Replicate Experiment

Table of contents

Introduction

Statistical process control (SPC) is the use of statistical methods to monitor and control the quality of processes. Though the tools were initially developed and used for manufacturing, they have found applications in many other areas where outcomes can be measured and quality is valued, including scientific laboratories. Some of the most common SPC tools include Run Charts, Control Charts, Experimental Design, and Replicate Experiment.

The Replicate-Experiment (Eastwood et al. 2006a) is based on Bland-Altman difference analysis (Bland and Altman 1986) between two sets of measurements and is used to assess either **repeatability** or **reproducibility**. These terms have specific meanings in SPC. **Repeatability** is the ability to produce comparable results within a group of conditions (e.g. people, instruments, reagents, …), while **reproducibility** refers to the ability to replicate results across different conditions (e.g. between two assayers or pieces of equipment). Initially repeatability should be demonstrated by having a scientist test the same set of samples with 2 independently prepared sets of working reagents/cells (note: these experiments can be performed at the same time or on different days) on the same set of equipment. Once repeatability has been established, reproducibility can be determined between different assayers, pieces of equipment, or lots of reagents or cells.

Rationale

Replicate-Experiment studies (Iversen et al. 2012) are used to formally evaluate the *within-run* assay variability and formally compare a new assay to the existing (old) assay. They also allow a preliminary assessment of the *overall* or *between-run* assay variability, but two runs are not enough to adequately assess overall variability. Post-production monitoring, such as Retrospective MSR (Haas et al. 2017) analysis and Control Charts (Beck et al. 2017) are used to formally evaluate the overall variability in the assay. Note that the Replicate-Experiment study is a diagnostic and decision tool used to establish that the assay is ready to go into production by showing that the endpoints of the assay are repeatable over a range of values. It is not intended as a substitute for post-production monitoring or to provide an estimate of the overall Minimum Significant Ratio (MSR).

It may seem counter-intuitive to call the differences between two independent assay runs as *within-run* variability. However, the terminology results from how assay runs are defined. Experimental variation is categorized into two distinct components: *between-run* and *within-run* sources. Consider the following examples:

- If there is variation in the concentrations of buffer components between 2 runs, then the assay results could be affected. However, assuming that the same buffer is used with all compounds within one run, each compound will be equally affected and so the difference will only show up when comparing one run to another run, i.e. in two runs, one run will appear higher on average than the other run. This variation is called *between-run* variation.
- If the concentration of a compound in the stock plate varies from the target concentration then all wells where that compound is used will be affected. However, wells used to test other compounds will be unaffected. This type of variation is called *within-run* as the source of variation affects different compounds in the same run differently.
- Some sources of variability affect both within- and between-run variation. For example, if assay cells are plated and then incubated for 24-72 hours to achieve a target cell

density taking into account the doubling time of the cells. If the doubling time equals the incubation time, and the target density is 30,000 cells/well, then 15,000 cells/well are plated. But even if exactly 15,000 cells are placed in each well there won't be exactly 30,000 cells in each well after 24 hours. Some will be lower and some will be higher than the target. These differences are *within-run* as not all wells are equally affected. But also suppose in a particular run only 13,000 cells are initially plated. Then the wells will on average have fewer than 30,000 cells after 24 hours, and since all cells are affected this is *between-run* variation. Thus cell density has both *within-* and *between*-run sources of variation.

The total variation is the sum of both sources of variation. When comparing two compounds across runs, one must take into account both the *within-run* and *between-run* sources of variation. But when comparing two compounds in the same run, one must only take into account the *within-run* sources, since, by definition, the *between-run* sources affect both compounds equally.

In a Replicate-Experiment study, the *between-run* sources of variation cause one run to be on average higher than the other run. However, it would be very unlikely that the differences between the two runs were exactly the same for every compound in the study. These individual compound "differences from the average difference" are caused by the *within-run* sources of variation. The higher the within-run variability the greater the individual compound variation in the assay runs.

The analysis approach used in the Replicate-Experiment study is to estimate and factor out between-run variability, and then estimate the magnitude of within-run variability.

Experimental Procedure

The Replicate-Experiment is intended to be easy to execute with a modest resource commitment. Most executions can be performed with 2-4 assay plates. It is most commonly run in the potency mode, though it can be run in efficacy mode to gain a better understanding of assay variability across the dynamic range of the assay or facilitate the interpretation of screening results.

The potency mode is ideally run with 20-30 active compounds with a broad range of potencies and the potencies should be well spaced across the range of values. If this number of active compounds is not available, then it can be run with a smaller set of compounds in replicate, with each replicate treated as an independent sample (e.g. 5 compounds with 5 separate replicate dilutions).

In the efficacy mode, it is particularly important to have samples where the activity spans the dynamic range of the assay. Often the variability of measurements will not be constant across the dynamic range of the assay. For this reason, efficacy studies should not be conducted with random screening plates, since most compounds will be inactive and could skew the assessment. It may be simpler to use a small number of active compounds in a dilution series, as if for a potency determination, but treat each dilution as an independent sample for the efficacy analysis. This will ensure that the data cover the entire dynamic range of the assay with just a few compounds.

Initially **repeatability** should be demonstrated with identical compounds tested with 2 independently prepared sets of reagents/cells. Once **repeatability** has been demonstrated for a protocol, it is ready for routine testing. The assay should be monitored to ensure it behaves as validated. This can include any or all of the following methods:

- Control charting reference compound(s) (Beck et al. 2017).
- Retrospective MSR analysis (Haas et al. 2017).
- periodic retests if the samples used in previous replicate experiments to compare to previous data (Eastwood et al. 2006b).

The replicate experiment is also used to validate minor assay changes such as new assayers, equipment substitutions, or changes to lots of reagents or cells. In this case, the data from identical samples is compared in the current and new formats. Historical data can be used for the current format, as long as it comes from a single experiment.

Data Analysis

The statistical analysis assumes that any measurement errors are normally distributed. While this is true for efficacy data, potency data is log-normal. This means that potency data must first be transformed to their $\log_{10}$ values before the analysis. After that transformation the analysis methods are the same:

1. For each pair of compound measurements, calculate the mean = (meas1 + meas2)/2 and the difference = (meas1 - meas2). *Note when these values are transformed back to the linear scale for potency data they will generate the geometric mean and the ratio, since log(meas1) - log(meas2) = log(meas1/meas2).*
2. Calculate the mean ($\bar{d}$) and standard deviation (sd) for the set of the difference values.
3. Calculate the difference limits $DLs = \bar{d} \pm 2sd/\sqrt{n}$ where n is the number of compounds tested. This is the 95% confidence interval for the mean difference.
4. Calculate limits of agreement $LSAs = \bar{d} \pm 2sd$. Most of the individual compound differences (~95%) should fall within these limits.
5. Samples outside the agreement limits are flagged as outliers. Since these values are 2sd outliers, it is expected that 5% of the data to be flagged if the data follow a normal distribution. This is not a flag for automatically excluding data.

6. Calculate the Minimum Significant Difference $MSD = 2sd$. This is the minimum difference between two compounds that is statistically significant.

7. For potency data all statistics are transformed back to linear scale and differences become ratios (e.g. Minimum Significant Ratio, $MSR = 10^{MSD}$).

Interpretation

The replicate experiment is based on two assumptions:

1. The variability in the measurements of all the samples is equivalent.
2. The variability is normally distributed.

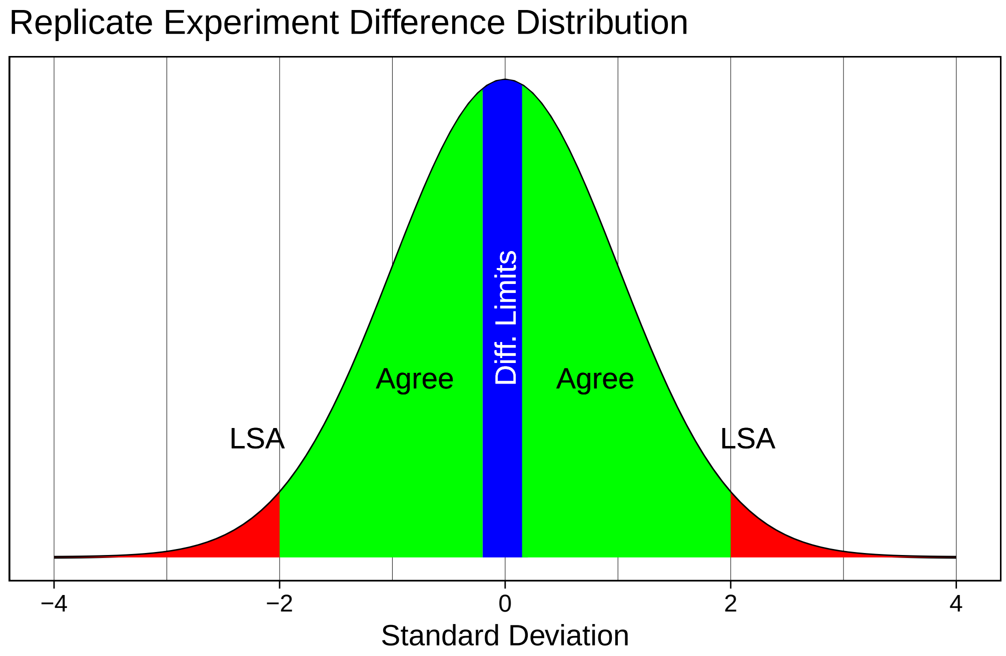

Figure 1: Difference Distribution for Replicate Experiment.

Figure 1 illustrates the distribution of difference values along with the associated analysis statistics. The difference limits are derived from the standard error of the mean (SEM) and represent a 95% Confidence Interval (CI) around the mean difference. If the difference limits include 0, then there is no discernible systematic bias between the first and second measurements. **Note: As the number of samples increases, a few samples may fall outside the LSA due to random variation, especially if they are just outside of the LSA. If outliers are to be excluded, start with the most extreme differences. Only obvious outliers or those with an assignable cause should be removed.**

Variability of Data Sets

The two most common types of data reported from *in vitro* bioassays are efficacy and potency. Understanding the variability of these data types is essential for the interpretation of replicate experiment results.

Efficacy

Efficacy measures the magnitude of an **effect** at a given concentration and is usually what is directly measured in an assay well. The variability in efficacy measurements is normally distributed at a given concentration, though this variability may change across the dynamic range of the assay. There are two primary factors that can contribute to this. The first is the detection method. For example fluorescence intensity often exhibits a constant coefficient of variation (cv) across the dynamic range, while the standard deviation (sd) of ratiometric fluorescence polarization measurements is more constant. Second, the sigmoidal nature of most dose-response curves means that the variability of individual efficacy measurements will be greatest around the inflection point in the curve, due to the effect that small differences in sample concentration can exhibit. While it is difficult to predict the overall effect in advance, the replicate experiment with samples across the full dynamic range of the assay should indicate whether the data can be treated as a single population or should be divided into more discrete populations (e.g. inactive, moderately active, highly active).

Potency Data

Potency data are the **concentration** of a substance that elicits a specified efficacy. These are generated by fitting concentration-response curves to the Hill equation (Y vs log(X)) to determine the binding constant and slope. While AC_{50} values are used for the examples, any potency measurement (K_i, LD_{80}, etc.) can be used.

Potency values generally follow a log-normal distribution (Elassaiss-Schaap and Duisters 2020). In other words, log transformed potency values are normally distributed. Potency values are determined from the efficacy plotted at log(concentration), which is often represented as an x-axis with a log scale. Therefore when doing a statistical analysis, which assumes the data are normally distributed, the analysis must be done with log transformed potency values.

A simulated data set of 10,000 potency measurements of a compound with a true potency of 100 nM was used to illustrate some of the properties of log-normal data. Figure 2 shows several views of the data distribution. Panel A represents the original data values plotted on a linear axis. Panel B is the $\log_{10}$ transformed potency data and exhibits the expected normal distribution centered on 2 ($\log_{10}(100) = 2$). Panel C is the original data plotted with a $\log_{10}$ transformed axis, which now appears to be normally distributed. Transforming the axis spacing for log-normal data facilitates visualization of the variability in the data. The true potency of 100 is the geometric mean of the data set and should be the median in the distribution of

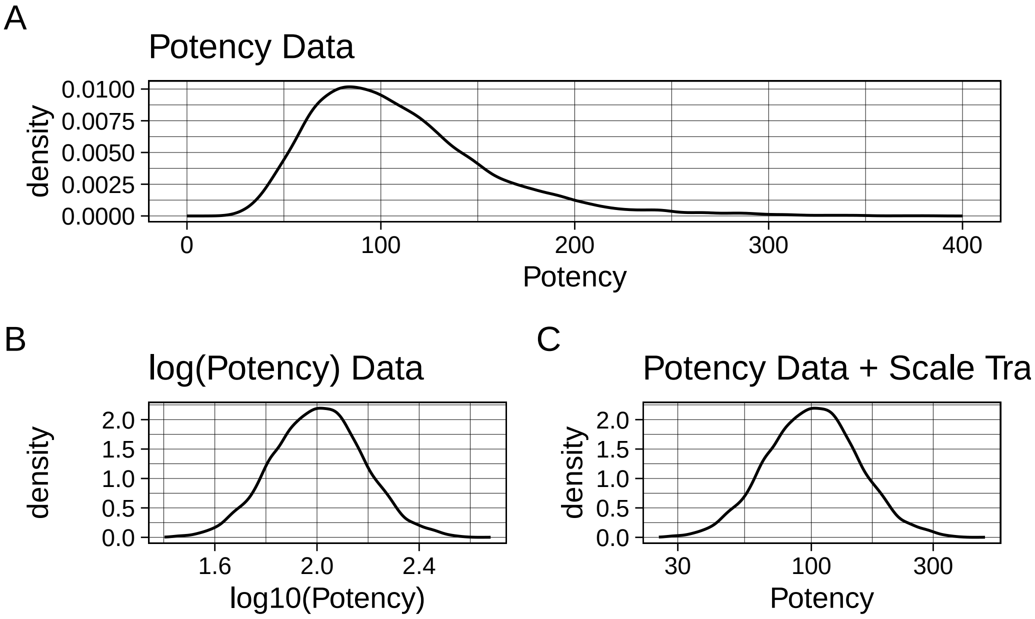

Figure 2: Potency data is log-normal.

the original data. Note that the peak of the original data is less than the geometric mean, since the data is right skewed. This skew in the linear potency data illustrates why geometric means, rather than arithmetic means, are used for summarizing potency data. The geometric mean corresponds to the center of the log-normal data distribution.

This log-normal data distribution means that the data analysis must be performed on log(Potency) data as described in Section . This ensures that the differences will be normally distributed as shown in Figure 1. These are then converted back to the linear scale for reporting, so differences are transformed to ratios and a difference of 0 becomes a ratio of 1. These transformed values are also log-normal so the axis scales are log transformed for graphing. **Note: If potency has already been transformed into a log scale (eg. $\mathbf{pK_a}$ or $\mathbf{log(IC_{50})}$ it should be analyzed in the Replicate-Experiment web tool with the Efficacy setting.**

Bland-Altman Plot

The Replicate-Experiment has three primary assumptions. The first assumption it that the differences between each pair of measurements and should be random and normally distributed, Second, the variability is consistent across the range of measured values. Finally, the mean difference for the set of paired measurements should be 0 (or 1 for ratios), if the measured values are equivalent. Bland-Altman plots (Bland and Altman 1986) can be used to examine

Table 1: Bland-Altman Data.

Bland-Altman Data

Sample	Exp1	Exp2	Mean	Difference
8712555	-31.693612	-18.471848	-25.082730	-13.221764
3168234	48.123361	59.983678	54.053520	-11.860317
6776201	57.330792	59.333070	58.331931	-2.002279
6396802	52.199878	43.796948	47.998413	8.402931
9634108	48.244736	55.571406	51.908071	-7.326670
1196678	3.678662	2.016909	2.847785	1.661753

the assumptions in the replicate experiment. In this plot, the x-axis is the mean of the measurement with the measurement difference on the y-axis.

A sample data set of 100 samples tested twice can be used to illustrate these concepts. Using just the first 2 steps in Section will produce a table with the measurements for each sample along with the mean an difference for each pair as shown in Table 1.

A histogram of the difference values for each pair of samples is centered around 0 and the distribution is roughly normal Figure 3

The summarized data for each pair can then be visualized using Bland Altman plot with the Mean on the x-axis and the Difference on the y-axis, Figure 4 Panel A shows a simulated data set with the y-axis is scaled in units of standard deviation similar to Figure 1 with the same color scheme. Panel B shows the same data bun the y-axis is now scaled to the actual difference values, rather than standard deviations. The x-axis spreads the data across the dynamic range of the measurements, so it's possible to visualize if the variation is distributed equally across this range. The difference values should cluster near 0 on the y-axis with fewer points as you increase the distance from 0. Panel B shows a more conventional representation. There is a reference line for 0 as well as lines to indicate the mean difference, the difference limits, and the limits of statistical agreement. About half of the data points should be in either side of the mean difference line.

While the standard Bland-Altman plot is appropriate for efficacy data or log transformed potency data, it must be modified to display potency data in the original concentration units. The x-axis is log transformed and the ratio of the two measurements is plotted on the y-axis (see Section).

Correlation plot

Correlation plots directly comparing the measurement values for each sample between the 2 experiments are also useful.

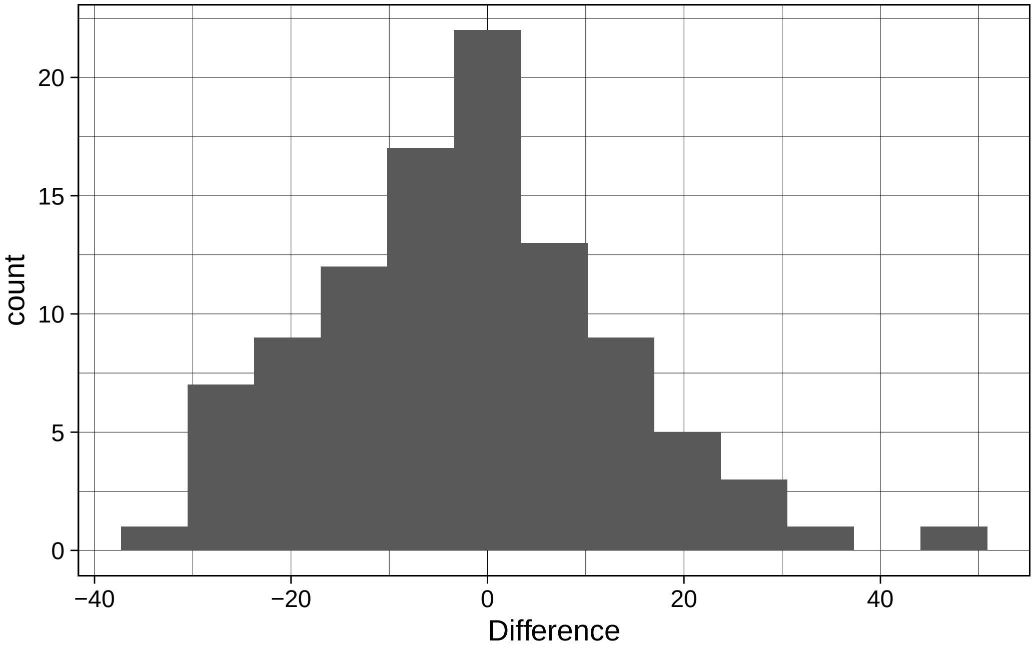

Figure 3: Histogram of Difference Values.

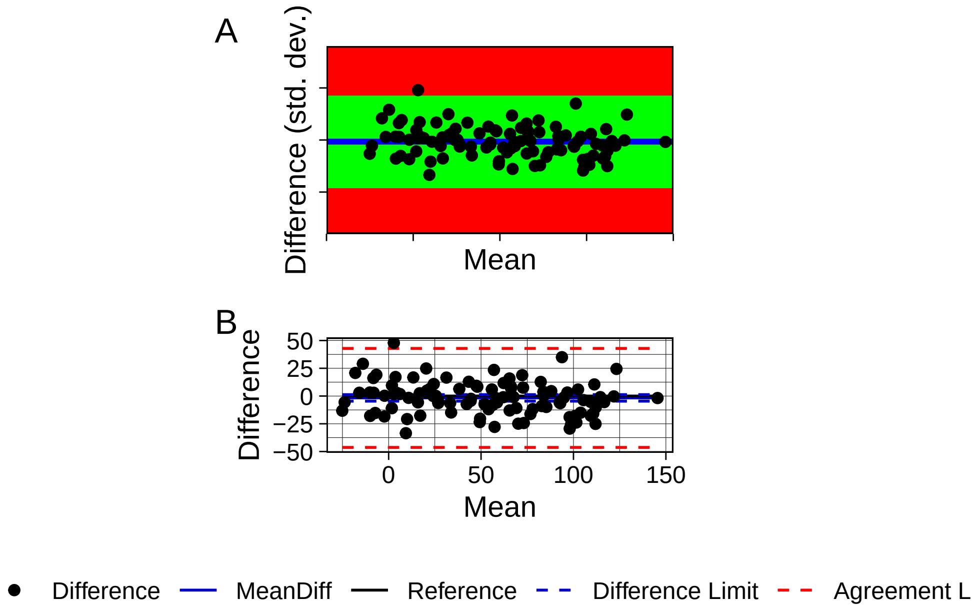

Figure 4: Bland-Altman Plots.

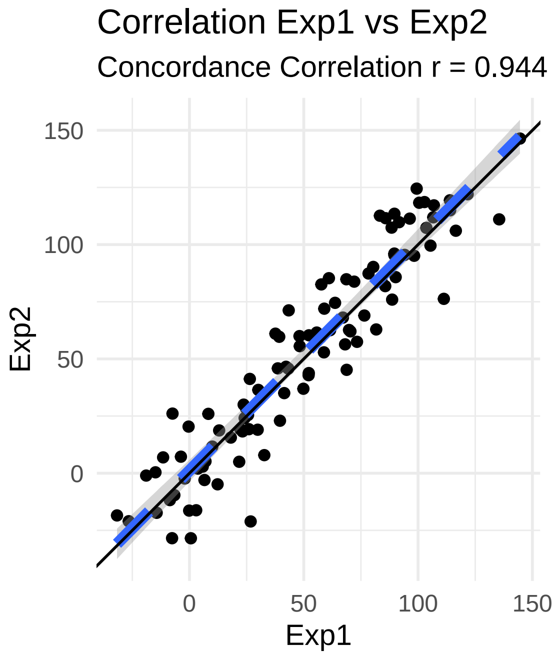

Figure 5: Correlation plot.

Figure 5 shows the correlation between the two experiments using Spearman's method. This linear correlation is appropriate for efficacy data and log transformed potency data. To view potency data in the original concentration units, both the x-axis and y-axis should be log transformed to see a linear correlation.

Examples

All the visualizations and tables below were created with the Replicate Experiment web tool, which also adds sample labels to any data outside of the Agreement Limits.

Efficacy - Constant SD

Efficacy measurements can be made with raw data from a detector or data which has been normalized (e.g. % Activity) using plate controls. Normalized data is preferred, since it provides biological context and facilitates comparisons across the lifetime of an assay. ***Note all the example data sets are % activity, where the vehicle control is set to 0.***

Generally, efficacy experiments can be performed with a single plate of samples tested twice. Even a 96-well plate with control wells will provide enough samples for a good statistical analysis. This data set represents 320 samples tested in 2 independent experiments.

Table 2: Sample efficacy data (sd constant).

Uploaded Data

Sample	Exp1	Exp2
7032194	115.044632	118.490663
2451491	-7.202141	1.521088
1771761	35.259537	44.730147
2769859	20.962247	23.343282
3779070	106.162994	108.951022
7853477	-7.261590	-19.891421

Table 3: Calculated Replicate Experiment Data (sd constant).

Calculated Data

Sample	Meas1	Meas2	MeasMean	MeasDiff	Class
7032194	1.15×10^{2}	1.18×10^{2}	1.17×10^{2}	-3.45	NA
2451491	-7.20	1.52	-2.84	-8.72	NA
1771761	3.53×10^{1}	4.47×10^{1}	4.00×10^{1}	-9.47	NA
2769859	2.10×10^{1}	2.33×10^{1}	2.22×10^{1}	-2.38	NA
3779070	1.06×10^{2}	1.09×10^{2}	1.08×10^{2}	-2.79	NA
7853477	-7.26	-1.99×10^{1}	-1.36×10^{1}	1.26×10^{1}	NA

This data set represents efficacy data with a constant sd, as described in Section .

Table 2 shows the data from the first 6 samples. The data is simply the sample identifiers (numeric or character) and the measured activity values for the 2 experiments.

Once the data have been uploaded, the replicate experiment calculations are displayed, Table 3. This data now includes the Mean and Difference values as well as Class, in addition to the original data. The Class column is used to identify flagged samples that fall outside the agreement limits. This table is fully sortable in the web tool. ***Note. While the input data may contain any number of digits after the decimal point and are used for the analysis, calculated values and statistics are displayed to 3 significant digits*** (Dahlin et al. 2019)***.***

The Bland-Altman plot, Figure 6 shows that the variation is consistent across the range of measured values with most of the difference values close to 0. The statistics in the plot are based on the distribution in Figure 1.

The statistics for the Bland-Altman plot are shown in Table 4. There are 18 samples with differences outside of the agreement limits (95% CI) which are flagged and labelled in the

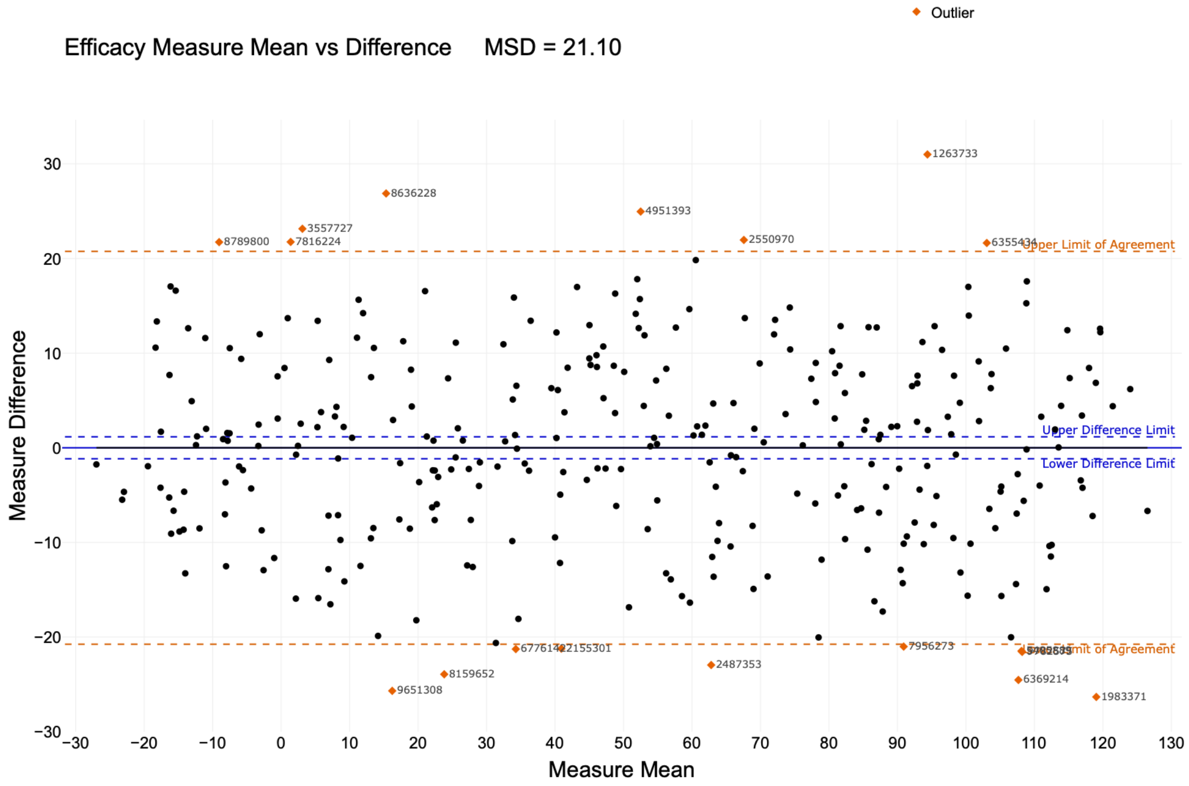

Figure 6: Mean-Difference plot of efficacy data (sd constant).

Table 4: Efficacy Replicate Experiment Statistics (sd constant).

Efficacy Stats

n	MeanDiff	MSD	Upper Difference Limit	Lower Difference Limit	Upper Agreement
320	-1.80 × 10^{-3}	2.11 × 10^{1}	1.16	-1.16	2.08 ×

Table 5: Efficacy (cv constant) Replicate Experiment Statistics.

Efficacy Stats (Constant cv

n	MeanDiff	MSD	Upper Difference Limit	Lower Difference Limit	Upper Agreement L
320	-1.05 × 10^{-1}	1.82 × 10^{1}	8.96 × 10^{-1}	-1.11	1.78 ×

report graphs. This is close to the expected value of 16 for 320 samples. The MSD is 21.1 and recall that MSD equals 2 standard deviations. It appears that all of the data are within about 3 standard deviations, so none of the flagged samples appear to be obvious outliers.

Figure 7 is a correlation plot of the two data sets with a unity reference line and Spearman's correlation coefficient. The data should overlay the equality reference line.

The calculated data, plots, and statistics can be downloaded for documentation and future reference.

CV Constant

Often the variability in efficacy data varies across the dynamic range of an assay and is better represented by the CV, Section . Unfortunately this violates a primary assumption in the analysis, making the replicate-experiment less useful. This is illustrated in the following example, where the CV is 10% across the assay range.

Figure 8 shows the variability increasing as the efficacy values increase. This occurs in signal increase assays, where the raw data values for the vehicle controls are smaller than those for active samples. The pattern would be reversed for a signal decrease assay.

Table 5 shows the statistics associated with Figure 8. Since the variability in the measurements is not constant across the dynamic range of the assay, it violates one of the assumptions of the Bland-Altman analysis, Section . Therefore these statistics are not meaningful.

The same pattern is observed in the correlation plot, Figure 9, with the spread between experiments increasing as the efficacy increases.

Assays which show this behavior should not use the Replicate-Experiment as a general measurement of uncertainty in the assay. However Replicate-Experiments with the vehicle and active controls analyzed separately can be useful to document that the assay is repeatable and help to define the limits for activity.

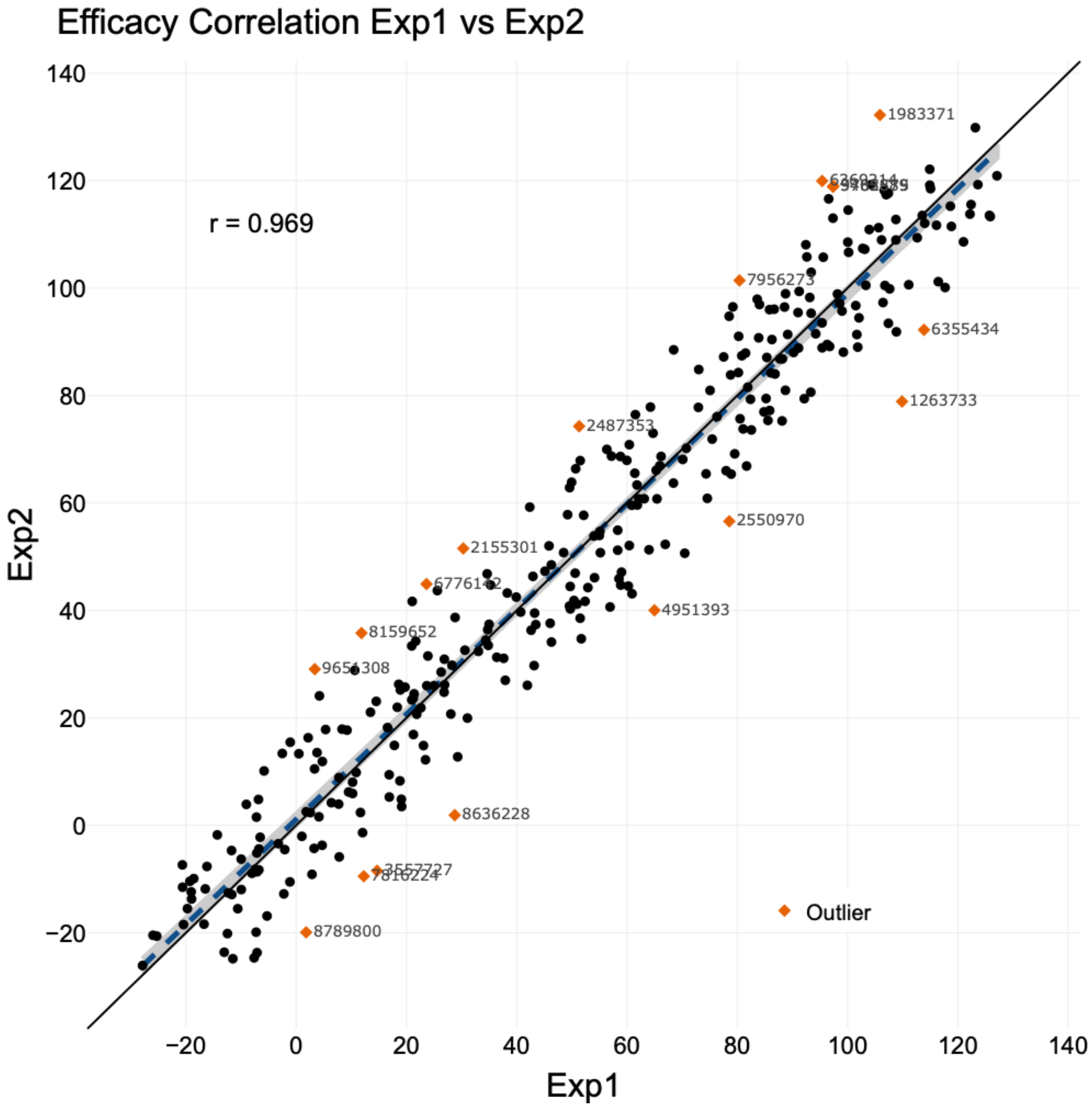

Figure 7: Efficacy Correlation plot (sd constant).

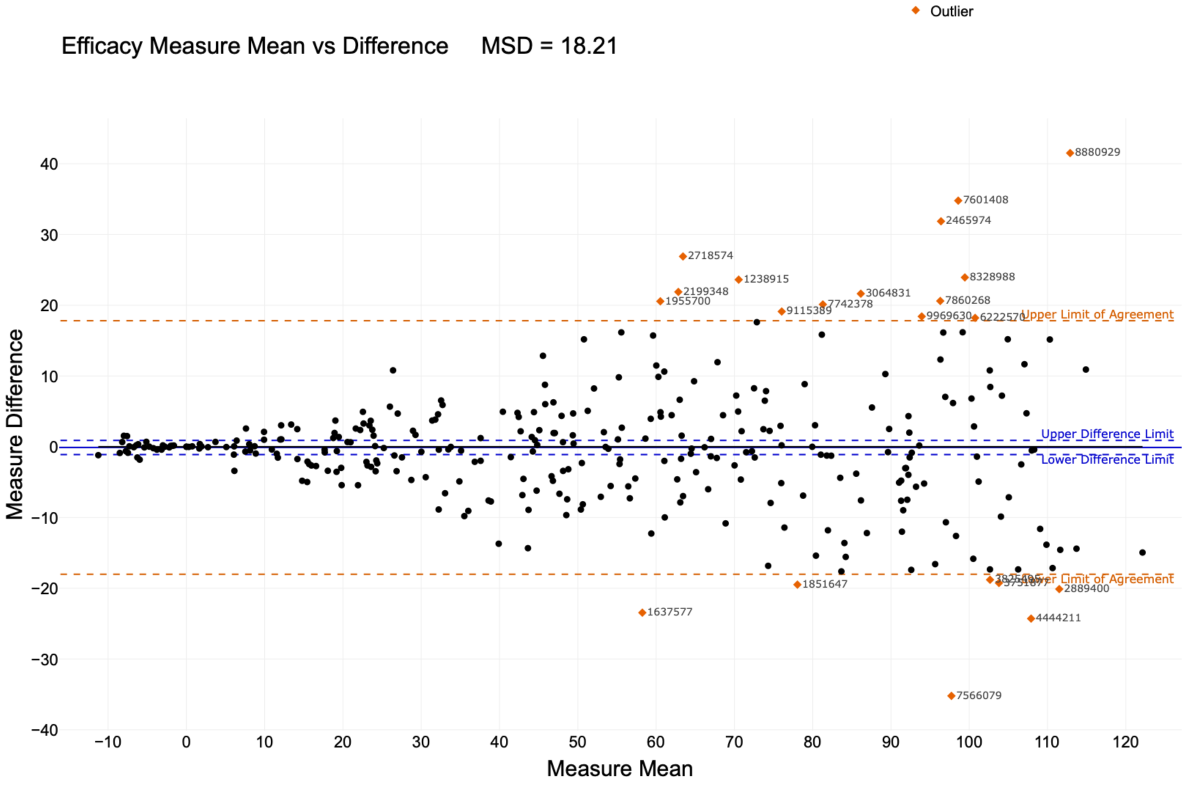

Figure 8: Mean Difference plot for an assay with 10% CV.

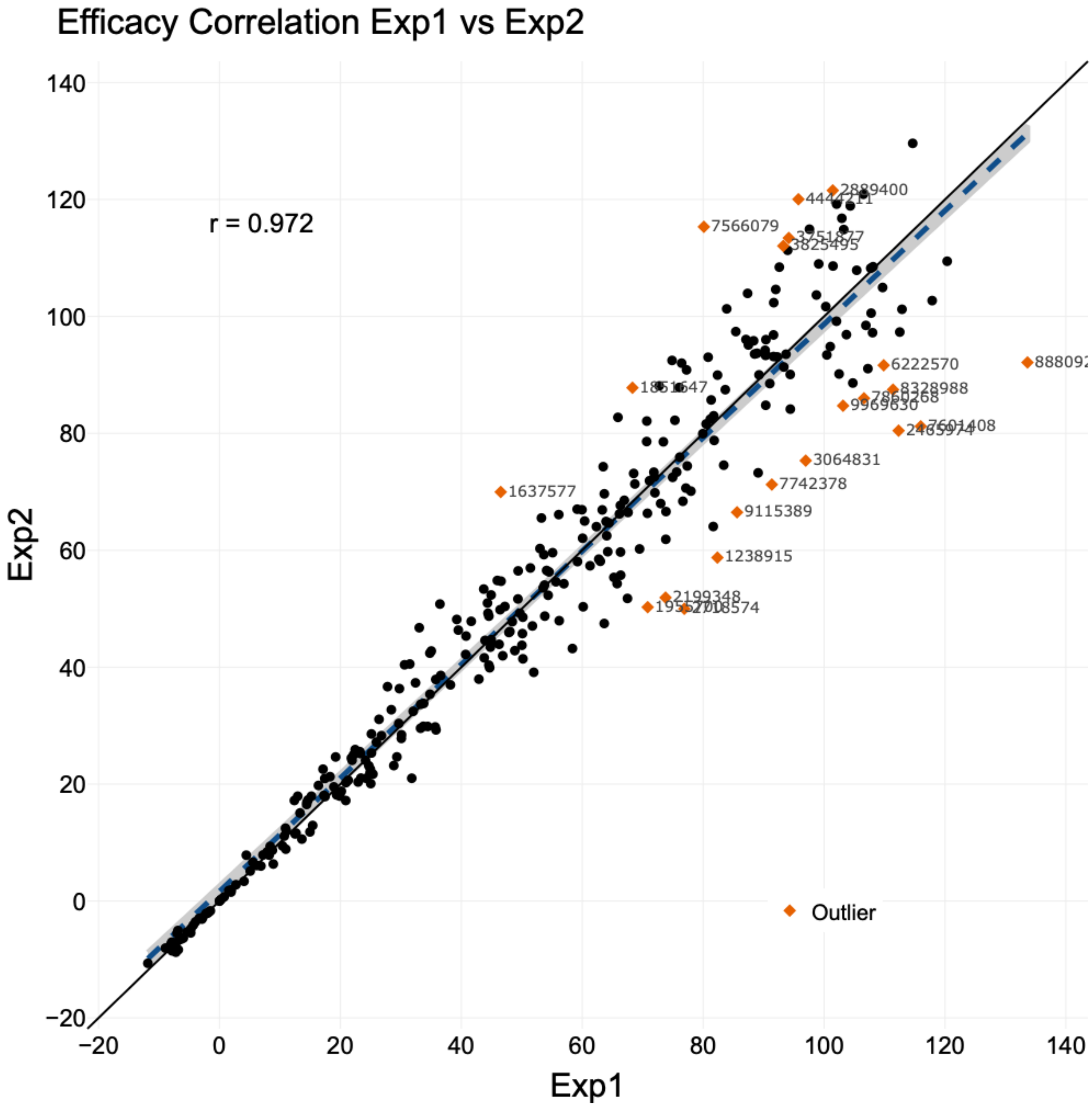

Figure 9: Correlation plot for an assay with 10% CV.

Table 6: Sample Potency data.

Uploaded Data

Sample	Exp1	Exp2
2455167	883.59032115	531.29130588
2225318	0.33218715	1.03798254
4625166	2.16832037	2.15084190
9852359	2.55900969	2.06438777
2107695	146.93978930	214.71751251
7465843	0.07546636	0.08069902

Table 7: Calculated Replicate Experiment Data.

Calculated Data

Sample	Exp1	Exp2	Geometric Mean	Ratio
2455167	8.84×10^{2}	5.31×10^{2}	6.85×10^{2}	1.66
2225318	3.32×10^{-1}	1.04	5.87×10^{-1}	3.20×10^{-1}
4625166	2.17	2.15	2.16	1.01
9852359	2.56	2.06	2.30	1.24
2107695	1.47×10^{2}	2.15×10^{2}	1.78×10^{2}	6.84×10^{-1}
7465843	7.55×10^{-2}	8.07×10^{-2}	7.80×10^{-2}	9.35×10^{-1}

Potency

Below are the first 6 samples from an example experiment with 32 total samples. The data is simply the sample identifiers (numeric or character) and the measured potency values for the 2 experiments. This is all that is needed to upload for analysis.

The data analysis is similar to that used with efficacy data, except that the potency values must first be transformed to their $\log_{10}$ values, so the variability will have a normal distribution for the statistical analysis. Then the means and differences for the pairs of $\log_{10}$(Potency) values are determined along with the associated statistics as described in Section .

Once the data have been uploaded, the Replicate-Experiment calculations are displayed, as shown in Table 7. This data now includes the Geometric Mean and Ratio values as well as Class, in addition to the original data. The Geometric Mean and the Ratio represent the mean and difference values for the log(Potency) data after they have been transformed back to the original linear scale. Recall that $log(A) - log(B) = log(A/B)$, so the difference between two log values becomes a ratio, when anti-logged. Similarly, the mean of logs becomes a geometric mean when anti-logged. The Class column is used to identify flagged samples that fall outside

Table 8: Potency Replicate Experiment Statistics.

Potency Stats

n	MeanRatio	MSR	Upper Ratio Limit	Lower Ratio Limit	Upper Agreement Limit	Lower Ag
32	1.06	2.53	1.25	8.98 × 10^{-1}	2.73	

the agreement limits. This table is fully sortable in the web tool. ***Note. While the input data may contain any number of digits after the decimal point and are used for the analysis. Calculated values and statistics are displayed to 3 significant digits*** (Dahlin et al. 2019).

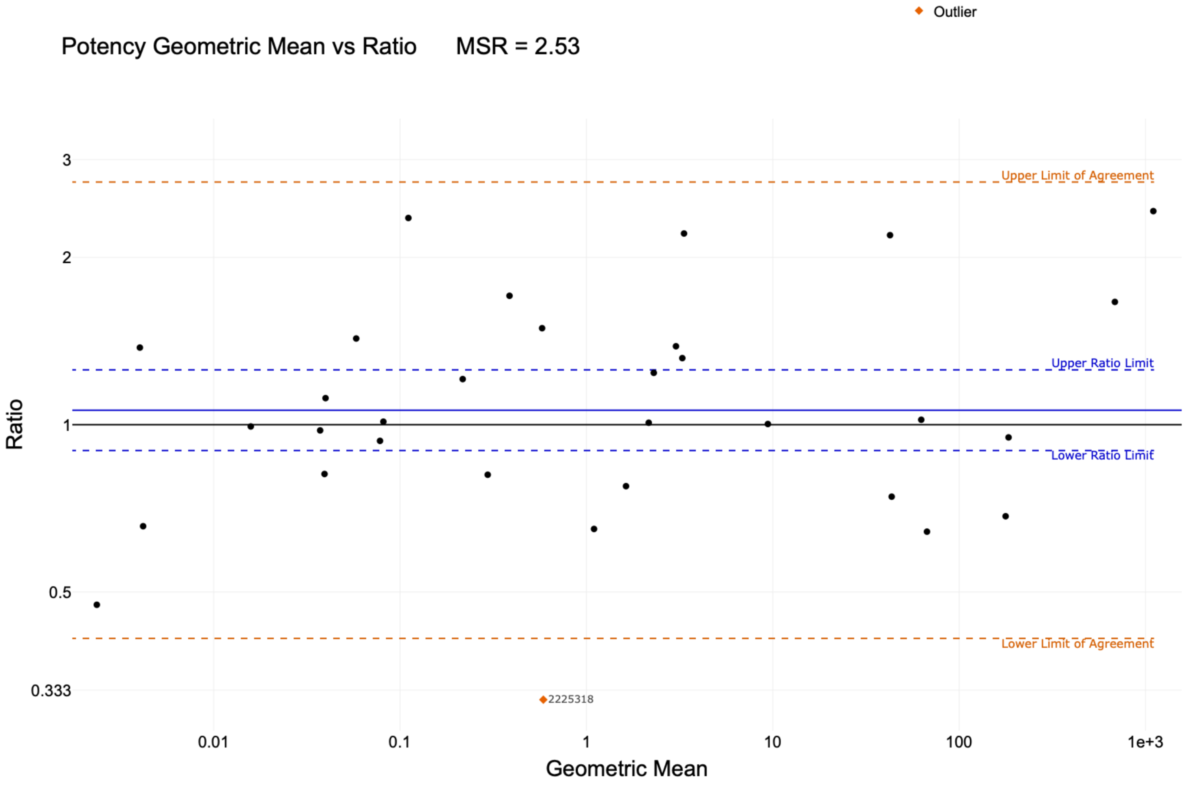

Figure 10: Mean-Ratio plot of potency data.

The Bland-Altman plot, Figure 10, for potency data uses the geometric mean value of each data pair on the x-axis with the ratio between the potencies is represented on the y-axis. Both axes are plotted on the log scale, as described in Section . The data points are centered around 1 with about half of the data on either side of the center. The variability is also evenly distributed across the range of x values.

Table 9: MSR values assuming different numbers of independent replicates.

n = 1	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7	n = 8
2.53	1.93	1.71	1.59	1.51	1.46	1.42	1.39

The corresponding statistics are shown in Table 8. There is only a single value labelled as an outlier in the Bland-Altman plot. This is consistent with a data set of this size. The ratio for this sample, while outside of the ratio limits, is close and thus not an obvious outlier.

Figure 11 is a correlation plot of the two data sets with a unity reference line and the Spearman correlation. The data should overlay the equality reference line.

The calculated data, plots, and statistics can be downloaded for documentation and future reference.

Acceptance Criteria

A Replicate-Experiment MSR $\leq$ 3.0 and limits of agreement between 1/3 and 3 is generally acceptable for most assays. An MSR $\leq$ 5.0 may be acceptable in a secondary assay, where the data will be used for more categorical decisions (e.g. > 100-fold selectivity vs. the primary assay). If more precision is needed, that can be achieved by increasing the number of times that each sample is independently tested. The MSR calculation for the Replicate-Experiment assumes that most samples tested will only be measured a single time. MSR is dependent upon the number of routine replicates (Haas et al. 2017):

$$MSR = 10^{2\sqrt{2(s/\sqrt{n})}}$$

Table 9 shows the calculated MSR for different numbers of independent replicates. The value of n indicates the number of different experiments each compound should be tested in to reach the calculated MSR.

Too Few Samples

Early in a project, it may be a struggle to identify a sufficient number of validated, active samples to perform a replicate experiment. Here is an example where there are only 5 unique samples. Unfortunately, this low sample number does not provide enough statistical power to obtain a good estimate of the assay's variability, so the Ratio Limits and MSR can appear to be greater than their true values.

The Bland-Altman plot, Figure 12, and the associated statistics, Table 10, illustrate the problem. While the MSR appears to be high, we also see that n is only 5, so there is probably not enough data to provide a good estimate of the MSR.

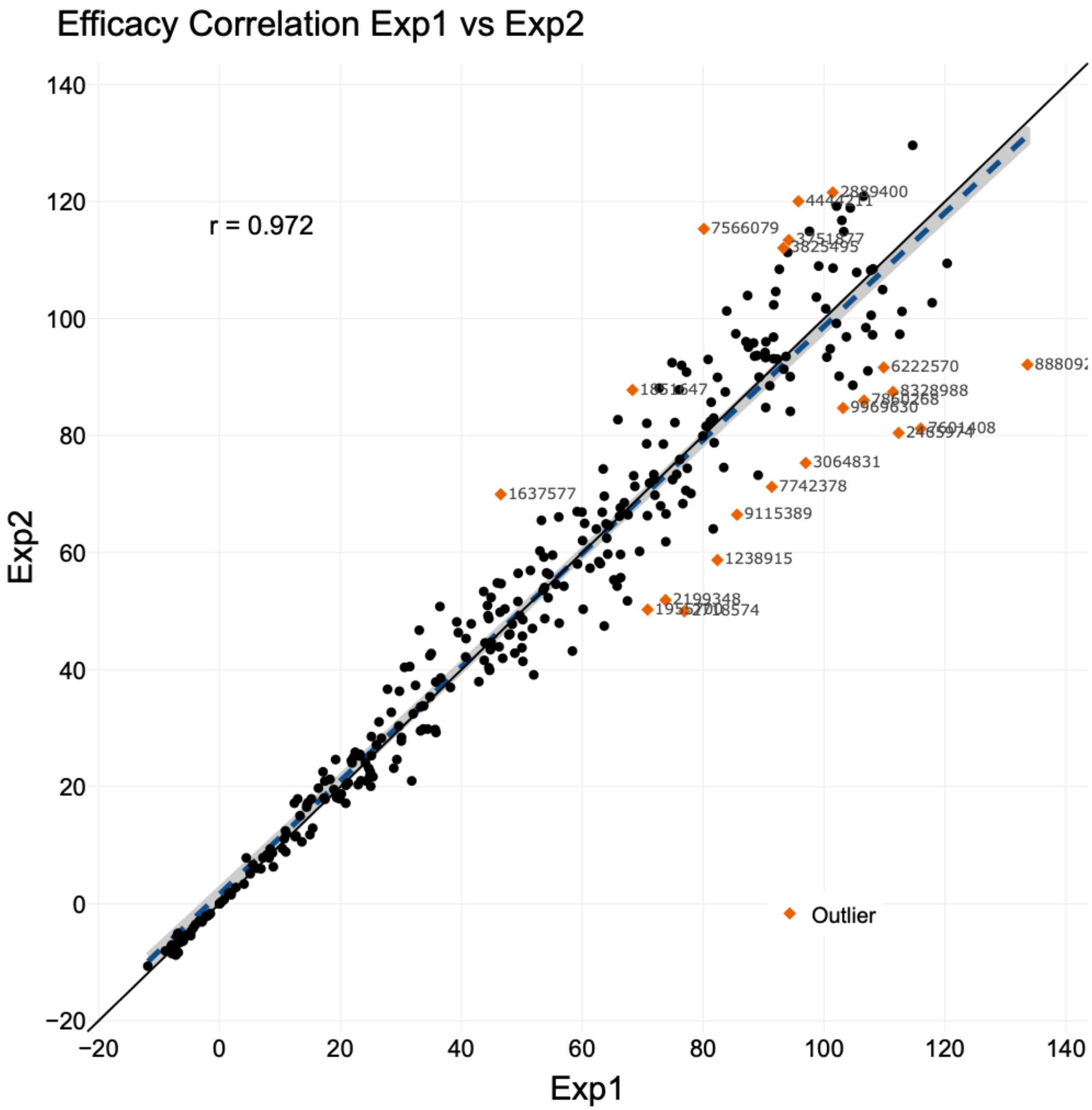

Figure 11: Potency Correlation plot.

Table 10: Replicate Experiment Statistics.

Potency Stats

n	MeanRatio	MSR	Upper Ratio Limit	Lower Ratio Limit	Upper Agreement Limit	Lowe
5	1.16	1.01×10^{1}	4.88	2.75×10^{-1}	2.89×10^{1}	

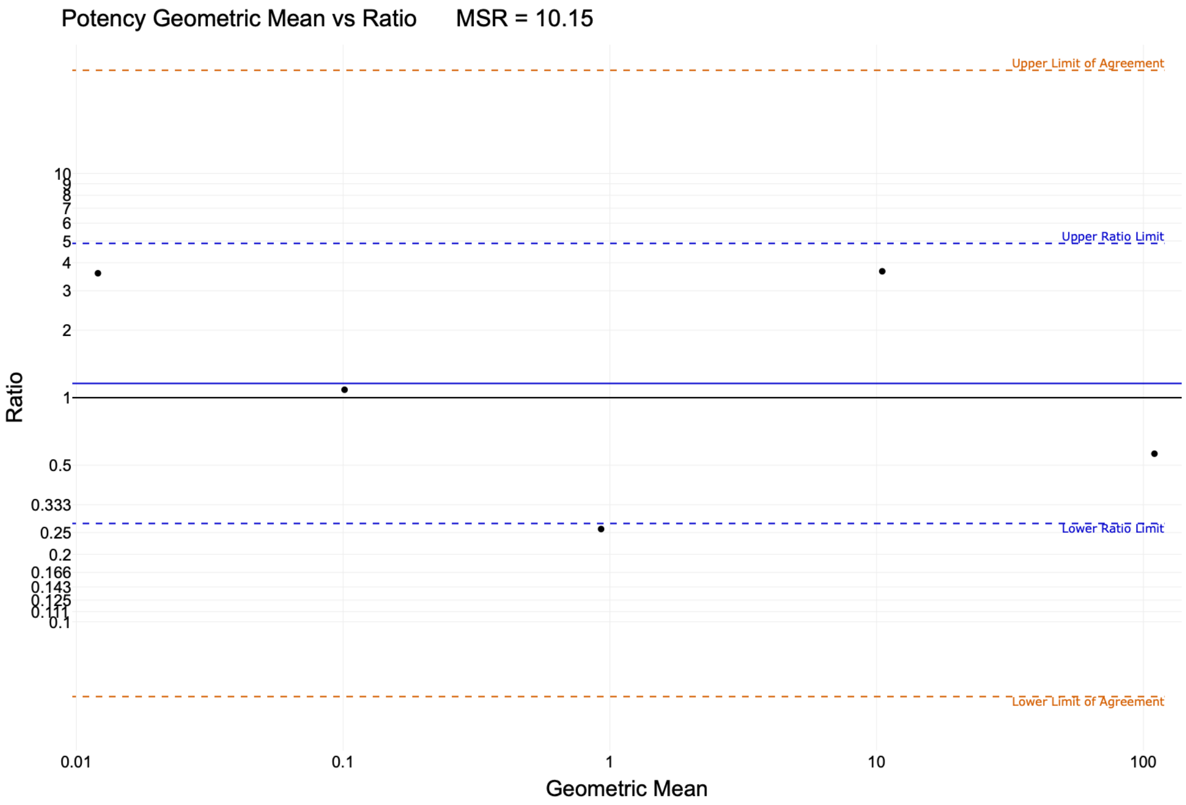

Figure 12: Mean Ratio plot.

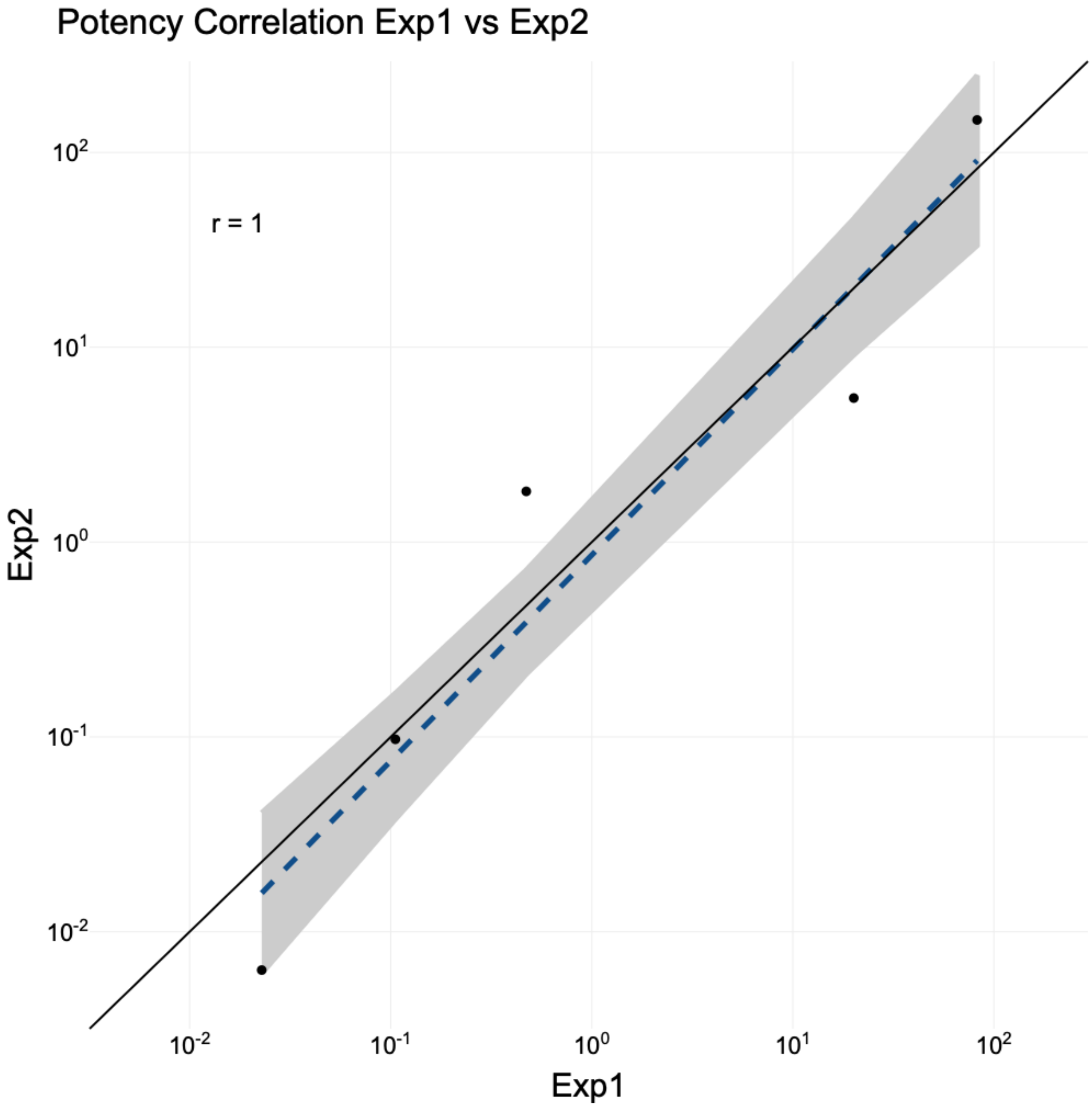

Figure 13: Correlation Plot

Table 11: Replicate Experiment Statistics.

Potency Stats

n	MeanRatio	MSR	Upper Ratio Limit	Lower Ratio Limit	Upper Agreement Limit	Lower Ag
30	9.90×10^{-1}	3.53	1.25	7.83×10^{-1}	3.59	

Fortunately, there is a simple solution to this problem. Each sample can be replicated multiple times on the plates. The replicate dose-response curves are then analyzed as if they were from different samples. Ideally, these replicate samples should be prepared independently, as if they were unique compounds with their own dilution series. In this example, the 5 unique samples were each replicated 6 times (e.g. sample 1 becomes 1A, 1B, 1C ...). This produces 30 pairs of potency values for the analysis

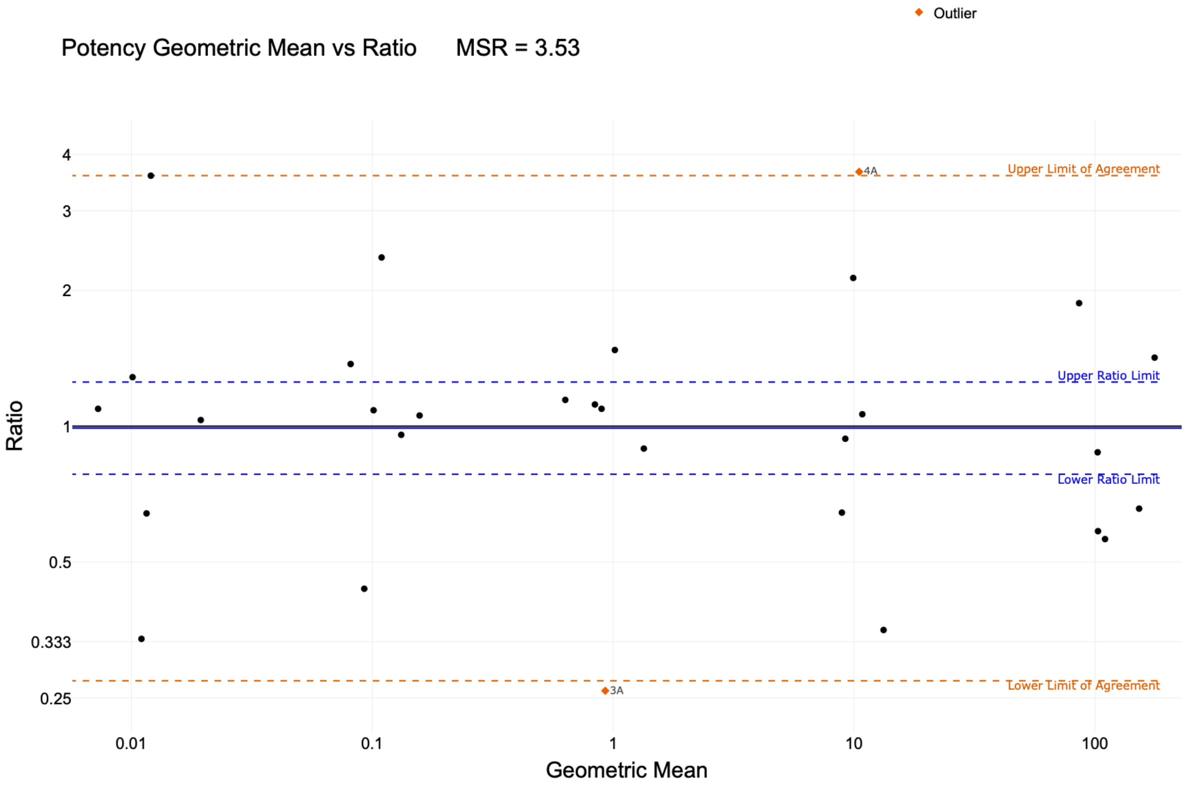

Figure 14: Replicated samples Mean Ratio plot.

The Bland-Altman plot, Figure 14, now shows 5 clusters of difference values corresponding to the 5 replicated samples.

The associated statistics, Table 11, show a much lower MSR.

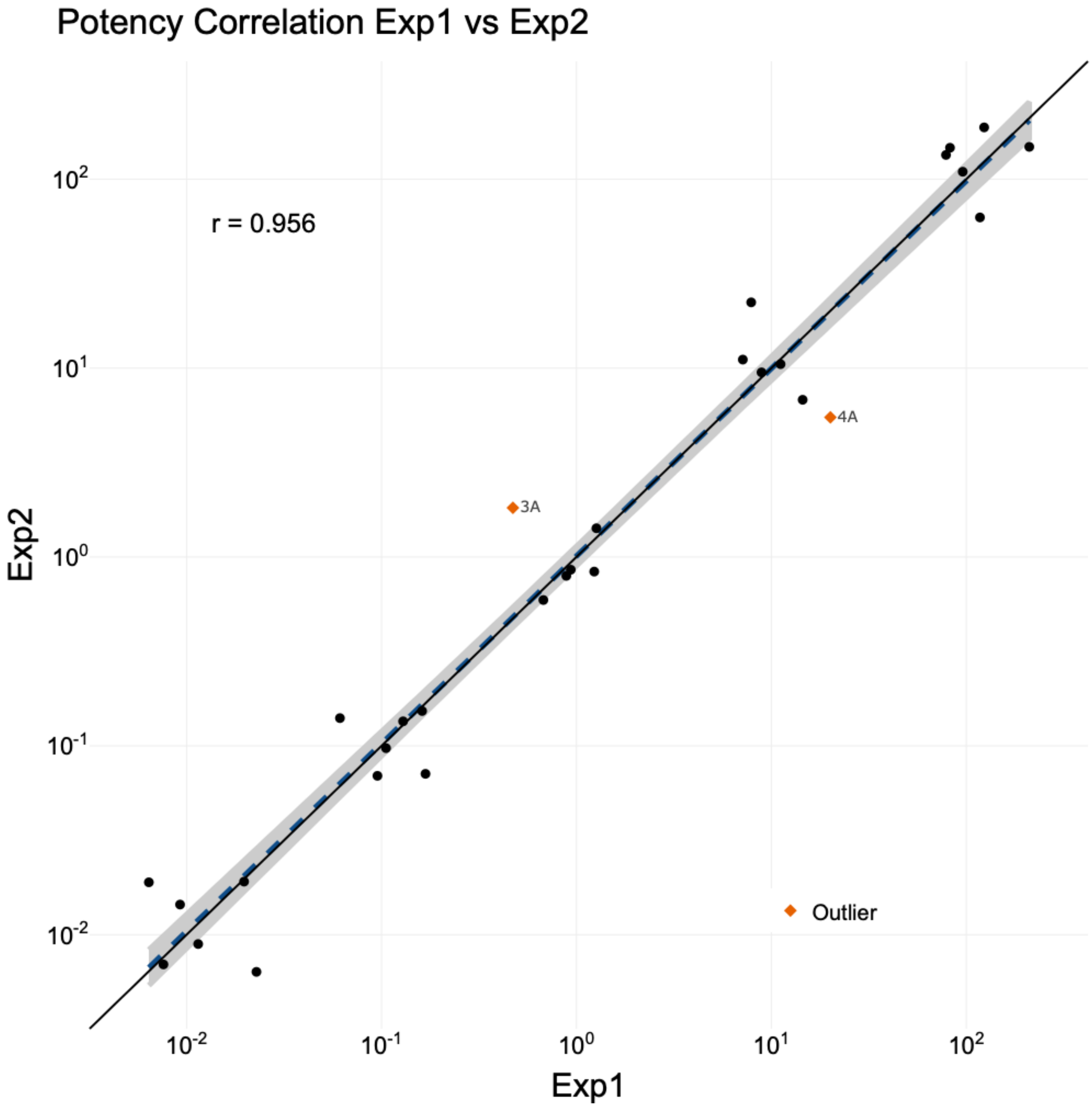

Figure 15: Replicated samples correlation plot.

Table 12: Replicate Experiment Potency Shift Data.

Potency Shift Data.

Sample	Exp1	Exp2	Geometric_Mean	Ratio
4649104	2.45	5.28	3.60	4.64×10^{-1}
2303569	1.68×10^{3}	2.33×10^{3}	1.98×10^{3}	7.20×10^{-1}
2002357	1.01×10^{2}	1.05×10^{2}	1.03×10^{2}	9.62×10^{-1}
8605428	1.25×10^{1}	5.99×10^{1}	2.73×10^{1}	2.08×10^{-1}
6472482	1.13×10^{2}	2.03×10^{2}	1.51×10^{2}	5.56×10^{-1}
1210635	2.04×10^{-3}	3.97×10^{-3}	2.85×10^{-3}	5.13×10^{-1}

Table 13: Potency Shift Replicate Experiment Statistics.

Potency Stats

n	MeanRatio	MSR	Upper Ratio Limit	Lower Ratio Limit	Upper Agreement Limit	Lower Ag
32	1.06	2.53	1.25	8.98×10^{-1}	2.73	

Similar clustering is observed in the correlation plot, Figure 15. While it is easy to identify the clusters due to the sample replication in the graphs, the improvement in statistical power from n=30 instead of n = 5 provides a much better estimate of the true assay variability. Both the MSR and the ratio limits are significantly lower.

Systematic Difference between Runs

Systematic differences between two experiments could be indicated when the mean ratio line is displaced from the ratio = 1 reference line. If the reference line is outside of the Ratio Limits for the Center Line, the difference is statistically significant and should be investigated to determine the cause. However, if the reference line is within the Ratio Limits, it could simply be random variation or to a cause with a small effect.

The data in Table 12 is used to illustrate this situation.

Now the black reference line (Ratio = 1) is clearly outside of the observed Ratio Limits around the blue Center Line for the data. This indicates that there is a systematic shift in the data between the first and second measurements of the compound potencies.

The associated statistics, Table 13, also shows that the expected ratio of 1 is not within the ratio limits.

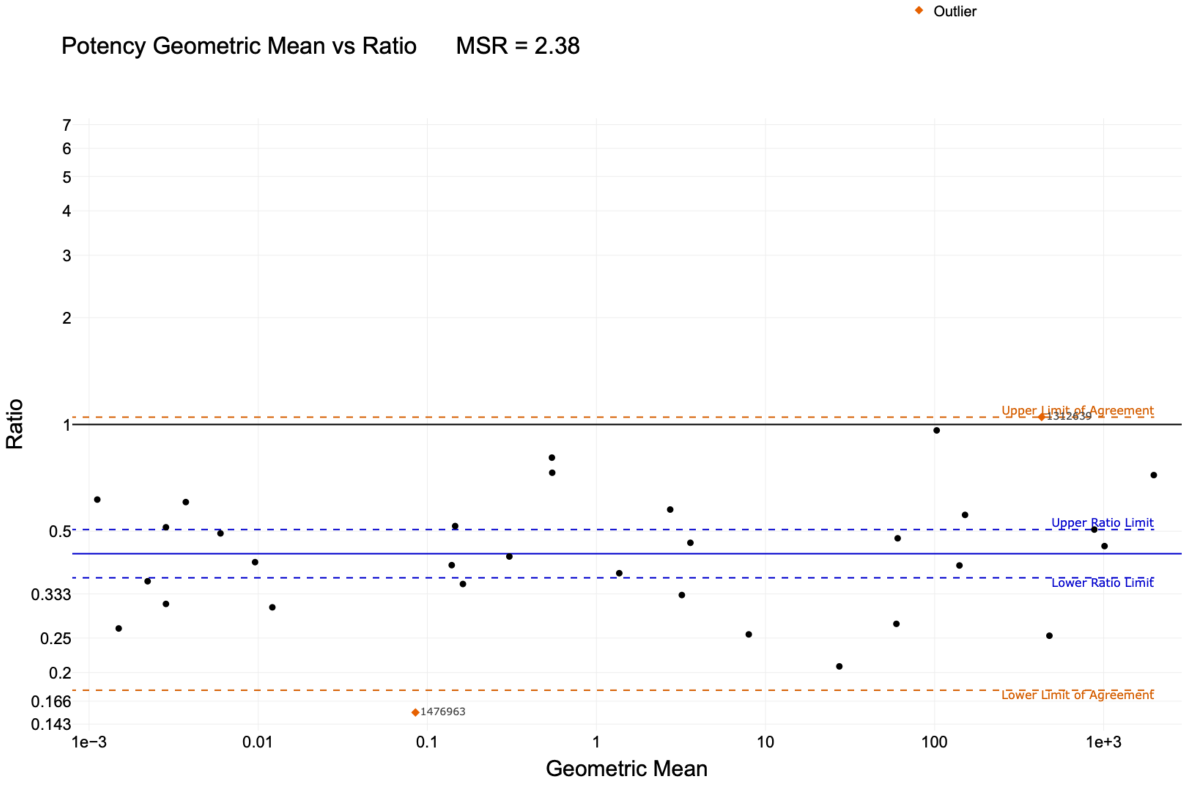

Figure 16: Systematic error Bland-Altman plot.

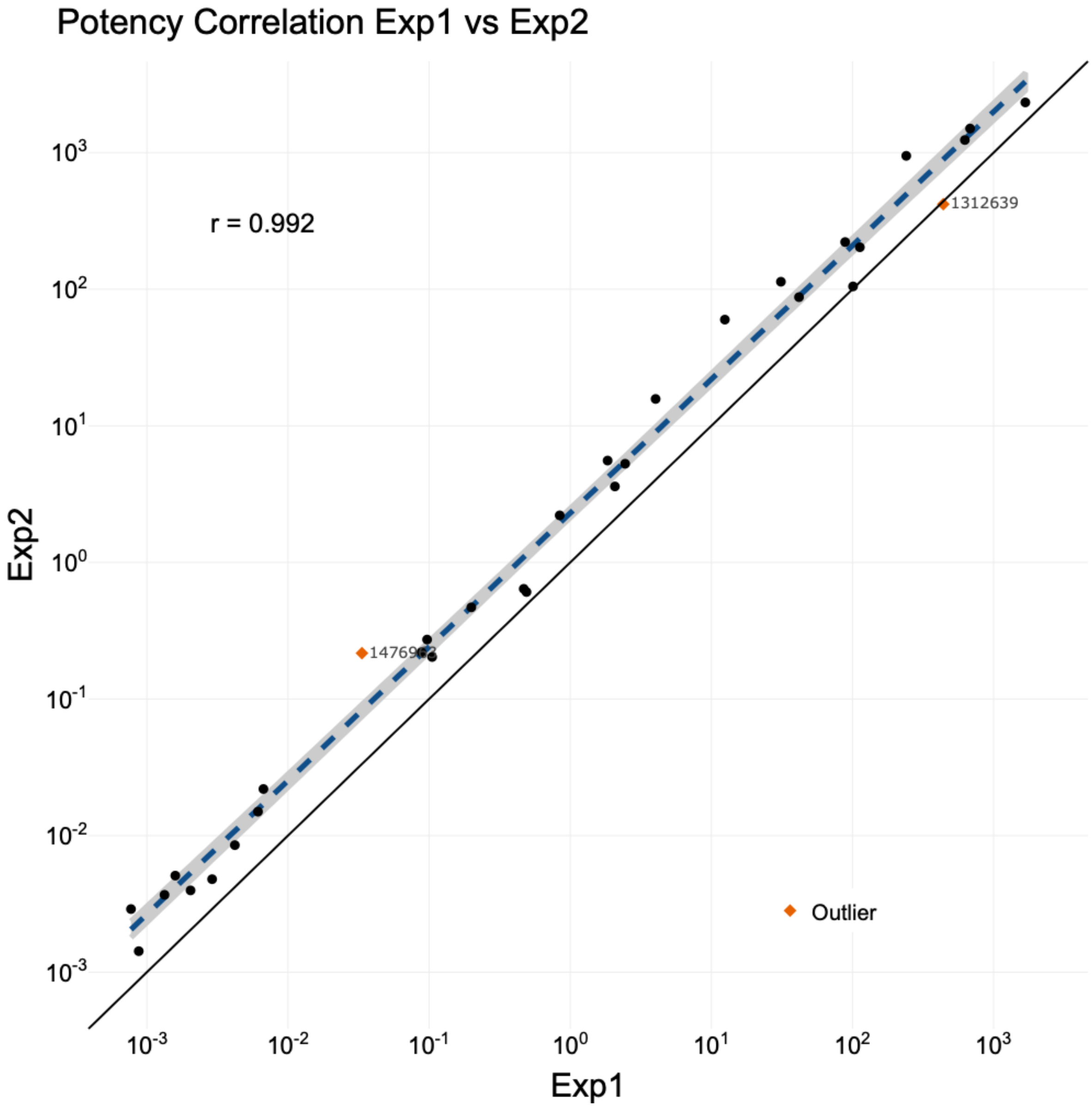

Figure 17: Systematic Shift Correlation Plot.

The correlation plot, Figure 17, also shows separation between the reference identity line and the correlation line.

Common causes can include any of the following:

1. Differences in the samples between experiments (e.g. sample lots, or solubilizations).
2. Reagents or cells used in the assay.
3. Equipment changes.
4. Environmental changes (e.g. temperature).
5. Personnel

If a cause is identified, it should be mitigated (if possible) and the replicate experiment repeated. Sometimes a cause may be identified which includes historical data that can't be changed. If this happens, it may be significant enough to require retesting of key compounds to assess the impact.

Acknowledgement

The Assay Guidance Manual Webtools are supported by the Intramural Research Program of the National Center for Advancing Translational Sciences at the National Institutes of Health (ZIA TR000340).

References

Beck, Benoit, Yun-Fei Chen, Walthere Dere, Viswanath Devanarayan, Brian J. Eastwood, Mark W. Farmen, Stephen J. Iturria, et al. 2017. "Assay Operations for SAR Support." In, edited by Sarine Markossian, Abigail Grossman, Michelle Arkin, Douglas Auld, Chris Austin, Jonathan Baell, Kyle Brimacombe, et al. Bethesda (MD): Eli Lilly & Company; the National Center for Advancing Translational Sciences. http://www.ncbi.nlm.nih.gov/books/NBK91994/.

Bland, J. M., and D. G. Altman. 1986. "Statistical methods for assessing agreement between two methods of clinical measurement." *Lancet (London, England)* 1 (8476): 307–10.

Dahlin, Jayme L., G. Sitta Sittampalam, Nathan P. Coussens, Viswanath Devanarayan, Jeffrey R. Weidner, Philip W. Iversen, Joseph V. Haas, et al. 2019. "Basic Guidelines for Reporting Non-Clinical Data." In, edited by Sarine Markossian, Abigail Grossman, Michelle Arkin, Douglas Auld, Chris Austin, Jonathan Baell, Kyle Brimacombe, et al. Bethesda (MD): Eli Lilly & Company; the National Center for Advancing Translational Sciences. http://www.ncbi.nlm.nih.gov/books/NBK550206/.

Eastwood, Brian J., Mark W. Farmen, Philip W. Iversen, Trelia J. Craft, Jeffrey K. Smallwood, Kim E. Garbison, Neil W. Delapp, and Gerald F. Smith. 2006a. "The Minimum Significant Ratio: A Statistical Parameter to Characterize the Reproducibility of Potency Estimates from Concentration-Response Assays and Estimation by Replicate-Experiment Studies." *SLAS Discovery* 11 (3): 253–61. https://doi.org/10.1177/1087057105285611.

———. 2006b. "The Minimum Significant Ratio: A Statistical Parameter to Characterize the Reproducibility of Potency Estimates from Concentration-Response Assays and Estimation by Replicate-Experiment Studies." *SLAS Discovery* 11 (3): 253–61. https://doi.org/10.1177/1087057105285611.

Elassaiss-Schaap, Jeroen, and Kevin Duisters. 2020. "Variability in the Log Domain and Limitations to Its Approximation by the Normal Distribution." *CPT: pharmacometrics & systems pharmacology* 9 (5): 245–57. https://doi.org/10.1002/psp4.12507.

Haas, Joseph V., Brian J. Eastwood, Philip W. Iversen, Viswanath Devanarayan, and Jeffrey R. Weidner. 2017. "Minimum Significant Ratio – A Statistic to Assess Assay Variability." In, edited by Sarine Markossian, Abigail Grossman, Kyle Brimacombe, Michelle Arkin, Douglas Auld, Chris Austin, Jonathan Baell, et al. Bethesda (MD): Eli Lilly & Company; the National Center for Advancing Translational Sciences. http://www.ncbi.nlm.nih.gov/books/NBK169432/.

Iversen, Philip W., Benoit Beck, Yun-Fei Chen, Walthere Dere, Viswanath Devanarayan, Brian J. Eastwood, Mark W. Farmen, et al. 2012. "HTS Assay Validation." In, edited by Sarine Markossian, Abigail Grossman, Kyle Brimacombe, Michelle Arkin, Douglas Auld, Chris Austin, Jonathan Baell, et al. Bethesda (MD): Eli Lilly & Company; the National Center for Advancing Translational Sciences. http://www.ncbi.nlm.nih.gov/books/NBK83783/.