Chapter 8. Performance of the method

Key message: *In vitro method developers need to ensure that in vitro methods they design will produce good quality data, i.e. fit for purpose, thanks to a stringent assessment of the performance of the method.*

Key content: *Elements of experimental design and how to determine the performance of a method are detailed, including aspects of plate layout, data analysis, and in-house method validation, including the assessment of linearity, range, accuracy, etc.*

Guidance for improved practice: *Details are given to increase the reliability of endpoint calculations when multiple independent experiments are run and to use tools to quantify performance characteristics.*

Recommendations *are given to in vitro method developers on how to increase the possibility of adoption of their method for regulatory purposes.*

In vitro method development and in-house validation should be considered as continuous and inter-dependent. Early in the development stage, the choice of instrumentation and methodology are selected based on the intended purpose and scope of the *in vitro* method. Once the development and optimisation of the method in the laboratory has been finalised it is recommended to perform an in-house validation (Section 8.3) of the method prior to routine use. This will provide documented evidence of the method performance in the laboratory and also prescribes on-going measures to ensure quality monitoring for the lifetime of the method. It will also check the feasibility of the method before the costly exercise of a formal collaborative trial (OECD , 2005[1]). The validation process, as described in the OECD Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment (OECD , 2005[1]), is not address this chapter (Section 1.3).

When discussing the in-house validation of *in vitro* methods, it is important to distinguish the analytical (measurable) endpoint (e.g. spectrophotometric, fluorimetric, mass-spectrometry and/or luminometric) from the *in vitro* method endpoint (e.g., proliferation, differentiation or viability) which refer to the biological concept being evaluated (Aschner et al., 2016[2]); (Schmidt et al., 2016[3]). The in-house validation of *in vitro* method analytical endpoint(s) is discussed in detail in (Section 8.3).

The assessment to be performed will largely depend on the type of *in vitro* method (i.e., qualitative or quantitative). Very few guidelines exist concerning the validation of qualitative methods (NATA, 2013[4]). There are several international guidelines available addressing the validation of quantitative methods (EMA, 2011[5]); (FDA, 2001[6]); (ICH, 2005[7]) which describe the general parameters required to assess the performance of the method analytical endpoint(s). International efforts have been made to discuss and harmonise the validation guidelines, however different interpretations still exist and no consensus regarding acceptance criteria or the actual validation process (Rozet *et al*., 2011) has yet been achieved. Most of these differences stem from the different regulatory frameworks in place. For a detailed comparison of the various international guidelines see (Kollipara et al., 2011[8]).

8.1. Acceptance criteria

As most *in vitro* methods are generally intended to predict a qualitative and/or a quantitative response, predictive of the degree of human or environmental hazard, it is essential that the *in vitro* method performs consistently over time and between laboratories.

Acceptance criteria should be developed based on historical data for all critical components and aspects of the method. Criteria should be defined for the test system (e.g., passage number, growth curve, cell recovery) and test system performance (e.g., positive, negative, and vehicle controls where applicable). Acceptance criteria should also be set for the analytical endpoint determination (e.g., linearity, accuracy, range) and also include data analysis (e.g., line fitting). These criteria should be developed and detailed in the *in vitro* method SOP(s).

Acceptance criteria should primarily be established based on information from historical data. When available, these can be then supplemented by data from validation studies, or from relevant bibliographic data including guidance documents. Historical data should be collected using the unchanged method, unless it can be shown that any changes have not affected the values. Data should only be rejected when there is a clear, valid and

scientifically justified reason to do so (Hayashi et al., 2011[9]), and the reasons for rejecting said data should be clearly and accurately documented.

For (transformed) data, which follows an approximate normal distribution, the mean and standard deviation (SD), e.g., for the positive control historical data, are calculated and the acceptance criteria are set at for instance ± 2 SD. For example the Bovine Corneal Opacity and Permeability (BCOP) *in vitro* method (Figure 6.1) uses 100% ethanol as the positive control. It has a mean published *in vitro* score (opacity + 15×permeability) of 51.6 ± 6.2 (mean ± Standard Deviation SD), which would set the acceptance criteria (mean ± 2 SD) to 39.2 to 64.0 (n=1171 trials) (Harbell and Raabe, 2014[10]).

For dose-response methods it is important to test multiple concentrations of the test item that fall within the linear dynamic range (Section 8.3.2) of the method, so as to narrowly define the 50% activity point. The 50% activity point (concentration) for the positive control may be used for e.g., establishing the acceptance criteria for a dilution-based cytotoxicity assay. This approach allows increased and decreased sensitivity to be readily identified.

Establishing acceptance criteria for the negative control is important to assure that the test system performs normally, and is just as important as for the positive control and can be done in the same manner (e.g., within +/- 2 or 3 SD of the historical mean response, or within the 95% control limits of the distribution of the historical data). Other acceptance criteria may be established such as criteria for the variability of the (quantifiable) data (e.g., OECD TG 431, 439, 492) or criteria for the minimum level of cell viability (e.g., OECD TG 442E).

Finally, it is also important to establish the cut-off value of the acceptance criteria, i.e., clear rules whether the response of a reference/control item is accepted or not, also taking into consideration the number of significant digits. The preferred approach is to specify the same number of significant figures both for the acceptance criteria and the measured result.

8.2. Experimental design

The number of replicates for each testing condition, including concentration level(s) used for the reference and control item(s), and test items etc., should be specified. During *in vitro* method development the number of replicates must be chosen using appropriate statistical methods. For example, a statistical power analysis (Crawley, 2015[11]) can be used to calculate the desirable number of replicates to detect a defined difference between treatments with pre-set levels of confidence (Krzywinski and Altman, 2013[12]). However, one should be aware that this number may be too high to be useful in practice. Alternatively the statistical power is provided for the chosen number of replicates.

Additionally, when multiple concentrations of a test item are tested, the mathematical model (e.g., dose-response curve) can be fitted to the experimental data using increasing number of tested concentrations and/or replicates, but generally it is better to increase the number of concentrations than the number of replicates. The lowest number of replicates that gives satisfactory variability of the parameter of interest (e.g., IC_{50} within acceptable limits) can be used in future studies (Assay Guidance Manual[1], High-Throughput Screening (HTS) Assay Validation[2]). Apart from these statistical considerations, sometimes practicalities such as cost and availability of replicates may also play a role in the selection process. However, the impact of reducing replicates should always be subjected to careful analysis and the corresponding statistical power should be given.

Similarly, the number of independent experiments needs to be evaluated. For instance, *in vitro* methods with a high degree of inter-experimental variability, such as those using primary tissues, may need a higher number of independent experiments compared to *in vitro* methods employing continuous cell lines.

Statistical methods (e.g., factorial design) can be very useful in the process of optimising new *in vitro* methods. To obtain an *in vitro* method that leads to accurate, reliable and robust readouts, the results of several combinations of any changes in the *in vitro* method would have to be assessed. Factorial design of experiments is often used where there are a large number of variables to be assessed, as it is nearly impossible to approach all possible combinations experimentally. It is efficient at evaluating the effects and possible interactions of several factors (independent variables). A statistical approach predicting the effect of changes in the *in vitro* method steps on the observed readout (known also as method robustness assessment) would allow for the development of an efficient *in vitro* method design, since the experimental robustness check can be based on a much smaller subset of combinations (Box, Hunter and Hunter, 2005[13]; Groten et al., 1997[14]).

8.2.1. Plate layout

When developing an *in vitro* method care should be taken to minimise any potential systematic effects. Many cell-based assays employ cell culture plates (6, 12, 96, or 384 well plates), so care must be taken to ensure cell seeding, treatment and measurement is performed in a uniform fashion across the whole plate (well-to-well), between plates and across multiple runs. Plate effects may occur in the outer wells (e.g., due to evaporation), across columns or rows or even within the actual wells (within well effects). Plating cell density is also crucial if exposure is taking place during the log phase such as for inhibition of cell growth endpoint. In this case if the control cells reach the stationary phase during the exposure time, the effect of the test item may be underestimated. Plate effects should be evaluated e.g., by using the same conditions/treatments across a complete plate.

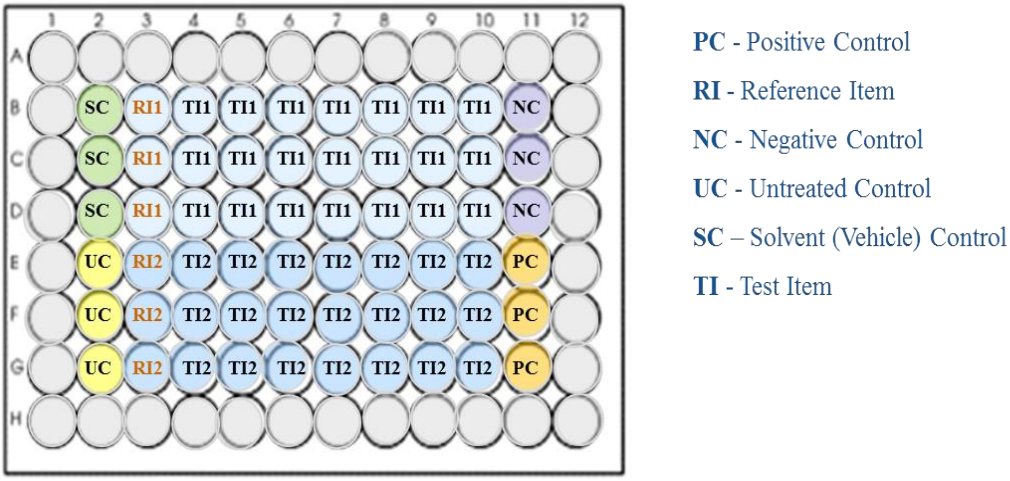
Drift can be due to seeding density variation during the process of initial cell seeding in plates, e.g., cells may be settling down in the master vessel which is used to store a cell suspension used to seed a particular plate. Additionally, using the same set of tips on a multichannel pipette while pipetting cells in media compositions prone to foaming, may compromise the accuracy of the seeding. Higher variability, which cannot be resolved via technique optimisation may require increased number of replicates/concentrations used to calculate the dose-response, or a higher numbers of independent experiments (Iversen et al., 2004[15]).

Randomisation of treatment wells in the cell culture plate is a strategy used to minimise inherent plate bias due to edge effects[3], drift, etc., and is particularly effective in an automated dosing setup. However randomisation may introduce other unforeseen errors, such as increased pipetting errors (usually only single wells can be pipetted), or data transfer/analysis errors (data may need to be rearranged for data analysis). It may also take significantly longer to treat the whole plate and so inadvertently introduce timing errors.

The plate layout will depend on the specific needs of the in vitro method, e.g., the numbers of controls, concentrations tested and replicates required. It should be such that cross-contamination (e.g., between test items) can be controlled for by checking variability between replicates. The plate design should also take into account how to perform comparison across plates so as to check between run or plate variability, by using

appropriate reference and control items. An example of an experimental 96-well plate layout using reference and control items is shown in Figure 8.1 (Coecke et al., 2014[16]).

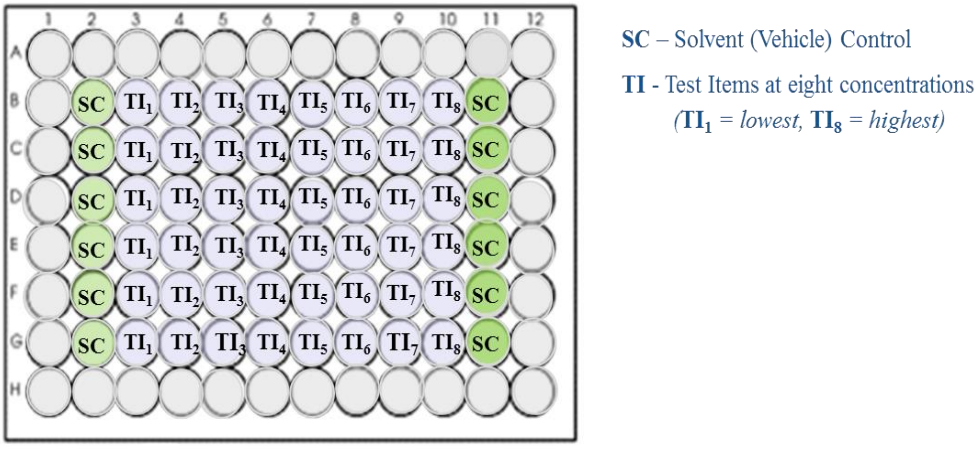
Figure 8.1. Example of plate layout including reference and control items, and solvent and untreated controls



Source: (Coecke et al., 2014[16])

The example plate layout minimises potential edge effects (difference between outer and inner wells due to evaporation). A way to assess plate drift is to include solvent controls (SC) on both the left and right side of the plate (Figure 8.2). Left and right SCs should not differ by more than a certain percentage for the plate to be accepted, e.g., a test meets acceptance criteria if the left and the right mean of the SCs do not differ by more than 15% from the mean of all SCs (NIH, 2001[17]).

Figure 8.2. Plate layout for systematic cell seeding errors



In addition, certain test or reference items may be volatile (e.g., solvents) or may contaminate neighbouring wells by capillary action, known as the wicking effect (Sullivan, 2001) and this may need to be taken into account in designing plate layouts. For instance, the commonly used cell lysis surfactant Triton X can affect cell viability in neighbouring wells and should be used at low concentrations or separated from cell-

containing wells by placing wells containing media or buffer in-between. Covering the plates with a foil prior to incubation, may also be employed, to avoid evaporation of volatile test items (e.g., OECD TG 442D).

Different effects found in the outer row of wells compared to the inner wells, are often due to uneven evaporation rates or plate stacking and can be a source of variation, as outer wells can often present as outliers compared to inner wells. Often the outermost wells contain a sterile, water-based solution and are not used for control, reference or test items as evaporation may take place during opening the door of the incubator. The cell-free wells may also be used for controls in the assay (e.g., background OD/FI, test item interference with the assay). Modern incubators are able to compensate much better for the change in humidity and so limit evaporation inside the incubator. If these outer wells are used for test, reference or control items, it should be clearly stated in the SOP to check for potential variation prior to use as not all laboratories may be equipped with appropriate incubators.

The inclusion of relevant reference and control items, and setting of acceptance criteria on the basis of historical data, is essential for regulatory applicability of *in vitro* methods and should be considered when developers decide on their plate layout. By including the correct reference and control items, the data set obtained from the *in vitro* method will demonstrate the correct functioning of the test system and the method used for analysis and therefore the validity of the experiments executed.

8.2.2. Data analysis

Transformation of data, e.g. normalisation or fitting to model equations, should be defined prior to data acquisition, and should be described in a SOP (OECD, 2017[18]) or in the relevant study plan. Formulas for normalisation (checked for accuracy) should be documented, validated (when implemented in electronic format) and disclosed along with a description and justification of the controls used in the calculation. It is recommended that computer scripts used to process raw data (e.g., Excel spreadsheets, scripts, macros etc.) should be validated and fully documented. The OECD Advisory Document Number 17 on the Application of GLP Principles to Computerised Systems provides guidance on validation of computerised systems (OECD, 2016[19]).

When a relationship is assumed between the tested concentrations and the response, a dose response curve, if required, can be fitted to obtain summary data such as the EC_{50} or IC_{50}. When fitting mathematical models, such as a dose-response curves or standard curves, to the data the models and reasoning behind their choice need to be documented. For example, when fitting a dose-response curve, the type of the equation used should be documented (e.g., a four parameter logistic curve) together with any constraints (e.g., top constrained to 100% in normalised data), limitations (e.g., assumption of monotonicity) and weightings (e.g., by inverse data uncertainty) applied (Motulsky, H.J.; Christopoulos, A., 2004[20]). Furthermore, the software name and version used to fit the equations should be documented, as well as the confidence interval of the parameters of interest, e.g., EC_{50}, RPC_{Max}, PC_{Max}, PC_{50} and/or PC_{10} (OECD, 2016[21]) and the relevant goodness of fit parameters (R-square, sum of squares etc.) so as to justify the selection of the model. In some cases it may be preferable to test multiple models (e.g., include non-monotonic curve) and select the best fit after the curve fitting.

8.2.3. Outlier detection and removal

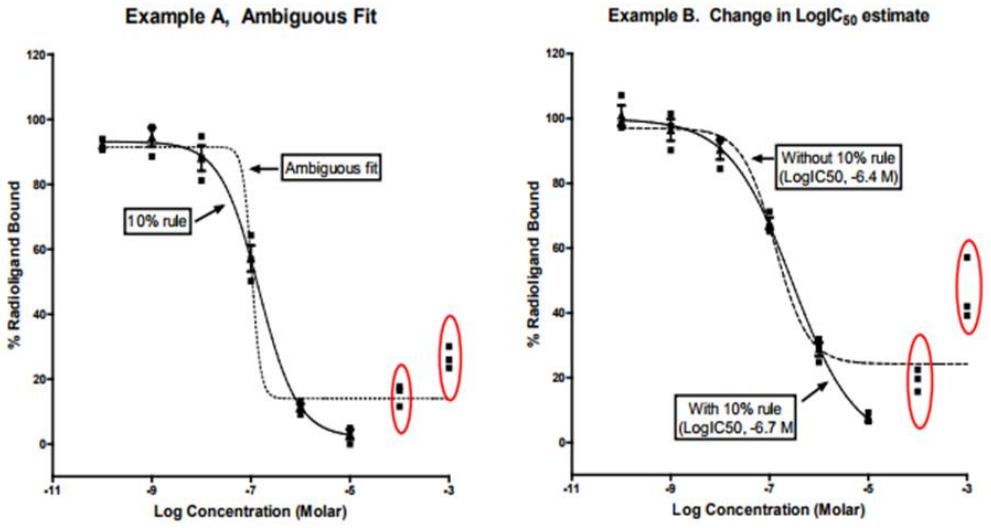
Criteria to detect/remove outliers should be stated and the reasoning should be given (Motulsky and Brown, 2006[22]; Pincus, 1995[23]). Outlier tests, such as Grubbs' Test (for single outlier) or using visual tools like boxplot (e.g., Mahalanobis distance), may be used to rule out outlier when it is difficult to judge whether the data should be regarded as an outlier or not, or when the reason for the occurrence of the outlier is unknown.

8.2.4. Non-monotonic dose and U-shaped curves

While dose response curves for relatively simplistic models, like most *in vitro* assays, are monotonic (i.e., they increase or decrease over the entire dose response range), sometimes a non-monotonic dose response can be observed. This potentially can be due to superimposition of separate effects that would individually elicit monotonic dose responses, or could indicate a perturbation of the system, e.g., due to chemical insolubility and/or precipitation at high concentrations (Section 6.5). Usually, such curves are U-shaped and can occur in many types of *in vitro* and binding assays.

An example of such a U-shaped curve is given below, where the analysis and interpretation of competitive binding data (Figure 8.3) can be complicated by an upturn of the percent binding when testing chemicals at the highest concentrations. As the concentration of the test item approaches the limit of solubility, the displacement of [^{3}H]17β-estradiol begins to generate a "U-shaped response curve" (OECD, 2015[24]). Such U-shaped curves are typically considered artefacts of the test conditions rather than relevant descriptors of the binding affinity of the test item. Retaining such data points (circled in red) when fitting competitive binding data to a sigmoid curve can inappropriately raise the perceived bottom of the curve Figure 8.3B, and can sometimes lead to a misclassification of the Estrogen Receptor (ER) binding potential for a test item (Figure 8.3A). This problem can be further controlled by excluding from the analyses all data points where the mean of the replicates for the % specific bound show 10% or more radioligand binding than the mean at a lower concentration (i.e., 10% rule), but it may not always be appropriate to include such a rule, as unintended and unforeseen consequences have been observed.

Figure 8.3. Example, Analysis of Competitive Binding Data, with and without use of 10% rule



If there is reason to believe that the test item is a non-binder, it might be appropriate to include a subjective waiver, so that the laboratories are allowed to use their judgement with regard to the use of the "10% rule", however this needs to be justified in the study report. It is important to note that a subjective waiver of the 10% rule may not be considered statistically appropriate.

8.3. In-house validation of the measurement endpoint(s)

This section focuses on the in-house validation of the measurement (analytical) endpoint(s), as described in the facility *in vitro* method SOPs. It is intended mainly for quantitative methods, however some of the performance characteristics relevant for qualitative methods are also described.

1. There are no specific regulations concerning *in vitro* method in-house validation, however guidance documents (e.g., FDA) and guidelines (e.g., EMA, ICH Q2(R1)) for analytical endpoint(s) validation describe the elements required in most validation studies in order to characterise the methods in terms of e.g., their reliability and reproducibility. The assessment to be performed will vary depending on whether the method has been developed in-house, whether it has been transferred from another laboratory, whether it is commercially available or whether it has previously undergone full validation. Many guidelines classify method validations into full validation, partial validation, and cross-validation (EMA, 2011[5]; FDA, 2001[6]).
2. A full validation includes all relevant aspects of a method, and should be performed for all new in-house developed methods or when major changes are made to an existing method.
3. A partial validation should be performed when a previously validated method undergoes minor modification(s), such as change in calibration concentration range or when published methods, often modified to the facility's requirements, are been transcribed into facility SOP(s).

4. A cross-validation is performed when two or more methods are used to generate data within the same study or across different studies. When performing cross-validation the same set of reference, control and/or test items should be analysed by both analytical methods.

While not all aspects of the in-house validation might be applicable to all methods, the guidance documents describe several acceptable approaches to be taken depending on e.g. the purpose of the method and whether the *in vitro* method is quantitative or qualitative. In general, quantitative methods should address at least, where applicable, the method's accuracy, reproducibility, linearity, limits of detection and range of measurement, while for qualitative methods specificity and sensitivity are key criteria.

Many of these guidance documents describe the principles of validation and when and how to apply them. Regardless of the type of validation (full, partial or cross-validation), it is recommended to develop a validation plan, preferably written as step-by-step instructions, where the appropriate performance characteristics to be assessed are defined up-front before beginning the validation. Some GLP concepts, such as the requirements for a study plan and a final report can be applied when performing an in-house validation study.

The validation should be performed by trained and qualified laboratory personnel, to minimise operator variability and only calibrated/validated (Section 4.1) equipment should be used, so as to reduce equipment related issues. The results of the validation should be compared with the acceptance criteria for the performance characteristics described in the validation plan and/or SOP(s) and any deviation should be recorded and documented in a summary report (validation report) together with conclusions on the outcome of the validation.

Care should be taken to select appropriate and meaningful performance characteristics, and not to miss potentially critical ones such as reagent variability when drafting the validation plan, as these will depend on the nature and type of the method being validated. A comparison of some of the quality parameters for both quantitative and qualitative methods is shown in Table 8.1.

Table 8.1. Common quality parameters for quantitative and qualitative analytical methods

Quantitative Method	Qualitative Method
Accuracy: trueness, precision	Sensitivity and specificity
Uncertainty	Unreliability region
Sensitivity and specificity	False positive and negative rates
Selectivity: interferences	Selectivity: interferences
Range and linearity	Cut-off limit
Detection limit	Detection limit
Ruggedness or robustness	Ruggedness or robustness

Source: (Trullols, Ruisánchez and Rius, 2004[25])

Caution must be applied when comparing these quality parameters as similar terms are used for different concepts and their evaluation may be different, e.g., for quantitative methods sensitivity should be a numerical value that indicates how the response changes whenever there is a variation in the concentration of the analyte. However for qualitative methods often reported as true/false or positive/negative, sensitivity is evaluated differently and as such may not be comparable (Section 8.3.4). The same applies to

specificity, detection limits, cut-off value and uncertainty or unreliability region (Trullols et al., 2005[26]) Some of the most common performance characteristics are discussed in the following sections.

8.3.1. Detection Limits and Cut-off values

The response of the instrument and the *in vitro* method with regard to the readouts of interest should be known, and should be evaluated over a specified concentration range, usually of the reference item. Various approaches may be used to determine the Limit of Detection (LOD) and Limit of Quantitation (LOQ) (ICH, 2005[7]).

1. Based on Visual Evaluation
2. Based on Signal-to-Noise ratio
3. Based on the Standard Deviation of the Blank[4]
4. Based on the Calibration Curve

Other approaches, described in the validation plan may also be employed. The LOD determines the lowest actual concentration or signal that can be consistently detected with acceptable precision, but not necessarily quantified. For normally distributed data, the LOD is often determined as the concentrations at the average response + 3 SD of the negative control range, as this gives only 1% chance of a false positive. LOQ is frequently calculated based on acceptable accuracy and precision of the reference item/reference item[5].

The Signal to Noise (S/N) ratio is frequently applied for methods which exhibit background noise (observed as the variation of the blanks) as baseline. It is calculated by comparing measured signals from samples with the reference item/positive control item with those of blank samples. A S/N ratio of 3 is generally accepted for estimating LOD, and S/N ratio of 10 is used for estimating LOQ (ICH, 2005[7]). Alternatively, assay acceptance can be determined using a signal window calculation[6].

For qualitative methods detection limits cannot be calculated as the SD can only be calculated when the response is a numerical value. For these methods a cut-off value, i.e., the minimum concentration of a substance needed to ascertain detection with a certain probability of error (usually 5%), can be calculated. The cut-off value is usually determined by establishing the false positive and negative rates at a number of levels below and above the expected cut-off concentration, and as such is related to the sensitivity of the method (Section 8.3.4).

8.3.2. Linearity and dynamic range

The response of the instrument detector can be expressed either as dynamic range or as linear dynamic range. The dynamic range is the ratio of the maximum and minimum concentration over which the measured property (e.g., absorbance) can be recorded. The linear dynamic range, i.e., the range of solute (e.g., reference item) concentrations over which detector response is linear, is more commonly used.

To quantify the amount of analyte in a sample a calibration curve is prepared, often assessed using a dilution series of the reference item. The results are plotted and a curve, usually linear, is fitted to the data. However not all *in vitro* methods will be linear for their full range, so a linear range (dynamic range) will need to be defined within the method's range.

For quantitative measurements, the boundaries of the dynamic range are determined by the lowest and highest analyte concentrations that generate results that are reliably produced by an *in vitro* method. The lower limit of linearity is frequently referred to as the Lower Limit of Quantification (LLOQ) and the upper limit of linearity as the Upper Limit of Quantification (ULOQ). The upper limit of linearity may be restricted by the highest available concentration in a sample or by the saturation of the signal generated by the instrument, while the lower limit is often limited by the instrument specifications.

The range is normally derived from the linearity and is established by confirming that the procedure provides an acceptable degree of linearity, accuracy and precision when applied to samples containing analyte (e.g., reference item) within or at the extremes of the specified range of the analytical method.

If a linear relationship exists statistical methods can be employed such as fitting of a regression line using the least squares and calculating the linear regression parameters (correlation coefficient, slope, y-intercept as well as residual sum of squares). Regression calculations on their own are usually considered insufficient to establish linearity and objective tests, such as goodness-of-fit may be required.

A correlation coefficient (r) of 0.99, based on a Goodness of Fit test, is often used as an acceptance criterion for linearity, however depending on the method lower r values may also be acceptable. Where a non-linear relationship exists, it may be necessary to perform a mathematical transformation of the data prior to the regression analysis. As linear regression is easy to implement, compared to other regression models (e.g., non-linear), a straight-line calibration curve will always be preferred.

For certain assays/methodologies, equations other than the linear can be fit as a standard curve, provided that the user is operating within the range of the assay/equipment (Section 4.1). However, it is recommended that the simplest model that adequately describes the concentration-response relationship is used. Selection of weighting and use of a complex regression equation should be justified. (Burd, 2010[27]; EMA, 2011[5]; FDA, 2001[6]; Viswanathan et al., 2007[28]). A minimum of 5 concentrations is recommended when assessing linearity, however other approaches may be used if justified (ICH, 2005[7]).

Subsequently, to facilitate efficient *in vitro* method transfer, the calculated linear regression parameters should be submitted along with a plot of the data. When the upper limit is exceeded (i.e., samples fall outside of the linear range), they may need to be diluted if possible. Where samples give a result below the lower limit of the linear range, it may be necessary to adapt the sample preparation to higher concentrations or change to a more sensitive apparatus.

8.3.3. Accuracy and precision

Assessment of accuracy and precision of a method will depend on whether the method is a quantitative or a qualitative method. The precision of a quantitative method describes the closeness of individual measures of an analyte (e.g. reference item) and is expressed as the coefficient of variation (CV). The FDA Bioanalytical Method Validation guidance document recommends that precision should be measured using a minimum of five determinations per concentration and a minimum of three concentrations in the range of expected concentrations (FDA, 2001[6]). Within-run and between-run precision should be reported.

For small molecules the within-run and between-run precision should not exceed 15% (20% at the LLOQ and ULOQ) while for large molecules (e.g. peptides and proteins) the within-run and between-run precision should not exceed 20% (25% at the LLOQ and ULOQ). The total error (i.e., sum of absolute value of the % relative error and % coefficient of variation) should not exceed 30% (40% at LLOQ and ULOQ) (EMA, 2011[5]).

For quantitative methods accuracy is usually determined using certified reference materials, if available, or by comparison to a reference method or to other methods. The accuracy of a method describes the closeness of mean test results obtained by the method to the actual value (or nominal) value (concentration) of the reference item. Accuracy is determined by replicate analysis of validation samples containing known amounts of the analyte (FDA, 2001[6]). The preparation of validation samples should mimic that of the study samples, and measurements should be made across at least 6 independent assay runs over several days (EMA, 2011[5]).

Accuracy should be reported as a percentage of the nominal value. When assessing the within-run and between-run accuracy for small molecules the mean concentration should be within 15% of the nominal value at each concentration level (20% at the LLOQ and ULOQ). For large molecules, both for within-run and between-run accuracy, the mean concentration should be within 20% of the nominal value at each concentration level (25% at the LLOQ and ULOQ) (EMA, 2011[5]).

Qualitative *in vitro* methods (e.g., as strong, weak), depend on accuracy and reliability to correctly classify chemicals according to its stated purpose (e.g., sensitivity, specificity, positive and negative predictivity, false positive and false negative rates). In such cases cut off values are used and their impact on the accuracy and reliability should be taken into account. The use of confidence bounds based on the distance from these cut-off values may not always be determined and therefore it may be preferable to conclude that the result is inconclusive (i.e., neither clear positive nor negative). The false positive rate is the probability that a test item which is actually negative being classified as positive by the method (Trullols et al., 2005[26]).

8.3.4. Sensitivity and specificity

For quantitative methods sensitivity may be defined as the capacity of the *in vitro* method to discriminate small differences in concentration or mass of the test item, while specificity may be defined as the ability of the *in vitro* method to identify, and where appropriate quantify, the analyte(s) of interest in the presence of other substances, i.e., the extent to which other substances may interfere with the identification/quantification of the analyte(s) of interest (ICH, 2005[7]). It may not be always possible to demonstrate that the method is specific for a particular analyte (complete discrimination). Specificity is concentration-dependent and is usually determined by adding materials which might be encountered in samples for which the method was developed. Specificity should be determined at the low end of the working range and ensure that the effects of impurities, cross-reacting substances, etc., are known.

Sensitivity in relation to qualitative methods may be defined as the ability of the method to detect the true positive rate while specificity is the ability of the method to correctly identify the true negative rate. The performance of a qualitative method can also be assessed with positive predictive values (PPV) and negative predictive values (NPV). PPV is the proportion of correct positive responses testing positive while NPV is the proportion of correct negative responses testing negative by an *in vitro* method

(Table 8.2). When calculating parameters such as sensitivity, specificity, false positive rate, false negative rate it is important that a balanced dataset is used (approximately an equal number of positive and negative compounds), otherwise these parameters will not reflect the true situation. The level of sensitivity, specificity, etc. which is acceptable is not standardised and is dependent on the list of items with which they are determined. Therefore, strict boundaries in acceptable levels for these accuracy parameters are not realistic. Generally though, sensitivities below 75% should not be accepted.

Table 8.2. Possible outcomes of an *in vitro* method result of a test item in a validation

		Condition		Prediction
		True	*False*	
Test Outcome	Positive	True positive (TP)	False positive (FP)	PPV
	Negative	False negative (FN)	True Negative (TN)	NPV
		↓ Sensitivity	↓ Specificity	Accuracy

$$\textbf{\textit{Sensitivity}}\ (\%) = \mathbf{100} \cdot \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FN})} \qquad \textbf{\textit{Specificity}}\ (\%) = 100 \cdot \frac{TP}{(TP + FN)}$$

$$\text{PPV} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}} \qquad \text{NPV} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false negatives}}$$

8.3.5. Repeatability

Repeatability is defined as the closeness of the results between a series of measurements of a single sample obtained by the same study personnel, usually a single person, under the same operating conditions over a short interval of time, and is also called intra-assay precision. The most suitable means of expressing repeatability of an assay should be established following e.g., biostatistical evaluations. It is often expressed as % CV of a series of measurements, but may depend on the specific *in vitro* method and analytical methods being used.

8.4. Proficiency chemicals

For complex test methods, a number of proficiency chemicals are defined post-validation on the basis of the applicability domain and the dynamic range (i.e., spread of responses in the dataset) of the test method. These proficiency chemicals can be used by laboratories to demonstrate proficiency prior to the routine use of a validated test method or a test method falling within an adopted OECD test guideline. They usually represent either a subset of the reference chemicals included in the Performance Standards relating to the OECD TG, or chemicals used in the validation studies of the test method falling within the OECD TG.

Criteria used to select the proficiency chemicals for the *in vitro* method typically include chemicals that: i) represent the range of responses to be predicted, ii) have high quality reference data available; iii) cover the method's dynamic range of responses; iv) were correctly predicted by the test method during its validation study; v) cover a wide and representative range of relevant physical states, chemical classes, organic functional

groups and structures falling within the applicability domain of the in vitro method; vi) are commercially available and vii) are not associated with prohibitive acquisition and/or disposal costs.

It is also useful to include chemicals suspected of potentially interfering with the specific *in vitro* method format to better understand potential interference from unknown test items (examples include cytotoxic and cytostatic agents, fluorescent compounds and luciferase inhibitors and MTT interfering chemicals).

The number of proficiency chemicals will depend on the *in vitro* method type and purpose and should be chosen in such a way that a new laboratory can be confident that their results will be acceptable and robust. Since this greatly depends on the properties of the method, some methods may require 5 proficiency chemicals while for others up to 20 compounds should be tested.

In vitro method users should test the proficiency chemicals prior to routine testing for regulatory purposes and to formally comply with the OECD TG. The number of runs needed to correctly predict the proficiency chemicals is not limited. In this way, laboratories can demonstrate their proficiency in the *in vitro* method. Proficiency chemicals can also be used for training purposes, e.g., study personnel can demonstrate their ability to perform the method within the laboratory.

8.5. Data-intensive *in vitro* methods

The 21st century brought a paradigm shift in toxicity testing of chemical substances, relying more on higher throughput and/or high-content screening *in vitro* methods (National Research Council, 2007[29]). These allow the processing of hundreds or thousands of compounds simultaneously enabling the identification of mechanisms of action, and ultimately facilitating the development of predictive models for adverse health effects in humans. Furthermore, image analysis and omics-based *in vitro* method read-outs are getting more popular for *in vitro* method developers due to the data rich information obtained with such methods. The documentation and validation requirements for "data-intensive" approaches do not materially differ from those outlined in Section 8.3, but there may be additional specific aspects to address (e.g., the validity of the image-analysis approach used). It is recommended that the performance of high-throughput methods should be compared to "gold standard(s)", if available, or well-established methods (e.g., qPCR validation of key microarray/RNA-Seq findings). Further standardisation work will be required to achieve transferability and reproducibility of these *in vitro* methods.

The utility of "big data" for regulatory safety assessment has been addressed, e.g., omics (ECETOC, 2013[30]) or high throughput screening (Judson et al., 2013[31]). These data may be used in various contexts, such as supporting evidence for read-across, defining categories or to allow the design of Integrated Testing Strategies (ITS). Still, most applications have focused on screening and prioritisation as in the US EPA ToxCast program (Judson et al., 2010[32]; Judson et al., 2013[31]).

Although some technologies have been extensively used for decades (e.g., microarrays), debate is still ongoing about the interpretability and comparability of data generated from different sites and/or platforms. For many omics technologies consensus is still to be achieved concerning best practices in many critical aspects such as the experimental design and protocols for sample preparation and handling, data processing, statistical analysis and interpretation, and quality control (Bouhifd et al., 2015[33]). For some

technologies such as transcriptomics first respective frameworks have been proposed (Bridges et al., 2017[34]; Gant et al., 2017[35]; Kauffmann et al., 2017[36]).

The maintenance of high standards is essential for ensuring the reproducibility, reliability, acceptance, and proper application of the results generated. A certain level of standardisation is also needed since "big data" are generated using diverse technological platforms and various biochemical, analytical and computational methods, producing different data types and formats. Also to be addressed is the issue of "black box" validation for complex data processing routines.

Notes

1. See: https://www.ncbi.nlm.nih.gov/books/NBK53196/
2. See: https://www.ncbi.nlm.nih.gov/books/NBK83783/
3. See: http://labstats.net/articles/randomise_spatial.html
4. Blank: A sample of a biological matrix to which no analytes have been added that is used to assess the specificity of the bioanalytical method (FDA). An untreated control could be considered as a blank.
5. The reference item is often also used as a positive control
6. See: https://www.ncbi.nlm.nih.gov/books/NBK83783/#htsvalidation.Plate_Uniformity_and_Signa

References

Aschner, M. et al. (2016), "Upholding science in health, safety and environmental risk assessments and regulations", *Toxicology*, Vol. 371, pp. 12-16, http://dx.doi.org/10.1016/j.tox.2016.09.005. [2]

Bouhifd, M. et al. (2015), "Quality assurance of metabolomics", *ALTEX*, Vol. 32/4, pp. 319-326, http://dx.doi.org/10.14573/altex.1509161. [33]

Box, G., J. Hunter and W. Hunter (2005), *Introduction to fractional factorial experimentation*. [13]

Bridges, J. et al. (2017), "Framework for the quantitative weight-of-evidence analysis of 'omics data for regulatory purposes", *Regulatory Toxicology and Pharmacology*, Vol. 91, pp. S46-S60, http://dx.doi.org/10.1016/j.yrtph.2017.10.010. [34]

Burd, E. (2010), "Validation of Laboratory-Developed Molecular Assays for Infectious Diseases", *Clinical Microbiology Reviews*, Vol. 23/3, pp. 550-576, http://dx.doi.org/10.1128/cmr.00074-09. [27]

Coecke, S. et al. (2014), "Considerations in the Development of In Vitro Toxicity Testing Methods Intended for Regulatory Use", in *Methods in Pharmacology and Toxicology, In Vitro Toxicology Systems*, Springer New York, New York, NY, http://dx.doi.org/10.1007/978-1-4939-0521-8_25. [16]

Crawley, M. (2015), *Chapter 1. Fundamentals*, Wiley. [11]

ECETOC (2013), *Omics and Risk Assessment Science, Workshop Report No. 25*, European Centre for Ecotoxicology and Toxicology of Chemicals. [30]

EMA (2011), *Guideline on bioanalytical method validation*, European Medicines Agency, London. [5]

FDA (2001), *Guidance for Industry Bioanalytical Method Validation*, U.S. Department of Health and Human Services Food and Drug Administration, United-States. [6]

Gant, T. et al. (2017), "A generic Transcriptomics Reporting Framework (TRF) for 'omics data processing and analysis", *Regulatory Toxicology and Pharmacology*, Vol. 91, pp. S36-S45, http://dx.doi.org/10.1016/j.yrtph.2017.11.001. [35]

Groten, J. et al. (1997), "Subacute Toxicity of a Mixture of Nine Chemicals in Rats: Detecting Interactive Effects with a Fractionated Two-Level Factorial Design", *Fundamental and Applied Toxicology*, Vol. 36/1, pp. 15-29, http://dx.doi.org/10.1006/faat.1996.2281. [14]

Harbell, J. and H. Raabe (2014), "In Vitro Methods for the Prediction of Ocular and Dermal Toxicity", in *Handbook of Toxicology, Third Edition*, CRC Press, http://dx.doi.org/10.1201/b16632-6. [10]

Hayashi, M. et al. (2011), "Compilation and use of genetic toxicity historical control data", *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, Vol. 723/2, pp. 87-90, http://dx.doi.org/10.1016/j.mrgentox.2010.09.007. [9]

ICH (2005), *Validation of Analytical Procedures : Text and Methodology*. [7]

Iversen, P. et al. (2004), *Assay Guidance Manual*, Eli Lilly & Company and the National Center for Advancing Translational Sciences. [15]

Judson, R. et al. (2010), "In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project", *Environmental Health Perspectives*, Vol. 118/4, pp. 485-492, http://dx.doi.org/10.1289/ehp.0901392. [32]

Judson, R. et al. (2013), *Perspectives on Validation of High-Throughput Assays Supporting 21st Century Toxicity Testing*, ALTEX, https://doi.org/10.14573/altex.2013.1.051. [31]

Kauffmann, H. et al. (2017), "Framework for the quality assurance of 'omics technologies considering GLP requirements", *Regulatory Toxicology and Pharmacology*, Vol. 91, pp. S27-S35, http://dx.doi.org/10.1016/j.yrtph.2017.10.007. [36]

Kollipara, S. et al. (2011), "International Guidelines for Bioanalytical Method Validation: A Comparison and Discussion on Current Scenario", *Chromatographia*, Vol. 73/3-4, pp. 201-217, http://dx.doi.org/10.1007/s10337-010-1869-2. [8]

Krzywinski, M. and N. Altman (2013), "Power and sample size", *Nature Methods*, Vol. 10/12, pp. 1139-1140, http://dx.doi.org/10.1038/nmeth.2738. [12]

Motulsky, H.J.; Christopoulos, A. (2004), *Fitting Models to Biological data using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting*, Oxford University Press, Oxford (United Kingdom). [20]

Motulsky, H. and R. Brown (2006), , *BMC Bioinformatics*, Vol. 7/1, p. 123, http://dx.doi.org/10.1186/1471-2105-7-123. [22]

NATA (2013), *Guidelines for the validation and verification of quantitative and qualitative test methods*, National Association of Testing Authorities, Australia. [4]

National Research Council (2007), *Toxicity Testing in the 21st Century: A Vision and a Strategy*. [29]

NIH (2001), *Guidance Document on Using In Vitro Data to Estimate In Vivo Starting Doses for Acute Toxicity,*, National Institute of Health, United States. [17]

OECD (2017), *Guidance Document for Describing Non-Guideline In Vitro Test Methods*, OECD Series on Testing and Assessment, No. 211, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264274730-en. [18]

OECD (2016), *Application of Good Laboratory Practice Principles to Computerised Systems*, OECD Series on Principles on Good Laboratory Practice and Compliance Monitoring, No. 17, OECD Publishing Paris. [19]

OECD (2016), *Test No. 455: Performance-Based Test Guideline for Stably Transfected Transactivation In Vitro Assays to Detect Estrogen Receptor Agonists and Antagonists*, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264265295-en. [21]

OECD (2015), *Integrated Summary Report: Validation of Two Binding Assays Using Human Recombinant Estrogen Receptor Alpha (hrERa)*, OECD Publishing, Paris. [24]

OECD (2005), *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment*, OECD Publishing, Paris.. [1]

Pincus, R. (1995), "Barnett, V., and Lewis T.: Outliers in Statistical Data. 3rd edition. J. Wiley & Sons 1994, XVII. 582 pp., £49.95", *Biometrical Journal*, Vol. 37/2, pp. 256-256, http://dx.doi.org/10.1002/bimj.4710370219. [23]

Schmidt, B. et al. (2016), "In vitro acute and developmental neurotoxicity screening: an overview of cellular platforms and high-throughput technical possibilities", *Archives of Toxicology*, Vol. 91/1, pp. 1-33, http://dx.doi.org/10.1007/s00204-016-1805-9. [3]

Trullols, E., I. Ruisánchez and F. Rius (2004), "Validation of qualitative analytical methods", *TrAC Trends in Analytical Chemistry*, Vol. 23/2, pp. 137-145, http://dx.doi.org/10.1016/s0165-9936(04)00201-8. [25]

Trullols, E. et al. (2005), "Validation of qualitative methods of analysis that use control samples", *TrAC Trends in Analytical Chemistry*, Vol. 24/6, pp. 516-524, http://dx.doi.org/10.1016/j.trac.2005.04.001. [26]

Viswanathan, C. et al. (2007), "Quantitative Bioanalytical Methods Validation and Implementation: Best Practices for Chromatographic and Ligand Binding Assays", *Pharmaceutical Research*, Vol. 24/10, pp. 1962-1973, http://dx.doi.org/10.1007/s11095-007-9291-7. [28]