SSIM 2023 – PM Dight (ed.)

doi:10.36487/ACG_repo/2335_19

The risk of confusing model calibration and model validation with model acceptance

D Elmo *The University of British Columbia, Canada*

Abstract

This paper examines the meaning of calibration and validation in rock engineering design, highlighting several challenges and limitations associated with these processes. There exist two fundamental limitations: i) the inability to rely on engineering judgement as a substitute for proper calibration and validation, and ii) the use of qualitative characterisation methods introduces subjectivity in the data subsequently used for calibration and validation. Furthermore, varying modelling conceptualisations result in a paradoxical situation whereby the same problem analysed using different numerical models requires a different set of parameters, which can all be claimed to be calibrated. The author acknowledges that some of the points raised in this paper may encounter objections. However, by ignoring the epistemic limits of calibration and validation, there is the risk of letting engineering faith become the excuse behind the tendency to replace model validation with model acceptance.

Keywords: *model calibration, model validation, model acceptance*

1 Introduction

Calibration and validation are critical in assessing the accuracy and reliability of numerical models used in engineering design processes. While these terms are often used interchangeably, it is important to understand their distinct meanings and implications. Calibration refers to adjusting a model's parameters or inputs to agree with observed data. Calibration aims to minimise the discrepancy between model outputs and observed data, making the model more reliable for future predictions. Validation involves testing the model's predictive capability against independent data not used during calibration.

Both calibration and validation are vital for accurate decision-making and permitting in mining developments. Calibrated models allow engineers to make informed decisions based on field observation predictions. Validated models, on the other hand, establish the model's credibility by demonstrating its ability to generalise to unseen scenarios, instilling confidence in the model's application to new mining projects.

While engineers often focus on the technical aspects of the design process, it is essential to communicate the nuances of calibration and validation to the broader community. This is particularly crucial in mining projects, where the risks associated with such developments can have significant environmental, social and economic impacts. Accordingly, this paper focuses on presenting something other than visually captivating numerical results. Instead, this paper will discuss the meaning and limitations of calibration and validation. It is understood that some of the questions raised in this paper will encounter strong resistance. However, openly discussing calibration and validation processes can help bridge the gap between technical expertise and community understanding. It enables effective communication of the model's accuracy, limitations and associated uncertainties to rightsholders and stakeholders, fostering trust and facilitating informed discussions about the risks and benefits of mining projects.

1.1 Towards a holistic approach to slope design

As shown in Figure 1, the technical scope of open pit design includes advanced numerical modelling and analytical methods to evaluate the behaviour of rock masses. This approach employs tools like finite element analysis, discrete element modelling or other hybrid forms of numerical simulations to study the rock mass's response to proposed design conditions. Numerical models allow for detailed analyses of complex

geotechnical problems and provide *quantitative* insights into the performance of open pits. Likewise, they enable engineers to optimise designs, assess different scenarios and predict potential failure modes. Note that we have purposely used italic for the term 'quantitative' since many of the input parameters used in open pit design are qualitative (Yang & Elmo 2022; Harrison 2017). The quantification problem will be raised in several parts of this paper since it is intrinsic to calibration and validation.

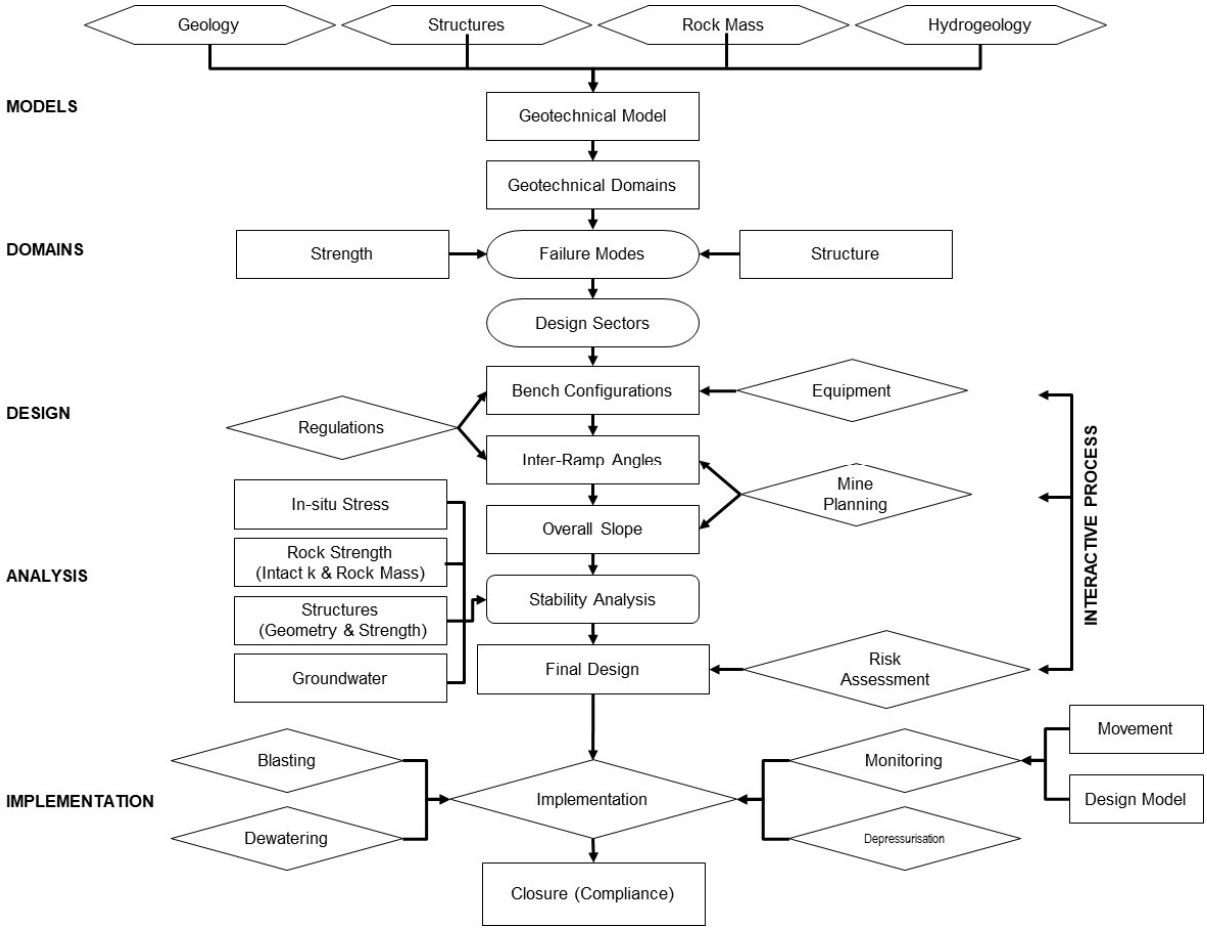

Figure 1 Slope design process (adapted from Read & Stacey 2009, Figure 1.4)

Figure 1 suggests a seamless flow of quantitative information from models to analysis, reinforcing an incorrect narrative of a quantitative design approach that can be easily calibrated and validated. However, there exist significant limitations to this approach, including:

- The processes of calibration and validation are only possible with independent data. However, new open pit mines are designed before developing access to pit walls. Under these conditions, we cannot access all the independent data required to calibrate and validate our rock mass models (e.g. fracture size). Furthermore, our decision-making process is more likely to be influenced by 'anchoring' bias that occurs when we place too much emphasis on the first piece of information we receive (known as the 'anchor') and use it as a reference point for all subsequent information (e.g. rock mass strength parameters). The problem of anchoring bias applies indistinctly to green sites and existing operations (e.g. pushbacks, pit optimisation, double benching etc.).
- The current design approach transforms qualitative assessments (e.g. classification of Geology, Structures and Rock Mass data, Figure 1) into calibrated quantitative measurements (Final Design, Figure 1). This transformation occurs by manipulating the natural variability of rock masses and the subjective nature of characterisation/classification methods to justify using specific input properties that fit the assumed modelling approach. This aspect leads to two significant consequences:

- Engineering judgement cannot be used as a substitute for proper calibration and validation processes; and
- The labels *calibrated* and *validated* apply to the modelling output, not the input. An incorrect practice exists of describing rock mass strength parameters (input) as *calibrated* or *validated* simply because the modelled deformations match field measurements. Differences in models' conceptualisation lead to different modelling inputs and the paradox of rock mass strength parameters that remain calibrated despite changing from model to model.

- Pit slopes are excavated, not built. This semantic problem results in technical differences between what we define as an active design process (building – controlled by known knowns and known unknowns) and a reactive design process (excavating – controlled by unknown conditions). The excavation process is complicated by the need to manage different forms of uncertainty, including unknown uncertainty (Elmo & Stead 2020). The insistence on using qualitative characterisation methods complicates the matter as it introduces a degree of subjectivity in the way we collect and manage the same data used to calibrate and validate our numerical models. There is also a compound effect when qualitative data are used to validate assumptions based on engineering judgement since both terms of the validation process (data and judgement) remain subjective. The resulting validation process gives rise to confirmation bias.
- The overall design process treats pit slopes as large and complex engineered (i.e. built) structures that obey specific mathematical rules. The commonly accepted wisdom is that knowing the rules describing the configuration of engineered structures leads to the conclusion that their behaviour is entirely predictable. While mathematical rules control rock mass behaviour, unpredictable behaviour remains possible. This raises the question of whether models calibrated and validated using observed data can capture unpredictable behaviour. Quoting Taleb (2010), '*How can we know the future, given knowledge of the past, or more generally, how can we figure out properties of the infinite unknown based on the finite knowns?'.* Two major anthropogenic slope failures confirm the conclusions by Taleb (2010):
  - The Manefay landslide (Bingham Canyon mine, Ross 2017), where attempts to predict the failure run-out were based on experience and understanding of other failures at the mine. As a result, the models were not trained to simulate unknown conditions. Therefore, they could not predict the actual run-out velocity of the failed material and that the failure eventually consisted of two separate significant events.
  - In 1961 and 1962, physical experiments were conducted to study the potential tsunami that might be caused by a large landslide falling in the Vajont reservoir. Indeed, a large relic landslide was discovered as early as 1959 along one of the flanks of the reservoir. The dam was completed in 1962. The experiments considered scenarios designed by engineers, based on their knowledge and experience. None of the experiments included the scenario that eventually unfolded on 9 October 1963, leading to one of the most significant engineering disasters in history (more than 2,000 people perished as massive flooding destroyed several villages and towns). The Vajont tragedy teaches us that when discussing the process of calibration and validation, we need to acknowledge that there is no assurance that engineering judgement may lead to clearly identifying conclusions that are not correct (Elmo et al. 2022).
- In Figure 1, mine closure is the last step in the overall pit design process. On this basis, it reflects a *'limited use'* approach to mine design opposite to the '*continued use'* of the land by the rightsholders. A mine operation represents a temporary alteration of the original land use, which must be re-established upon closure. In this context, the problem of model calibration and validation goes beyond a pure engineering scope and is translated into the need to communicate our results to communities who want to be informed about the long-term stability and impact of pit slopes. Therefore, pit design should include *time* as an important aspect of stability analysis. The best-designed open pit is not the one that fails just after mine closure but the one that does

not fail within a measurable time window that different generations can experience (e.g. 100–150 years into the future). In Figure 2 we propose a modified version of Figure 1 that sees mine closure elevated to a component of the stability analysis phase to include technical, permitting, social and environmental constraints as part of the stability analysis.

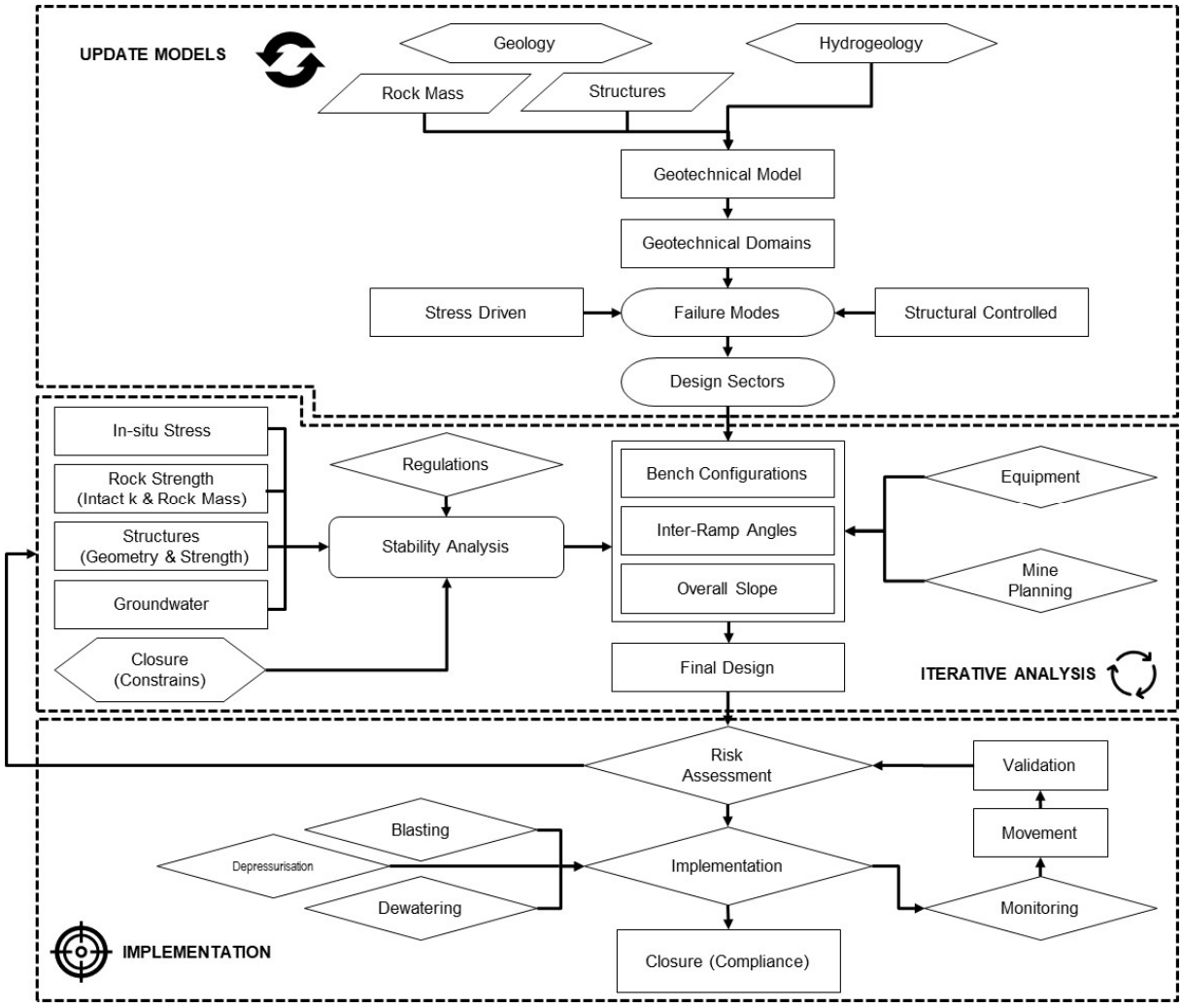

Figure 2 Revised slope design process to better include the importance of mine closure considerations

2 Calibration and validation

Calibration involves adjusting model parameters to match observed data, while validation focuses on evaluating the model's predictive capabilities against independent data. This distinction is crucial, as it impacts the models' applicability and the conclusions derived based on the modelling results. A calibrated model is not a validated model, but a validated model is a calibrated model.

Model calibration entails fine-tuning the model parameters to achieve a close agreement between simulated results and observed data. Calibration is a necessary step to improve the representativeness of a model. While in principle it is possible to calibrate (and validate) a numerical model using data from an existing mine operation (Site A), the same model cannot be said to be calibrated (nor validated) when using it to study the stability of a new mine operation (Site B) since independent data for Site B will not become available until the implementation and monitoring stages.

As stated above, calibration alone does not guarantee the model's predictive capabilities or its ability to capture the underlying physical mechanisms. The model may reproduce observed data (e.g. slope deformations), but it may fail to replicate the fundamental behaviours or mechanisms that govern the system

being studied. For example, when one considers the problem shown in Figure 3, any intervening simplification process adopted to reduce run times and improve mesh quality impacts the observed results (Shapka-Fels & Elmo 2022). As a result, we must accept that four potentially different sets of calibrated rock mass properties exist because of four different rock mass model representations. Claiming, for example, that models (A) and (D) are both calibrated (and validated) results in the following two corollaries, which could have significant ramifications when communicated outside of a rock engineering context:

- Rock mass behaviour exists in an undefined state, and it can simultaneously be continuum and discontinumm.
- The simplification assumptions we make in our models concerning the rock mass structural character do not impact the predictive capabilities of our numerical models.

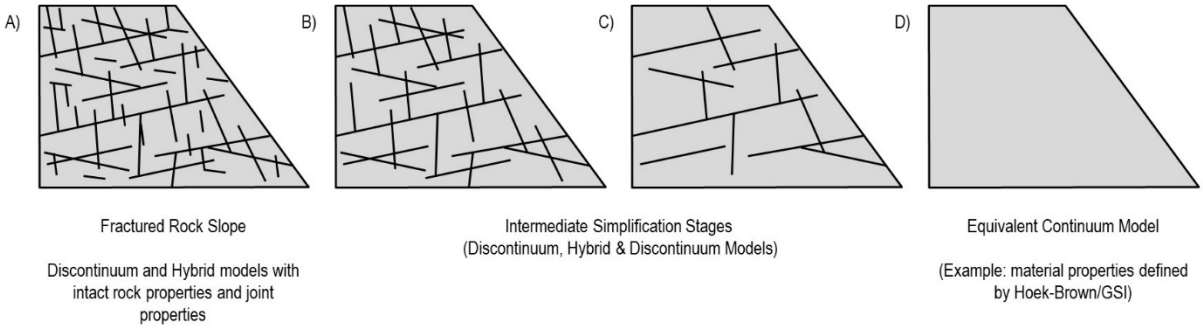

Figure 3 Example of how modelling trades structural complexity (inserts C and D) for geometrical simplicity (inserts A and B). Modified from Shapka-Fels & Elmo (2022)

For the behaviour of Model (A) to agree with that of Model (D), we would need to assume that either rock mass behaviour is isotropic or somehow incorporate anisotropic effects in Model (D). In contrast, we can justify using intact rock properties in Model (A) for the rock matrix. Models (B) and (C) would require deriving some form of intermediate rock mass upscaling. However, this raises important questions about the simplification process, including how to decide which joints to keep and which joints to simulate implicitly. This problem was demonstrated by Karimi et al. (2020), who showed that jointed configuration independently simplified by a senior engineer, a junior engineer and an automated algorithm yielded very different results in terms of modelled rock mass strength. Note that the main difference between the three models in Karimi et al. (2020) was produced by the adopted simplification process, which resulted in the models displaying different degrees of network connectivity while the overall fracture intensity did not change.

More recently, Elmo (2023) demonstrated that two models with an equivalent rock mass quality (i.e. the same geological strength index (GSI) – Hoek 1994) but different degrees of network connectivity could yield drastically different rock mass responses. These results confirm that the calibration of numerical methods should place more emphasis on mechanisms rather than the process of correcting qualitative input data to match observations. More importantly, the author believes we must abandon the idea of rock mass ratings as quantifiable properties that can be calibrated since it is impossible to calibrate a qualitative condition that cannot be measured.

This brings us to model validation, which aims to assess the model's ability to predict independent data or scenarios. Successful validation imparts confidence in the model's predictive capabilities and enhances its credibility for decision-making purposes. Validation is a crucial test of the model's reliability and generalisability beyond the calibration dataset. It involves comparing model predictions against observations or experiments not used during calibration. However, more than this process is required to warrant a guarantee of universal validation, and the model validation process is not without questions. The validation process should guarantee that the model is not overfitting or tailoring itself too closely to the calibration data, which could result in unreliable predictions when applied to new scenarios. This problem ties in with the availability of independent validation data and their representativeness of the system under

investigation. Another question arises from the uncertainty inherent in the validation process itself. Real-world slope problems are complex and governed by numerous factors, many of which are not fully quantifiable. This inherent uncertainty introduces a degree of unpredictability, making it difficult to validate a model against all possible scenarios definitively. Consequently, model validation is often viewed as an ongoing process rather than conclusive, requiring continuous refinement and adaptation to incorporate new knowledge and observations.

Additionally, there is a need for clear criteria to assess the success of model validation. Determining the level of agreement or acceptable deviation between model predictions and independent data introduces subjectivity. Developing robust validation metrics and guidelines considering quantifiable factors is crucial for ensuring consistency and transparency in the validation process. To overcome these challenges, several approaches can be considered. First, sensitivity analysis can be employed to assess the influence of different parameters and assumptions on model predictions. This analysis helps identify critical factors and sources of uncertainty that require further investigation and refinement. Furthermore, integrating multiple lines of evidence, such as field measurements, laboratory experiments and monitoring data, can strengthen the validation process. Combining diverse data sources and knowledge domains enhances the robustness and reliability of model predictions.

3 The risk of confusing model acceptance with model validation

We must acknowledge that cognitive biases in our engineering design methods often confuse model acceptance with model validation (Yang et al. 2021; Elmo et al. 2022). What can we learn from history that applies to calibrating and validating rock engineering mechanisms? For almost 1,400 years, the geocentric model (Ptolemy 100 to c. 170 AD) was used to predict planetary motions (and did that reasonably well, despite the wrong underlying assumptions). The heliocentric model proposed by Copernicus (1473–1543 AD) did not significantly improve predictions of planetary motions over the geocentric model. Still, it broke away from a dogmatic view that placed Earth at the centre of the known universe. While the heliocentric and geocentric models meet the condition for validation (they predict planetary motions), Kepler (1571–1630) showed both to be mechanically wrong in 1609–1616, since neither is Earth at the centre of the solar system (heliocentric model) nor are the orbits of the planets in the solar system circular (geocentric model).

Regarding methods and numerical models used in rock engineering applications, we cannot assume they are mechanistically correct (i.e. validated) just because they match observed deformations. Rock engineering design is thus exposed to the same non-validation paradox of the heliocentric and geocentric models. Indeed, we often rely upon an *a priori* assumption of the failure mechanisms when calibrating our rock engineering models (susceptibility models – Kalenchuk 2019).

By ignoring the epistemic limit of numerical modelling principles we increase our exposure to risk by replacing model validation with model acceptance. When Galileo tried to assert the use of the heliocentric model over the geocentric model, his work was not questioned based on the underlying mechanics. Instead, it was rejected based on the geocentric model countering cosmological orthodoxy. Likewise, allowing expressions like engineering judgement, educated guess and realistic assumptions to infiltrate and somehow control numerical analysis confirms the role that cognitive biases play in calibration and validation processes.

Using behavioural science concepts (Kahneman 2011), it is possible to subdivide design methods used in rock engineering into heuristics (System 1) and rational thinking (System 2) methods (Figure 4). Empirical methods belong to System 1, while numerical models belong to System 2. Empirical methods are common among rock engineering practitioners because of their fast and practical nature, which suits working (everyday) decisions. Biases are more prevalent when System 1 is convinced of its correctness and System 2 fails to correctly filter or use data and conclusions received from System 1. Furthermore, input for our numerical models is often determined using empirical methods that rely on somewhat subjective qualitative assessments. Self-acceptance of empirical methods and over-reliance on qualitative assessments is responsible for the overlapping between System 1 and System 2 and the transferring of cognitive bias in our numerical modelling assumptions and results.

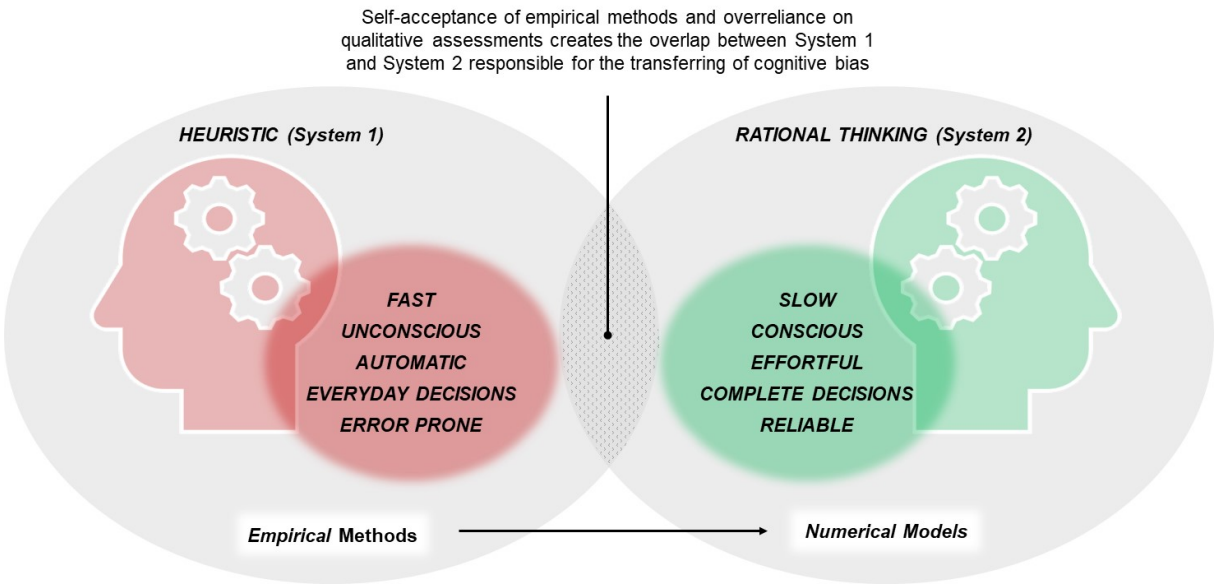

Figure 4 Relationship between design methods used in rock engineering in terms of heuristics (System 1) and rational thinking (System 2)

4 The temporal dimension of calibration and validation

The objective of predictions is to anticipate future conditions. However, there cannot be calibration and validation without data. Therefore, calibration and validation inherit the temporal dimensions of the data used in the process. Because data cannot exist in the future (except synthetic data), calibrating and validating the results of numerical models becomes a challenge when the models are expected to represent a reality that either does not physically exist yet (e.g. pre-feasibility and feasibility studies) or cannot be directly measured (e.g. intact rock strength versus rock mass strength). This temporal dimension presents a significant hurdle in assessing the accuracy and reliability of complex numerical models.

The process used in rock engineering practice to calibrate numerical models is primarily based on the principle of back-analysis (Figure 5). As such, the process is limited by the lack of separate consideration of calibration and testing data (Figure 5). Validation of the models is only possible with independent testing data. As a result, calibrated models are used to provide design recommendations even though they are not technically validated. The issue is often compounded by an *a priori* assumption about failure mechanisms or because data from another site have been used in the calibration process.

In principle, these limitations may hypothetically be reduced by continuously repeating the calibration process as new data becomes available. Under these conditions, the new data become the independent testing data, and a continuous calibration process is performed in lieu of the validation process (indirect validation). However, two challenges remain: i) while the indirect validation is likely to be conducted using data specific to the mine site under consideration, engineers may still decide not to carry out an independent mechanistic validation; and ii) if the indirect validation is negative (i.e. the modelling results do not agree with the field data), the entire calibration process must be repeated.

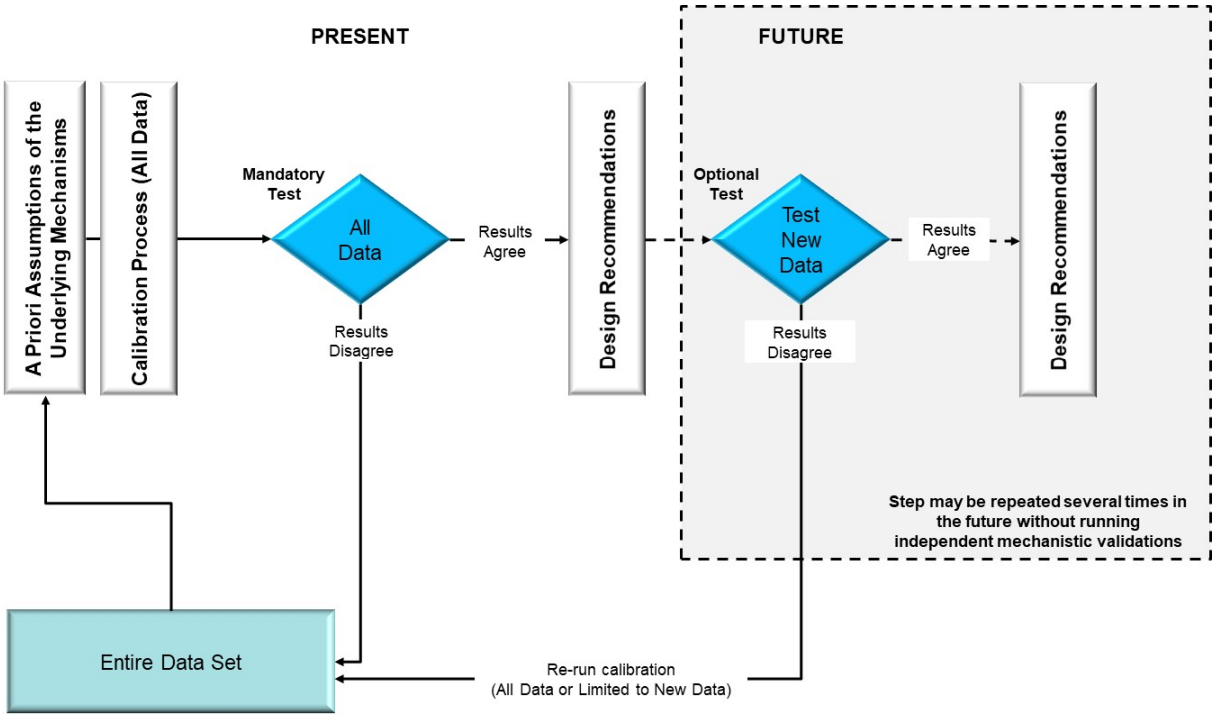

Figure 5 Flow chart describing the calibration approach generally used in rock engineering design. The process is missing the validation step due to the deliberate choice by engineers not to split the available information into calibration and testing data

Figure 6 proposes a different process, which requires the available data to be randomly split between a calibration set (C_{Old}) and a validation set (V_{Old}). The calibration set is used to train the model, and the calibrated model is later tested against the validation set. Should the model fail the validation step, the calibration process is repeated by shuffling and subdividing the data into two new calibration and training sets. Note that this new and improved process still leads to a validated model with clear temporal limits. As discussed earlier, the design assumption that what occurred in the past will repeat itself in the future conflicts with the challenge of rock mass predictability. Only by continuously testing the model using newly acquired independent data can we extend the temporal boundary of the model's predictions.

There is another temporal and cognitive barrier to the problem of calibration and validation. Should the predictions made by the model recommend a significant redesign, the new design would become the focus of the validation process. It is no longer possible to ground truth the *failure* of the old design since the conditions behind it would have been removed from the project's temporal dimensions. For example, imagine a calibrated and validated model recommending reducing the overall slope angle from 55 to 52°. Once the mine proceeds with the new design, verifying whether the slope would have remained stable at 55° becomes impossible. The need to avoid unpredictability and the lack of prototypes leads to an obedient engineering faith in the modelling results.

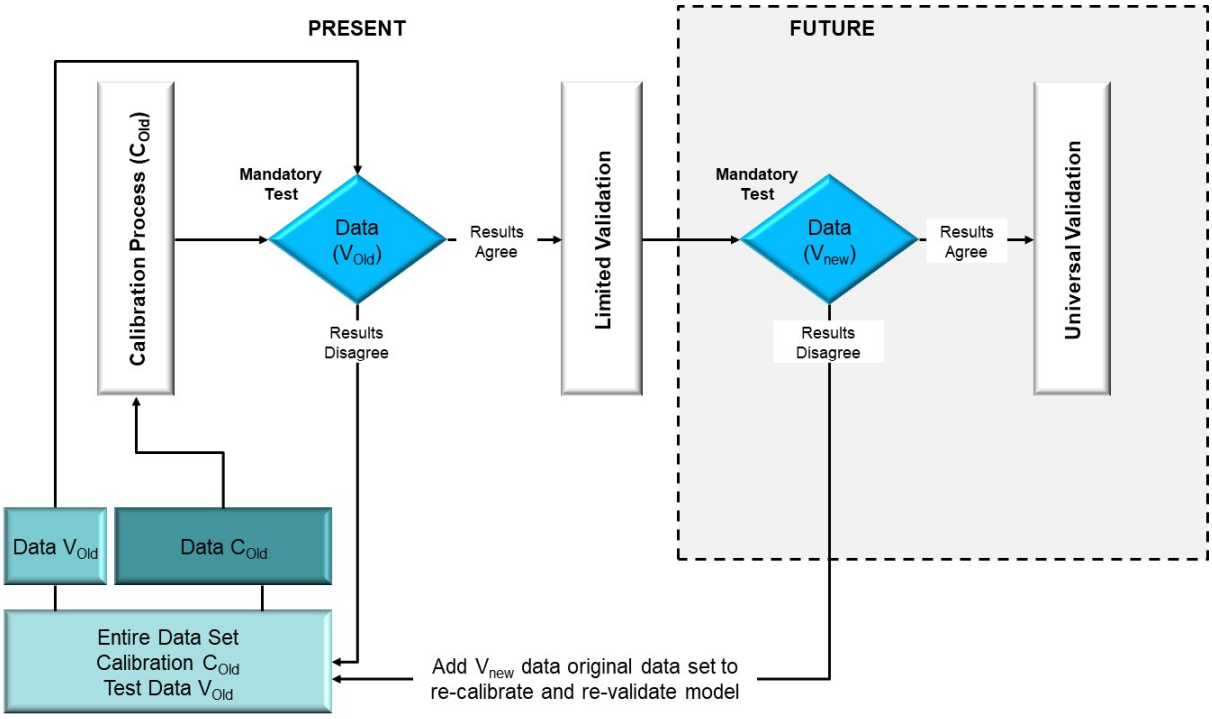

Figure 6 Recommended calibration and validation flow chart

5 Conclusion

The terms calibration and validation are widely used when discussing the results of numerical models. In this paper we have discussed the meaning of calibration and validation in the specific context of rock engineering design. In the discussion, several challenges and limitations were identified, including:

- Engineering judgement cannot be used as a substitute for proper calibration and validation processes.
- The insistence on using qualitative characterisation methods introduces a degree of subjectivity in the way we collect and manage the data used for calibration and validation.
- Differences in models' conceptualisation lead to different modelling inputs and the paradox of rock mass strength parameters that remain calibrated despite changing from model to model.
- Using the principle of back-analysis as the basis to calibrate numerical models is such that the models are not validated against an independent dataset.
- Using inductive principles to drive calibration and validation makes the models vulnerable to unpredictability (i.e. the models learn from the data and the experience available to the engineers at that time).

Acknowledgement

The authors acknowledge that some of the points raised in this paper may encounter objections. However, by ignoring the epistemic limits of calibration and validation there is the risk of letting engineering faith become the excuse behind the tendency to replace models' calibration and validation with model acceptance.

References

Elmo, D 2023, 'The Bologna Interpretation of rock bridges', *Geosciences*, vol. 13, no. 2, https://doi.org/10.3390/geosciences13020033

Elmo, D, Mitelman, A, & Yang, B 2022, 'An examination of rock engineering knowledge through a philosophical lens', *Geosciences*, vol 12, 174. doi.org/10.3390/geosciences12040174

Elmo, D & Stead, D 2021, 'The role of behavioural factors and cognitive biases in rock engineering', *Rock Mechanics and Rock Engineering*, vol. 54, no. 1, https://link.springer.com/article/10.1007/s00603-021-02385-3

Elmo, D & Stead, D 2020, 'Disrupting rock engineering concepts: is there such a thing as a rock mass digital twin and are machines capable of learning rock mechanics?', in PM Dight (ed.), *Slope Stability 2020: Proceedings of the 2020 International Symposium on Slope Stability in Open Pit Mining and Civil Engineering*, Australian Centre for Geomechanics, Perth, pp. 565–576, https://doi.org/10.36487/ACG_repo/2025_34

Harrison, JP 2017, 'Rock engineering design and the evolution of Eurocode 7', *Proceedings of EG50 Engineering Geology and Geotechnics Conference*.

Hoek, E 1994, 'Strength of rock and rock masses', *International Society for Rock Mechanics News Journal*, vol. 2, no. 2, pp. 4–16.

Kahneman, D 2011, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Kalenchuk, KS 2019, 'Canadian geotechnical colloquium: mitigating a fatal flaw in modern geomechanics: understanding uncertainty, applying model calibration, and defying the hubris in numerical modelling', *Canadian Geotechnical Journal*, vol. 59, no. 3, pp. 315–329, https://doi.org/10.1139/cgj-2020-0569

Karimi, L, Elmo, D & Stead, D 2020, 'An investigation of the factors controlling the mechanical behaviour of slender naturally fractured pillars', *Rock Mechanics and Rock Engineering*, vol. 53, no. 11, pp. 5005–5027, https://link.springer.com/article/10.1007/s00603-020-02203-2

Ross, B 2017, *Rise to the Occasion: Lessons From the Bingham Canyon Manefay Slide*, Society for Mining, Metallurgy & Exploration, Littleton.

Read, J & Stacey, P 2009, *Guidelines for Open Pit Slope Design*, CSIRO Publishing, Melbourne.

Shapka-Fels, T, & Elmo, D 2022, 'Numerical modelling challenges in rock engineering with special consideration of open pit to underground mine interaction', *Geosciences*, vol. 12, no. 5, https://doi.org/10.3390/geosciences12050199

Taleb, N 2010, *The Black Swan: The Impact of the Highly Improbable*, Random House, New York.

Yang, B & Elmo, D 2022, 'Why engineers should not attempt to quantify GSI', *Geosciences*, vol. 12, no. 11.

Yang, B, Mitelman, A, Elmo, D & Stead, D 2021, 'Why the future of rock mass classification systems requires revisiting its empirical past', *Quarterly Journal of Engineering Geology and Hydrogeology*, vol. 55, https://doi.org/10.1144/qjegh2021-039