Accuracy Is Not One Number

A Measurement Assurance Framework for Camera-Based Human Movement

Hossein Mokhtarzadeh, PhD
PoseIQ, Arvinovate Pty Ltd, Melbourne, Australia
Scientific Note · Version 1.0 · September 2026
https://assurance.poseiq.com/

Abstract

Camera-based human movement systems can produce fundamentally different outputs, including event detections, repetition counts, temporal measures, two-dimensional kinematics, and reconstructed three-dimensional trajectories. Asking for a single platform-level “accuracy percentage” therefore conflates measurands with different measurement chains, sources of uncertainty, and validation requirements.

This scientific note presents the **PoseIQ Measurement Assurance Framework**, in which evidence is attached to a defined **measurand, task, measurement configuration, and operating envelope**, rather than assumed to generalise across an entire platform. Performance should be evaluated against an appropriate reference using metrics suited to the measurement question, including bias, absolute error, limits of agreement, repeatability, reliability, classification performance, and failure rate where applicable. Model confidence is distinguished from measurement accuracy, and unsuccessful observations are treated as part of performance evidence rather than silently excluded.

The framework introduces measurement-specific **Evidence Cards** that document the validation protocol, operating envelope, quantitative performance, failure criteria, limitations, and evidence status of a configured measurement. The objective is not to weaken the requirement for validation, but to make validation claims specific, reproducible, and bounded. Evidence established for one configuration should not be extrapolated to untested measurements, populations, or capture conditions without further evaluation.

Keywords: markerless motion capture; biomechanics; computer vision; measurement error; validation; reliability; human movement analysis

1. Accuracy is a property of a defined measurement

A camera-based movement platform does not directly measure an abstract quantity called "movement accuracy." It acquires images from which increasingly derived quantities may be estimated:

Physical movement → image acquisition → visual observability → detection/tracking → geometric processing → derived biomechanical quantity → interpretation

Each stage introduces different assumptions and potential sources of uncertainty. Image acquisition may depend on frame rate, exposure, resolution, timing, and camera placement. Landmark estimation may be affected by visibility, occlusion, clothing, and model behaviour. Derived two-dimensional or three-dimensional quantities may additionally depend on scale, camera geometry, calibration, synchronisation, reconstruction, anatomical definitions, and subsequent processing.

Consequently, different outputs require different validation strategies. A repetition count based on a landmark crossing predefined boundaries is fundamentally different from estimating a three-dimensional knee trajectory.

The relevant scientific question is therefore not:

> **How accurate is PoseIQ?**

but:

> **How accurately does a defined PoseIQ configuration measure a specified quantity, during a specified task and operating envelope, relative to an appropriate reference?**

Consider a repetition defined by a deterministic sequence such as:

below boundary A → cross boundary B → return below boundary A = one repetition

If a validation dataset contained 1,000 manually verified repetitions and the system correctly counted all 1,000, the defensible statement would be:

> **Observed repetition-count agreement was 100% (1,000/1,000) under the tested protocol and conditions.**

It would not establish that "PoseIQ is 100% accurate." The former is a measurement-specific empirical result; the latter improperly extrapolates that result across different measurands and operating conditions.

The same principle applies to temporal measurements. Nominal recordings at 30, 60, and 120 frames·s^{-1} correspond to sampling intervals of approximately 33.3, 16.7, and 8.3 ms, respectively. These intervals describe temporal sampling, not guaranteed event-timing accuracy. Actual measurement performance may additionally depend on timestamp integrity, dropped frames, synchronisation, event definition, and algorithm behaviour.

Table 1. Examples of measurement-specific evidence

Measurand	Example	Principal evidence
Event detection	Target crossing	Sensitivity, specificity, false-trigger rate
Count	Exercise repetitions	Count agreement, absolute error, failure rate
Timing	Event or repetition duration	Bias, MAE/RMSE, limits of agreement
2D kinematics	Sagittal knee flexion	Bias, error, repeatability
3D kinematics	Joint trajectory	Spatial/angular error, agreement
Decision output	Threshold exceeded	Sensitivity, specificity, decision error

A single statistic cannot adequately characterise all of these measurement properties.

2. Measurement configuration and operating envelope

For this framework, a **measurement configuration** comprises the measurand, task, acquisition system, processing method, and operating envelope under which a measurement claim is made.

The **operating envelope** describes the conditions over which supporting evidence has been established. Depending on the measurement, these may include camera number and placement, frame rate and resolution, subject-to-camera distance, calibration requirements, lighting, visibility and occlusion, movement characteristics, population, and synchronisation requirements.

An operating envelope is a **boundary on evidence, not necessarily a restriction on use**. A consumer camera operating in a gym and a calibrated multi-camera laboratory system can each constitute defined operating envelopes. Their measurement uncertainty may differ and should be quantified accordingly.

Optical motion-capture performance is itself dependent on factors such as capture volume, object geometry and speed, ambient conditions, camera characteristics, calibration, and processing algorithms [3]. The same principle applies to markerless camera-based measurement.

Evidence demonstrated within one operating envelope should therefore not automatically be extrapolated beyond it.

2.1 Model confidence is not measurement accuracy

Computer-vision systems frequently expose quantities described as confidence, visibility, detection probability, or tracking quality. These values can be useful components of quality control, but they are not automatically measures of anatomical or biomechanical accuracy.

For example, a model may assign high confidence to a detected knee landmark while the estimated location still differs from a reference anatomical joint centre because of perspective, anatomical ambiguity, occlusion, or model behaviour.

A landmark confidence or visibility score therefore characterises an aspect of model output; it does not, by itself, quantify error relative to an anatomical or biomechanical reference.

Such quantities should not be represented as “percentage biomechanical accuracy” unless their relationship with measurement error has itself been empirically established.

2.2 Reference methods

The appropriate reference method depends on the measurand.

Manual video annotation may be the most direct reference for repetition counts or visually defined events. High-frame-rate or hardware-timestamped recordings may be appropriate for temporal measurements. Calibrated marker-based optical motion capture is commonly used as a reference for spatial three-dimensional kinematics.

Marker-based motion capture is an appropriate comparator for many biomechanical validation studies, but it is itself a measurement system. Instrumental uncertainty, marker placement, soft-tissue artefact, and biomechanical modelling can contribute to disagreement between methods.

Consequently, method-comparison results should distinguish **agreement with a reference method** from claims of error relative to an assumed error-free anatomical ground truth.

3. Quantifying performance and uncertainty

Published markerless-motion literature illustrates why measurement performance must be characterised specifically rather than through a universal platform-level number.

A 2025 systematic review covering 53 studies of lower-extremity and trunk markerless kinematics reported differences from marker-based systems ranging approximately from 0.2° to 28.6°, depending on the system, task, joint, and movement plane [4]. Reliability likewise varied according to the variable and task.

A 2024 systematic review and meta-analysis of markerless three-dimensional gait measurement reported generally stronger performance for several spatiotemporal variables, while kinematic validity varied across joints and movement planes. Sagittal hip and knee measures were generally stronger than several frontal-plane, transverse-plane, and ankle measures [5].

Similarly, a 2024 validation of OpenCap across 437 return-to-sport trials reported an overall joint-angle mean absolute error (MAE) of 3.85° and root-mean-square error (RMSE) of 4.34°, with performance varying substantially by joint and movement plane [6].

These findings do not establish that markerless measurement is inherently accurate or inaccurate. Rather, they demonstrate that performance is **measurand-specific, task-specific, plane-specific, system-specific, and protocol-specific**.

Correlation alone is also insufficient to establish agreement between measurement methods. Bland and Altman demonstrated that two methods can be strongly correlated while exhibiting systematic or clinically relevant disagreement [7]. Method-comparison studies should therefore examine the magnitude and distribution of differences using statistics appropriate to the measurement question.

Table 2. Measurement properties and their interpretation

Quantity	**Question addressed**
Bias	Is there systematic over- or underestimation?
MAE / RMSE	What is the magnitude of disagreement?
Limits of agreement	How variable is disagreement between individual observations?
Repeatability / measurement error	How stable is the measurement under repeated conditions?

Reliability	How well are subjects or conditions distinguished relative to measurement variability?
Sensitivity / specificity	How well is a defined event or class identified?
Failure rate	How often does the system fail to return an acceptable measurement?

No single statistic substitutes for the others. Established measurement frameworks distinguish related but non-equivalent properties such as trueness, precision, reliability, measurement error, and validity [1, 2].

3.1 Failed and invalid observations

Failure handling is part of measurement performance.

Validation protocols should define, before results are reported, what constitutes a valid observation, an invalid observation, and a measurement failure; how missing outputs are handled; and whether unsuccessful observations are included in summary statistics.

Failure criteria should be specific to the measurand. For a repetition counter, failure may mean an incorrect count or inability to return a count. For a continuous joint-angle trajectory, failure may involve loss of valid tracking or violation of predefined quality criteria. For a multi-camera reconstruction, failure may relate to calibration, synchronisation, or reconstruction quality.

> **Performance calculated only after undocumented exclusion of unsuccessful trials is not a complete description of measurement performance.**

Failure frequency should therefore be reported alongside quantitative performance where relevant.

4. The PoseIQ Measurement Assurance Framework

The framework proposes a measurement-specific Evidence Card as a standard structure for documenting validation evidence.

An Evidence Card links a defined measurement configuration to the evidence supporting it. It may be completed as part of a PoseIQ validation study or by researchers and other evaluators assessing a PoseIQ configuration for a particular application.

Figure 1. PoseIQ Measurement Assurance Framework

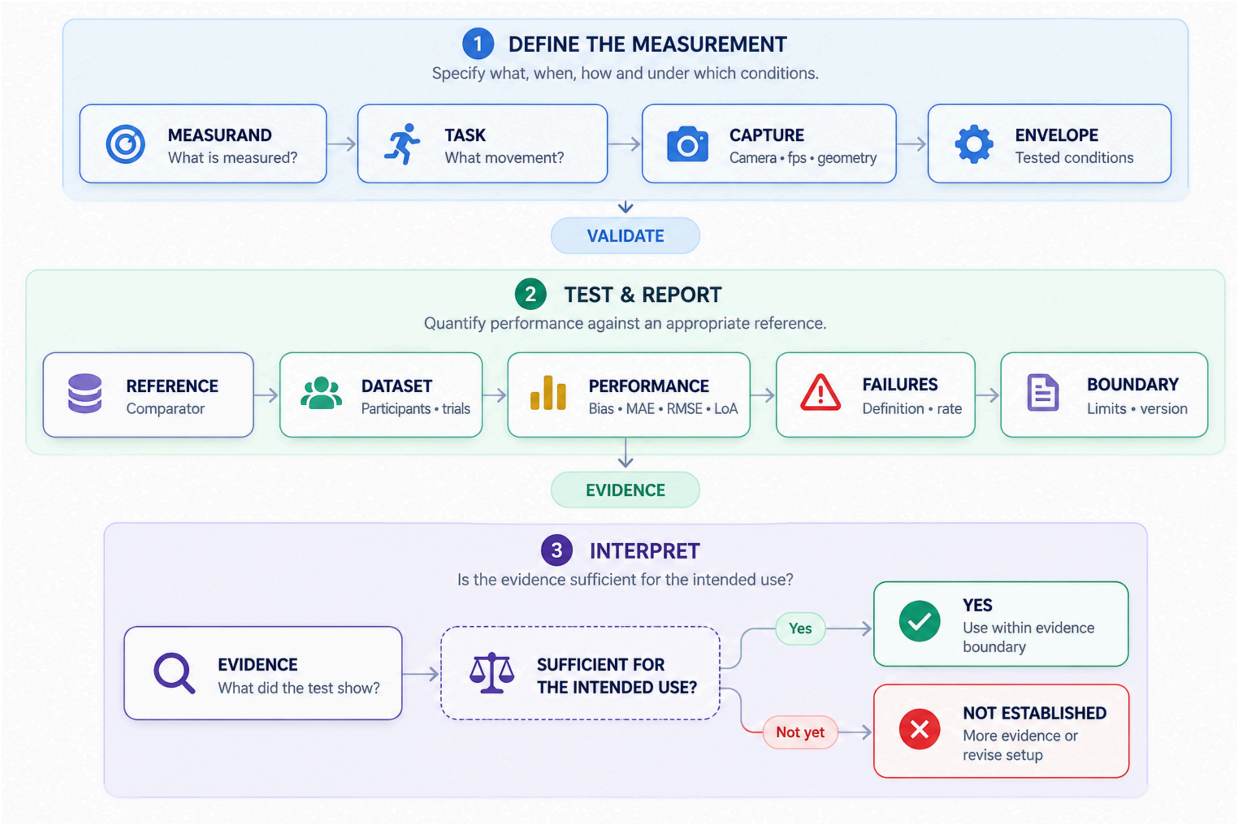

Figure 1. PoseIQ Measurement Assurance Framework. Measurement assurance is attached to a defined measurement configuration. Quantitative evidence is established against an appropriate reference and interpreted within its operating envelope before suitability for an intended use is considered.

Table 3. Proposed Evidence Card structure

Field	Information reported
Measurand	Quantity being measured
Task	Movement or activity evaluated
Configuration	Camera geometry, acquisition, and processing setup
Operating envelope	Conditions over which evidence has been established
Reference method	Comparator used for validation
Dataset / protocol	Population, observations, and experimental procedure
Performance	Bias, MAE, RMSE, LoA, reliability, or other appropriate metrics

Failure definition / rate	Definition and observed frequency of unsuccessful measurements
Limitations	Conditions to which evidence should not be extrapolated
Evidence version	Software, algorithm, and protocol version associated with the evidence
Evidence status	Current state of validation

Evidence status is reported using three categories:

Status	**Interpretation**
Validated	Quantitative evidence is available for the defined configuration
Under evaluation	Formal evaluation is underway and evidence is not yet considered complete
Exploratory research	No established quantitative validity claim is made for this configuration

These categories describe the **state of evidence**, not overall product quality.

Exploratory configurations should remain inspectable where technically appropriate. Relevant information may include source video or frames, detected landmarks, confidence or quality indicators, processing rules, and intermediate measurements. This enables researchers to examine the measurement pipeline rather than relying solely on an opaque derived output.

Importantly, the framework does not imply that all possible PoseIQ measurements have already been validated. Where quantitative evidence exists, it should be reported. Where it has not yet been established, the evidence boundary should be explicit.

5. Evidence boundaries and intended use

Validation of one measurement configuration does not establish validity for materially different measurands, tasks, populations, camera geometries, processing methods, or operating conditions.

A researcher or practitioner may therefore use the framework in two ways. First, an existing measurement configuration may be adopted where its Evidence Card and operating envelope are sufficiently aligned with the proposed application. Second, a new configuration may be evaluated using an appropriate reference and validation protocol, generating new measurement-specific evidence.

This makes validation **extensible rather than universal**.

Before interpreting a PoseIQ measurement, five questions should therefore be established:

1. Measurand: What quantity is being estimated?
2. Configuration: During what task and under what operating envelope?
3. Reference: Against what appropriate comparator is performance established?
4. Evidence: What are the error, agreement, reliability, classification, and failure characteristics relevant to this measurement?
5. Intended use: Is the demonstrated performance sufficient for the proposed application?

The final question is application-specific. The uncertainty acceptable for a low-risk interaction or coaching trend may differ from that required for a quantitative research outcome or higher-consequence decision.

Crucially:

> **Suitability should be interpreted from measured performance; the intended use should not be retrospectively redefined to accommodate an unfavourable validation result.**

The purpose of the operating envelope is therefore not to explain away poor performance after testing. It is to define the boundary within which evidence was generated and to prevent unsupported extrapolation beyond that boundary.

6. Conclusion

A universal accuracy percentage is not an adequate description of a camera-based human movement platform because different outputs represent different measurands, measurement chains, and sources of uncertainty. This does not remove the requirement for validation; it makes the required evidence more specific.

The PoseIQ Measurement Assurance Framework therefore attaches evidence to a defined **measurand, task, measurement configuration, and operating envelope**. Performance is quantified using metrics appropriate to the measurement question, unsuccessful observations are explicitly accounted for, and evidence is not extrapolated beyond the conditions in which it was established.

> **Specify the measurement. Define the conditions. Quantify performance. Report failures. State the evidence boundary.**

This framework provides a structured approach for designing, evaluating, and reporting measurement-specific validation of camera-based human movement applications.

Scope statement

This scientific note defines the PoseIQ measurement-assurance framework. It is **not itself a validation study** and does not establish numerical performance for a PoseIQ measurement unless that performance is linked to a defined protocol and dataset.

Numerical results from external systems or published studies are included to illustrate measurement-science principles and should not be interpreted as PoseIQ performance claims. Current and future quantitative PoseIQ claims should be associated with measurement-specific validation records or Evidence Cards.

References

1. International Organization for Standardization. *ISO 5725-1:2023. Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions.* Geneva: ISO; 2023.

2. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63(7):737–745. doi:10.1016/j.jclinepi.2010.02.006.

3. Massey T. *Vicon Study of Dynamic Object Tracking Accuracy.* Vicon Motion Systems; 2021.

4. Yoma M, Llurda-Almuzara L, Herrington L, Jones R. Reliability and validity of lower extremity and trunk kinematics measured with markerless motion capture during sports-related and functional tasks: a systematic review. *J Sports Sci.* 2025;43(17):1703–1730. doi:10.1080/02640414.2025.2518359.

5. Scataglini S, Abts E, Van Bocxlaer C, Van den Bussche M, Meletani S, Truijen S. Accuracy, validity, and reliability of markerless camera-based 3D motion capture systems versus marker-based 3D motion capture systems in gait analysis: a systematic review and meta-analysis. *Sensors.* 2024;24(11):3686. doi:10.3390/s24113686.

6. Turner JA, Chaaban CR, Padua DA. Validation of OpenCap: a low-cost markerless motion capture system for lower-extremity kinematics during return-to-sport tasks. *J Biomech.* 2024;171:112200. doi:10.1016/j.jbiomech.2024.112200.

7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1(8476):307–310. doi:10.1016/S0140-6736(86)90837-8.