

The Hidden Error Budget: Measurement-System Failure, False Precision and Experimental Design in Laboratory Science

Juan Seguroloa

3Dresyns by Resyner Technologies S.L.

© 2026 Juan Seguroloa. All Rights Reserved.

Correspondence: info@3Dresyns.com · www.3Dresyns.com

HEB-001 · Independent Technical Article · Version v1.0 · 20 September 2026

Abstract

Modern laboratory instruments can display many numerical digits while the complete measurement system remains insufficiently controlled. Instrument calibration alone does not establish the validity of a measurement when uncertainty is dominated by sample preparation, reference instability, positioning, environmental coupling, temporal drift, operator-dependent procedures, inadequate replication, or numerical over-reporting. This methodological review develops the concept of a **hidden error budget** through documented and retrospective case studies drawn principally from photochemical and photopolymerisation research carried out in the late 1990s, together with observations from other laboratory settings.

The cases include direct UV/visible measurement of very low absorbance, relative phosphorescence quantum-yield measurements at 77 K, frost formation on cryogenic optical cells, time-dependent reference drift, which becomes bias when reference and sample are not measured at the same time, thin-film interference introduced by polyethylene films in double-beam real-time infrared spectroscopy, non-uniform coating application in photoyellowing measurements, vibration affecting Vickers hardness measurements, and thermal-environment coupling in flow microcalorimetry. A contemporary photopolymer example is also used to distinguish degree of conversion from residual free monomer, extractables/leachables and biological exposure, illustrating that a precisely measured proxy is not automatically the intended measurand. The review distinguishes explicitly between observations documented in contemporaneous sources, retrospective first-person observations by the author, and present-day metrological interpretation.

A second part develops constructive practice: replication, means and standard deviations, significant figures matched to uncertainty, estimation of pure error, factorial design, response-surface methods, and mixture design. The author's 1999 photoinitiator-mixture work provides a documented example of moving beyond one-factor-at-a-time experimentation toward designed experiments capable of estimating interactions and balancing multiple responses. The central proposition is that the measurement system is not the instrument alone but the coupled system formed by the instrument, specimen, reference, environment, operator, protocol and statistical treatment. Sophisticated instrumentation cannot compensate for an uncontrolled weakest component, and statistical sophistication cannot create information that the experiment never generated.

Keywords: measurement uncertainty; repeatability; reproducibility; significant figures; false precision; measurand definition; degree of conversion; extractables; protocol validation; environmental coupling; photochemistry; RTIR; flow microcalorimetry; factorial design; mixture design; Design of Experiments

1. Purpose and evidence boundary

This article is a methodological review built around a set of historical laboratory cases. Its purpose is not to adjudicate misconduct, invalidate entire bodies of published work, or assign blame to named individuals or institutions. Its purpose is to examine how apparently elementary experimental weaknesses can survive within technically sophisticated laboratories and how those weaknesses can dominate the uncertainty of the final result.

Three evidence classes are used throughout:

1. **Contemporaneously documented evidence.** A method, numerical limitation, or corrective procedure appears in the author's 1999 doctoral thesis or a resulting publication.
2. **Retrospective author observation.** The author directly recalls observing a condition or practice, but a contemporaneous record has not yet been recovered for this review.
3. **Current metrological interpretation.** A present-day interpretation of why the observed condition matters for measurement validity. This interpretation should be assessed independently against modern metrology sources before publication.

These classes must not be collapsed. A retrospective observation is not treated as documentary proof of another researcher's complete methodology, and a methodological weakness does not by itself prove that every historical conclusion obtained with that method was wrong.

The article advances one central proposition:

A sophisticated instrument does not turn an uncontrolled procedure into a reliable measurement.

A second proposition follows:

The measurement system is not the instrument. It is the instrument, specimen, reference, environment, operator, protocol and statistical treatment acting together.

2. A number is not yet a measurement

Modern instruments routinely provide outputs with several decimal places. This creates a strong visual impression of precision. The number of displayed digits, however, is a property of the data-acquisition and calculation chain; it is not automatically a property of the physical knowledge generated by the experiment.

A useful conceptual separation is:

display resolution $\neq$ repeatability $\neq$ reproducibility $\neq$ trueness $\neq$ measurement uncertainty.

A result can therefore have high displayed numerical resolution yet remain metrologically weak. If repositioning a cell changes an absorbance by an amount comparable with the absorbance being measured, five digits on the display do not make those five digits informative. If a calorimetric response varies when a nearby heating system switches on or off, the existence of an instrument calibration certificate does not establish that the complete measurement environment is stable. If a coating has non-uniform thickness, a highly repeatable colorimeter can still produce values that are not attributable solely to formulation chemistry.

The practical question is not simply whether an instrument produces a number. It is whether the effect of interest is distinguishable from the combined experimental variability of the complete method.

2.1 The wrong measurand can be measured precisely: conversion, residual monomer and extractables

A further source of experimental error arises when a precisely measured quantity is used as a proxy for a different quantity without demonstrating the relationship between them. Vat-photopolymerised materials provide a useful contemporary example.

A 2025 study of vat-photopolymerised Formlabs Clear V4 [11] measured degree of conversion by ATR mid-infrared spectroscopy from the change in the aliphatic C=C signal relative to an internal reference. Because ATR interrogates a shallow near-surface region rather than the complete bulk volume, the spectroscopic conversion itself remains a sampling-dependent proxy for the conversion state of the whole printed article. The study used ten polymerised specimens per conversion group and reported that post-cure time and temperature significantly affected conversion. Across the reported groups, the lowest mean degree of conversion was 80.99% $\pm$ 0.61 and the highest was 97.20% $\pm$ 0.13 [11]. The tested resin was described as UDMA-based, providing a directly relevant example of a multifunctional network in which residual unsaturation cannot be equated numerically with free residual monomer. These data demonstrate that the chemical state of a printed photopolymer remains strongly process-dependent even after the nominal printing step.

The metrological point is not that the unconverted percentage can automatically be read as the same mass percentage of free residual monomer. Degree of conversion is a bond-based measurand: it reports the fraction of monitored reactive groups that have reacted. Residual mobile monomer, extractables and leachables are composition- and transport-based measurands. They are related, but they are not interchangeable.

Molecular functionality determines part of that relationship. In an idealised system composed only of monofunctional monomers, failure of a molecule's sole reactive group to react leaves that molecule unincorporated, so incomplete conversion can correspond more directly to a population of potentially mobile residual monomer. In a multifunctional network, by contrast, a molecule may react through one functionality and become incorporated into the network while retaining another unreacted double bond. Residual C=C therefore does not imply that the same fraction of material remains as free monomer.

Conversely, where a formulation contains substantial low-molar-mass monofunctional reactive diluent, incomplete conversion increases the plausible population of molecules that can remain unbound and potentially extractable. The realised extractable fraction will still depend on molecular size, polarity, network affinity, diffusion, washing, post-cure history, geometry and extraction conditions.

The distinction can be written as:

degree of conversion $\neq$ free residual monomer fraction $\neq$ extractables/leachables $\neq$ biological exposure.

These quantities may be causally connected, but one cannot be substituted numerically for another without an experimentally supported measurement model.

A precise proxy is not the measurand unless the relationship between proxy and measurand has been demonstrated.

This example extends the hidden-error concept beyond instrument performance. Experimental validity can fail not only because a quantity is measured badly, but because the wrong quantity is measured well and then interpreted as if it answered a different question.

3. Case study I: UV/visible absorbance near the practical measurement floor

3.1 Documented method

During the doctoral study, relative phosphorescence quantum-yield measurements required low optical densities at the excitation wavelength. The thesis states that similar low absorbance values, around 0.05 at 290 nm in methylcyclohexane, were used for reference and sample. It also states that direct measurement near this absorbance could carry a relative error of approximately 20%, "considering the detector sensitivity around 0.01" [1].

Rather than relying on the direct low-absorbance reading, absorbance values were obtained from more concentrated solutions of known absorbance, around 1, followed by dilution by a known factor. The thesis explicitly states that this gave higher accuracy than direct measurement of the low absorbance [1].

3.2 Metrological interpretation

The important point is not that low absorbance can never be measured directly. The point is that the practical measurement should be performed where the ratio between useful signal and measurement error is adequate for the intended conclusion.

Where Beer-Lambert linearity is demonstrated over the relevant concentration range, moving the primary measurement into a more robust absorbance region and using a controlled dilution can reduce relative photometric error. That benefit must be balanced against volumetric uncertainty introduced by the dilution itself.

Whenever the underlying physical law permits it, measure in a region where the instrument performs reliably rather than close to the practical noise floor.

In modern metrological language, the thesis phrase "detector sensitivity around 0.01" should not be treated automatically as a complete uncertainty specification; it identifies the scale of the instrumental limitation that motivated the dilution strategy. This case also illustrates why the number of significant figures must follow the uncertainty of the procedure rather than the numerical precision of the spectrometer display.

4. Case study II: Relative quantum yield when the reference is measured at the wrong time

4.1 Documented basis

The thesis describes relative phosphorescence quantum-yield determination using benzophenone as reference and states an assumed reference value of 0.74 in absolute ethanol at 77 K [1]. The experimental photoinitiator measurements were carried out in methylcyclohexane, so transfer of a reference value between solvent conditions, refractive-index correction and low-temperature glass behaviour are themselves quantities that require explicit control in a modern implementation. The relative method depends on absorbed fraction, excitation intensity, refractive index, integrated emission area, constant path length, reproducible cell positioning, equivalent reflection losses, and other optical assumptions [1,7].

The reported experimental programme used a Perkin-Elmer LS-5 luminescence spectrometer and low-temperature measurements at 77 K [1].

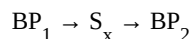
4.2 Retrospective author observation: temporal invalidation of the reference

The central problem observed by the author was not simply that the benzophenone signal had finite repeatability. The author recalls that the Perkin-Elmer luminescence response changed progressively during a measurement session. A benzophenone reference measured substantially earlier or later than an unknown sample could therefore represent a different instrumental state from that in which the sample was measured.

Because the phosphorescence quantum yield was determined relative to benzophenone, temporal separation between reference and sample propagated directly into the calculated quantum yield. A reference could be chemically correct, measured correctly, and still be metrologically inappropriate for a sample measured hours later if the instrument response had drifted in the interval.

The author recalls that the benzophenone signal tended to increase as the experimental session progressed. At the time, this was attributed to progressive thermal equilibration of the measurement system during repeated liquid-nitrogen operation. The observed temporal change is a retrospective author observation; the proposed thermal cause remains a retrospective interpretation because the original time-series data are not available here.

The practical correction adopted by the author was to bracket each unknown measurement with benzophenone measurements made immediately before and after the sample:



The purpose of this sequence is not merely to average two references. It is to estimate the reference response corresponding to the time at which the sample was measured. If the drift is approximately linear over the short bracketing interval, the local reference response at sample time (t_x) can be estimated by temporal interpolation:

$$R(t_x) = R_1 + (t_x - t_1)/(t_2 - t_1)(R_2 - R_1).$$

If the sample is measured approximately midway between the two references, this reduces to:

$$R(t_x) \approx (R_1 + R_2)/(2).$$

This correction is valid only to the extent that the drift is sufficiently smooth over the bracketing interval and that the relevant drift acts on the response term being normalised. A ratio correction is appropriate for predominantly multiplicative sensitivity drift; an additive baseline or offset drift requires a different correction model.

4.3 General principle: a reference has a time validity

A correct reference measured at the wrong time is an incorrect reference for a drifting measurement system.

The validity of a reference measurement is therefore conditional not only on reference identity, preparation and traceability, but also on temporal proximity to the unknown whenever instrument response is not demonstrably stationary.

Reference measurements must be temporally matched to the sample whenever the measurement system is not demonstrably stable over the sequence.

A single benzophenone measurement made at the beginning of a long session cannot correct a sample measured hours later if the luminescence response has changed in the intervening period. In that situation, the relevant ratio no longer compares sample and reference under the same instrumental state.

4.4 Interaction with low-absorbance uncertainty and signal variability

This temporal-reference error was compounded by other uncertainty contributions. The thesis states that similar low absorbance values around 0.05 at 290 nm were used for reference and sample, and that direct measurement near this absorbance could carry a relative error of approximately 20% under the stated detector-sensitivity assumption [1]. To improve accuracy, the thesis obtained these low absorbances from more concentrated solutions of known absorbance, around 1, by controlled dilution [1].

The relative quantum-yield determination also depends on integrated emission areas and other optical factors. In simplified form,

$$\Phi_x = \Phi_R ((Area_x)/(Area_R)) ((B_R)/(B_x)) ((I_R)/(I_x)) ((n_x^2)/(n_R^2)),$$

where the exact terms and conventions depend on the implementation. The thesis explicitly identifies absorbed fraction, excitation intensity, refractive index, integrated emission area and reproducible geometry among the dependencies of the relative determination [1].

In this expression, $(Area_x)$ and $(Area_R)$ are the integrated areas under the corrected emission spectra of sample and reference, (B_x) and (B_R) are the fractions of incident light absorbed by sample and reference, and (I_x) and (I_R) are the excitation-light intensities for sample and reference at the excitation wavelength. These definitions are stated in the thesis method [1, Chapter V, printed p. 50; PDF p. 61].

The author additionally recalls that the integrated luminescence response could vary substantially during the experimental sequence. When the reference was not measured locally in time, this variation was not a small secondary uncertainty: it could invalidate the normalising denominator used to calculate the sample quantum yield. The resulting error is therefore better described as **time-dependent reference bias** than as random noise alone.

4.5 When uncertainty approaches the physical range of the measurand

Quantum yield is a bounded quantity: for the conventional definition relevant here, physically admissible values lie between 0 and 1. A result inside that interval is not automatically informative merely because the calculation returns a plausible number.

If uncertainty from low absorbance, integrated-area variability, reference drift, positioning and cryogenic optical instability becomes comparable with the magnitude of the estimated quantum yield, numerical discrimination can collapse. An illustrative result such as

$$\Phi = 0.5 \pm 0.5$$

would be compatible, after respecting the physical bounds, with almost the entire interval from 0 to 1. Such a result would not constitute a well-resolved estimate of 0.5. The numerical example is illustrative and is not assigned retrospectively to any specific published value because the complete historical uncertainty budget cannot now be reconstructed from the surviving material.

The key metrological point is that the failure need not arise from random detector noise. It can result from combining a low-absorbance measurement with a reference obtained in a different instrumental state. In a relative method, a drifting denominator can dominate the inferred value even when each individual reading appears plausible.

When the propagated uncertainty spans most of the physically admissible range of a measurand, the experiment has not resolved that measurand quantitatively.

For historical datasets, this does not prove that every qualitative ordering was wrong. It does mean that claims depending on numerical separation smaller than the unresolved uncertainty require re-evaluation against surviving raw data, replicate structure, temporal reference controls and method qualification.

5. Case study III: Frost became an uncontrolled optical component

5.1 Retrospective author observation

Low-temperature phosphorescence measurements required cooling the sample cell with liquid nitrogen. The author recalls that ambient humidity produced condensation and frost on the cold cell. The amount and morphology of the frost varied with room conditions and time.

An informal operational response was to rotate the cell so that contact with the holder mechanically removed part of the frost. This procedure did not define the initial frost amount, the fraction removed, or the residual optical state after manipulation.

5.2 Metrological interpretation

A variable frost layer can change transmission, scattering, reflection and effective optical geometry. It therefore becomes part of the optical system even though it is not part of the intended experiment.

The robust design response is prevention rather than manual correction: isolate the cryogenic optical zone from ambient moisture or use an appropriate dry-gas purge so that condensation is suppressed.

If an environmental contaminant enters the optical path, it becomes part of the measurement whether or not the protocol acknowledges it.

This case also illustrates a broader hierarchy developed later in this article: remove a source of variability physically when feasible before attempting to correct it numerically.

6. Case study IV: The RTIR reference beam generated its own baseline structure

6.1 Documented original configuration

Real-time infrared spectroscopy was used to follow photopolymerisation through the decrease in the acrylate band at approximately 812 cm⁻¹. The thesis describes a reactive sample positioned between sodium chloride plates covered with two pieces of low-density polyethylene film, using a Teflon separator to give a film thickness of 50 micrometres. Two polyethylene films were placed in the reference beam to subtract polyethylene absorption [1].

The thesis also states that the baseline for the 812 cm⁻¹ band was calculated by drawing a line from the shoulders of the peak [1]. Thus, instability in those shoulders directly affects the calculated peak area or height and therefore the derived conversion.

6.2 Retrospective author observation

The author recalls that the shoulders surrounding the 812 cm⁻¹ band oscillated unpredictably when the polyethylene films were present in the reference beam. A colleague at the time explained the phenomenon in terms of a thickness-dependent optical interference condition. In present terminology, the most plausible interpretation is thin-film or Fabry-Perot interference fringes associated with parallel films of finite optical thickness. This physical interpretation is retrospective and is not required to establish the empirical observation that removal of the reference-beam films reduced the structured baseline variation.

When the polyethylene films were removed from the reference beam, the oscillation of the shoulders was greatly reduced, allowing a more stable and reproducible baseline. That removal is methodologically defensible only where polyethylene absorption around the analytical region does not itself create a material subtraction error; this must be checked for the exact film grade and spectrum. A later measurement geometry simplified the sample arrangement further by applying material directly onto sodium chloride plates.

6.3 General principle

The original reference arrangement had a reasonable conceptual purpose: subtract the spectral contribution of polyethylene. The failure was that the reference component was not optically neutral and could create structured spectral variation of its own.

A reference intended to remove an artefact can itself become a source of artefact.

This motivates a broader rule:

Do not correct mathematically what can first be eliminated physically.

Numerical subtraction is reliable only if the quantity being subtracted is itself stable and reproducible.

7. Case study V: Non-uniform specimens can dominate photoyellowing measurements

7.1 Retrospective observation of an early preparation problem

The author recalls an early photoyellowing experiment in which a viscous resin was applied using a blade. The resulting coatings showed visible blade lines and non-uniform film thickness. Initial colour results were therefore affected by a specimen-preparation variable that was not part of the intended photoinitiator comparison.

7.2 Documented final method

The final doctoral method was more controlled. Yellowing of commercial photoinitiators was measured with a Datacolour colorimeter using the CMC system. Coatings containing 2% photoinitiator were applied at a nominal 75 micrometres with a Brown K wired bar No. 7 over transparent glass, cured under defined conditions, measured after 24 h, and replicated. Results were reported as mean values, with the thesis describing the relative error as below 5% for the tested samples [1]. Because replication estimates precision rather than trueness by itself, that historical wording should not be read as a modern demonstration of measurement accuracy. The corresponding photoyellowing study was published in *Polymer Degradation and Stability* [2].

7.3 Metrological interpretation

For an optically measured transparent or translucent coating, thickness and surface topography can alter optical path length, absorption, reflection and local cure state. A precise colorimeter cannot remove variability that was already built into the specimen.

If the specimen is not reproducible, the instrument cannot make the measurement reproducible.

Sample preparation is therefore not a preliminary convenience. It is part of the measurement system.

8. Case study VI: Environmental thermal coupling in flow microcalorimetry

8.1 Historical boundary

During the same institutional period, the author observed a flow microcalorimetry installation used in a separate research programme. This case is included only as a first-person observation about measurement-system design. It is not linked here to a named researcher, company, thesis or publication, and it is not used to infer the validity of any historical dataset.

8.2 Retrospective author observation

The author recalls that the flow microcalorimetry equipment was positioned close to a hot-water heating source. The local thermal environment changed when the building heating system switched between operating and non-operating states, and the author observed corresponding instability in the calorimetric response. The suspected mechanism was thermal coupling between the room environment and a measurement system intended to resolve small heat effects.

The author proposed separating or otherwise controlling the thermal disturbance. No contemporaneous installation log, raw time series or full uncertainty budget has yet been recovered for this review. The case therefore remains a retrospective observation, not a documentary finding about another researcher's final method.

8.3 Metrological interpretation

Flow microcalorimetry is used precisely because thermal changes associated with adsorption or desorption carry the information of interest. A changing thermal environment can therefore couple directly to the measurement channel.

When both the measurand and the environmental disturbance are expressed through heat, thermal isolation and environmental stability are part of the measurement method. Replication and randomised run order can reveal or reduce confounding by time-dependent disturbances, but they do not substitute for controlling a large known thermal perturbation.

A thermal measurement should not be interpreted independently of the thermal environment in which it is made.

9. Case study VII: Vibration and Vickers hardness measurements

9.1 Retrospective author observation

During work in an industrial research laboratory in 1995-1996, the author recalls that a Vickers hardness instrument was installed in a small room together with other equipment capable of transmitting mechanical vibration. The measurements showed substantial variability. Mechanical vibration was suspected as one contributor, and physical separation of vibration-generating and vibration-sensitive equipment was proposed. Other possible contributors, including specimen preparation, indentation load and dwell conditions, and optical diagonal reading, were not independently excluded in the surviving evidence.

This case is reported as a retrospective observation. Contemporaneous records of the room layout, load range and individual measurement series have not yet been recovered for the present review.

9.2 Metrological interpretation

A hardness measurement may be performed by a calibrated device while still being degraded by its installation environment. The relevant measurement system is therefore better represented as:

instrument + support + floor + nearby equipment + operator + specimen.

A calibrated instrument installed in an unsuitable environment is not necessarily a calibrated measurement system.

The principle is not specific to Vickers hardness. It applies whenever vibration, temperature, humidity, air movement, electrical noise or another environmental variable couples to the measurand or detection system.

10. The laboratory is part of the instrument

In the author's retrospective observations of the cryogenic, Vickers and flow-microcalorimetry cases, the surrounding laboratory entered the measurement chain.

Relevant variables can include:

- vibration and mechanical coupling;
- temperature and temperature gradients;
- humidity and condensation;
- air flow;
- stray light;
- electrical and electromagnetic noise;
- surface contamination;
- warm-up, cool-down and long-term drift;
- sample conditioning and handling history.

The manufacturer's instrument specification usually describes performance under defined conditions. The laboratory measurement, however, is carried out in a specific room, with a specific sample, reference, fixture and operator. Qualification of the measurement therefore remains a responsibility of the experimental method.

A useful conceptual representation is:

$$M = f(I, S, R, E, O, P, A)$$

where (I) is the instrument, (S) the specimen, (R) the reference, (E) the environment, (O) the operator, (P) the protocol, and (A) the analysis and statistical treatment.

The expression is conceptual rather than a prescribed uncertainty equation. Its purpose is to prevent the instrument from being mistaken for the complete measurement system.

11. Significant figures are an experimental claim

One of the most visible forms of weak metrology is false numerical precision.

Consider a calculated phosphorescence quantum yield:

$$\Phi = 0.67123.$$

The calculation may legitimately return five decimal places. That does not mean the experiment resolved the quantum yield to five decimal places.

If an uncertainty evaluation supports an expanded uncertainty of, for example, ($U=0.1$) with coverage factor ($k=2$), then a representation such as

$$\Phi = 0.7 \pm 0.1 \text{ (} U, k=2 \text{)}$$

communicates the experimental information more honestly than (0.67123). The numerical example is illustrative; it is not assigned retrospectively to a specific thesis value without an uncertainty budget.

The author's thesis reported many phosphorescence quantum yields to two decimal places, including values such as 0.68, 0.75, 0.61 and 0.12 for studied compounds; 0.74 also appears as the benzophenone reference value used in the method [1]. In retrospect, where the total practical uncertainty was substantially larger than the last reported digit, even two decimal places could imply more numerical resolution than the complete measurement system justified.

This is not a retrospective claim that the reported trends were invalid. It is a distinction between **computed numerical resolution** and **defensible metrological resolution**. The distinction becomes especially severe for a bounded measurand. If the uncertainty interval occupies a large fraction of the full 0-to-1 quantum-yield range, additional decimal places do not merely add cosmetic precision; they can visually imply discrimination that the experiment did not achieve.

Significant figures should describe the experiment, not the calculator.

A practical reporting convention is to state the type of uncertainty being reported and, where appropriate, the coverage factor or coverage probability. The uncertainty is commonly rounded to one or two significant digits and the reported value to the corresponding decimal position, consistent with GUM guidance [5].

12. From single numbers to experimental evidence

12.1 Mean and standard deviation

A single measurement provides a value but no empirical estimate of the variability of the method under repeated conditions.

For independent replicate measurements (x_i), the sample mean is

$$\bar{x} = (1/n) \sum_{i=1}^n x_i,$$

and the sample standard deviation is

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

The mean and standard deviation answer different questions. The mean estimates a central response under the tested conditions; the standard deviation quantifies observed dispersion among the replicates.

The doctoral RTIR procedure explicitly states that replication was performed to estimate pure errors and prevent gross errors. Results were expressed as mean values and standard deviations, and reactivity plots were averaged [1]. A related published kinetic study from the same research programme provides additional publication-level context [3]. In the colour measurements, the thesis states that replicated samples were measured to obtain an estimate of "accuracy" and that results were expressed as mean values [1]. In modern metrological terminology, replication by itself estimates a precision component; it does not establish trueness or accuracy without an appropriate reference or bias evaluation [6].

The same thesis notes that at least five replicate pendulum-hardness measurements are desirable because local film imperfections can affect individual readings [1].

Without replication, experimental scatter is unknown; without experimental scatter, numerical precision cannot be justified.

12.2 Repeated readings are not automatically independent replicates

Replication must be defined at the level relevant to the uncertainty being estimated.

Five repeated instrument readings without removing the specimen may estimate short-term readout repeatability but not repositioning error. Five points on one coating may characterise local surface heterogeneity but not batch-to-batch coating preparation. Five separately prepared specimens can capture a broader component of experimental variation.

Accordingly, publications should state what was replicated: readings, positions, specimens, preparations, experimental days, operators, or combinations of these.

12.3 Mean without spread is incomplete

Suppose two conditions each have ($n=5$) independent replicates and give means of 10.2 and 10.8. The numerical difference is 0.6.

If the corresponding standard deviations are 0.1 and 0.1, the standard error of the difference is small relative to 0.6 under a simple independent-samples model. If the standard deviations are 1.8 and 2.1, the standard error of the difference is much larger. Formal inference depends on sample size, variance structure and the design, not on the standard deviations alone.

The numbers 10.2 and 10.8 have not changed. The interpretation has.

The magnitude of a difference must be judged relative to the variability of the method that produced it.

12.4 Pure error and lack of fit

The experimental-design chapter of the thesis states that replicated observations allow an estimate of pure error and therefore enable a lack-of-fit test. It also identifies Root MSE as an estimator of experimental error [1]. More precisely, residual Root MSE can represent the common error standard deviation only when the fitted model is adequate; otherwise the residual term can contain both pure error and lack of fit.

This is more than a statistical detail. Without an empirical error estimate, a poor model and a noisy measurement system can be difficult to distinguish. Replication provides the baseline against which model inadequacy can be judged.

13. One-factor-at-a-time versus factorial Design of Experiments

13.1 Why OFAT is attractive

One-factor-at-a-time (OFAT) experimentation is intuitive: select a baseline condition, change factor (A), identify an apparently favourable level, hold it fixed, then change factor (B), and continue sequentially. Designed experiments can extract interaction information that such OFAT paths do not estimate efficiently [8].

For simple monotonic systems with negligible interactions, OFAT can be useful for preliminary exploration. Its limitation is structural: it does not efficiently map interactions between factors.

13.2 Interaction is information

For two factors, a simple model can be written as

$$Y = \beta_0 + \beta_{AA} + \beta_{BB} + \beta_{AB} + \epsilon$$

The (AB) term represents interaction: the effect of changing (A) depends on the level of (B).

An OFAT path can miss this dependence because it explores only selected lines through the experimental space. A factorial design deliberately combines factor levels so that main effects and interactions can be estimated within the design assumptions.

For three two-level factors (A), (B), and (C), a full factorial design contains

$$2^3 = 8$$

factor combinations. Subject to an adequate error and replication structure, those runs can provide information on the three main effects, the three two-factor interactions and the three-factor interaction.

OFAT asks what one factor does while the others are frozen. Factorial design asks how the system responds when factors operate together.

13.3 A simple conceptual example

Imagine photopolymerisation response depending on temperature and photoinitiator concentration. An OFAT sequence might first examine temperature at 2% photoinitiator, choose 30 °C, and then examine concentration only at 30 °C. If the effect of photoinitiator concentration changes strongly with temperature, the optimum may lie at a combination that the OFAT path never tests.

The missed phenomenon is not experimental noise. It is an interaction that the design was not structured to resolve.

An interaction is not noise; it is information that an OFAT design may be structurally incapable of seeing.

14. Factorial design, response surfaces and mixture design are different tools

DOE is not a single experimental layout. The design must match the structure of the factors.

14.1 Factorial designs

Factorial designs are appropriate when factors can be varied independently within defined ranges. Examples might include exposure time, temperature, irradiance, additive concentration, or post-cure duration, provided the chosen factors are physically independent in the specified design space.

14.2 Response-surface methods

When curvature is expected and the objective includes optimisation within a continuous region, a two-level factorial alone cannot estimate pure quadratic terms. Centre points can reveal curvature, while second-order response-surface designs such as central composite or Box-Behnken designs can estimate quadratic behaviour [9].

14.3 Mixture designs

Formulation proportions require a different geometry because component fractions are constrained:

$$\sum_i x_i = 1$$

Increasing one component necessarily decreases at least one other. The factors are therefore not independent in the conventional factorial sense.

The author's 1999 work on eutectic photoinitiator blends explicitly used Design of Experiments and mixture designs to optimise reactivity and bulk stability of blends containing benzophenone, benzil dimethyl ketal, isopropyl thioxanthone and ethyldimethylaminobenzoate [1,4]. The thesis documents regression modelling, ANOVA-based evaluation, residual analysis and parameter standard errors, while the published study reports the mixture-design optimisation and the reactivity-stability trade-off [1,4].

The study showed an important trade-off: relatively high ethyldimethylaminobenzoate (EPD) proportions were associated with enhanced reactivity but reduced stability, while benzophenone contributed strongly to blend stability. Multiple-response optimisation was therefore used to identify compositions balancing more than one performance objective [1,4].

This example illustrates why formulation optimisation is poorly represented by independent one-factor-at-a-time adjustments. The experimental space is constrained, components interact, and the useful answer may be a compromise region rather than an isolated maximum of one response. Mixture-design geometry and analysis are treated specifically in the mixture-experiment literature [10].

15. DOE cannot rescue an invalid measurement method

DOE increases the information extracted from an experiment only if the response being modelled is itself an adequate representation of the intended measurand.

Randomisation and blocking are central protections against nuisance variation. Randomised run order helps prevent time-dependent drift or environmental cycles from becoming aligned with factor settings, while blocking can separate known nuisance sources such as day, batch or operator from the factor effects of interest [9]. These tools reduce confounding, but they do not make the disturbance disappear; large uncontrolled variation still inflates the error variance and reduces power.

If response values are contaminated by drift, variable specimen thickness, thermal coupling, vibration or an unstable reference **and those disturbances are aligned with factor settings**, for example because run order was not randomised, a statistical model can attribute nuisance structure to the experimental factors. Conversely, a constant systematic bias common to every run, an incorrect reference value or a response that does not measure the intended property cannot be repaired by randomisation or blocking.

A robust sequence is therefore:

qualify measurement method → estimate repeatability → define replication → design experiment, including
randomisation/blocking as appropriate → fit and validate model → optimise.

DOE increases information from a qualified measurement process; it does not convert an invalid response into a valid measurand.

Statistics cannot manufacture information that the experiment never generated.

16. Protocol inheritance as a failure mode

Laboratory procedures are often transmitted through demonstration and local practice as well as through formal written methods. Such transmission can be efficient, but a procedure can become socially stable before every relevant assumption has been challenged metrologically.

The cases in this review should not all be described as inherited procedures. Some are documented methods, some are the author's own early practices, and some are retrospective observations of laboratory conditions. They nevertheless illustrate procedures of a kind that can persist when their assumptions are not deliberately tested. Examples include assuming that a reference remains stable throughout a long sequence, manually removing condensation instead of preventing it, adding a reference-beam component without testing whether it creates spectral structure, accepting non-uniform specimen preparation, or installing sensitive equipment without qualification of the local environment.

The problem is not that such choices are irrational. Each can have an operational rationale. The problem is that a plausible rationale is not the same as demonstrated measurement validity.

Reproducibility of a procedure is not the same as validity of a procedure.

A method can be reproduced consistently while preserving the same uncontrolled influence.

17. The hidden error budget

A conventional error budget contains sources of uncertainty that have been recognised and estimated. A hidden error budget consists of sources that have not yet been recognised as part of the measurement.

Conceptually:

$$Y_{\text{measured}} = f(Y_{\text{true}}, E_I, E_S, E_R, E_E, E_O, E_T, E_A),$$

where the error terms denote instrument, specimen, reference, environment, operator, time and analysis contributions.

These contributions are not necessarily independent or random. A frost layer may change both transmission and scattering. A film-thickness variation may alter both optical density and cure state. Temperature may affect both instrument electronics and the specimen. Repositioning may alter alignment and baseline simultaneously.

The most dangerous error is therefore not necessarily the largest known component. It may be the component that the experimenter has not yet identified.

This suggests a practical strategy: deliberately perturb the measurement system before trusting it.

Examples include:

- remove and reinsert the same cell repeatedly;
- repeat references throughout the day;
- measure before and after instrument thermal equilibration;
- turn suspected vibration-producing equipment on and off;
- test different room-temperature or humidity conditions where safe and relevant;
- remove an optical element suspected of producing interference;
- have a second operator prepare specimens independently;
- repeat the entire specimen-preparation process rather than only repeat the instrument reading.

A robust method should remain within predefined acceptance limits under perturbations that are expected in normal use.

18. The complementary-failure principle

Experimental controls are commonly introduced to suppress a known failure mode. Yet every control can introduce a new assumption.

The polyethylene films in the RTIR reference beam illustrate the point. Their purpose was to subtract polyethylene absorption, but the films themselves could generate interference structure that compromised the shoulders used to construct the 812 cm⁻¹ baseline.

The appropriate question is therefore not only:

What failure does this control prevent?

but also:

What new failure can this control introduce?

This review proposes the following complementary-failure principle:

For every experimental control introduced to suppress one failure mode, explicitly test whether the control creates a second failure mode.

The principle applies to references, blanks, fixtures, correction algorithms, environmental controls and even statistical preprocessing.

19. Retrospective implications for historical data

Identification of an uncontrolled variable raises a difficult question: what should be concluded about older data obtained under the same or similar procedure?

The answer cannot be global. A methodological weakness does not justify the statement that all historical data are false.

At least three cases should be distinguished.

19.1 Qualitative conclusion robust

The observed effect is much larger than plausible experimental variability. The exact numerical value may be uncertain, but the qualitative conclusion remains supported.

19.2 Numerical precision questionable

The trend may be real, but the number of significant figures or the magnitude of the reported difference exceeds what the method can support.

19.3 Conclusion potentially vulnerable

The claimed effect is of the same order as the uncontrolled variability. Re-measurement under a qualified protocol may be necessary before the conclusion can be regarded as quantitatively secure.

This framework is intentionally conservative. It separates criticism of a method from allegations concerning intent. Historical reassessment should be based on surviving protocols, raw data, replicate structure and effect size wherever those materials remain available.

20. A practical measurement-system qualification framework

Before a quantitative method is used to compare formulations, materials or treatments, the following questions should be answered.

20.1 Instrument

- Is the instrument calibrated for the intended range?
- What is its practical, not merely nominal, resolution?
- Does the response drift with warm-up, cool-down or time?

20.2 Specimen

- Is specimen preparation reproducible?
- Are thickness, geometry, surface state and conditioning controlled?
- Are independent specimens used where specimen variability matters?

20.3 Reference and blank

- Is the reference more stable than the signal it is intended to correct?
- Does repositioning change the reference response?
- Can the reference itself introduce interference or another artefact?

20.4 Environment

- Can vibration, temperature, humidity, condensation, airflow, light or electrical noise affect the measurement?
- Are acceptance limits defined for relevant environmental variables?

20.5 Operator

- Does the result depend on subjective manipulation?
- Does removal and reinsertion materially change the result?
- Has inter-operator reproducibility been tested where relevant?

20.6 Time

- Does the instrument or specimen drift during the measurement sequence?
- Is calibration or reference bracketing frequent enough to detect that drift?

20.7 Replication

- What constitutes an independent replicate?
- Is the number of replicates sufficient to estimate the variability relevant to the claim?

20.8 Significant figures

- Does the number of reported digits match the observed uncertainty?
- Are uncertainty or dispersion statistics reported beside the mean?

20.9 Experimental design

- Can multiple factors interact?
- Is run order randomised where time-dependent nuisance variation could confound factor effects?
- Should known nuisance variables such as day, batch, instrument or operator be blocked?
- Is OFAT sufficient for the narrow question, or is factorial or response-surface DOE needed?
- Are the variables mixture components subject to a sum constraint, requiring mixture-design methods?

20.10 Validation

- Are model residuals and lack of fit examined?
- Is the predicted optimum checked experimentally rather than accepted solely from the model?
- Can the result be reproduced on a different day, specimen preparation or operator where the claim requires that level of robustness?

21. Discussion

The individual cases described here span spectroscopy, photochemistry, photopolymerisation, colorimetry, calorimetry and mechanical testing. The cases suggest that the nominal instrument need not be the dominant source of uncertainty; none of the historical cases presented here provides a complete quantitative decomposition of all uncertainty components.

Instead, the limiting contribution could arise from cell positioning, condensation, reference drift, thin-film interference, specimen preparation, room temperature, mechanical vibration or insufficient replication.

This produces a recurring asymmetry in experimental work. The more sophisticated the instrument, the easier it is to assume that the difficult part of the experiment has already been solved by the manufacturer. But instrument manufacturers qualify instruments under defined conditions. Researchers must qualify measurements under the conditions in which they actually use them.

The constructive response is not methodological pessimism. It is to design the experiment so that uncertainty becomes visible.

Replication exposes dispersion. Means and standard deviations prevent isolated readings from masquerading as stable quantities. Significant figures can be matched to the experimentally resolved scale. Factorial designs reveal interactions that OFAT can miss. Mixture designs respect the mathematical dependence of formulation proportions. Residual analysis, pure-error estimation and lack-of-fit testing distinguish model structure from experimental scatter.

However, these tools form a sequence rather than substitutes for one another. Replication does not remove systematic bias. DOE does not repair an unstable instrument environment. ANOVA cannot separate an error source that is perfectly confounded with an experimental factor. Randomisation and blocking are therefore part of measurement-system design, not optional statistical decoration. A highly significant model can still describe a systematically biased response.

The strongest experimental architecture therefore starts before the first dataset is collected: identify plausible error sources, perturb the system deliberately, simplify the physical measurement where possible, quantify repeatability, and only then use statistical design to extract factor effects efficiently.

22. Conclusions

The cases examined in this review support a set of general methodological conclusions.

1. A numerical output is not automatically a valid measurement.
2. Calibration of an instrument does not validate the complete measurement system.
3. A reference can increase uncertainty if it is less stable than the signal it is intended to correct.
4. Environmental variables become part of the instrument whenever they couple to the measurement channel.
5. Sample preparation is part of the measurement system.
6. Replication is required to estimate experimental dispersion at the level relevant to the claim.
7. Means should normally be accompanied by an appropriate measure of spread.
8. Significant figures should reflect experimental uncertainty rather than calculator output.
9. OFAT can miss interactions that factorial DOE is specifically designed to estimate.
10. Mixture DOE is required when formulation proportions are constrained and cannot vary independently.
11. DOE can increase information from reliable measurements but cannot make an uncontrolled measurement valid.

12. Laboratory procedures, whether written or transmitted through local practice, should be treated as methods to be tested rather than assumptions to be accepted without qualification.

The broader conclusion is:

The reliability of an experiment is limited not by the sophistication of its best component, but by the uncontrolled behaviour of its weakest relevant component.

A defensible experimental result therefore requires qualification of the whole measurement system: instrument, specimen, reference, environment, operator, protocol and statistics.

References

- [1] Segurola, J. *Photochemical Characteristics and Design of Synergistic Photoinitiator Combinations for UV Curable Inks and Coatings*. Doctoral thesis, Manchester Metropolitan University, Department of Chemistry and Materials, in collaboration with Swale Process, Manchester, UK, 1999.
- [2] Segurola, J.; Allen, N. S.; Edge, M.; McMahon, A.; Wilson, S. Photoyellowing and discolouration of UV cured acrylated clear coatings systems: influence of photoinitiator type. *Polymer Degradation and Stability* 64(1) (1999) 39-48. DOI: 10.1016/S0141-3910(98)00169-4.
- [3] Allen, N. S.; Segurola, J.; Edge, M.; McMahon, A.; Roberts, I. A comparative kinetic study of UV-curable resins: the relationship between structure and rheology. *Surface Coatings International* 82(6) (1999) 285-292. DOI: 10.1007/BF02720124.
- [4] Segurola, J.; Allen, N. S.; Edge, M.; McMahon, A. Design of eutectic photoinitiator blends for UV/visible curable acrylated printing inks and coatings. *Progress in Organic Coatings* 37(1-2) (1999) 23-37. DOI: 10.1016/S0300-9440(99)00052-1.
- [5] Joint Committee for Guides in Metrology. *Evaluation of measurement data - Guide to the expression of uncertainty in measurement*. JCGM 100:2008. DOI: 10.59161/JCGM100-2008E.
- [6] Joint Committee for Guides in Metrology. *International Vocabulary of Metrology - Basic and general concepts and associated terms (VIM)*, 3rd ed. JCGM 200:2012. DOI: 10.59161/JCGM200-2012.
- [7] Brouwer, A. M. Standards for photoluminescence quantum yield measurements in solution (IUPAC Technical Report). *Pure and Applied Chemistry* 83(12) (2011) 2213-2228. DOI: 10.1351/PAC-REP-10-09-31.
- [8] Czitrom, V. One-Factor-at-a-Time versus Designed Experiments. *The American Statistician* 53(2) (1999) 126-131. DOI: 10.1080/00031305.1999.10474445.
- [9] Heckert, N. A.; Filliben, J. J.; Croarkin, C. M.; Hembree, B.; Guthrie, W. F.; Tobias, P.; Prinz, J. *Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods*. National Institute of Standards and Technology, Gaithersburg, MD, 2002. Relevant DOE material includes §5.3, "Choosing an experimental design"; §5.3.3.3 on full factorial designs; §5.3.3.3 on blocking of full factorial designs; and the process-modelling/response-surface sections. Official NIST record: <https://www.nist.gov/publications/handbook-151-nistsematech-e-handbook-statistical-methods> (accessed 20 September 2026).
- [10] Cornell, J. A. *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, 3rd ed. Wiley, 2002. DOI: 10.1002/9781118204221.
- [11] Cheadle, A. M. G.; Maier, E.; Palin, W. M.; Tomson, P. L.; Poologasundarampillai, G.; Hadis, M. A. The impact of modifying 3D printing parameters on mechanical strength and physical properties in vat photopolymerisation. *Scientific Reports* 15 (2025) 12592. DOI: 10.1038/s41598-025-97294-8.

Knowledge boundary

The documentary historical claims in this manuscript are limited to the 1999 thesis and the cited publications. First-person recollections are explicitly labelled and are not used as documentary evidence about another person's complete method or dataset. Modern metrological interpretation is grounded in the cited GUM, VIM, IUPAC and DOE sources but has not been retrospectively applied as a quantitative uncertainty budget to the historical datasets.

Limitations

No raw historical time series were reanalysed for this review. No complete uncertainty budget was reconstructed for the 1990s measurements. The flow-microcalorimetry and Vickers cases remain retrospective observations. The RTIR thin-film-interference mechanism is a present-day interpretation consistent with the observed baseline behaviour, not a surviving contemporaneous diagnosis. Historical data should therefore be reassessed only at the level supported by surviving methods, raw records, replicate structure and effect sizes.

Source control

Version 1.0 adopts the independent identifier HEB-001 and retains the scientific content independently approved in v0.7 without alteration. HEB-001 is a standalone technical article outside the ER-001 to ER-500 sequence. Its subject is complementary to the related Engineering Reviews identified in the ownership screen and does not replace, renumber or absorb any ER roadmap topic. Final rendering, author countersignature and repository actions remain separate controlled steps.

© 2026 Juan Segurola. All Rights Reserved.