo, L.M., Garfinkel, S.B., et al., Standard Reference Materials: Half Lives of Materials Used in the Preparation of Standard Reference Materials of Nineteen Radioactive Nuclides Issued by the National Bureau of Standards, NBS Misc. Publ. 260-9 (November 1965). COM74-11065**

Yakowitz, H., Vieth, D.L., Heinrich, K.F.J., et al., Standard Reference Materials: Homogeneity Characterization of NBS Spectrometric Standards II: Cartridge Brass and Low-Alloy Steel, NBS Misc. Publ. 260-10 (December 1965). COM74-11064**

Napolitano, A., and Hawkins, E.G., Standard Reference Materials: Viscosity of Standard Lead-Silica Glass, NBS Misc. Publ. 260-11** (November 1966).

Yakowitz, H., Vieth, D.L., and Michaelis, R.E., Standard Reference Materials: Homogeneity Characterization of NBS Spectrometric Standards III: White Cast Iron and Stainless Steel Powder Compact, NBS Misc. Publ. 260-12 (September 1966).

Spijkerman, J.J., Snediker, D.K., Ruegg, F.C., et al., Standard Reference Materials: Mossbauer Spectroscopy Standard for the Chemical Shift of Iron Compounds, NBS Misc. Publ. 260-13** (July 1967).

Menis, O., and Sterling, J.T., Standard Reference Materials: Determination of Oxygen in Ferrous Materials (SRMs 1090, 1091, 1092), NBS Misc. Publ. 260-14** (September 1966).

Passaglia, E. and Shouse, P.J., Standard Reference Materials: Recommended Method of Use of Standard Light-Sensitive Paper for Calibrating Carbon Arcs Used in Testing Testiles for Colorfastness to Light, NBS Spec. Publ. 260-15 (July 1967). Superseded by SP 260-41.

Yakowitz, H., Michaelis, R.E., and Vieth, D.L., Standard Reference Materials: Homogeneity Characterization of NBS Spectrometric Standards IV: Pre-paration and Microprobe Characterization of W-20% Mo Alloy Fabricated by Powder Metallurgical Methods, NBS Spec. Publ. 260-16 (January 1969). COM74-11062**

Catanzaro, E.J., Champion, C.E., Garner, E.L., et al., Standard Reference Materials: Boric Acid; Isotopic, and Assay Standard Reference Materials, NBS Spec. Publ. 260-17 (February 1970). PB189457**

Geller, S.B., Mantek, P.A., and Cleveland, N.G., Calibration of NBS Secondary Standards Magnetic Tape Computer Amplitude Reference Amplitude Measurement "Process A," NBS Spec. Publ. 260-18 (November 1969). Superseded by SP 260-29.

Paule, R.C., and Mandel, J., Standard Reference Materials: Analysis of Interlaboratory Measurements on the Vapor Pressure of Gold (Certification of SRM 745). NBS Spec. Publ. 260-19 (January 1970). PB190071**

260-20: Unassigned

Paule, R.C., and Mandel, J., Standard Reference Materials: Analysis of Interlaboratory Measurements on the Vapor Pressures of Cadmium and Silver, NBS Spec. Publ. 260-21 (January 1971). COM74-11359**

Yakowitz, H., Fiori, C.E., and Michaelis, R.E., Standard Reference Materials: Homogeneity Characterization of Fe-3 Si Alloy, NBS Spec. Publ. 260-22 (February 1971). COM74-11357**

Napolitano, A., and Hawkins, E.G., Standard Reference Materials: Viscosity of a Standard Borosilicate Glass, NBS Spec. Publ. 260-23 (December 1970). COM71-00157**

Sappenfield, K.M., Marinenko, G., and Hague, J.L., Standard Reference Materials: Comparison of Redox Standards, NBS Spec. Publ. 260-24 (January 1972). COM72-50058**

Hicho, G.E., Yakowitz, H., Rasberry, S.D., et al., Standard Reference Materials: A Standard Reference Material Containing Nominally Four Percent Austenite, NBS Spec. Publ. 260-25 (February 1971). COM74-11356**

Martin, J.F., Standard Reference Materials: NBS-U.S. Steel Corp. Joint Program for Determining Oxygen and Nitrogen in Steel, NBS Spec. Publ. 260-26 (February 1971). PB 81176620**

Garner, E.L., Machlan, L.A., and Shields, W.R., Standard Reference Materials: Uranium Isotopi Standard Reference Materials, NBS Spec. Publ. 260-27 (April 1971). COM74-11358**

Heinrich, K.F.J., Myklebust, R.L., Rasberry, S.D., et al., Standard Reference Materials: Preparation and Evaluation of SRMs 481 and 482 Gold-Silver and Gold-Copper Alloys for Microanalysis, NBS Spec. Publ. 260-28 (August 1971). COM71-50365**

Geller, S.B., Standard Reference Materials: Calibration of NBS Secondary Standard Magnetic Tape (Computer Amplitude Reference) Using the Reference Tape Amplitude Measurement "Process A-Model 2," NBS Spec. Publ. 260-29 (June 1971). COM71-50282** Supersedes Measurement System in SP 260-18.

Gorozhanina, R.S., Freedman, A.Y., and Shaievitch, A.B., (translated by M.C. Selby), Standard Reference Materials: Standard Samples Issued in the USSR (A Translation from the Russian), NBS Spec. Publ. 260-30 (June 1971). COM71-50283**

Hust, J.G., and Sparks, L.L., Standard Reference Materials: Thermal Conductivity of Electrolytic Iron SRM 734 from 4 to 300 K, NBS Spec. Publ. 260-31 (November 1971). COM71-50563**

Mavrodineanu, R., and Lazar, J.W., Standard Reference Materials: Standard Quartz Cuvettes for High Accu-racy Spectrophotometry, NBS Spec. Publ. 260-32 (December 1973). COM74-50018**

Wagner, H.L., Standard Reference Materials: Comparison of Original and Supplemental SRM 705, Narrow Molecular Weight Distribution Polystyrene, NBS Spec. Publ. 260-33 (May 1972). COM72-50526**

Sparks, L.L., and Hust, J.G., Standard Reference Material: Thermoelectric Voltage of Silver-28 Atomic Percent Gold Thermocouple Wire, SRM 733, Verses Common Thermocouple Materials (Between Liquid Helium and Ice Fixed Points), NBS Spec. Publ. 260-34 (April 1972). COM72-50371**

Sparks, L.L., and Hust, J.G., Standard Reference Materials: Thermal Conductivity of Austenitic Stainless Steel, SRM 735 from 5 to 280 K, NBS Spec. Publ. 260-35 (April 1972). COM72-50368**

Cali, J.P., Mandel, J., Moore, L.J., et al., Standard Reference Materials: A Reference Method for the Determination of Calcium in Serum NBS SRM 915, NBS Spec. Publ. 260-36 (May 1972). COM72-50527**

Shultz, J.I., Bell, R.K., Rains, T.C., et al., Standard Reference Materials: Methods of Analysis of NBS Clay Standards, NBS Spec. Publ. 260-37 (June 1972). COM72-50692**

Richard, J.C., and Hsia, J.J., Standard Reference Materials: Preparation and Calibration of Standards of Spectral Specular Reflectance, NBS Spec. Publ. 260-38 (May 1972). COM72-50528**

Clark, A.F., Denson, V.A., Hust, J.G., et al., Standard Reference Materials: The Eddy Current Decay Method for Resistivity Characterization of High-Purity Metals, NBS Spec. Publ. 260-39 (May 1972). COM72-50529**

McAdie, H.G., Garn, P.D., and Menis, O., Standard Reference Materials: Selection of Differential Thermal Analysis Temperature Standards Through a Cooperative Study (SRMs 758, 759, 760), NBS Spec. Publ. 260-40 (August 1972) COM72-50776**

Wood. L.A., and Shouse, P.J., Standard Reference Materials: Use of Standard Light-Sensitive Paper for Calibrating Carbon Arcs Used in Testing Textiles for Colorfastness to Light, NBS Spec. Publ. 260-41 (August 1972). COM72-50775**

Wagner, H.L., and Verdier, P.H., eds., Standard Reference Materials: The Characterization of Linear Polyethylene, SRM 1475, NBS Spec. Publ. 260-42 (September 1972). COM72-50944**

Yakowitz, H., Ruff, A.W., and Michaelis, R.E., Standard Reference Materials: Preparation and Homogeneity Characterization of an Austenitic Iron-Chromium-Nickel Alloy, NBS Spec. Publ. 260-43 (November 1972). COM73-50760**

Schooley, J.F., Soulen, R.J., Jr., and Evans, G.A., Jr., Standard Reference Materials: Preparation and Use of Superconductive Fixed Point Devices, SRM 767, NBS Spec. Publ. 260-44 (December 1972). COM73-50037**

Greifer, B., Maienthal, E.J., Rains, T.C., et al., Standard Reference Materials: Development of NBS SRM 1579 Powdered Lead-Based Paint, NBS Spec. Publ. 260-45 (March 1973). COM73-50226**

Hust, J.G., and Giarratano, P.J., Standard Reference Materials: Thermal Conductivity and Electrical Resistivity Standard Reference Materials: Austenitic Stainless Steel, SRMs 735 and 798, from 4 to 1200 K, NBS Spec. Publ. 260-46 (March 1975). COM75-10339**

Hust, J.G., Standard Reference Materials: Electrical Resistivity of Electrolytic Iron, SRM 797, and Austenitic Stainless Steel, SRM 798, from 5 to 280 K, NBS Spec. Publ. 260-47 (February 1974). COM74-50176**

Mangum, B.W., and Wise, J.A., Standard Reference Materials: Description and Use of Precision Thermometers for the Clinical Laboratory, SRM 933 and SRM 934, NBS Spec. Publ. 260-48 (May 1974). Superseded by NIST Spec. Publ. 260-113. COM74-50533**

Carpenter, B.S., and Reimer, G.M., Standard Reference Materials: Calibrated Glass Standards for Fission Track Use, NBS Spec. Publ. 260-49 (November 1974). COM74-51185**

Hust, J.G., and Giarratano, P.J., Standard Reference Materials: Thermal Conductivity and Electrical Resistivity Standard Reference Materials: Electrolytic Iron, SRMs 734 and 797 from 4 to 1000 K, NBS Spec. Publ. 260-50 (June 1975). COM75-10698**

Mavrodineanu, R., and Baldwin, J.R., Standard Reference Materials: Glass Filters As a SRM for Spectrophotometry-Selection, Preparation, Certification, and Use-SRM 930 NBS Spec. Publ. 260-51 (November 1975). COM75-10339**

Hust, J.G., and Giarratano, P.J., Standard Reference Materials: Thermal Conductivity and Electrical Resistivity SRMs 730 and 799, from 4 to 3000 K, NBS Spec. Publ. 260-52 (September 1975). COM75-11193**

Durst, R.A., Standard Reference Materials: Standardization of pH Measurements, NBS Spec. Publ. 260-53 (December 1978). Superseded by SP 260-53 Rev. 1988 Edition. PB88217427**

Burke, R.W., and Mavrodineanu, R., Standard Reference Materials: Certification and Use of Acidic Potassium Dichromate Solutions as an Ultraviolet Absorbance Standard, NBS Spec. Publ. 260-54 (August 1977). PB272168**

Ditmars, D.A., Cezairliyan, A., Ishihara, S., et al., Standard Reference Materials: Enthalpy and Heat Capacity; Molybdenum SRM 781, from 273 to 2800 K, NBS Spec. Publ. 260-55 (September 1977). PB272127**

Powell, R.L., Sparks, L.L., and Hust, J.G., Standard Reference Materials: Standard Thermocouple Material, Pt-67: SRM 1967, NBS Spec. Publ. 260-56 (February 1978). PB277172**

Cali, J.P., and Plebanski, T., Standard Reference Materials: Guide to United States Reference Materials, NBS Spec. Publ. 260-57 (February 1978). PB277173**

Barnes, J.D., and Martin, G.M., Standard Reference Materials: Polyester Film for Oxygen Gas Transmission Measurements SRM 1470, NBS Spec. Publ. 260-58 (June 1979). PB297098**

Chang, T., and Kahn, A.H., Standard Reference Materials: Electron Paramagnetic Resonance Intensity Standard: SRM 2601; Description and Use, NBS Spec. Publ. 260-59 (August 1978). PB292097**

Velapoldi, R.A., Paule, R.C., Schaffer, R., et al., Standard Reference Materials: A Reference Method for the Determination of Sodium in Serum, NBS Spec. Publ. 260-60 (August 1978). PB286944**

Verdier, P.H., and Wagner, H.L., Standard Reference Materials: The Characterization of Linear Polyethylene (SRMs 1482, 1483, 1484), NBS Spec. Publ. 260-61 (December 1978). PB289899**

Soulen, R.J., and Dove, R.B., Standard Reference Materials: Temperature Reference Standard for Use Below 0.5 K (SRM 768), NBS Spec. Publ. 260-62 (April 1979). PB294245**

Velapoldi, R.A., Paule, R.C., Schaffer, R., et al., Standard Reference Materials: A Reference Method for the Determination of Potassium in Serum, NBS Spec. Publ. 260-63 (May 1979). PB297207**

Velapoldi, R.A., and Mielenz, K.D., Standard Reference Materials: A Fluorescence SRM Quinine Sulfate Dihydrate (SRM 936), NBS Spec. Publ. 260-64 (January 1980). PB80132046**

Marinenko, R.B., Heinrich, K.F.J., and Ruegg, F.C., Standard Reference Materials: Micro-Homogeneity Studies of NBS SRM, NBS Research Materials, and Other Related Samples, NBS Spec. Publ. 260-65 (September 1979). PB300461**

Venable, W.H., Jr., and Eckerle, K.L., Standard Reference Materials: Didymium Glass Filters for Calibrating the Wavelength Scale of Spectrophotometers (SRMs 2009, 2010, 2013, 2014). NBS Spec. Publ. 260-66 (October 1979). PB80104961**

Velapoldi, R.A., Paule, R.C., Schaffer, R., et al., Standard Reference Materials: A Reference Method for the Determination of Chloride in Serum, NBS Spec. Publ. 260-67 (November 1979). PB80110117**

Mavrodineanu, R., and Baldwin, J.R., Standard Reference Materials: Metal-On-Quartz Filters as a SRM for Spectrophotometry SRM 2031, NBS Spec. Publ. 260-68 (April 1980). PB80197486**

Velapoldi, R.A., Paule, R.C., Schaffer, R., et al., Standard Reference Materials: A Reference Method for the Determination of Lithium in Serum, NBS Spec. Publ. 260-69 (July 1980). PB80209117**

Marinenko, R.B., Biancaniello, F., Boyer, P.A., et al., Standard Reference Materials: Preparation and Characterization of an Iron-Chromium-Nickel Alloy for Microanalysis: SRM 479a, NBS Spec. Publ. 260-70 (May 1981). SN003-003-02328-1*

Seward, R.W., and Mavrodineanu, R., Standard Reference Materials: Summary of the Clinical Laboratory Standards Issued by the National Bureau of Standards, NBS Spec. Publ. 260-71 (November 1981). PB82135161**

Reeder, D.J., Coxon, B., Enagonio, D., et al., Standard Reference Materials: SRM 900, Anti-epilepsy Drug Level Assay Standard, NBS Spec. Publ. 260-72 (June 1981). PB81220758

Interrante, C.G., and Hicho, G.E., Standard Reference Materials: A Standard Reference Material Containing Nominally Fifteen Percent Austenite (SRM 486), NBS Spec. Publ. 260-73 (January 1982). PB82215559**

Marinenko, R.B., Standard Reference Materials: Preparation and Characterization of K-411 and K-412 Mineral Glasses for Microanalysis: SRM 470, NBS Spec. Publ. 260-74 (April 1982). PB82221300**

Weidner, V.R., and Hsia, J.J., Standard Reference Materials: Preparation and Calibration of First Surface Aluminum Mirror Specular Reflectance Standards (SRM 2003a), NBS Spec. Publ. 260-75 (May 1982). PB82221367**

Hicho, G.E., and Eaton, E.E., Standard Reference Materials: A Standard Reference Material Containing Nominally Five Percent Austenite (SRM 485a), NBS Spec. Publ. 260-76 (August 1982). PB83115568**

Furukawa, G.T., Riddle, J.L., Bigge, W.G., et al., Standard Reference Materials: Application of Some Metal SRMs as Thermometric Fixed Points, NBS Spec. Publ. 260-77 (August 1982). PB83117325**

Hicho, G.E., and Eaton, E.E., Standard Reference Materials: Standard Reference Material Containing Nominally Thirty Percent Austenite (SRM 487), NBS Spec. Publ. 260-78 (September 1982). PB83115576**

Richmond, J.C., Hsia, J.J., Weidner, V.R., et al., Standard Reference Materials: Second Surface Mirror Standards of Specular Spectral Reflectance (SRMs 2023, 2024, 2025), NBS Spec. Publ. 260-79 (October 1982). PB84203447**

Schaffer, R., Mandel, J., Sun, T., et al., Standard Reference Materials: Evaluation by an ID/MS Method of the AACC Reference Method for Serum Glucose, NBS Spec. Publ. 260-80 (October 1982). PB84216894**

Burke, R.W., and Mavrodineanu, R., Standard Reference Materials: Accuracy in Analytical Spectrophotometry, NBS Spec. Publ. 260-81 (April 1983). PB83214536**

Weidner, V.R., Standard Reference Materials: White Opal Glass Diffuse Spectral Reflectance Standards for the Visible Spectrum (SRMs 2015 and 2016), NBS Spec. Publ. 260-82 (April 1983). PB83220723**

Bowers, G.N., Jr., Alvarez, R., Cali, J.P., et al., Standard Reference Materials: The Measurement of the Catalytic (Activity) Concentration of Seven Enzymes in NBS Human Serum (SRM 909), NBS Spec. Publ. 260-83 (June 1983). PB83239509**

Gills, T.E., Seward, R.W., Collins, R.J., et al., Standard Reference Materials: Sampling, Materials Handling, Processing, and Packaging of NBS Sulfur in Coal SRMs 2682, 2683, 2684, and 2685, NBS Spec. Publ. 260-84 (August 1983). PB84109552**

Swyt, D.A., Standard Reference Materials: A Look at Techniques for the Dimensional Calibration of Standard Microscopic Particles, NBS Spec. Publ. 260-85 (September 1983). PB84112648**

Hicho, G.E., and Eaton, E.E., Standard Reference Materials: A SRM Containing Two and One-Half Percent Austenite, SRM 488, NBS Spec. Publ. 260-86 (December 1983). PB84143296**

Mangum, B.W., Standard Reference Materials: SRM 1969: Rubidium Triple-Point - A Temperature Reference Standard Near 39.30° C, NBS Spec. Publ. 260-87 (December 1983). PB84149996**

Gladney, E.S., Burns, C.E., Perrin, D.R., et al., Standard Reference Materials: 1982 Compilation of Elemental Concentration Data for NBS Biological, Geological, and Environmental Standard Reference Materials, NBS Spec. Publ. 260-88 (March 1984). PB84218338**

Hust, J.G., Standard Reference Materials: A Fine-Grained, Isotropic Graphite for Use as NBS Thermophysical Property RMs from 5 to 2500 K, NBS Spec. Publ. 260-89 (September 1984). PB85112886**

Hust, J.G., and Lankford, A.B., Standard Reference Materials: Update of Thermal Conductivity and Electrical Resistivity of Electrolytic Iron, Tungsten, and Stainless Steel, NBS Spec. Publ. 260-90 (September 1984). PB85115814**

Goodrich, L.F., Vecchia, D.F., Pittman, E.S., et al., Standard Reference Materials: Critical Current Measurements on an NbTi Superconducting Wire SRM, NBS Spec. Publ. 260-91 (September 1984). PB85118594**

Carpenter, B.S., Standard Reference Materials: Calibrated Glass Standards for Fission Track Use (Supplement to NBS Spec. Publ. 260-49), NBS Spec. Publ. 260-92 (September 1984). PB85113025**

Ehrstein, J.R., Standard Reference Materials: Preparation and Certification of SRM for Calibration of Spreading Resistance Probes, NBS Spec. Publ. 260-93 (January 1985). PB85177921**

Gills, T.E., Koch, W.F., Stolz, J.W., et al., Standard Reference Materials: Methods and Procedures Used at the National Bureau of Standards to Certify Sulfur in Coal SRMs for Sulfur Content, Calorific Value, Ash Content, NBS Spec. Publ. 260-94 (December 1984). PB85165900**

Mulholland, G.W., Hartman, A.W., Hembree, G.G., et al., Standard Reference Materials: Development of a 1mm Diameter Particle Size Standard, SRM 1690, NBS Spec. Publ. 260-95 (May 1985). PB95-232518/AS**

Carpenter, B.S., Gramlich, J.W., Greenberg, R.R., et al., Standard Reference Materials: Uranium-235 Isotopic Abundance Standard Reference Materials for Gamma Spectrometry Measurements, NBS Spec. Publ. 260-96 (September 1986). PB87108544**

Mavrodineanu, R., and Gills, T.E., Standard Reference Materials: Summary of the Coal, Ore, Mineral, Rock, and Refactory Standards Issued by the National Bureau of Standards, NBS Spec. Publ. 260-97 (September 1985). PB86110830**

Hust, J.G., Standard Reference Materials: Glass Fiberboard SRM for Thermal Resistance, NBS Spec. Publ. 260-98 (August 1985). SN003-003-02674-3*

Callanan, J.E., Sullivan, S.A., and Vecchia, D.F., Standard Reference Materials: Feasibility Study for the Development of Standards Using Differential Scanning Calorimetry, NBS Spec. Publ. 260-99 (August 1985). PB86106747**

Taylor, J.K., Trahey, N.M., ed., Standard Reference Materials: Handbook for SRM Users, NBS Spec. Publ. 260-100 (February 1993). PB93183796**

Mangum, B.W., Standard Reference Materials: SRM 1970, Succinonitrile Triple-Point Standard: A Temperature Reference Standard Near 58.08° C, NBS Spec. Publ. 260-101 (March 1986). PB86197100**

Weidner, V.R., Mavrodineanu, R., Mielenz, K.D., et al., Standard Reference Materials: Holmium Oxide Solution Wavelength Standard from 240 to 640 nm - SRM 2034, NBS Spec. Publ. 260-102 (July 1986). PB86245727**

Hust, J.G., Standard Reference Materials: Glass Fiberblanket SRM for Thermal Resistance, NBS Spec. Publ. 260-103 (September 1985). PB86109949**

Mavrodineanu, R., and Alvarez, R., Standard Reference Materials: Summary of the Biological and Botanical Standards Issued by the National Bureau of Standards, NBS Spec. Publ. 260-104 (November 1985). PB86155561**

Mavrodineanu, R., and Rasberry, S.D., Standard Reference Materials: Summary of the Environmental Research, Analysis, and Control Standards Issued by the National Bureau of Standards, NBS Spec. Publ. 260-105 (March 1986). PB86204005**

Koch, W.F., ed., Standard Reference Materials: Methods and Procedures Used at the National Bureau of Standards to Prepare, Analyze, and Certify SRM 2694, Simulated Rainwater, and Recommendations for Use, NBS Spec. Publ. 260-106 (July 1986). PB86247483**

Hartman, A.W., and McKenzie, R.L., Standard Reference Materials: SRM 1965, Microsphere Slide (10 μm Polystyrene Spheres), NIST Spec. Publ. 260-107 (November 1988). PB89153704**

Mavrodineanu, R., and Gills, T.E., Standard Reference Materials: Summary of Gas Cylinder and Permeation Tube Standard Reference Materials Issued by the National Bureau of Standards, NBS Spec. Publ. 260-108 (May 1987). PB87209953**

Candela, G.A., Chandler-Horowitz, D., Novotny, D.B., et al., Standard Reference Materials: Preparation and Certification of an Ellipsometrically Derived Thickness and Refractive Index Standard of a Silicon Dioxide Film (SRM 2530), NIST Spec. Publ. 260-109 (October 1988). PB89133573**

Kirby, R.K., and Kanare, H.M., Standard Reference Materials: Portland Cement Chemical Composition Standards (Blending, Packaging, and Testing), NBS Spec. Publ. 260-110 (February 1988). PB88193347**

Gladney, E.S., O'Malley, B.T., Roelandts, I., et al., Standard Reference Materials: Compilation of Elemental Concentration Data for NBS Clinical, Biological, Geological, and Environmental Standard Reference Materials, NBS Spec. Publ. 260-111 (November 1987). PB88156708**

Marinenko, R.B., Blackburn, D.H., and Bodkin, J.B., Standard Reference Materials: Glasses for Microanalysis: SRMs 1871-1875, NIST Spec. Publ. 260-112 (February 1990). PB90215807**

Mangum, B.W., and Wise, J.A., Standard Reference Materials: Description and Use of a Precision Thermometer for the Clinical Laboratory, SRM 934, NIST Spec. Publ. 260-113 (June 1990). PB90257643**

Vezzetti, C.F., Varner, R.N., and Potzick, J.E., Standard Reference Materials: Bright-Chromium Linewidth Standard, SRM 476, for Calibration of Optical Microscope Linewidth Measuring Systems, NIST Spec. Publ. 260-114 (January 1991). PB91167163**

Williamson, M.P., Willman, N.E., and Grubb, D.S., Standard Reference Materials: Calibration of NIST SRM 3201 for 0.5 in. (12.65 mm) Serial Serpentine Magnetic Tape Cartridge, NIST Spec. Publ. 260-115 (February 1991). PB91187542**

Mavrodineanu, R., Burke, R.W., Baldwin, J.R., et al., Standard Reference Materials: Glass Filters as a Standard Reference Material for Spectrophotometry-Selection, Preparation, Certification and Use of SRM 930 and SRM 1930, NIST Spec. Publ. 260-116 (March 1994). PB94-188844/AS**

Vezzetti, C.F., Varner, R.N., and Potzick, J.E., Standard Reference Materials: Antireflecting-Chromium Linewidth Standard, SRM 475, for Calibration of Optical Microscope Linewidth Measuring Systems, NIST Spec. Publ. 260-117 (January 1992). PB92-149798**

Williamson, M.P., Standard Reference Materials: Calibration of NIST Standard Reference Material 3202 for 18-Track, Parallel, and 36-Track, Parallel Serpentine, 12.65 mm (0.5 in), 1491 cpmm (37871 cpi), Magnetic Tape Cartridge, NIST Spec. Publ. 260-118 (July 1992). PB92-226281**

Vezzetti, C.F., Varner, R.N., and Potzick, Standard Reference Materials: Antireflecting-Chromium Linewidth Standard, SRM 473, for Calibration of Optical Microscope Linewidth Measuring System, NIST Spec. Publ. 260-119 (September 1992)

Caskey, G.W., Philips, S.D., Borchardt., et al., Standard Reference Materials: A Users' Guide to NIST SRM 2084: CMM Probe Performance Standard, NIST Spec. Publ. 260-120 (1994)

Rennex, B.G., Standard Reference Materials: Certification of a Standard Reference Material for the Determination of Interstitial Oxygen Concentration in Semiconductor Silicon by Infrared Spectrophotometry, NIST Spec. Publ. 260-121 (1994) PB95-125076/AS

Gupta, D., Wang, L., Hanssen, L.M., Hsai, J.J., and Datla, R.U., Polystyrene Films for Calibrating the Wavelength Scale of Infrared Spectrophotometer (SRM 1921). NIST Spec. Publ. 260-122 (1995) PB95-226866/AS

Development of Technology and the Manufacture of Spectrometric SRMs for Naval Brasses (MC62 M63). NIST Spec. Publ. 260-123 (IN PREP)

Strouse, G.F., SRM 1744: Aluminum Freezing Point Standard. NIST Spec. Publ. 260-124 (1995) SN003-003-03342-1

Guenther, F.R., Dorko, W.D., Miller, W.R., et al., Standard Reference Materials: The NIST Traceable Reference Material Program for Gas Standards, NIST Spec. Publ. 260-126 (IN PREP)

*Send order with remittance to: Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20102. Remittance from foreign countries should include an additional one fourth of the purchase price for postage.

**May be ordered from: National Technical Information Services (NTIS), Springfield, VA 22161.
For information phone (703-487-4650)
To Place an Order with PB# phone (800-553-6847)

Statistical Aspects of the Certification of Chemical Batch Standard Reference Materials

Susannah B. Schiller
Statistical Engineering Division

The accurate determination of chemical analytes in batches of material is the principal requirement in chemical Standard Reference Material (SRM) certification. There are many issues dealing with material sampling, experiment design, and data analysis which must be addressed. Many measurements made for chemical constituent batch SRMs are destructive, and batches are usually large, so selecting a random but representative sample from the batch is vital to inference about the material. The experiments need to be designed to use as few measurements as possible while minimizing bias in the results. Homogeneity assessment must be done to verify the material's suitability for sale, and to determine what type of statistical interval will make an appropriate summary for the certificate. Finally, results from more than one independent chemical method often must be combined in a statistically meaningful way to arrive at a realistic estimate of the uncertainty of the results achieved. This paper provides guidelines for addressing these statistical issues. The motivations behind those guidelines are also explained to facilitate understanding of them.

1. Introduction

The accurate determination of chemical analytes in batches of material is the principal requirement in chemical Standard Reference Material (SRM) certification. Natural matrix materials, such as PCBs and pesticides in whale blubber (SRM 1945) or trace elements in estuarine sediment (SRM 1646a), and synthetic sample or "calibration solution" SRMs, such as PAHs in acetonitrile (SRM 1647c) are just a few examples of this class of SRM.

"Natural matrix" SRMs are prepared from material found in nature. After the material is collected, it is blended, dried (for some materials), and bottled. (Note: ampules or vials may be the appropriate type of container, depending on the material. However, the word "bottle" is used throughout this text for convenience.) Natural matrix SRMs are sold for use primarily as control samples, although they may also be used as calibrants for analytical methods which directly analyze solid samples. When a natural matrix SRM is similar in chemical composition and concentration to an unknown sample, it provides a good check for the accuracy of an analytical method because interferences and dissolution problems that might cause biases in analysis of the unknown sample will also occur for the SRM. Each analyte in a natural matrix SRM is certified using data from a single definitive method or data from at least two chemically independent methods if no definitive method is available. Material homogeneity is an important issue for these SRMs.

"Calibration solution" SRMs are often prepared gravimetrically by weighing quantities of pure materials and dissolving each in an appropriate solvent. After the solution has been well blended, it is bottled. Measurements from one analytical method are usually compared with the results of the gravimetric preparation to verify the certified value for each analyte. Purity assessment of the materials used to prepare the solution is an important factor in the final quality of the certification.

There are many issues dealing with material sampling, experiment design, and data analysis which must be addressed for this type of SRM. Many measurements made for chemical constituent batch SRMs are destructive (i.e., the sample is consumed), and batches are usually large, so selecting a random but representative sample from the batch is vital to inference about the material. Homogeneity assessment must be done to verify the material's suitability for sale, and to determine what type of statistical interval will make an appropriate summary for the certificate. When designing and analyzing the actual measurements, the goal is to use as few measurements as possible while minimizing bias and determining a realistic estimate of the uncertainty of the results achieved. Designs are also affected by the degree of uncertainty acceptable to the intended users of the SRM.

The goal of any SRM certification is to tell the user the "right" answer (i.e., the true concentration) and how well it is known. In the absence of systematic error, material heterogeneity, or sources of uncertainty determined by nonstatistical means, this information is summarized by a confidence interval for each mean concentration. If the material is heterogeneous, there is no single "right" answer, but rather there is a statistical population of "right" answers (i.e., true concentrations) corresponding to each unit (or sample) of the SRM. In this case a statistical prediction interval or a tolerance interval (depending on the degree of conservatism that is desired) is an appropriate summary of the population of "right" answers in the batch of material. Systematic error and uncertainties based on experience rather than data introduce additional wrinkles into assessing the final uncertainty.

This paper provides guidelines for addressing the many statistical issues in producing an SRM: sampling from a batch of material, designing and evaluating a homogeneity study, designing experiments to determine the certified value, analyzing the data, and combining the results from multiple methods. The motivations behind those guidelines are also explained to facilitate understanding of them. When experiments are carefully designed, the resulting data are easy to analyze and the resulting SRM is useful to the users it is intended to serve.

2. Expression of Uncertainty

When expressing the uncertainty of certified values on SRMs, it is important to follow a consistent format whenever possible. This facilitates comparisons among NIST SRMs and between NIST SRMs and similar reference materials sold by other producers around the world. Following the NIST Uncertainty Policy [1] provides this standardization, since it is based upon the CIPM approach to expressing uncertainty in measurements, given in the ISO "Guide to the Expression of Uncertainty in Measurement" [2].

Each component of uncertainty is expressed as a standard-deviation-like quantity called a "standard uncertainty" and categorized by the way in which it was estimated. Type A standard uncertainties are those which were evaluated by statistical methods, and Type B standard uncertainties are those which were evaluated by "other means." Each standard uncertainty also has degrees of freedom associated with it. When Type B standard uncertainties are derived from a known range, they are ascribed infinite degrees of freedom.

The standard uncertainties are combined by root sum of squares to form the "combined standard uncertainty." In order to define an interval which should cover the true concentration of the SRM, the combined standard uncertainty is multiplied times a "coverage factor," k, to form the "expanded uncertainty." The coverage factor depends upon the degrees of freedom of the combined standard uncertainty and the desired level of confidence to be associated with the interval, and typically comes from the Student's t distribution.

3. Material Sampling

Batches of chemical SRMs typically consist of from 500 to 3000 units. In this paper, it is assumed that the bottling is complete before the measurements begin, which is usually the case. This way, the samples used for analysis are truly representative of the samples to be sold.

Stratified random sampling is used to select bottles for analysis. The sampling is done at random to ensure validity of inference to the rest of the batch. It is stratified to ensure that bottles are selected from the entire range of the population. Because material heterogeneity may be a problem for each SRM, it is essential to include bottles from the entire preparation process and fill sequence, which strictly random sampling does not guarantee. For example, if the blending is done in sub-batches, these sub-batches must be represented in the group of bottles to be analyzed so that the null hypothesis of no difference between sub-batches can be tested once the measurements are made. Also, since bottle filling can introduce trends, either because a solvent evaporates, a dry material absorbs moisture, or an analyte volatilizes, the chosen bottles must span the fill sequence.

When multiple chemical methods are used, the bottles are assigned to each method so that each gets a group spanning the entire fill sequence. This way, if there is a trend with fill sequence, this trend will not be confounded with between-method differences.

To select a stratified random sample of n bottles, the population of bottles is divided into n equal-sized groups corresponding to sequential intervals of the fill sequence. One bottle is selected at random from each of the n groups, and this provides the group of bottles to be analyzed. Sometimes, the chemists feel strongly that the first and last bottles filled should be analyzed. In this case, after the first and last bottles are pulled, the remaining bottles are divided into n-2 equal-sized groups and a stratified random sample of n-2 bottles is chosen.

4. The Homogeneity Study

One vital component of the certification process is homogeneity testing. The null hypothesis of no differences between samples is tested; if this hypothesis is rejected, then the material variability must be quantified and incorporated into the final uncertainty. The final uncertainty can be based on a statistical tolerance interval, if a conservative interval is desired. Alternatively, the standard uncertainty due to material variability can be included in the uncertainty according to the CIPM approach. This is less conservative, and is the same as a statistical prediction interval in the absence of Type B uncertainty. This paper will address prediction intervals. The way in which the standard uncertainty for material variability is estimated depends on the nature of the heterogeneity.

The nature of heterogeneity, when it exists, varies from SRM to SRM. It might appear as a trend with fill sequence, or as random variation within bottles, or as random variation between bottles over and above any variation within bottles. Trends with fill sequence usually occur in materials that are liquids which evaporate during filling, materials with components which volatilize during filling, or dried materials which absorb moisture during filling. Random within- and between-bottle variations are often caused by an element which is bound to some particles but not to others--the concentration in a sample will depend on how many bound particles it receives, and when those particles are rare, this can produce noticeable heterogeneity. It may occur for some analytes being certified even when others are homogeneous. This type of heterogeneity occurs in solid natural materials such as powdered coal, where one chemical form is fairly rare and widely dispersed throughout the material, and in viscous liquids such as fuel oil, which has suspended solids.

The way in which homogeneity studies are designed and evaluated depends on the anticipated nature of the heterogeneity. If the expected heterogeneity is a trend with fill sequence, then the experiment should be designed to get as much information as possible about the trend. However, if random between-bottle variation is more likely, duplicate samples must be prepared from each of several bottles, so that an analysis of variance can be used to test for a between-bottle effect in the data. In all cases, a chemical method should be chosen which gives the best precision practical. In some cases, it may be advisable to exclude steps and corrections which would improve accuracy but degrade precision.

The samples used for homogeneity testing should be at the minimum weight recommended for use on the certificate, since a material may appear homogeneous when large quantities (e.g., 1 g to 5 g) of material are analyzed, but heterogeneity may be detectable in small samples (e.g., 50 mg).

4.1. Heterogeneity as a Trend

In the case of calibration solutions, the most likely heterogeneity to expect is a trend with fill sequence. Thus, for a homogeneity study, the sampling plan consists of taking a stratified random sample, so that samples are spread throughout the fill sequence for homogeneity measurements. These samples are then analyzed by a precise chemical method. A linear trend in fill sequence is most often the concern. However, there are a myriad of other possible material trends, such as a step function at some unidentified point in the fill sequence.

For homogeneity studies designed to detect a trend, only one sample from each bottle is prepared and analyzed. This way, if n measurements are to be made, they can be made on n unique bottles, giving a more detailed view of a possible trend than duplicate samples from each of n/2 bottles would.

It is vital to randomize the order in which the bottles are measured. That way, the measurement order will not be confounded with a potential material problem, and if a trend is observed, it will be possible to determine whether it was caused by the fill sequence or the measurement process.

Linear regression (or analysis of covariance, if other factors are involved) is used to test the null hypothesis that the slope of the response variable (usually measured concentration) as a function of fill sequence is zero. If that hypothesis is rejected, then the slope is estimated and the standard uncertainty due to the fill sequence trend is estimated.

To estimate the standard uncertainty due to the fill sequence trend, each bottle is treated as having a concentration which comes at random from a uniform distribution. The uniform distribution is centered at the mean of the certification measurements and has a range given by the slope multiplied by the number of bottles. Although a bottle's concentration is determined by its serial number in the fill sequence, the customer receives a bottle at random, so the random variable representation has a natural interpretation. Under this model, the standard uncertainty due to material is

$$\frac{b\ N}{\sqrt{12}} \qquad (1)$$

where b is the estimated slope and N is the number of bottles in the population. For simplicity, the uncertainty of the slope has not been included here; the material standard uncertainty is assigned infinite degrees of freedom.

Example 4.1.1 Purity of Lithium Carbonate, SRM 924a
Using coulometric titration, duplicate samples from each of eight bottles (selected according to a stratified random sample) were analyzed, since random between-bottle differences, not a trend, were anticipated. The first sample from each bottle was analyzed before the second sample from any bottle was begun, but the measurement sequence was randomized within each group. There was no trend in the results with measurement order (see fig. 4.1.1), but there was a statistically significant trend in the results with fill sequence (see fig. 4.1.2). If the samples had been measured in bottling order, it would have been impossible to separate a fill sequence trend from a measurement sequence trend. The measurement mean, 99.867 %, and the slope of the material trend, 0.000036587 %, were estimated for this SRM. Since there were 666 bottles of the material, the material standard uncertainty is 0.00703 %.

Other Type A sources of uncertainty and several Type B sources of uncertainty were included as well. The combined standard uncertainty (including all Type A and Type B sources) was 0.00838 %. Clearly, the material trend was the dominant source of uncertainty.

4.2. Between-Bottle Heterogeneity

Random between-bottle heterogeneity tends to occur in solid natural materials and in viscous liquids which have suspended solids. In testing for between-bottle heterogeneity, it is assumed that the material is homogeneous within bottles, so that any observed variability between samples within a bottle is ascribed to the measurement process. When designing an experiment which will measure n samples to test for between-bottle heterogeneity, n/2 bottles are selected according to a stratified random sample from the batch. Duplicate samples are taken from each bottle.

If the homogeneity study will only be used to assess heterogeneity, but not to determine an accurate concentration estimate, then as many factors as possible (such as calibration, day, etc.) should be held constant. (Of course, if the homogeneity study will also be used to estimate concentration, then these sources of variability should be designed to vary in the experiment, as described in sec. 5). Whenever possible, it is important to avoid any extraneous measurement variation to maximize the power of the test of hypothesis of no between-bottle variance. However, unavoidable variation should be explicitly designed into the experiment so that variation due to the measurement process is observed as within-bottle variability, and is not confounded with between-bottle variability. To this end, two completely separate samples are prepared and analyzed from each bottle. Often, one sample is analyzed from each bottle before the second sample from any bottle is begun so that measurement trends won't be confounded with between-bottle variation.

If the homogeneity study was designed to test for between-bottle variability, an analysis of variance is used to test the null hypothesis of no between-bottle variability. If the null hypothesis is rejected, then the variance component for bottle and its degrees of freedom are estimated from the homogeneity study. The square root of the between-bottle variance component is the estimate of material standard uncertainty, and the material will be certified using a prediction or tolerance interval, at least for that analyte. If the homogeneity study is based on relative (as opposed to absolute) measurements because it was only intended to assess homogeneity, then the relative material standard uncertainty is estimated.

Example 4.2.1 Iron in Moderately Elevated Trace Element Soil, SRM 2711
X-ray fluorescence spectrometry was used to check the homogeneity of several elements in SRM 2711. Duplicate samples from each of 12 bottles were prepared and analyzed. The graph for Iron in fig. 4.2.1 supports the results of the analysis of variance, which shows a statistically significant bottle effect. The relative

material standard deviation (the square root of the between-bottle variance component) is 0.18 %.

When a precise measurement method is used for the homogeneity study, it is possible to get a statistically significant bottle effect whose size is of no practical significance. This does not affect the treatment of the data, since the estimate of material variability will also be small relative to the practical needs for the SRM, and a prediction or tolerance interval will still be useful.

When the entire certification process is designed, only one chemical method is chosen to assess between-bottle homogeneity for each analyte. If two methods with differing precisions were used, the lack of power of the test for the less precise method could easily produce contradictory results. Therefore, the most precise method is selected for the homogeneity study. Another reason for selecting only one method for the homogeneity study is so that other chemical methods can use one sample per bottle instead of two. This way, the number of bottles measured by each additional method is twice as large as if the method was being used to assess homogeneity (assuming the same number of measurements would have been made either way). This reduces the uncertainty of the mean concentration, if the material turns out to be heterogeneous, without increasing the number of measurements required.

4.3. Within-Bottle Heterogeneity

Even though within-bottle variability is the most likely form in which to find material heterogeneity, it is difficult or impossible to test for in this class of SRM. The problem is that, when destructive analytical methods must be used, it is impossible to perform the complete measurement process (including sample preparation or dissolution) on the same sample twice. As a result, the between-sample variance of the measurements estimates the sum of the material variance and the variance due to the sample preparation process.

Instrumental neutron activation analysis (INAA), a nondestructive method, is sometimes used to detect within-bottle variation. Virtually no sample preparation is required for this method, and since the concentration is determined by calibrating counts, the Poisson nature of the counts gives an estimate of measurement variability. A Chi-squared test of the null hypothesis that the observed between-sample variance is no larger than expected based on the within-sample (Poisson-based) variance can be used to decide whether or not the material is homogeneous. However, this can lead to an overestimate of material variability when other aspects of the analysis contribute to the observed between-sample variability, which is often the case. Therefore, the within-sample standard deviation is often inflated to account for other suspected sources of variability before the test of hypothesis is done.

The Paule-Mandel algorithm, described in Section 8, can be used to estimate between-sample heterogeneity. The degrees of freedom are taken to be the number of samples minus one.

> **Example 4.3.1** Chlorine in Level III of Lubricating Base Oil, SRM 1818a
> Six samples, one per bottle from a stratified random sample, were measured by INAA for chlorine in each of five levels of this material. For Level III (this example), the within-sample standard deviation, as obtained from Poisson counting statistics, was 0.31 mg/kg. The observed standard deviation of the data (which contains contributions from counting statistics and material heterogeneity, if it is present), was 0.74 mg/kg. To determine whether or not there were statistically significant differences between samples, the null hypothesis that the variance between the samples was equal to the known within sample variance, σ_0^2, was tested against the alternative that it was greater. Because the observed between-sample variance, 0.55 (mg/kg)^2, was larger than the appropriate cutoff,
>
> $$\frac{\sigma_0^2 \chi_{n-1}^2}{n-1} = 0.22\ (\text{mg/kg})^2,$$
>
> the null hypothesis was rejected. The estimate of the between-sample standard uncertainty, which is the difference, in quadrature, of the observed standard deviation and that predicted by the counting statistics, was 0.68 mg/kg.

Isotope dilution mass spectrometry (IDMS) is also used occasionally to detect within-bottle variation. This method is quite precise, and the analysts frequently have *a priori* information about the variability of the measurement process. The same approach is followed as for INAA, except that the *a priori* information is used instead of the Poisson-based estimate of measurement process variability.

5. Selecting Chemical Methods

The selection of chemical methods is based on chemical judgment about which methods give the most accurate and precise data for an analyte, the cost of the analysis, and the availability of analysts and equipment. The usual approach to certification of chemical SRM batches is either to use a definitive chemical method or to use two or more independent methods for each analyte.

According to Uriano and Gravatt [3], "Definitive methods of chemical analysis are those that have a valid and well-described theoretical foundation, have been experimentally evaluated so that reported results have negligible systematic errors, and have high levels of precision." They are not just considered definitive by scientists at NIST, but by chemists nationwide. For some inorganic elements, IDMS with thermal ionization is considered definitive [3]; a gas chromatographic mass spectrometry (GC-MS) method for cholesterol is another example of a definitive method. Often, a second method is used in conjunction with the definitive method as a blunder check.

However, for many analytes which are certified in SRMs, no definitive method is available; this situation is the focus of this paper. In this case, the design usually involves at least two chemically independent methods. Epstein [4] includes a table of references describing the development of the independent method concept at NIST. Chemical independence is required so that the potential sources of bias in each method are different. For example, incomplete dissolution might be a problem for atomic absorption spectrometry (AA), but it would not be for INAA since no dissolution is done. The use of at least two chemically independent methods is based on the assumption that if methods with different potential bias sources agree well, then it is likely that neither method is significantly biased. However, defining "chemical independence" is difficult, since many factors such as the source of calibration standards, sample preparation, sample analysis, and data evaluation are involved [4].

6. Experiment Design

When designing the measurement process for an analyte, the goal is to use as few measurements as possible while minimizing bias and determining a realistic and useful uncertainty estimate for the results. In order to do this in a cost effective manner, the measurements should be as independent as possible. Two measurements are independent if they do not share the same level of any significant factor. For example, if sample preparation introduces variability into the measurements (which is often the case), aliquots must come from separately prepared samples for their measurements to be independent.

The goal of most classical experiment designs discussed in statistics and engineering literature is to determine which factors have a significant effect on the response of interest. Designs with few (often two) levels per factor, carefully set up so that each main effect and many interactions can be estimated, are popular in those situations. However, designs used for SRM analyses are different!

For one thing, any chemical method being used to analyze an SRM for certification should already be well-understood, even if it is not a definitive method. The chemists should already know which factors introduce the largest sources of variability, and they should have already optimized the method so that the effect of these factors has been minimized. For another thing, the goal of the SRM measurements is to learn about the material, not the analytical method.

To this end, measurements for SRM certification are designed to maximize independent information about the concentration of the analyte in the material. The focus is not on determining important sources of variability, but on ensuring that all unavoidable sources of variability are replicated so that their effects will be reflected in the variability of the measurements, allowing their contributions to the uncertainty of the method to be (empirically) estimated. It is very important that the uncertainty of each measurement is not underestimated, because ultimately at least two methods will be compared. When two methods don't agree according to a statistical test (which is often the case), understanding the differences between them can be very difficult. Having realistic estimates of the variances of the methods ensures that as little as possible of the observed method differences are attributed to something poorly understood.

6.1. Underlying Assumptions

In order to understand SRM designs, one must first understand the underlying assumptions about the measurement process, which are described in terms of a random effects model. Typically, one assumes that each replicate of a factor (sample dissolution, instrument calibration, etc.) introduces a random error into the measurement process, and that for each factor, the random error is independent and identically distributed according to a normal distribution with mean 0 and some variance. It is also assumed that these effects are independent across factors.

If only one replicate of a factor was observed (for example, only one sample of the material was dissolved) and the variance due to that factor was unknown from previous experiments, there would be no empirical basis on which to estimate it. As a result, instead of including a variance component equal to the variance due to that factor in the uncertainty, a variance component of 0 would be included. If that factor was a large source of variability, this could result in a gross underestimate of the uncertainty of the measurement.

6.2. Replication of Factors

An ideal design (for ease of statistical analysis and amount of independent information) might be to replicate each factor in each measurement. An example of such a design is: each day, dissolve a sample from one bottle, calibrate the equipment, and measure one aliquot of the sample; repeat the entire process, with different bottles and calibrations, for n days. This type of design is only efficient if all factors are equally important. Certainly, if one of these factors, such as instrument calibration, turns out to introduce little variability into the results, then lots of unnecessary effort would have been expended. Although this ideal design is never seen in practice, it is a good place to start when explaining replication requirements for more realistic designs.

Each factor must be replicated in the design enough times so that, if it is the largest source of variability, enough degrees of freedom are available to estimate it reasonably well. For example, suppose only two samples were dissolved and eight aliquots from each dissolved sample were measured. If sample dissolution was a significant source of variation, then effectively, the 16 aliquot measurements would have resulted in two independent observations (the mean of each group of eight measurements). The correct standard deviation of the grand mean would be the standard deviation of these two means divided by the square root of two, which has one degree of freedom.

A useful rule of thumb is to be sure that each factor is replicated at least four times in a design (more if possible), especially if the factor is known to be an important contributor to measurement variability. If one factor introduces more variability into the measurement process than any other factor, then the variance of the mean concentration will depend primarily on that factor, and a confidence interval for the method mean will use a multiplier based mostly on that factor's degrees of freedom. Although the multiplier for a confidence interval continues to shrink with each additional degree of freedom, the change is most dramatic for low degrees of freedom, as is shown in the following table of multipliers for two-sided 95% confidence intervals:

n	t_{n-1}
2	12.706
3	4.303
4	3.182
5	2.776
6	2.571
.	
.	
.	
30	2.045

With four observations, the multiplier for a two-sided 95% confidence interval is 3.2 (which most chemists find tolerable), and chemists can often be persuaded to make four replications.

If the chemical method is not being used in a homogeneity study, only one sample per bottle is analyzed, not two. This helps to maximize the amount of independent information available from a fixed number of measurements, since if between-bottle variation exists in the material, duplicate samples from each bottle will not be independent.

6.3. Run Sequence and Blocking

The measurement run sequence must always be randomized. If all factors are replicated in each measurement, randomization may be the only apparent design feature that statistical methodology introduces. If multiple measurements are made within a single replicate of a factor (such as day), then the randomization may be constrained. The achievable degree of randomization depends upon the measurement process.

> **Example 6.3.1** A simple experiment design.
> As a simple example, consider a measurement process in which n samples are dissolved, and all n samples are measured on each of 4 days. This type of design is often used for atomic absorption spectrometry. The n samples come from a stratified random sample of n bottles, one sample per bottle. On each of the measurement days, the instrument is calibrated and the samples are measured in a random order which is different each day. This design permits checking for between-day and between-sample variability, although since only one sample per bottle is dissolved, there is no way to distinguish between material differences and sample preparation effects, if between-sample variability is statistically significant. If between-sample variability dominates, then there are n-1 degrees of freedom; if between-day variability dominates, then there are 3 degrees of freedom. If neither between-day nor between-sample variability is statistically significant, then there are 4n-1 degrees of freedom.

When steps are done in batches (for example, if sample preparation is done over several days, with a subset of the samples prepared each day), the batches should be balanced. When multiple steps must be done in batches (e.g., the samples, which were prepared in batches, cannot all be measured in the same day), balanced designs are sought to define all of the batches. Within a batch, randomization of the run sequence is always done. The order in which batches are analyzed is also randomized, if possible.

> **Example 6.3.2** An experiment design involving batches for preparation and measurement.
> Another example design involves the measurement of 12 samples with duplicate injections. Only three samples can be prepared per day, so four preparation days are needed; the samples are assigned to preparation day at random. Each sample will be injected twice; since four injection days are desired, six injections will be made each day. The duplicate injections from a single sample take place on separate days.
>
> A variety of possibilities exist for regrouping the samples to do the injections. The simplest is to inject the samples from preparation days 1 and 2 on injection days 1 and 2, and the samples from preparation days 3 and 4 on injection days 3 and 4 (the order of the injection days would, of course, be randomized). However, since pairs of preparation days are associated with pairs of injection days, a degree of freedom is lost for each of these effects (i.e., there will only be 2 degrees of freedom for preparation day and 2 degrees of freedom for injection day). Alternatively, for each injection day, one sample from each of two preparation days and two samples from each of the other two preparation days are injected. This confounds portions of the effects, but preserves degrees of freedom (i.e., there will still be 3 degrees of freedom for preparation day and 3 degrees of freedom for injection day).

Unfortunately, given realistic constraints about chemical analyses, designs are seldom optimal in any sense. However, every effort is made to ensure that desired variance components are estimable and that the design is as balanced as possible.

6.4. Calibration

Calibration is done for each chemical method. For some methods, a separate spike addition (of an extremely pure material) is added to an aliquot from each sample. In this way, each sample is calibrated individually and all uncertainty in the calibration process is automatically incorporated into the observable measurement process uncertainty.

For other methods, calibration is done for all samples at once. Several calibrants may be prepared to span a range around the concentration of the analyte, two calibrants tightly bracketing the concentration may be prepared, or a single calibration solution that is close in concentration to the analyte may be used. In these cases, a single linear model is fit to the calibrants and the fitted curve used to calibrate all the samples.

One source of uncertainty in the calibration is measurement of the calibrants. Each response for a calibrant is measured with some imprecision, so no calibration equation, whether it is a straight line or single-point calibration, can be fit without uncertainty. This variability is reflected in the standard deviation of the response factor if a single-point calibration was done, or the standard deviations of both parameters (slope and intercept) and their covariance if a straight line, not forced through the origin, was used.

Another source of uncertainty in the calibration is preparation of the calibrants. Often, previously certified SRMs which were designed for calibration use are used as calibrants; in this case, their uncertainty is well characterized (and usually quite small relative to other sources of uncertainty in the measurement process). However, calibrants often must be prepared gravimetrically, and uncertainty in the true concentration may have a large effect. In the case of straight line calibrations, it may be possible to see this effect when replicate measurements are made for each calibrant, and the deviations of the observations from the fitted line display a pattern (i.e., all measurements for one calibrant fall above the fitted line, while all measurements for another fall below the line).

When calibrant preparation is likely to be a significant source of uncertainty, it should be replicated, just like every other factor in the entire measurement process. If gravimetric preparation of the calibrants is required, the pure materials should be weighed into each calibrant separately, if possible--dilutions of a stock solution do not replicate errors in weighing the original material. For a straight line calibration, at least four concentrations should be selected at which to prepare calibrants. If replicate measurements are made and the residuals from a straight line show lack of fit, then the average response for each calibrant is used to fit the calibration line; the degrees of freedom for calibration are the number of calibrants minus two. When single-point calibration is to be used, several (at least four) calibrants should be made up at approximately the same concentration. If there is a statistically significant difference between the response factors for each calibrant, then the average response factor for each calibrant is its summary, the grand average of the response factors is used to compute the mean concentration, and the degrees of freedom for the response factor uncertainty are equal to the number of calibrants minus one.

Example 6.4.1 GC-ECD Calibration of PCB 52 in Marine Sediment, SRM 1941a

Four calibrants were prepared at concentrations spanning the concentration of PCB 52 in the SRM. Each was measured twice, and a straight line through the origin was fit to the data (a straight line through the origin was selected based on knowledge of the measurement process). As shown in Figure 6.4.1, the observed deviations from the line fit are not independent for each calibrant. Therefore, to assess the uncertainty in the calibration, the duplicate measurements were averaged and the fit was redone. Since the same number of measurements had been made for each calibrant, this did not affect the slope of the line, but it did increase the standard deviation of the slope from 0.072 to 0.104 and it decreased the degrees of freedom from 7 to 3.

Often, a single calibration is done for the entire experiment, which means that its uncertainty is not reflected in the variability of the calibrated responses. The calibration is effectively a constant, and if that constant is uncertain, the final results of the experiment are just as uncertain. When a single calibration is used for the whole experiment, the uncertainty in the fitting process, as described above, must be explicitly incorporated into the total uncertainty of the calibrated measurements. This can be done by propagation of errors. Of course, if the calibration were done several times, (which is the case with spike additions done separately for each sample, or a different calibration on each of several days) then the variability introduced by the calibration is already evident in the data, and an analysis of variance can be used to check for its significance.

6.5 Sample Preparation Blanks

Particularly when measuring trace elements, sample preparation blanks are checked in each experiment. A sample preparation blank is the background signal that is not due to analyte in the material of interest. Samples containing no material are processed in the same way and at the same time as real samples through the entire measurement process. If the blanks are significantly greater than zero (according to chemical significance, not necessarily statistical significance), then their mean is subtracted from the total analyte found in each sample; the variance of the mean blank must be included in the total uncertainty. Guidance about replication for blanks is the same as for other factors, although blanks usually have a small impact on the final uncertainty unless the measurements are near the detection limit of the method.

6.6 Controls

Whenever an appropriate material is available, at least two control samples are included in the measurement process. The selection of appropriate controls is determined by the chemists. Controls are used in a go/no go capacity--if bias is detected for the control (i.e., the new measurements on the control are "too far" from the certified value, where the definition of "too far" depends upon the requirements of the analysis), the measurements on the unknown are questioned because the measurement process itself was assumed to be biased. The uncertainty of the control measurements is not included in the uncertainty of the new SRM material, however, since the control is almost never used to correct analytical results for SRM certification.

6.7 Assessment of Purity

When calibration SRMs are prepared by dissolving "pure" materials in a solvent, the purity of each material plays an important role in estimation of the concentration. Since the uncertainty of the purity estimate must be included in the total uncertainty of the SRM, replicate assessments of the purity must be designed into the study so that a reliable estimate of purity (with an appropriate uncertainty) is available. If purity assessment methods are suspected of bias, then multiple, independent purity assessment methods should be used, for the same reason that multiple independent analytical methods are used to determine concentration in natural matrix SRMs.

7. Analysis of Individual Method Results

It is important that the variability of the mean for each method is estimated correctly, because more than one method will be compared to determine the certified value. When two methods don't agree according to a statistical test (which is often the case), understanding the difference between them can be very difficult. Having realistic estimates of the variance of each method mean ensures that as little of the observed method differences as possible are attributed to something poorly understood and difficult to handle statistically.

The statistical outputs from analysis of an individual method's results are the mean, the standard uncertainty of the mean, and its degrees of freedom. In order to estimate the standard uncertainty of the mean, the data must be modeled correctly.

With the experimental data, analysis of variance is used to determine which design factors have a statistically significant effect on the measurements. A parsimonious model is developed by eliminating effects which are not statistically significant. Many of the factors in the design are factors which were not expected to be significant, but were included in the design as insurance in case they were problematic. Therefore, eliminating them from the model is reasonable. Once an appropriate model has been found, variance components are estimated and combined to determine the variance of the mean.

The variance components are estimated as weighted sums of mean squares (the weights depend upon the expected values of those mean squares), since this provides a simple mechanism for estimating degrees of freedom, using Satterthwaite's approximation. Expected values of mean squares are discussed in many introductory statistics texts which discuss analysis of variance, such as Mendenhall and Sincich [5]. Satterthwaite's approximation gives an estimate of degrees of freedom for a weighted sum of independent mean squares (when the weights are positive). If the variance component is the weighted sum of I mean squares:

$$\text{Variance} = \sum_{1}^{I} w_i \, MS_i \qquad (2)$$

where MS_i is the i^{th} mean square, then the effective degrees of freedom, according to Satterthwaite [6], are:

$$df = \frac{\left(\sum_{1}^{I} w_i \, MS_i\right)^2}{\sum_{1}^{I} \frac{w_i^2 \, MS_i^2}{df_i}} \qquad (3)$$

Satterthwaite's approximation does not perform well when one or more of the weights is negative, especially if the mean square associated with the negative weight is relatively large.

Therefore, eliminating nonsignificant effects from the model is important for getting useful estimates of the variance of the mean and its degrees of freedom. For one thing, these variance component estimates can be negative if the observed F statistic for testing the statistical significance of that effect is less than 1. For another, when several effects are in the model, the variance of the mean usually involves subtraction of at least one mean square, which means that Satterthwaite's estimate of degrees of freedom can fail. Although these are problems with the methodology, not the data, they can be minimized by reducing the model to a parsimonious, statistically significant model in the first place.

Example 7.1.1 Arsenic by FIA-HAAS in Estuarine Sediment, SRM 1646a
Eight samples were analyzed in each of three runs by FIA-HAAS (see fig. 7.1.1). An analysis of variance showed that both sample preparation and run were statistically significant:

Source	DF	Mean Square	F Value	p Value
Sample Prep	7	0.3472	11.59	0.0001
Run	2	0.2376	7.93	0.0050
Error	14	0.0299		

The mean is 6.405 µg/g and the variance of the mean is

$$\mathrm{Var}(\bar{y}) = \frac{\sigma^2_{\mathrm{sample}}}{8} + \frac{\sigma^2_{\mathrm{run}}}{3} + \frac{\sigma^2_{\mathrm{error}}}{24} .$$

Since the expectations for these three mean squares are:

$$\begin{aligned} E[\mathrm{MSsample}] &= \sigma^2_{\mathrm{error}} + 3\,\sigma^2_{\mathrm{sample}} \\ E[\mathrm{MSrun}] &= \sigma^2_{\mathrm{error}} + 8\,\sigma^2_{\mathrm{run}} \\ E[\mathrm{MSerror}] &= \sigma^2_{\mathrm{error}} \end{aligned}$$

the variance component estimates are:

$$\begin{aligned} \hat{\sigma}^2_{\mathrm{error}} &= \mathrm{MSerror} \\ \hat{\sigma}^2_{\mathrm{run}} &= \frac{\mathrm{MSrun - MSerror}}{8} \\ \hat{\sigma}^2_{\mathrm{sample}} &= \frac{\mathrm{MSsample - MSerror}}{3} \end{aligned}$$

and the estimate of $\mathrm{Var}(\bar{y})$ is

$$\hat{\mathrm{Var}}(\bar{y}) = \frac{\mathrm{MSsample + MSrun - MSerror}}{24} = 0.023119.$$

The Type A combined standard uncertainty is $\sqrt{\hat{\mathrm{Var}}(\bar{y})} = 0.152$ µg/g and its degrees of freedom are 6.76, using the Satterthwaite approximation described above.

Uncertainty due to "side experiments" must be propagated into the standard deviation of the mean. If a global calibration was applied to the data, its uncertainty must be included; likewise, uncertainty due to blanks must be included. This can be done by propagation of errors, as described in references [1] and [2].

Example 7.1.2 Sulfur in Level III of Lubricating Base Oil, SRM 1819a
One sample from each of six bottles was analyzed by ID-TIMS for S in this material. Two blanks were carried through the process, and their mean was used to correct the total amount of sulfur in the samples. Since NIST has a 10-year history of sulfur blanks (approximately 200 blanks in all), the long-term variance was used to approximate the variance of a single blank; this variance was divided by 2 to estimate the uncertainty due to the blank correction (since the average of two blanks was used). This variance was normalized for the weight of the samples, since the blank correction was applied to the total amount of sulfur in the sample but results are reported as microgram per gram of sample. Twelve different mixes of the spike solution with two salts were used to calibrate the spike solution. Because the calibration enters the final concentration multiplicatively, the relative variance of the mean spike solution concentration was added to the relative variance of the mean measured concentration to incorporate its uncertainty. A breakdown of these Type A sources is given on mg/kg scale:

Type A Source	**Standard Uncertainty mg/kg**	**Degrees of Freedom**
Sample Measurements	6.21	5
Spike Calibration	2.20	11
Sample Blank	0.98	169
Combined (Type A)	6.66	6.57

Type B combined standard uncertainties and Type B degrees of freedom must also be assessed for each method. This should be done by the chemists, since Type B standard uncertainties are estimated using chemical judgment. The approach to combining them has been described in the ISO "Guide to the Expression of Uncertainty in Measurement" [2] and the NIST guide [1].

Example 7.1.2 (Continued) Sulfur in Level III of SRM 1819a
According to the analyst, additional, uncorrectable uncertainty could come from uncertainty in the purity of the spike solution and from mass fractionation. The analyst estimated the standard uncertainties due to these effects to be 0.0289 % and 0.0816 %, respectively. A complete table for Sulfur in this SRM, itemizing all sources of uncertainty on mg/kg scale, is:

Source	Standard Uncertainty mg/kg	Degrees of Freedom	Uncertainty Type
Sample Measurements	6.21	5	A
Spike Calibration	2.20	11	A
Sample Blank	0.98	169	A
Purity of Spike Cal.	1.16	∞	B
Mass Fractionation	3.28	∞	B
Combined:	7.51	11	
Mean Value of S	4022.00		

Once the individual method results have been summarized, the results of multiple methods can be combined to determine certified values and total uncertainties.

8. Combining the Results of Two or More Independent Methods

8.1 The Need for Weighting

When no definitive method is available, the NIST Standard Reference Material Program typically uses two or more chemically independent methods to determine a certified value and its uncertainty [4]. The independent methods are intended to confirm each other's results; good agreement suggests that neither method is biased. Even if there is a statistically significant difference between the methods, the agreement between methods may still be adequate for the intended use of the SRM, based on the subjective determination of the experts involved. In this case the data can still be combined in a statistical framework, and the property can be certified.

To combine the results, a weighted average of the method means is computed; a mean and standard uncertainty are never calculated from all of the observations from the different methods lumped together. For one thing, if variance components were needed or if side experiments were incorporated for an individual method, such a "lumped" standard deviation would not include them. For another thing, even if the observations within each method are independent, if the error variance is different for each method, then calculating a raw mean of all the observations from all methods would not give the minimum variance estimate of the grand mean.

8.2 Equal versus Unequal Weighting

When computing a weighted average of the method means, the weights can be equal, or they can be estimated using an algorithm developed by Paule and Mandel [7]. Equal weighting is the simplest choice. Equal weights are easy to compute, and since the weights are fixed, the variance of the weighted mean can be estimated correctly. However, if the method precisions are very different, equal weighting will not minimize the variance of the weighted mean. If the methods agree well, it makes intuitive sense to give the most precise method the most weight.

The weighting algorithm of Paule and Mandel is often implemented for multi-method SRMs. The weight for each method is inversely proportional to the sum of the variance of its mean and the between-method variance. When the methods agree well (i.e., the between-method variance is 0), the Paule-Mandel weighting scheme gives weights which are inversely proportional to the observed variances of the method means, and the estimated variance of the mean is smaller than it would have been with equal weights. However, when the methods agree poorly (i.e., the between-method variance is large compared to the variances of the method means) the weights are nearly equal. This makes sense when methods agree poorly because in that case within-method variation is not an adequate summary of precision and accuracy. This heuristic is intuitively satisfying. The disadvantage to this approach is that the weights are random, and since there is no good way to incorporate their uncertainty into the total uncertainty of the certified value, it is typically omitted. Also, the model and assumptions underlying this approach are most compelling when data are available from many chemical analysis methods, which is seldom the case.

Assuming that all of the data being combined come from NIST, the combined standard uncertainty (including both Type A and Type B sources) can be used instead of the standard uncertainty of the mean (which only includes Type A sources of uncertainty) for determining weights. It is desirable for Type B uncertainties to be included, since some methods have good precision but are known to have additional uncertainty that doesn't show up in observed measurement variability. Incorporating the Type B uncertainties into the weighting scheme prevents these methods from getting too much weight when the methods agree well.

8.3 Paule-Mandel Weights

The Paule-Mandel weighting scheme involves use of an algorithm for estimating the between-method variance, σ_b^2. To compute it assume that, for each of the M methods available, the mean

$\bar{Y}_i$ and the square of its combined standard uncertainty, S_i^2, are available. Then the method weights are defined implicitly as follows:

$$W_i = \left(\frac{1}{S_i^2 + \hat{\sigma}_b^2} \right) \tag{4}$$

the weights are

$$w_i = \frac{W_i}{\sum_1^M W_j} \tag{5}$$

and $\tilde{Y}$ is the weighted average of the $\bar{Y}$' s

$$\tilde{Y} = \sum_1^M w_i \, \bar{Y}_i. \tag{6}$$

The estimate, $\hat{\sigma}_b^2$, of the between-method variance is defined as the nonnegative value that satisfies:

$$\frac{\sum_1^M W_i(\bar{X}_i - \tilde{Y})^2}{M-1} = 1, \tag{7}$$

if such a nonnegative value exists; otherwise the estimate is defined to be 0, and the left hand side of (7) is less than 1. Paule and Mandel provide an iterative algorithm for solving this in [7].

8.4 Combining the Uncertainties

Once the weights have been determined, whether they are equal or unequal, and the weighted mean of the method means is computed for the certified value, the problem of how to assess the uncertainty of the certified value remains.

One possible estimate of the combined standard uncertainty of the weighted mean is the weighted root sum of squares of the combined standard uncertainties for the methods. This works well if the methods aren't too discrepant. The uncertainty reported on the certificate is the expanded uncertainty, which is found using this combined standard uncertainty and a t-multiplier based on its estimated degrees of freedom.

Let

A_i be the Type A combined standard uncertainty for method i

B_i be the Type B combined standard uncertainty for method i

Then S_i is the root sum of squares of these two quantities:

$$S_i = \sqrt{A_i^2 + B_i^2} \tag{8}$$

Using the approach just described, the combined standard uncertainty, S, for the weighted mean, is:

$$S = \sqrt{\sum_1^M w_i^2 \ S_i^2.} \tag{9}$$

If the material is heterogeneous, then the combined standard uncertainty is

$$S' = \sqrt{S_{mat}^2 + \sum_1^M w_i^2 \ S_i^2} \tag{10}$$

where S_{mat}^2 is the square of the material standard uncertainty (see sec. 4 for a description of determining the material standard uncertainty).

The degrees of freedom are:

$$\text{df (homogeneous)} = \frac{S^4}{\sum_1^M \frac{w_i^4 \ S_i^4}{df_i}} \tag{11}$$

or

$$\text{df (heterogeneous)} = \frac{(S')^4}{\frac{S_{mat}^4}{df_{mat}} + \sum_1^M \frac{w_i^4 \ S_i^4}{df_i}} \tag{12}$$

where df_i are the degrees of freedom for the combined standard uncertainty (S_i) for method i and df_{mat} are the degrees of freedom for the material standard uncertainty (S_{mat}).

Arsenic in Estuarine Sediment illustrates the application of this approach to computing the uncertainty in a homogeneous material.

Example 8.4.1 Arsenic in Estuarine Sediment, SRM 1646a
The summary statistics for As in SRM 1646a are:

		Type A		Type B		
Method	Mean	Std. Unc.	DF	Std. Unc.	DF	Weight
	μg/g	μg/g		μg/g		
FIA-HAAS	6.410	0.15205	6.76	0.074	∞	0.42
RNAA	6.095	0.03959	9	0.10362	3	0.58

It was determined that the material was homogeneous for this element. Using Paule-Mandel weights, the weighted mean is 6.2267 μg/g, the weighted combined standard uncertainty is 0.0957 μg/g, the degrees of freedom are 12.25, and the expanded uncertainty, which was reported on the certificate, is 0.2081 μg/g. The estimate of between-method standard deviation is:

$$\hat{\sigma}_b = 0.17 \ \ \mu g/g.$$

Note that if the material was heterogeneous and the homogeneity study data were also used as one of the methods for estimating the certified value, then the estimate of S^2_{mat} is dependent on the Type A standard uncertainty for one of the methods. If this is the case, the combined standard uncertainty (which includes material variability) must be written in terms of mean squares for the homogeneity data so that degrees of freedom can be computed correctly. This is illustrated by the following example:

Example 8.4.2 Heterogeneous material, homogeneity study data used to estimate material variance concentration.
Method 1, the homogeneity study, consists of duplicate samples measured on each of n_1 bottles, and shows a statistically significant bottle effect. The mean squares between (MSB) and within (MSE) bottles were computed. The estimates of material variability and var $(\bar{y}_1)$ are:

$$S^2_{mat} = \frac{MSB\text{-}MSE}{2} \qquad \hat{Var}\,(\bar{y}_1) = \frac{MSB}{2n_1}$$

Method 2 consists of a single measurement on each of n_2 bottles, so the estimate of var $(\bar{y}_2)$ is:

$$\hat{Var}\,(\bar{y}_2) = \frac{S^2_2}{n_2}.$$

Using weights w_1 and w_2 (whether equal or unequal) the certified value is

$$\tilde{Y} = w_1 \bar{Y}_1 + w_2 \bar{Y}_2$$

and

$$V\hat{a}r(\tilde{Y}) = w_1^2 \frac{MSB}{2n_1} + w_2^2 \frac{S_2^2}{n_2}$$

Including the estimate of material variability, the variance for prediction is:

$$V\hat{a}r(Pred) = (1 + \frac{w_1^2}{n_1}) \frac{MSB}{2} - \frac{MSE}{2} + w_2^2 \frac{S_2^2}{n_2}$$

Finally, this can be used with Satterthwaite's formula [6] to estimate degrees of freedom.

8.5 Including Between-Method Uncertainty

One goal often set for the expanded uncertainty is that the interval it defines around the certified value should cover all of the method means used to compute the certified value. When more than one method is used, they have been carefully chosen to represent different possible sources of bias. As a result, they are likely to represent extremes in measurement error and may disagree with each other statistically, although it it is difficult, if not impossible, to determine which method is "right" or "wrong" (or the "wrong" method wouldn't have been used). Therefore, the chemists feel that the certified interval should define a range which includes the method mean of each of the methods used.

The certified interval in Example 8.4.1 did cover both method means, but this is not always the case. If the propagation of within-method Type A and Type B uncertainties is not sufficient to give an expanded uncertainty that covers all method means, then an allowance for between-method differences must be added explicitly. Three approaches to adding this allowance are described below. Unfortunately, a good statistical approach for doing this is not known at this time.

A motivating example is magnesium in Estuarine Sediment, which was measured by Inductively Coupled Plasma Spectrometry (ICP), Isotope Dilution Inductively Coupled Plasma Mass Spectrometry (ICPMS-ID), and X-ray Fluorescence Spectrometry (XRF). The within-method assessment of uncertainty (Type A plus Type B) is not sufficient when the Paule-Mandel weighting scheme is used:

Example 8.4.3 Magnesium in Estuarine Sediment, SRM 1646a
The summary statistics for Mg in SRM 1646a (in weight percent) are:

		Type A		Type B		
Method	Mean	Std. Unc.	DF	Std. Unc.	DF	Weight
ICP	0.3830	0.0015411	7	0.0044225	∞	0.102
ICPMS-ID	0.3882	0.0007467	9.97	0.0004490	∞	0.851
XRF	0.3950	0.0009000	25	0.0069900	26	0.047

The weighted mean is 0.3880 %, the weighted combined standard uncertainty is 0.0009 %, the degrees of freedom are 46, and the expanded uncertainty is 0.0019 %. This summary is illustrated in Figure 8.4.3. The estimate of between-method standard deviation is:

$$\hat{\sigma}_b = 0.0015\ \%.$$

In this case, the interval defined by the weighted mean and the expanded uncertainty does not include either the ICP mean or the XRF mean.

One approach to including an allowance for between-method differences is to include the estimate $\hat{\sigma}_b^2$ in the combined standard uncertainty of the weighted mean. This would give a combined standard uncertainty of:

$$S = \sqrt{\sum_{1}^{M} w_i^2 \left(S_i^2 + \hat{\sigma}_b^2\right)} \qquad (13)$$

where S_i is the combined standard uncertainty (Type A plus Type B) for the i^{th} method. However, the joint distribution of $\hat{\sigma}_b^2$, the method means, and the method combined standard uncertainties is very complex and does not lead to simple summarization in an overall uncertainty statement. In particular, it is not known how to define degrees of freedom for $\hat{\sigma}_b^2$. If these issues are disregarded, 2 might be considered as the multiplier for the expanded uncertainty for the weighted mean. Even so, this still will not always solve the problem if Paule-Mandel weights are used:

Example 8.4.3 (continued) Magnesium in Estuarine Sediment, SRM 1646a If the estimate of $\hat{\sigma}_b^2$ divided by the number of methods is included, the new combined standard uncertainty from (13) is 0.00160 %. Thus, using a multiplier of 2, the new combined standard uncertainty is 0.0032 %, which still does not give a certified interval which includes either the ICP mean or the XRF mean.

If equal weights are used to combine the M methods, but $\hat{\sigma}_b^2$ was estimated without constraining the weights to be equal, the expanded uncertainty computed as:

$$U = 2\sqrt{\frac{1}{M^2}\sum_{1}^{M} S_i^2 + \frac{1}{M}\hat{\sigma}_b^2} \qquad (14)$$

will always cover all method means. In fact, if M=2, this uncertainty will always be exactly the difference between the two means, unless the S_i^2 (the squared combined standard uncertainties for the i^{th} method, including both Type A and Type B) are so large that $\hat{\sigma}_b^2 = 0$, in which case the uncertainty may be even larger.

Another approach to including an allowance for between-method differences is to reduce the data to only the method means, and treat each as a single independent observation. Under this approach, the certified value is the equally weighted mean of the method means. A confidence interval could be used to determine an uncertainty, but since there are usually only two methods, the interval would be too wide to be useful (with only two methods, the expanded uncertainty $U = 12.7 * |\bar{Y}_1 - \bar{Y}_2| / 2$). Also, this approach makes the most sense if the methods had been chosen at random from an infinite population of methods, which is scarcely the case. In fact, if a definitive method is not involved, the methods have been carefully chosen to represent different possible sources of bias, so they are more likely to represent extremes in measurement error. Finally, this ignores the wealth of information available from each method. However, if several methods provide data for which little information is available (such as when interlaboratory data is being used for certification), this approach is quite sensible.

A third approach, described in Schiller and Eberhardt [8], has been used extensively in the past. It does not conform to the format recommended by the ISO Guide [2], but it has produced useful intervals. Using the weights found either with the Paule-Mandel algorithm or equal weighting, the weighted combination of the Type A combined standard uncertainties from all the methods is computed:

$$A = \sqrt{\sum_{1}^{M} w_i^2\, A_i^2}. \qquad (15)$$

Only Type A sources are included at this stage, because the between-method difference which will be included below is an alternative estimate for the Type B sources. If the material is heterogeneous, then the material standard uncertainty is also included:

$$A' = \sqrt{S_{mat}^2 + \sum_{1}^{M} w_i^2\, A_i^2}. \qquad (16)$$

The degrees of freedom are:

$$\text{df (homogeneous)} = \frac{A^4}{\displaystyle\sum_{1}^{M} \frac{w_i^4\, A_i^4}{dfa_i}} \qquad (17)$$

or

$$\text{df (heterogeneous)} = \frac{(A')^4}{\displaystyle\frac{S_{mat}^4}{df_{mat}} + \sum_{1}^{M} \frac{w_i^4\, A_i^4}{dfa_i}} \qquad (18)$$

where dfa_i are the degrees of freedom for the Type A combined standard uncertainty for method i and df_{mat} are the degrees of freedom for the material standard uncertainty (S_{mat}). Using this, the Type A expanded uncertainty is computed. This accounts for the within-method random variation in the weighted mean. Then, to account for systematic between-method differences, the largest absolute deviation between any method mean and the weighted mean is computed:

$$\text{Bias allowance} = \max_i |\bar{X}_i - \tilde{X}| \qquad (19)$$

and is added linearly or in quadrature (root sum of squares) to the expanded uncertainty of the weighted mean. An interval computed this way always covers the mean of each method, but may be larger than necessary if within-method variation explains between-method differences.

Example 8.4.3 (Continued) Magnesium in Estuarine Sediment, SRM 1646a
To follow the third approach to including between-method differences, the weighted combined Type A standard uncertainty (0.00066 %), its degrees of freedom (11.56), and the expanded Type A uncertainty (0.00143 %) are computed. The maximum absolute difference between method means and the weighted mean is

0.0071 %, so the total uncertainty is 0.0085 %. The interval defined by the mean (0.388 %) and this uncertainty does include all three method means.

This section has described several approaches to including an allowance for between-method differences in the overall uncertainty of an SRM. While none of these approaches is completely satisfactory in all respects, it is hoped that this discussion will clarify issues and stimulate research that will lead to a better approach.

9. Conclusion

In order to certify SRMs, measurements must be made which are accurate, precise, and representative of the population of material which will be sold. Once the measurements are made, the data must be combined in a way which gives a fair assessment of measurement uncertainty, and includes material variability where relevant. Statistical design and analysis of material sampling and measurement processes are needed in order to achieve these goals while minimizing the cost.

Guidelines have been given in this paper for sampling from a batch of material, designing and evaluating a homogeneity study, designing and analyzing experiments to determine the certified value, and combining the results from multiple methods. It is hoped that these guidelines will be followed, or improved upon, to ensure the high quality of SRMs produced at NIST.

10. References

[1] Taylor, B.N. and Kuyatt, C.E. "Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results," NIST Tech. Note 1297, 1993.

[2] "Guide to the Expression of Uncertainty in Measurement," 1st Ed. ISO, Switzerland, 1993.

[3] Uriano, G.A. and Gravatt, C.C. *CRC Crit. Rev. Anal. Chem.* **6**, 362, 1977.

[4] Epstein, M.S. "The Independent Method Concept for Certifying Chemical-Composition Reference Materials," *Spectrochimica Acta,* Vol. 46B, No. 12, 1991.

[5] Mendenhall, W. and Sincich, T. *Statistics for Engineering and the Sciences*, 3rd Ed., Dellen Publishing Company, 1992.

[6] Satterthwaite, F. E. "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin* 2, 110-114.

[7] Paule, R. and Mandel, J., "Consensus Values and Weighting Factors," *NBS Journal of Research*, V. 87, No 5.

[8] Schiller, S.B. and Eberhardt, K.E. "Combining Data from Independent Analysis Methods," *Spectrochimica Acta,* Vol. 46B, No. 12, 1991.

Figure 4.1.1. Run Sequence Plot, Lithium Carbonate, SRM 924a

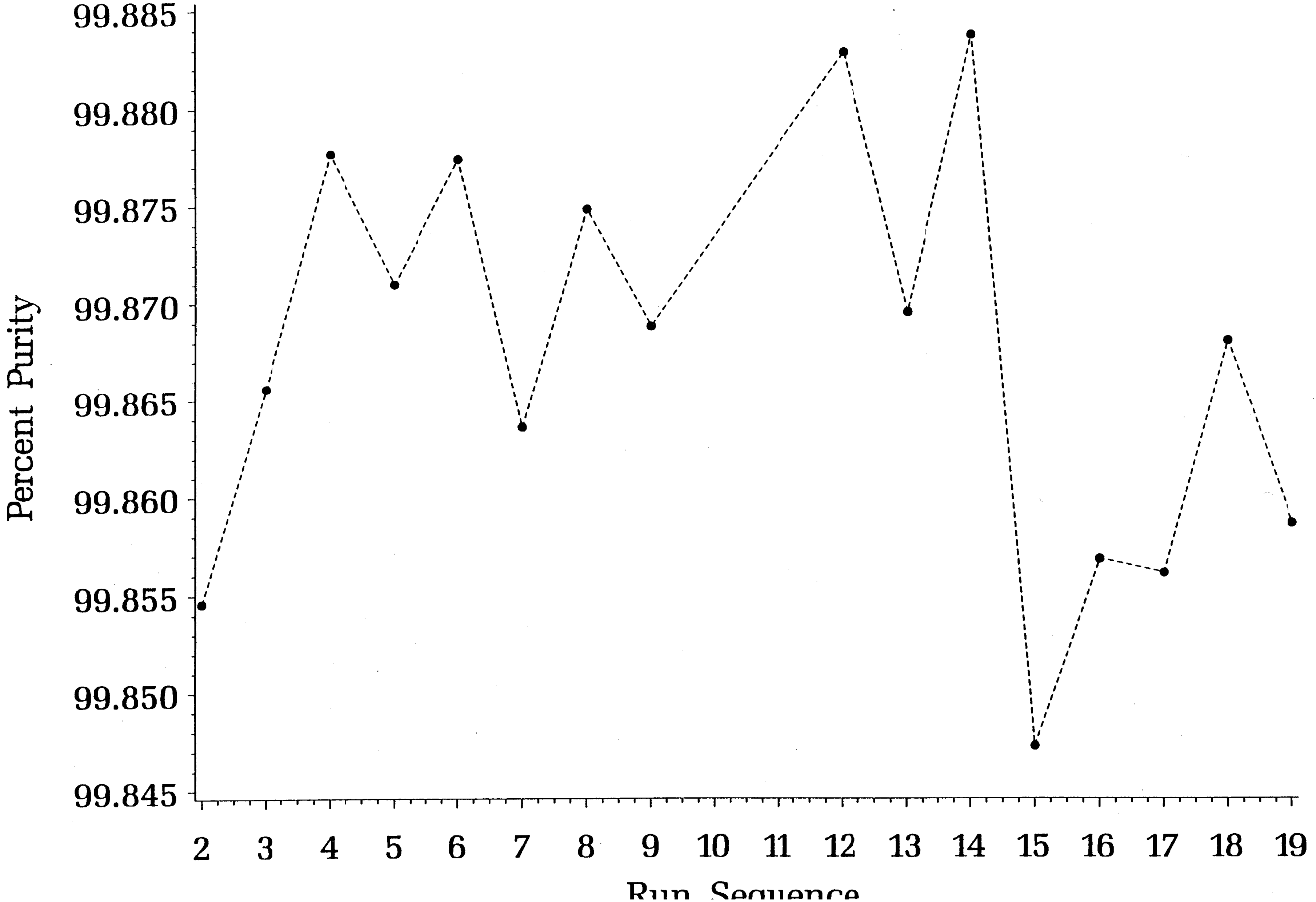


Figure 4.1.2. Fill Sequence Trend for Lithium Carbonate, SRM 924a

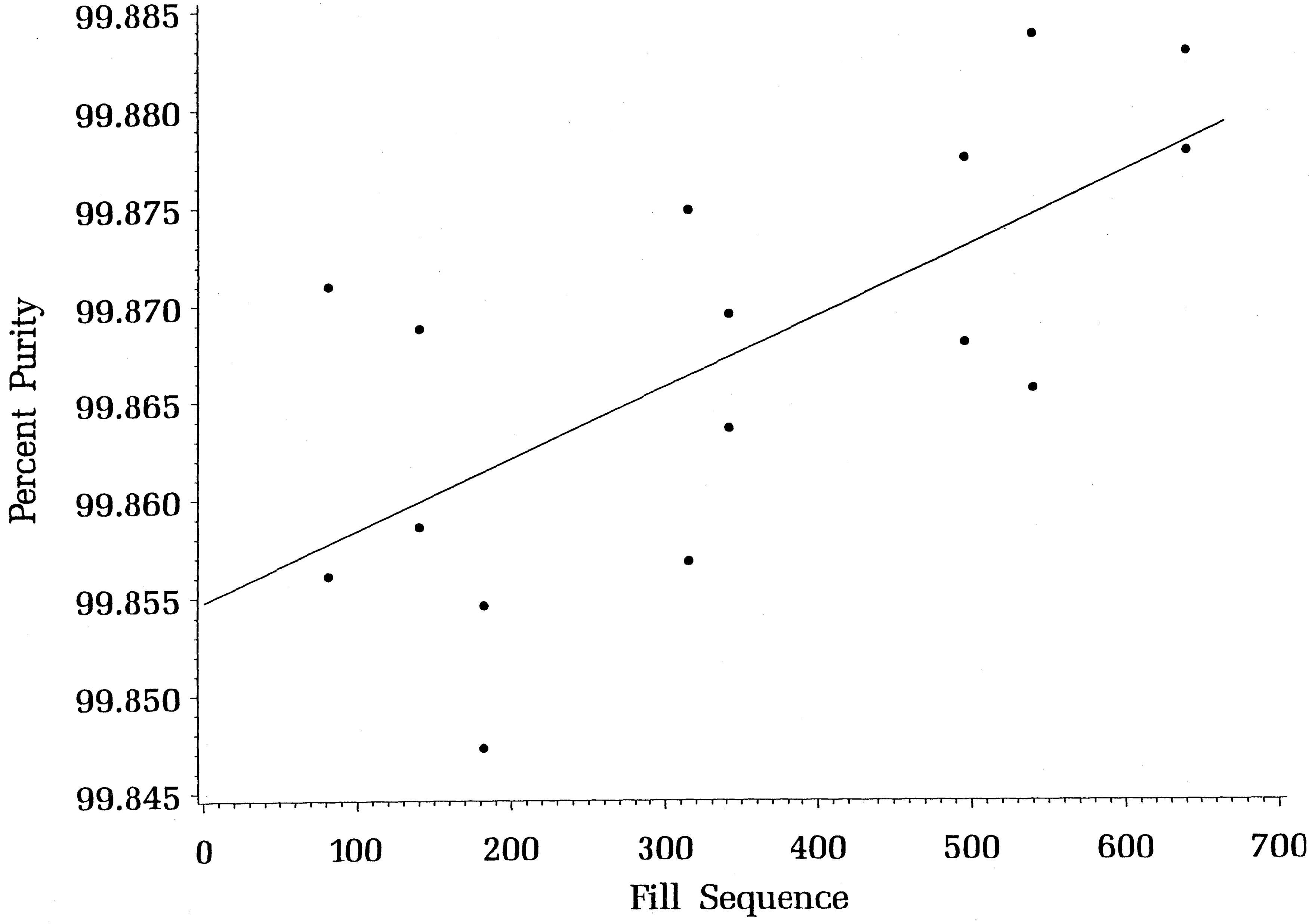


Figure 4.2.1. XRF Homogeneity Assessment
SRM 2711, Moderately Elevated Trace Element Soil

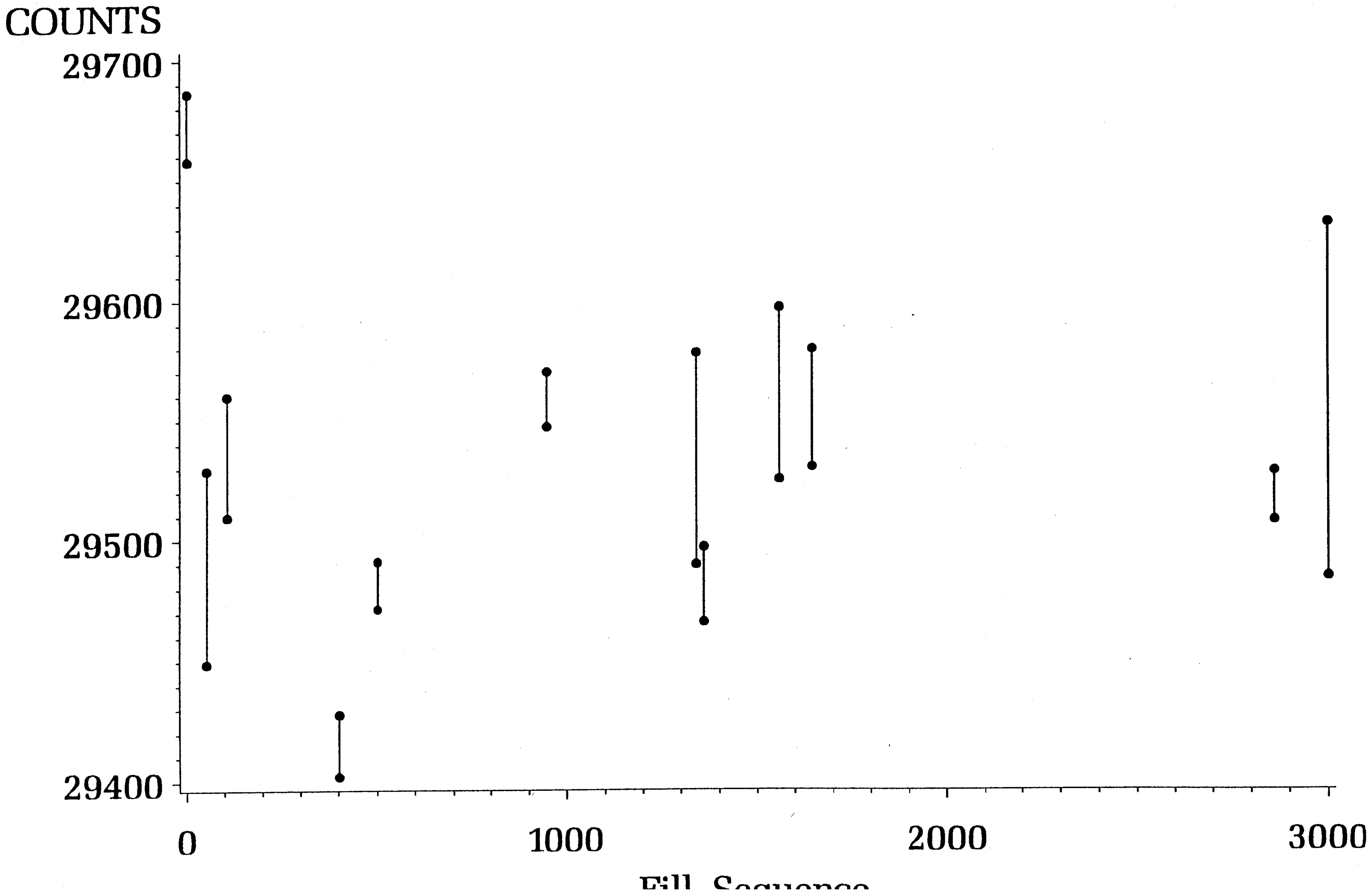


Figure 6.4.1. PCB 52 by GC−ECD in SRM 1941a
Straight Line Through the Origin for Calibration

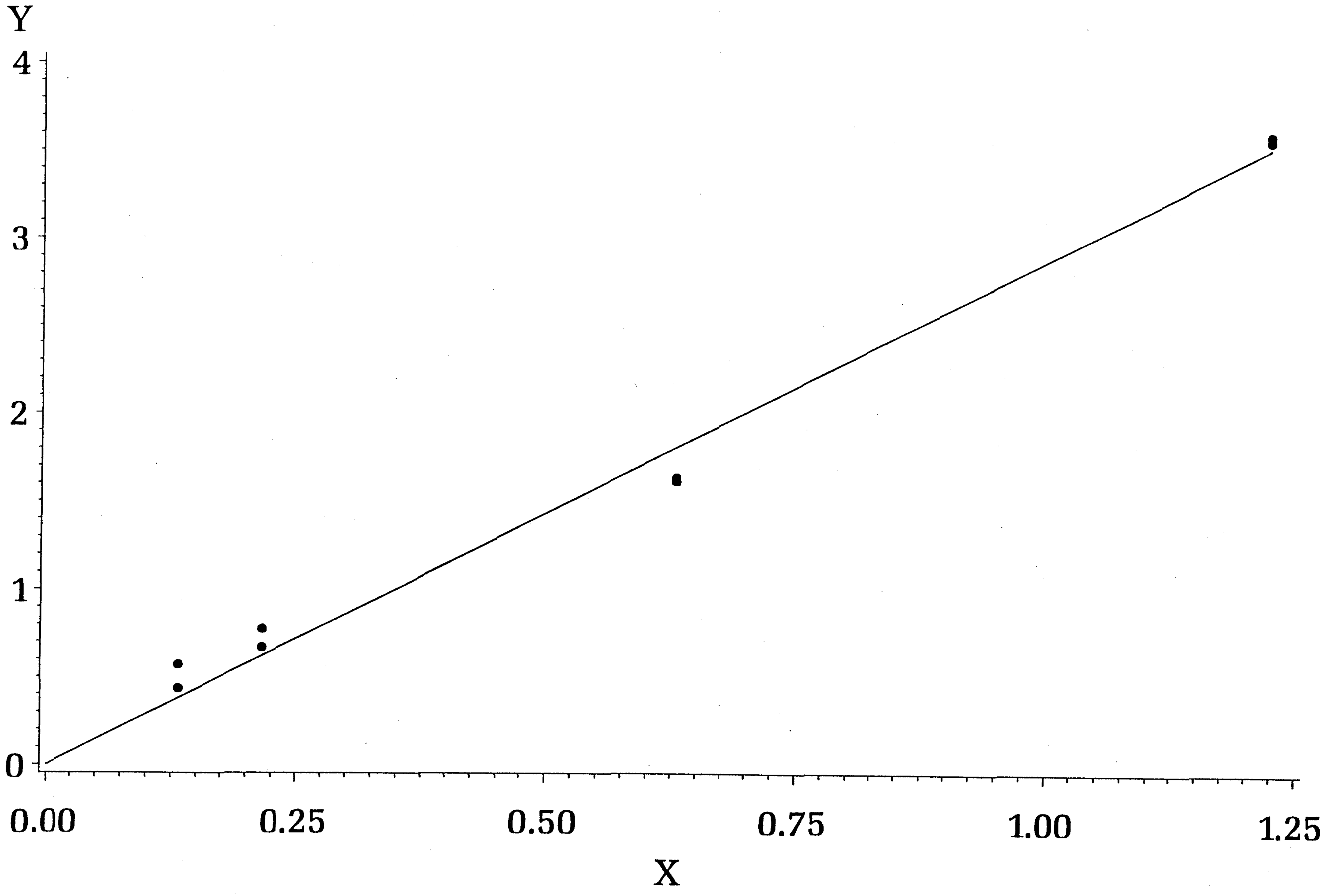


Figure 7.1.1. Run and Sample Effects for Arsenic by FIA–HAAS in SRM 1646a

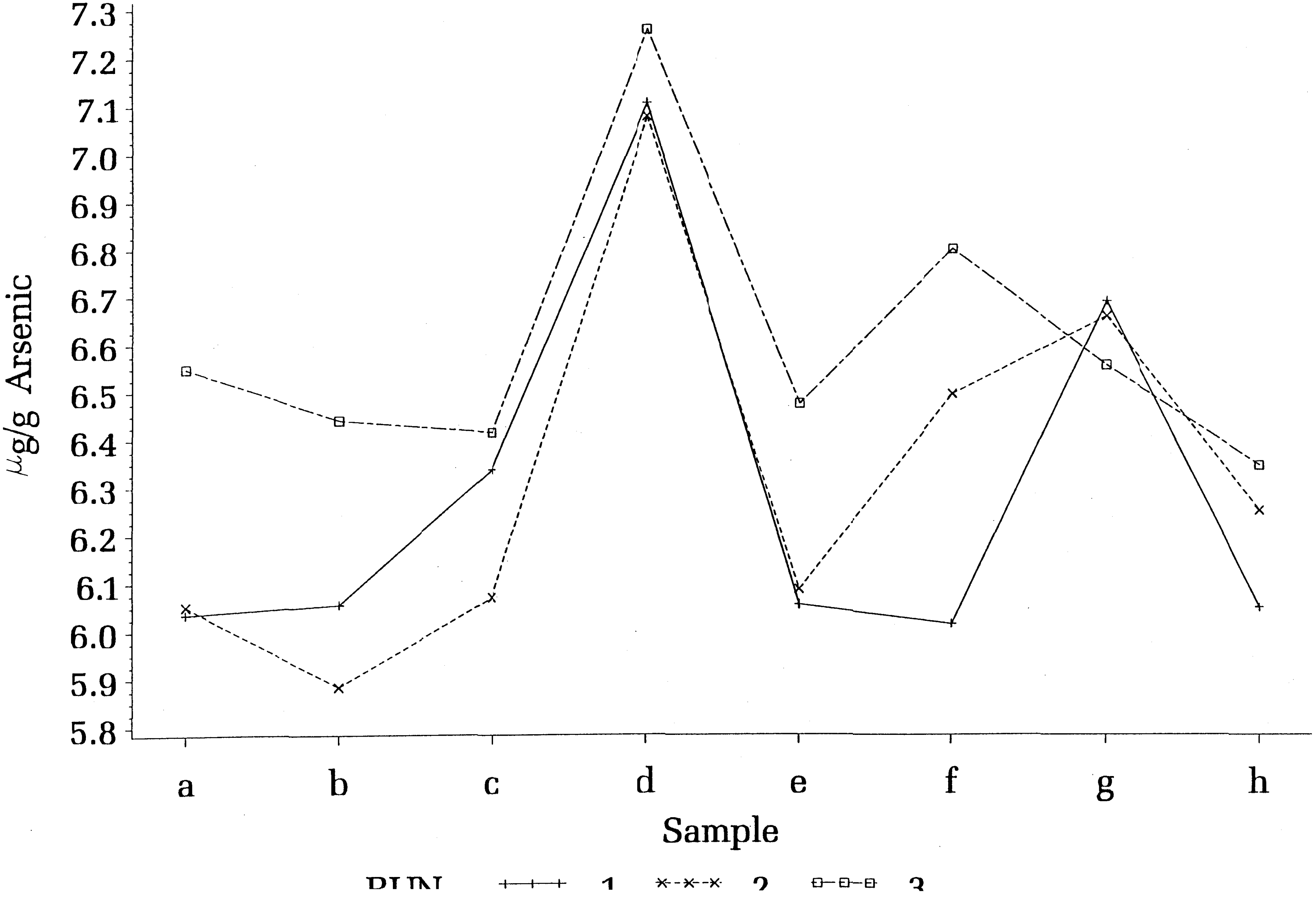
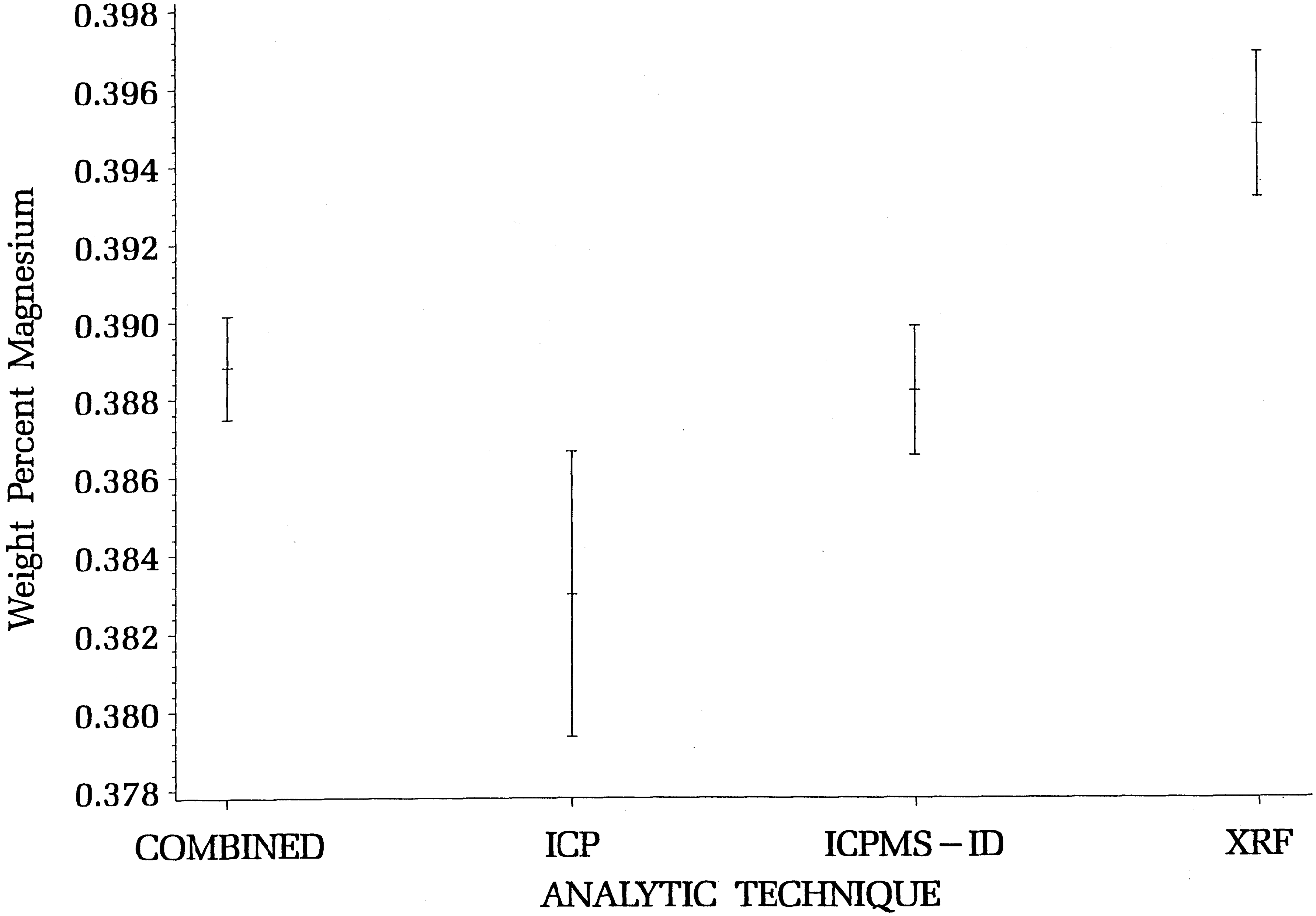


Figure 8.4.3. Magnesium in Estuarine Sediment, SRM 1646a


Weight Percent Magnesium
COMBINED
ICP
ICPMS – ID
XRF
ANALYTIC TECHNIQUE