

The Hidden Observer Problem: Subjective Endpoints, Expectation and the Unblinded Reader

Juan Seguro

3Dresyns by Resyner Technologies S.L.

© 2026 Juan Seguro. All Rights Reserved.

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

HEB-006 · Independent Technical Article · Version v0.2 · 22 September 2026

Abstract

Some experimental endpoints are completed by a human observer. A crack is judged present or absent, a surface is called acceptable, a colour match is declared, a specimen is assigned to a category, an image is segmented, a failure mode is classified or an ambiguous trace is accepted or rejected. In such cases the observer is not external to the measurement system. Perception, training, expectation, information available at the time of judgement, category definitions and stopping rules can all influence the recorded result. This article develops the hidden observer problem as a general methodological failure in which subjective or semi-subjective decisions are treated as though they were direct properties of the specimen. Historical colour matching provides a constructive counterexample: Maxwell's measurements treated perceptual equality as an empirical relation that could be repeated and compared across observers, rather than confusing perceptual equality with physical identity. Contemporary engineering examples show the complementary risks: "easy removal" is not transferable evidence, and visual inspection of hidden geometries is unsuitable unless validated. The article distinguishes legitimate human endpoints from uncontrolled observer dependence and proposes a practical architecture based on pre-defined criteria, blinding where feasible, randomisation, multiple observers, disagreement analysis, reference exemplars, repeat reads, objective anchoring and preservation of ambiguous cases. The central conclusion is that a human judgement can be a valid measurement component, but only when the observer's role is made explicit, bounded and testable.

Keywords: observer; subjective endpoint; blinding; visual inspection; inter-rater variation; expectation; classification; colour matching; measurement system; reproducibility; scientific inference

1. Purpose and evidence boundary

Scientific language often hides the observer inside verbs that sound objective: a surface "appeared" uniform, a specimen "showed" failure, an image "contained" a defect, a colour "matched", a trace was "acceptable", or a sample was "clearly" different. Sometimes these statements report robust observations. Sometimes they compress an unexamined human decision.

The goal of this article is not to eliminate human judgement. Many scientifically important endpoints are perceptual, morphological or classificatory and cannot be reduced to a single instrument output without losing information. The goal is to treat the observer as a component of the measurement system whenever the recorded result depends on human interpretation.

HEB-001 states that the measurement system includes the operator as well as instrument, specimen, reference, environment, protocol and statistical treatment [1]. HEB-006 develops the epistemic consequences of that inclusion.

The central proposition is:

When a human judgement completes the measurement, the conditions that shape that judgement are part of the method.

2. Observation is not the same as unrestricted subjectivity

Human observation can be highly informative. Microscopy, pathology, fracture-surface analysis, colour matching, defect classification and many visual inspections depend on trained observers. The methodological problem begins when a judgement is treated as self-validating because an experienced person made it.

A useful distinction is between perceptual access and inferential interpretation. An observer may reliably detect a visible discontinuity while being uncertain about its cause. A reader may identify a spectral feature while disagreeing about the mechanism that produced it. A colour match may be reproducible even though the two spectra are physically different. In colour science, such physically different stimuli that produce the same colour match under stated observing conditions are described as metamers; the phenomenon is metamerism.

The observer problem therefore contains at least three layers: detection, classification and explanation. These should not be collapsed.

3. Maxwell's colour work: the observer as part of the apparatus

The Maxwell project in this corpus highlights a useful historical case. Maxwell's colour experiments connected controlled physical stimuli to perceptual matches made by real observers. Two physically different spectra could be judged perceptually equal without being physically identical. The equality was empirical: it could be reproduced, quantified and compared across observers [2,5].

The significance is methodological. The observer was not treated as an embarrassment to be removed from the experiment. Perception was the endpoint, so the observer necessarily belonged to the apparatus. The scientific advance came from constraining that observer-dependent relation through controlled stimuli, adjustable proportions and comparison across people.

This example also shows why observer dependence does not make a result meaningless. A result can be observer-dependent and still be reproducible, bounded and scientifically powerful. The requirement is to define whose response, under what conditions, to which stimulus and with what repeatability.

4. Expectation can enter before the judgement is recorded

An observer may know which specimen is the control, which formulation is expected to perform better, which image belongs to the preferred process, or which data point would make a model fit. That information can influence attention, threshold selection, categorisation and the willingness to repeat or reject an observation.

This does not require misconduct. It is a structural property of an unblinded decision. If the observer has access to information correlated with the expected outcome, the measurement system includes that information pathway.

Blinding is therefore not a ritual. It is a method for removing an information channel that is irrelevant to the endpoint being judged. ER-209 requires randomisation and blind duplicates in interlaboratory work precisely because identity information can otherwise become entangled with interpretation and data handling [3].

5. Case study I: “easy removal” is not a measurement

ER-250 provides a contemporary engineering example. It states that where support-removal damage risk is significant, removal force or torque should be measured with representative coupons and that a subjective description such as “easy removal” is not transferable evidence [4].

The phrase “easy removal” compresses several variables: operator strength, tool choice, access angle, sequence, speed, expectation, prior experience, part geometry and the observer’s internal threshold for easy versus difficult. It can be useful as an exploratory note, but it is not a stable engineering measurand until the judgement is anchored to controlled conditions or an objective response.

The broader lesson is not that subjective descriptors are forbidden. It is that an uncalibrated adjective should not silently inherit the evidential status of a measurement.

6. Case study II: visual inspection is bounded by what can be seen

ER-250 also states that visual inspection alone is unsuitable for hidden geometries unless validated and calls for defined detection sensitivity with seeded defects or another bounded method [4]. This illustrates a different observer failure: the endpoint is limited not by judgement alone but by access.

An inspector can be perfectly consistent and still miss defects that are occluded, below resolution, poorly illuminated or confounded by surface texture. Repeating the same inspection does not recover information that never reached the observer.

Observer validation therefore requires both perceptual reliability and stimulus availability. A reproducible judgement on an inaccessible feature is still an invalid measurement of that feature.

7. Categories can create artificial certainty

Many observer-mediated endpoints are categorical: pass/fail, crack/no crack, acceptable/unacceptable, phase A/phase B, transparent/opaque, cured/uncured. The underlying reality may be continuous or multidimensional. A category boundary can therefore create a sharp recorded transition where the physical system changes gradually.

The method should define the threshold before the data are judged. If the boundary is moved after seeing the results, the observer becomes part of the model-fitting process. Borderline cases should be preserved rather than forced into a category solely to complete a table.

A useful practice is to record both the category and the underlying continuous or descriptive evidence where available. This allows later readers to see whether the conclusion depends on the threshold.

8. Training improves consistency but does not prove validity

Training can align observers around shared definitions and examples. Reference images, defect atlases, colour standards and worked examples can reduce avoidable variation. But agreement among trained observers can still be wrong if the shared criterion is not linked to the intended physical or functional endpoint.

Inter-observer agreement and validity answer different questions. High agreement shows that a rule can be applied consistently. Validity requires evidence that the rule identifies what the scientific claim says it identifies.

The same distinction applies to automated classifiers. Replacing a person with software can improve consistency while preserving the same biased labels, inaccessible features or inadequate ground truth.

9. Multiple observers turn disagreement into data

When an endpoint depends on judgement, a second observer is not merely a duplicate. Disagreement measures something about the ambiguity of the endpoint, the adequacy of the category definition, the visibility of the feature or the training state.

The Maxwell colour case demonstrates the constructive value of comparison across observers: a second observer helps separate a relation peculiar to one visual system from a relation reproduced in another [2,5]. In modern experiments, the same logic supports independent reads, blinded adjudication and explicit reporting of disagreement.

Consensus should not erase the original disagreement record. If adjudication is required, the initial independent judgements and the rule used to resolve them should remain recoverable.

10. Repeat reading can measure observer repeatability

An observer can be tested against their own previous decisions. Re-reading randomly ordered specimens or images after a delay can estimate within-observer repeatability. Repeating only the obvious cases is uninformative; the challenge set should include boundary cases and known negatives and positives where these can be established.

Repeat reading is especially valuable when the endpoint is used to train others or to create labelled datasets. If the reference observer is not self-consistent, the labels cannot be treated as a noise-free ground truth.

11. Objective anchoring should complement, not automatically replace, judgement

An objective measurement can anchor a subjective endpoint when the two address related parts of the same question: measured removal force can accompany perceived ease; profilometry can accompany surface appearance; image metrics can accompany visual defect ratings; spectral measurements can accompany colour matching.

The anchor should not be assumed to be the superior truth if the scientific endpoint is genuinely perceptual. In colour matching, for example, the perceptual equality is itself the phenomenon of interest. Physical spectra explain and constrain the stimulus, but they do not replace the perceptual endpoint.

The design question is therefore: which parts of the endpoint are intrinsically observer-dependent, and which can be independently measured to test the judgement?

12. A practical observer-control architecture

A robust observer-mediated measurement can be designed through a sequence of controls: define the endpoint before observation; specify the observable evidence; standardise viewing or reading conditions; remove irrelevant identity information where feasible; randomise order; provide reference exemplars; include ambiguous and challenge cases; obtain independent reads when consequence is high; quantify agreement and repeatability; preserve disagreements; and anchor the judgement to independent physical evidence where appropriate.

These controls should be proportional to the claim. A low-consequence exploratory note does not require the same architecture as an acceptance criterion, clinical endpoint or safety decision. But the more consequential the conclusion, the less defensible an undocumented single-reader judgement becomes.

13. Observer failure modes

Recurring failure modes include expectation leakage, unblinded identity, inconsistent thresholds, poor illumination or display conditions, fatigue, order effects, inaccessible features, category compression, training drift, consensus without independent reads, selective repetition of doubtful cases, and use of the same observer to generate and validate a reference label set.

A further failure occurs when interpretation is recorded as observation. “Crack present” and “failure caused by brittleness” are not the same statement. The first may be a visual classification; the second is a mechanistic inference requiring additional evidence.

The safest language names what the observer actually saw before naming what the observation may mean.

14. Relationship to the Engineering Review corpus

HEB-006 addresses observer dependence as a general scientific problem. HEB-002 addresses hidden conceptual boundaries, HEB-003 addresses visual reasoning as a route from structure to prediction, HEB-004 addresses proxy-measurand mismatch, and HEB-005 addresses reference validity. HEB-006 is distinct because the observer or reader participates directly in detection, classification or interpretation. ER-209 retains interlaboratory design, including randomisation and blinding. ER-250 retains support-removal measurement, operator effects, inspection sensitivity and functional qualification [3,4]. Other Engineering Reviews may contain application-specific visual or manual endpoints, but HEB-006 does not replace their qualification requirements.

The boundary is the same used throughout HEB: this article asks whether a conclusion follows from an observer-mediated measurement; the ER corpus asks how a particular engineering endpoint must be qualified.

15. Conclusions

A human observer can be an indispensable scientific instrument. The methodological error is to pretend that the observer is absent when perception, classification or interpretation completes the measurement.

The main conclusions are:

1. Observer-mediated endpoints are legitimate when the observer’s role is explicit and controlled.
2. Detection, classification and causal interpretation are distinct stages and should not be collapsed.
3. Blinding removes irrelevant information channels that can otherwise become part of the measurement system.
4. Subjective descriptors can support exploration but should not be treated as transferable quantitative evidence without anchoring.
5. Visual inspection is limited by feature visibility and detection sensitivity as well as by observer judgement.
6. Agreement among observers measures consistency, not necessarily validity.
7. Disagreement is informative and should be preserved rather than hidden by forced consensus.
8. Repeat reading can estimate within-observer reliability and expose unstable classification rules.
9. Objective measurements can anchor judgement, but they should not erase endpoints that are intrinsically perceptual.
10. The method should record what was observed before escalating to what the observation is claimed to mean.

The observer becomes dangerous to the experiment only when the experiment forgets that the observer was there.

References

- [1] Segurolo, J. The Hidden Error Budget: Measurement-System Failure, False Precision and Experimental Design in Laboratory Science. HEB-001. 3Dresyns Independent Technical Articles (2026). DOI: 10.5281/zenodo.22862642.

- [2] Segurola, J. Maxwell: Thinking the Invisible - James Clerk Maxwell and the Art of Reasoning Beyond What Could Be Seen. Version 1.3, 11 September 2026. DOI: 10.5281/zenodo.22714429.
- [3] Segurola, J. Interlaboratory Round-Robin Design for Vat-Photopolymer Materials: Specimen state, protocol transfer and reproducibility statistics. ER-209. 3Dresyns Engineering Publications (2026). Version DOI: 10.5281/zenodo.22862020.
- [4] Segurola, J. Support-Removal Effects on Fine Features and Internal Geometries: Access, Load Transfer, Surface Damage and Verification. ER-250. 3Dresyns Engineering Publications (2026). Version DOI: 10.5281/zenodo.22883403.
- [5] Maxwell, J. C. On the Theory of Compound Colours, and the Relations of the Colours of the Spectrum. Philosophical Transactions of the Royal Society of London 150, Part I (1860) 57-84.

Knowledge boundary

This article uses observer-related examples already documented within the controlled project corpus. The Maxwell example is limited to the evidential point supported there: perceptual matches were measured and compared across observers, and perceptual equality was not equated with physical identity. The engineering examples are used only to illustrate observer control and do not replace their original qualification contexts.

Limitations

The article is a methodological synthesis, not a quantitative review of observer-bias effect sizes or inter-rater statistics. No universal number of observers, agreement threshold or blinding procedure is prescribed. Appropriate controls depend on the endpoint, consequence, available ground truth and whether the observer-dependent response is itself the phenomenon of interest.