

Measurement Performance Assessment of Analytical Chemistry Analysis Methods using Sample Exchange Data

Tom Burr (Corresponding author)

Statistical Sciences, Los Alamos National Laboratory
Mail Stop F600, Los Alamos, NM, USA
Tel: 1-505-665-7865 E-mail: tburr@lanl.gov

Kevin Kuhn

Actinide Analytical Chemistry, Los Alamos National Laboratory
Mail Stop G740, Los Alamos, NM, USA
Tel: 1-505-665-50155 E-mail: kkuhn@lanl.gov

Lav Tandon

Analytical Chemistry, Los Alamos National Laboratory
Mail Stop G740, Los Alamos, NM, USA
Tel: 1-505-665-5458 E-mail: tandon@lanl.gov

Diane Tompkins

Statistical Sciences, Los Alamos National Laboratory
Mail Stop F600, Los Alamos, NM, USA
Tel: 1-505-667-3380 E-mail: dtompkins@lanl.gov

Received: September 13, 2011

Accepted: October 10, 2011

Published: December 1, 2011

doi:10.5539/ijc.v3n4p40

URL: <http://dx.doi.org/10.5539/ijc.v3n4p40>

The research is financed by the nuclear weapons pit sustainment program office at Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Los Alamos National Security, LLC for the U.S. Department of Energy under Contract number DE-AC52-06NA25396.

Abstract

Measurement error modeling is crucial to any assay method. Realistic error models prioritize efforts to reduce key error components and provide a way to estimate total (“random” and “systematic”) measurement error variances. This paper uses multi-laboratory data to estimate random error and systematic error variances for seven analytical chemistry destructive assay methods for five analytes (Gallium, Iron, Silicon, Plutonium, and Uranium). Because these variance estimates are based on multiple-component error models, strategies are described for choosing and then fitting error models that allow for lab-to-lab variation.

Keywords: Measurement error modeling, Sample exchange, Variance components

1. Introduction

A Pu metal exchange program has been coordinated at Los Alamos National Laboratory since 2001 (Tompkins, 2006). There are 6 Pu metals, approximately 1512 measurements distributed among seven methods (ICPAES, ICPMS, SPEC, TIMS, XRF, CORPEL, CPC) from 6 labs including ANL, AWE, INL, LANL, SRS, and LLNL, with 224, 152, 179, 675, 242, and 40 total measurements respectively. Each sample round consists of measurements by some or all of the labs for a subset of the 6 metals (usually 2 of the 6).

Assay quality is often qualitatively described using the terms “accuracy” and “precision.” When a quantitative description is needed, a model-based definition of the corresponding terms “random error variance” and “systematic error variance” is needed, as described below using multiple-component error models.

Strategies for choosing and then fitting error models are also described. Our context is to use error models to assess certified assay methods, without involving designed experiments to identify physical factors that impact the error. The main performance measure (which can be assessed by using results described in Burr et al., 2007) is the quality of the predicted difference between future measurements M (individual and multiple items) using the to-be-characterized assay and the gold-standard (or consensus) assay result T . In addition, goodness of fit tests can judge the suitability of Eq. (1) below as an adequate model-based summary of the data.

2. Background

A general conclusion regarding international target values (ITV) for uncertainty in nuclear assay methods is that the standard deviation of the random error σ_R and of the systematic error σ_S depend on material type and amount and assay method. It is generally hoped that an established and stable assay procedure will have consistent quality across analysts and laboratories over time. For example, in the 2000 ITVs (Aigner et al., 2002), the assumed error model is $d_{ij} = d + S_j + R_{ij}$, where d is the mean operator-inspector difference over all inspection periods, S_j is the systematic error of the operator-inspector difference during inspection j , and R_{ij} is the random error of the operator-inspector difference for item i during inspection j . The expected values of S_j and R_{ij} are assumed to be zero (i.e., they are both centered random variables in a statistical sense). An analysis of variance components of the operator-inspector differences, d_{ij} , gives estimates of the standard deviation of the random component, σ_R , and of the of the systematic component, σ_S . It is assumed that the σ_R and σ_S are the same for all inspections and that S_j is constant during inspection period j , but S_j varies randomly across periods.

In analyzing operator-inspector paired measurement comparisons for several inspection periods (assumed to not be impacted by diversion or any other anomaly), results on physical standards are not required. Analogously, although the current Pu metal exchange program assigns “consensus” values to the analyte quantities in each metal, consensus values are not always needed in order to estimate σ_R and σ_S , depending on what assumptions are made.

The ITV values were obtained by grouping the data pairs originating from one inspection period, j , by material balance areas (MBA), by strata of materials of similar characteristics and by measurement methods. For a given MBA and stratum, $d_{ij} = (X_{ij} - Y_{ij})/X_{ij}$ is the operator-inspector relative difference, d_{ij} , for item i in inspection j .

3. Model Selection

Because there are sample rounds, labs, and metals, we model the measurements M as

$$M = T(1 + B + M + L + S + R) \quad (1)$$

where

T is the true (consensus) value,

B is bias (systematic error) over a long specified time,

M is metal effect,

L is lab effect,

S is sample round effect, and

R is all other error sources.

Note that in Eq. (1), S is sample round, which could correspond to a random or systematic effect. Generally, “random” and “systematic” effects and definitions depend on the chosen error model (see below).

Fisher’s F test and inspection of relevant plots (see Figure 1) or both were used to choose which terms were needed. For example, some of the six metals can be pooled, thus dropping some terms. Generally, a lab effect and/or a sample round effect was appropriate. These considerations lead us to prefer reporting results by method for various pools of metals by lab, and to partition the variance into within-round and between-round variances. Here are two cautions:

(1) There are many approaches to measurement performance assessment. We follow the “type A” evaluation outlined in general form in the NIST guideline for expression of measurement uncertainties (Taylor, 1993; Taylor & Kuyatt, 1994). Any characterization of measurement uncertainty requires a measurement error mode.

(2) One typically assumes that a “gold-standard” method provides the true value T with either a negligible (and,

therefore, assumed to be 0) measurement error or an already-characterized method. In many cases, the uncertainty in the nominal value can be easily accommodated simply by subtracting the error variance assigned to the true value from the estimated systematic error variance of the method. Because our main goal is to compare methods, and because we do not in all cases yet have a defensible error variance to assign to the true value, we will ignore uncertainty in the true value. This overstates the error variance in all methods, but to the same extent so that the relative ranking of methods is unaffected. To some extent, there is always a concern regarding uncertainty in the standards or working reference materials.

Visual inspection of plots such as Figure 1 suggests:

- 1) It is sometimes reasonable to pool over sample round; however, our preferred model separately estimates within-round and between-round variance. In general, we compare “between round” to “within round” variation and use most of the metal items. However, some participating labs recalibrate before each sample round, so even if the between round variance is no larger than the within round variance would predict assuming there is no between-round effect, we can alternatively view each round as a calibration period. This results in two estimates of σ_B , one based on pooling over rounds and the other based on assuming each round is a separate calibration period.
- 2) It less reasonable to pool over labs (some lab results are quite far from the consensus and/or have large estimated $\hat{\sigma}_{R_{\text{effective}}}$ (see below for definition of $\hat{\sigma}_{R_{\text{effective}}}$) for some items).
- 3) The 6 metal items can be pooled in various ways, determined by purely statistical analysis or by metal types; as described in Burr et al. (2007), the metals could be pooled into these groups for Ga: {A, B, C}, {D}, and {442, 465}. Here, the metals were pooled on the basis on concentration level for all analytes except for Pu. For Pu, we used the same pools as for Ga.

A multiplicative model (Eq. (1)) was used, which was judged to be adequate after visual assessment of plots of the absolute and relative standard deviation (RSD) versus the consensus value. We note here that one of the smaller-valued consensus metals does not fit the “constant RSD” model very well. That is one reason that some metals are excluded in some summaries in Burr et al. (2007). We chose a model (such as which if any terms to omit from Eq. (1)) using Fisher’s F test (Miller, 1986). If the F test suggested that a term could be ignored, it is reasonable to pool assay results over that term as if the effect is absent. For example, if there is no obvious “metal effect” or “lab effect,” then we pool over all metals or groups of metals, or labs.

Recall that we ignore uncertainty in the consensus value; this slightly overestimates σ_B , but does so in a uniform manner across labs. It is simple to subtract off the effect of uncertainty in the consensus value to adjust downward the estimate of σ_B . Recall also that we can partition “random error” by lab, metal, sample round, depending on the context and model. One reason to do this is that it provides information, but the main reason is that to estimate all effects, we must partition the data into variance components.

We also caution that we typically have very few (one in some cases!) degree of freedom in $\hat{\sigma}_B$, which implies that its “error bars” are relatively wide. Because each lab recalibrates at least once between each sample round, we include a model that interprets the between-round variance as a “systematic error” variance associated with recalibration. This is analogous to “between inspections fluctuations” reported in international target values (Aigner et al., 2002). In that case, the systematic error variance associated with the recalibration procedure can be estimated with more degrees of freedom, resulting in a higher-quality estimate having smaller uncertainty. However, any between-round effects such as drifting in the true value would then be confounded with calibration effects. It could be argued that the procedure to decide when and whether to recalibrate should be part of the assay protocol. This is our “preferred model,” which we write as $M_{ij} = T_i(1 + B + S_i + R_{ij})$, where the metal and lab terms have been omitted; however, because we also prefer to report results by lab, in effect we retain the lab effect term in the preferred model. Lab and metal effects could be regarded as “fixed,” meaning that the particular labs and metals are of interest. Sample rounds should probably be regarded as “random,” meaning that they represent a sample of possible rounds.

Another way to decide which effects (such as metal, lab, or round) should be included in an error model such as (1) is to fit a linear model to the response $y = (M-T)/T$, and check which terms are statistically significant. We have done this for the seven analytes, and found that when outliers were omitted (outliers were defined on the basis of residuals using the lm function in Splus (2006), then in nearly all cases, all terms were statistically significant, but with inclusion of outliers, in many cases, only one of the four terms (B , M , L , and S) were significant. Example model-selection results are given in Burr et al., 2007. This issue deserves additional study, but a tentative conclusion is that the relatively small sample sizes make model choices quite erratic. In addition to linear regression fitting with and without outliers, we evaluated model selection methods such as the Bayesian

information criterion (Pinheiro & Bates, 2000) available using the varcomp function in Splus (2006). We opted to use any of three main models motivated by Eq. (1), to summarize performance, as described next.

The group averages, between-group variances, and scaled within-group variances can be used to estimate the model parameters (Burr et al., 2007) in Eq. (1). Following the procedure in, Burr et al. (2007), the estimated variances for analyte/method and/or for analyte/method/lab for model 3 among the following three models are calculated.

Model 1: Pool over metals by concentration level (metals {A, B, C}, {D}, {442, 465} for Ga, metals {{D}, {423, 465}, A, B, C} for Fe, metals {A, B, C, D, 442, 465} for Si and metals {C, D}, {442, 465, A, B} for Uranium) and labs and rounds. This results in a loss of the ability to separately estimate σ_R , σ_{Reff} and σ_S , where σ_{Reff} is the effective random error standard deviation, typically defined (Burr and Hemphill, 2006) as $\sigma_R = \sqrt{\sigma_{R1}^2 + \sigma_{R2}^2}$, where σ_{R1}^2 and σ_{R2}^2 are two components of random error, such as “within-round” and “between-round” in this case. However, one could argue that $\hat{\sigma}_{\text{Reff}}$ and $\hat{\sigma}_B$ could be separately estimated if this large amount of pooling were appropriate. Although we do not endorse this large pooling, it is useful if for no reason other than to see how much different the $\hat{\sigma}_{\text{Reff}}$ and $\hat{\sigma}_B$ values are compared to the other models.

Model 2: Pool over metals by concentration level and separately estimate the variance components associated with labs and rounds. Assume no recalibration during the entire study to produce estimates $\hat{\sigma}_R$, $\hat{\sigma}_S$ and $\hat{\sigma}_B$, which are interpreted according to the pooling described (and therefore have a different interpretation than those from any other model).

Model 3: Pool over metals by concentration level and separately estimate the variance components associated with labs and rounds. Assume a recalibration by each lab just prior to each sample round. Estimates $\hat{\sigma}_R$, $\hat{\sigma}_S$ and $\hat{\sigma}_B$ are interpreted according to the pooling described. Note that in this model, there are more degrees of freedom in the estimate $\hat{\sigma}_B$ (which implies that the estimate will have smaller uncertainty).

Results for Gallium by ICPAES for model 3 are given in Burr et al. (2007), regarding the inverse of within-round variance as “precision,” and the inverse of between-round variance as “accuracy,” as done in the ITV values (Aigner et al., 2002). Alternatively (model 2 and Table 1) a combination of within-round and between-round variance can define “precision,” and the long-term bias can define the “accuracy.”

Results for all seven analytes for model 2 are given in Burr et al., 2007, and selected cases for models 1 and 3 are also evaluated. If in Table 1 we interpret the pooled result for $\hat{\sigma}_B$ as the overall bias, then $\hat{\sigma}_B = 4.3\%$ is a one degree of freedom (approximately) estimate for σ_B . In all models, there is a choice regarding whether to include censored observations and/or outliers. Censoring arises because results less than the lower detection limit (LDL) are typically reported as the LDL (Tompkins, 2006). Results presented in Burr et al. (2007) omit outliers and censored data unless stated otherwise. Sample sizes and approximate effective degrees of freedom are available to estimate 95% confidence limits (shown in parentheses). Estimates are denoted $\hat{\sigma}_R$, $\hat{\sigma}_{\text{Reff}}$ and $\hat{\sigma}_B$ and are interpreted according to the pooling described.

The degrees of freedom (df) are related to sample size, but adjusted for the number of estimated parameters. Because sample sizes are small in many cases, these confidence limits can be relatively large, thus making it difficult to reliably “rank” the methods. Because the partition of total sample size by lab and sample round is unbalanced (unequal sample sizes per group), these estimated confidence limits are rough approximations (Miller, 1986), and are dependent, for example, on assuming that the various error components each have a Gaussian distribution. The df or the approximate df (Satterthwaite, 1946) are given, so that approximate lower and upper factors can be obtained.

It is well known that confidence interval (CI) quality for estimates of variance components arising from unbalanced designs (unequal sample sizes among the various groups of data) are more vulnerable to non-normality in the data (Miller, 1986; Pinheiro & Bates, 2000; Scheffe, 1959; Satterthwaite, 1946). This exchange data is moderately unbalanced as measured by the ratio of smallest to largest sample sizes. By “CI quality” we mean that the true coverage can sometimes be non-negligibly different from the nominal coverage. For example, we report nominal 95% CIs, which means that we estimate that in thousands of hypothetical repeats of this data collection, 95% of the CIs would contain the corresponding true parameter. The actual coverage could differ from 95% and this “CI quality” deserves further study. A tentative finding is that a simple

bootstrap procedure gives results that are close to those from the intervals function applied to the lme function (Splus function), which are based on assuming that the maximum likelihood estimates are close to Gaussian in distribution. Because the lme-based (linear mixed effects model) estimate is an asymptotic result and our sample sizes are fairly small, our CI quality is currently unknown. A small simulation study could help assess our CI quality and perhaps guide improvements to CI construction.

A related issue is the procedure for estimating variance components. Some of the variance component estimation results presented here use a combination of the Rankin approximation as described in Miller (1986) and a bootstrap procedure for the associated CI. The estimation procedure and the CI construction method both deserve further study.

For each assay method, our model selection strategy is:

- 1). Confirm using visual inspection of plots that the metal pooling groups are reasonable.
- 2). Confirm using F test that most results should be reported by lab. However, for completeness, we also report results that pool over labs.
- 3). Interpret “precision” as the inverse of “within-round variance plus the between-round variance.”
- 4). Interpret “accuracy” as the inverse of: the long-term average bias, or as the inverse of the between-round variance, depending on the model.
- 5). Estimate 95% CIs using the Satterthwaite approximation (Miller, 1986). Initial simulation results suggests that this approximation is adequate provided the data is approximately normally distributed, and is more accurate than a bootstrap method.

4. Discussion

We strongly advocate sample exchange programs as part of continuing quality improvement of any assay method. Exchange programs often reveal which aspects of assay procedures are truly necessary to tightly control. For example, “outlying” results or laboratories are usually correctable once the assay protocol is more narrowly defined.

Although there is no consensus among the laboratories, we believe models 1-3 are each defensible. At this stage, we report estimates of “precision” and “accuracy” for each reasonable model in Burr et al., 2007. Also, we note the following:

- a) It would also be acceptable to pool the six metals for Pu.
- b) All results included censored and “anomalous” data entries (Tompkins (2006), but for the data considered here, there are no “anomalous” entries) and also omitted outliers. Results for omitting censored data are available but are quite similar to the results shown.
- c) This paper’s focus is on estimates of precision and accuracy as in Table 1. In some situations, there is interest in options to calculate consensus values, and there are several reasonable ways to calculate consensus values. Results here use the unweighted average value from all “qualified” labs, as described in Tompkins (2006). Results based on other methods to calculate consensus value are available. For example, individual laboratory estimates can be weighted in inverse proportion to their respective within-laboratory variances. Another option could be based on semiparametric models (Burr & Doss, 2005) where the lab-effect term L in Eq. (1) is assumed to arise from an arbitrary probability distribution, concentrated if appropriate around a target distribution such as the Gaussian.
- d) Table 1 includes an average relative bias column, which can be negative or positive. However, in order to use such an estimate for variance propagation, we use its absolute value (a one degree-of-freedom estimate of systematic error variance). The confidence intervals included in the average relative bias column refer to the absolute value of the bias, with no attempt to adjust for effective random error variance. Therefore, if we use the absolute estimated bias as a one degree-of-freedom estimate of systematic error standard deviation, then it is straightforward to adjust for effective random error variance by subtraction, as shown for example in Burr and Hemphill (2006).

There are several open statistical issues, including the following. First, unbalanced designs, involve more difficult calculations and interpretation. Also, one could view this as a nested design with rounds nested within labs for example.

- a) Simultaneous consideration of censoring and outliers. Currently, we used an arbitrary but reasonable “decoupled” approach in which we first used standard methods for censored data which result in modifying “less

than” results to a revised-downward value while leaving the uncensored data unchanged. Simple outlier detection based on the median $\pm 3 \times \text{MAD}$ is then used on the resided data to identify outliers.

b) Calculation of consensus values. We have evaluated consensus values using two options. Several other options are possible, including options that first apply outlier rejection. The various options are distinguished by the type of weights used in combining estimates.

c) Estimation of CIs. CI quality is judged by comparing “nominal” to “actual” coverage. For example, a nominal 95% CI might have actual coverage of 98% or 90%. Several studies have shown that CI quality is impacted by the underlying distribution of the data. If the data has approximately a normal distribution, then CIs constructed using standard methods based for example on the χ^2 distribution are adequate (Miller, 1986). However, our variance component estimates for $\hat{\sigma}_S$ and $\hat{\sigma}_B$ involve linear combinations having complicated distributions. Satterwhaite’s approximation (Miller, 1986) remains a competitive option, and on the basis of a small simulation experiment, it appears that Satterwhaite’s approximation is better than the bootstrap method we implemented. Therefore, results reported here are based on Satterwhaite’s approximation. There is considerable related literature on CI construction in variance components problems, which suggests that CI quality is typically degraded by unequal sample sizes by group (“unbalanced designs”) and by non-normal data (Rankin, 1974; Pionheiro and Bates, 2000; Satterthwaite, 1946; Scheffe, 1959).

d) Measurement error model choice. Error modeling has been approached using a simple cross validation approach that was shown to be effective in a different but quite similar setting (Burr & Hemphill, 2006). Results presented here are a combination of informal assessment of plots and Fisher’s F test for omitting terms from Equation 1.

Sample exchange programs provide a valuable data source for assessing measurement uncertainty and developing assay protocols. Associated statistical challenges such as those just described should not be an obstacle for using sample exchange programs to advance understanding of assay methods.

References

- Aigner, H. et al. (2011). International Target Values 2000 for Measurement Uncertainties in Safeguarding Nuclear Materials. *Journal of Nuclear Materials Management*, 30(2), updated 2010, available: <http://www.inmm.org/topics/publications.htm>, 2011.
- Burr, D., & Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, 100(469), 242-251. <http://dx.doi.org/10.1198/016214504000001024>
- Burr, T., Kuhn, K., Tandon, L., & Tompkins, D. (2007). *Measurement performance assessment*. LAUR07-373.
- Burr, T., & Hemphill, G. (2006). Multi-component radiation measurement error models. *Applied Radiation and Isotopes*, 64(3), 379-385. <http://dx.doi.org/10.1016/j.apradiso.2005.09.002>
- Burr, T., Sampson, T., & Vo, D. (2005). Statistical evaluation of FRAM gamma ray isotopic analysis data, *Applied Radiation and Isotopes*, 62, 931-940. <http://dx.doi.org/10.1016/j.apradiso.2005.01.002>
- Miller, R. (1986). *Beyond ANOVA*, New York: Wiley, (Chapter 3).
- Pinheiro, J., & Bates, D. (2000). *Mixed-Effects Models in S and S-Plus*, New York: Springer, (Chapters 1-3). <http://dx.doi.org/10.1007/978-1-4419-0318-1>
- Satterthwaite, F. (1946). An approximate distribution of estimates of variance components, *Biometrics Bulletin*, 2(6), 110-114. [Online] Available: <http://www.soph.uab.edu/Statgenetics/People/MBeasley/Courses/Satterthwaite> (Dec. 1946).
- Scheffe, H. (1959). *The Analysis of Variance*, New York: Wiley, (Chapter 3).
- Splus version 6.0. (2006). Language for Statistical Computing, *Insightful Corporation*.
- Taylor, B., & Kuyatt, C. (1994). Guidelines for evaluating and expressing the uncertainty of NIST measurement results, *National Institute of Standards and Technology Technical Note 1297*. <http://physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf>.
- Taylor, J. (1993). *Handbook for SRM Users*. NIST Publication 260-100.
- Tompkins, D. (2006). Statistical analysis of plutonium sample exchange data, LA-CP, 07-0071.

Table 1. Estimates $\hat{\sigma}_R$, $\hat{\sigma}_{Reff}$ and $\hat{\sigma}_B$ for Gallium for metals A, B, and C

		Within-round RSD in percent	Between-round RSD in percent	Effective random RSD in percent	Average relative bias in percent	Sample sizes	Date Range
Lab	Method	$\hat{\sigma}_R$ (% relative)	$\hat{\sigma}_S$ (% relative)	$\hat{\sigma}_{Reff}$ (% relative)	$\hat{\sigma}_B$ (% relative)	n_{round}, n_{total} $n_{censor}, n_{outlier}$	
Lab1	ICP-AES	3.3 (2.0,4.7)	2.2 (0.1,4.9)	4.0 (0.1,9.0)	1.1 (0,2.4)	5,17 0,0	3/02-6/04
Lab2	ICP-AES	2.9 (1.5,4.3)	4.1 (0.1,9.1)	5.0 (0.2,11.2)	8.8 (0.3,19.8)	4,12 0,2	3/02-6/04
AVG	ICP-AES	3.3 (2.0,4.7)	2.2 (0.1,4.9)	4.0 (0.1,9.0)	4.6 (0,15.5)	9,29 0,2	3/02-6/04
POO L	ICP-AES	6.0 (4.3,7.7)	0	6.0 (4.3,7.7)	4.3 (0.1,9.6)	5,29 0,2	3/02-6/04

Outliers were identified by individual lab, and for the pooled case, by first pooling all results over labs and then checking for outliers using the “median $\pm 3 \times MAD$ ” (mean absolute deviation) test. The 0 entries are less than 0.1%. When available, approximate 95% CIs are listed in parentheses.

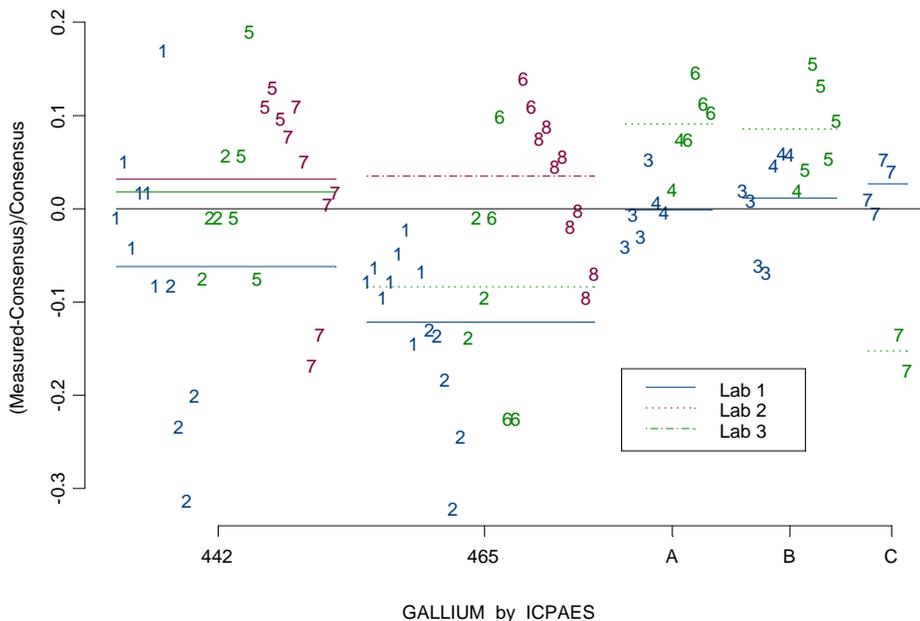


Figure 1. Example sample exchange data

Example plot of $(M-T)/T$ versus the metal standard for Gallium by ICPAES to motivate an error model such as Eq. (1). Sample rounds are indicated by the integers 1-8. Lab averages by metal are given by the dotted horizontal lines, with different line types (and color if available) for labs 1, 2, and 3. Inspection of such a plot can provide an informal way to choose which terms should be included in the selected model. If, for example, it appears reasonable to omit the lab effect, then the L term is omitted from Eq. (1) and assay results are pooled over labs. More formally, the vertical distances between pairs of lab means compared to within-lab variation can be used to test (using Fisher’s F test) whether there is significant between-lab variation.