SCIENTIFIC TEST & ANALYSIS TECHNIQUES
CENTER OF EXCELLENCE

# On the Construction of a Standard Scale for Model Validation Referent Authority

July 2023

Nicholas Jones, Ctr


To develop and apply independent, tailored Scientific Test & Analysis Techniques solutions to Test and Evaluation that deliver insight to inform better decisions.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, AFIT.ENS.STATCOE@us.af.mil
Call, 937-255-3636 x4736

Technical Reviewers:
Corinne Weeks
Aaron Ramert
Wayne Adams


Scientific Test & Analysis Techniques Center of Excellence
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio



Version: 1, FY23

## Abstract

A key problem of creating a standardized set of Model Validation Levels (MVLs) is the establishment of a standard scale for referent authority. Referent authority is the second pillar of validation and refers to the strength of credibility of a referent's claim to be a high-fidelity representation of reality. To objectively state the level of authority which the model can be said to have on the basis of comparison to the referent, the authority level of the referent(s) which form basis of trust must be objectively understood. This paper will examine the key requirements for a scale of the relative authority of referents to be useful, examine three options for constructing a scale, and recommend the most practical option for constructing an objective, standardized referent scale that facilitates the evaluation and comparison of MVLs.

*Keywords: referent, authority, trust, model validation, Digital Engineering*

## Table of Contents

## Introduction: The Need for Referent Authority

One of the key problems of creating a standardized set of Model Validation Levels (MVLs) is the establishment of a standard scale for referent authority (Ahner, et al., 2021). Referent authority is the second pillar of validation and refers to the strength of credibility of a referent's claim to be a high-fidelity representation of reality. The MVL provides a rigorous, data-driven means of comparing a model to a body of referent data on the basis of similarity in fidelity and scope. But to objectively state the level of trust we can justifiably place in the model based on this comparison, the authority level of the referents which form basis of trust must be objectively understood (Provost, Weeks, Jones, & Sieck, 2022).

Model validation is defined by Department of Defense (DOD) Instruction (DoDI) 5000.61 as "the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model" (2018). To achieve this, the process of model validation aims to establish the representativeness of a model relative to a referent. A referent is a body of data that serves as a trusted representation of reality, against which the model is validated. To objectively establish a standard of comparison between validations, the relative authority between referents (i.e., their relative trustworthiness as representations of reality) must also be determined. This requires a referent authority scale that relates different referents to each other. Such a scale would allow model validity to be quantified not only by how closely the model represents one or more referents, but also in terms of how representative those referents are of reality. This paper will examine the key requirements for a scale of the relative authority of referents to be useful, examine three options for constructing a scale, and recommend the most practical option for constructing an objective, standardized referent scale that facilitates the evaluation and comparison of MVLs.

## Background: The Foundations of Shared Trust

For validation to serve as a means for building trust in models, it must be based on a comparison of the model to authoritative referents. Because a referent is assumed to be a representation of reality, we believe it has some authority. However, not all referents are equally authoritative. For example, a set of recorded performance data for a system and the judgement of a subject matter expert (SME) may both be drawn from observation of the same real-world event, but the data is more objective than the SME, and is considered to be a closer representation of the real world. The amount of trust placed in a referent (i.e., it's level of authority) determines the maximum amount of trust that can be placed in a model validated against it.

The establishment of trust is essential both for a standardized referent scale and for the establishment of authority via validation in general. Establishing trust is well-studied as a key problem in numerous fields, from abstract philosophy to computer security. A practical example is certificate authentication of network communications. Authentication is the process of verifying the identity of a network entity or originator of a message and is a critical concern in cybersecurity. Typically for network connections, authentication is done using key-based encryption architectures, wherein unique keys are associated with users or network entities and used to encrypt messages as proof of identity (Stewart, Chapple, & Gibson, 2012; Hall, 2013). To ensure these keys are unique and only associated with a particular entity, and are therefore a trustworthy indicator of identity, they are backed by a central certificate authority. The certificate authority vouches for the uniqueness and authenticity of keys and their association to the users. Despite the mathematical complexity and rigor of the protocols involved in encryption and key exchanges, the ultimate question of whether an authentication is trustworthy comes down to trust in the certificate authority: the users of a given key architecture must choose to

trust the certificate authority, and the certificate authority must conduct its business honestly. This example also illustrates another critically important concept: shared trust in the authentication architecture is dependent on users trusting a *common* certificate authority. This commonality is key because it enables the certificate authority to act as a standard baseline of trust in the architecture for every user, so that every user can trust the certificates of any other user whose certificates also come from that certificate authority.

This practical example contains three core aspects of the establishment of shared trust, which are just as central to MVL referent authority as to authentication in cybersecurity: the chain of authority, the necessity of trust, and the necessity of standardization. Trust in an entity may be established by comparison to some other entity, but regardless of the number of steps in the chain of comparison, the first step must always be a choice to take some baseline as truthful and trustworthy; and that baseline of trust must be standardized, so that *all* users of a system can trust both the baseline and any claims to authority made by comparison to it.

## Methods and Alternatives for Building a Referent Authority Scale

In the MVL framework, trusted referents are the baseline of authority. However, we have noted that there are numerous types of referents, and some are generally considered more trustworthy or authoritative than others. For the MVL framework to objectively handle referent authority, referents must be assigned a quantifiable level of trust. Furthermore, in order for the validity of a model to hold from one user and use case to the next, the referent scale must be standardized throughout the DOD. One possible method to establish a standard scale for quantified trust in a referent is to leverage an established and widely accepted scale for data authority or maturity, such as Technology Readiness Levels (TRLs). Alternatively, a scale of referent authority could be solicited from the community of interest, on their own authority. Finally, the authority of referents could be derived on a per-referent basis by data comparison to each other, as sources of authority. The rest of this paper will explore these three options.

### *TRL-based Referent Scale*

TRLs are a standard tool for assessing the maturity of new technologies in Government acquisition and development programs (Government Accountability Office, 2020). TRLs originated in NASA in the mid-1970s on the basis of engineering subject matter expertise and were first formally standardized in a NASA publication in 1995 (Mankins, 1995; Mankins, 2009). Typically, TRLs are applied to technologies to indicate their technical maturity (e.g., a lab-scale prototype wing design demonstrated in a wind tunnel might have TRL 5). On the other hand, a referent is typically a body of data, not a technology (e.g., the data collected from testing the prototype wing in a wind tunnel), and so should not be thought of as having a TRL itself. However, a referent's authority level could be inferred based on the TRL of the system that produced it (e.g., the wind tunnel prototype test was sufficient to decide if the wing design met TRL 5, so the data collected that supported that conclusion would be a referent with an authority level of 5). This line of inference supports the assignment of authority levels to referents based on the maturity (or nearness to real-world operational expectations) of the systems from which they were derived. Extending this line of reasoning across all TRLs allows easy generation of an ordered referent scale from a source that is already widely used and accepted as standard in the Test & Evaluation (T&E) community. Such a scale is shown in Table 1.

Ultimately, if models will be used to make decisions about real-world operations for DOD systems, then operational data, while potentially noisier than more controlled referents, is the most authoritative referent for model validation because it is what the warfighter has experienced.

**Table 1**

*Referent Authorities from Relevant TRLs*

| Authority Level | Relevant Referent |
|---|---|
| 1 | SME Judgement |
| 2 | First Principles/Physics Predictions |
| 3 | Component Lab Test Data |
| 4 | Integrated Component Lab Test Data |
| 5 | Lab-Scale System Test Data |
| 6 | HWIL & SWIL Data |
| 7 | Prototype Field Test Data |
| 8 | Live System Test Data |
| 9 | Operational Real-World Data |

While the simplicity of the TRL-derived method for establishing a referent scale is convenient, it does come with some shortfalls. One of the most readily apparent is the vagueness of the referent descriptions. This vagueness comes from the imprecise definitions of technologies at a given TRL level, which create ambiguity about the most realistic type of referent that might be derived for that technology. While the TRL-derived referent scale is consistent with TRLs, not all listed referents may be applicable to every use case, limiting the interpretability of some values in certain cases. The completeness of the scale may also be a concern, as it might be argued that some referents exist that are difficult or impossible to categorize according to the nine types that might be derived from TRLs. However, a potential solution to this problem comes from the same logic which allows the derivation of the TRL-based scale, as any referent that might at first be difficult to classify could be assigned a rank within the same scale on the basis of the TRL it would most likely support.

A further point of difficulty comes from the ordinal nature of the TRL ranking scale: TRLs have a specific order but do not have quantitative relationships between the ordered values, and the same is true of a TRL-derived referent scale. To enable the scale to be mathematically meaningful, an additional quantitative weighting scale must be introduced to enable the referent ranking levels to be related to one another. As there is little data on what the relationships should actually be, the definition of such a weighting scheme is somewhat arbitrary.

***Community-Informed Referent Scale***

A referent scale may be defined by soliciting feedback from the T&E, Modeling & Simulation (M&S), and Digital Engineering (DE) communities of interest regarding types of referents commonly used, and their perceived degree of authority. This method has the advantage that the scale generated may have immediate buy-in from the community of interest because it was built from their inputs. Assuming the community members providing input are knowledgeable across the scope of interest, a scale derived by this method can be expected to accurately reflect the T&E community's perceived trustworthiness of the referents they commonly encounter in practice.

However, this method also comes with several disadvantages. One is the potential for the scale to become very large and unwieldy if users are unwilling to allow similar referents to be grouped together. While a highly specific referent scale with many levels may satisfy claims from the community regarding small differences between referents, over-specificity may cause difficulties in actually building a standardized scale. In particular, disagreements about the relative rankings of similar referents may prevent the creation of a standard scale, or result in an unwieldy scale that is difficult to use. Thus, utility calls for a trade-off in simplicity vs. specificity. If users are

unable to agree on what referents should be included or on reasonable groupings by preestablished type, then multiple scales may result leading to ambiguity when comparing the validity of models. The existence of multiple referent scales, potentially pertaining to the same subjects, could lead to subjectivity in selecting a referent scale for validating a given model. This could lead to a loss of standardization of the referent scales, with each community or user building their own, which would defeat the purpose of building a referent scale by eliminating the ability of the scale to build shared trust in models.

A further practical difficulty is securing the input of the communities of interest. STAT COE has attempted to solicit feedback on potential referent scales with minimal response from the T&E community. Even if the community were to provide stronger feedback or assist with building referent scales, it would still be difficult to ensure that input had been solicited from or provided by all interested parties. At a minimum, the scale would require broad input from those performing validation of M&S for T&E purposes, testers and test range owners, and program personnel associated with the broad range of defense technologies which might be modeled. As a final difficulty, a scale (or scales) built using this method would again only be ordinal in nature, so as with the TRL-derived scale it would still require the creation of a somewhat arbitrary, secondary weighting scheme to enable its use in calculating an MVL.

***Direct Referent Comparison***

The final method by which one might compile a referent authority scale is by collecting data from potential referents and directly comparing them to each other. This has the advantage of being the most objective mechanism for the establishment of referent authority, because it is completely data-driven. In essence, the process would involve validating each referent against some other referent which represents the same scope. Ideally this validation would use the MVL method of objective fidelity and scope comparison to maintain consistency and comparability of results. The result would be an individual authority score for each referent in each given scope of use, which would then serve as that referent's authority level when calculating an MVL using that referent.

However, this method would also introduce several issues, some of which stem from the problem of establishing trust and authority, as discussed in the Background section. First, to validate referents, we would still require some set of absolutely trustworthy "golden" referents to validate all others against. As in the previous options, the most authoritative referents for model validation would be data collected on real-world operations, but rather than simply being the most authoritative referents, the "golden" referents would then be the *only* source of authority, from which all other referent authority would be derived. This would create the issue that much more real-world data (or other high-authority "golden" referent data) would be required, because it would be necessary to have a sufficient pool of authoritative data against which to establish the authority level of other referents before they can be used. This in turn might render other referents superfluous, since the existence of sufficient quantities of higher authority data would mean that models could simply be validated against the "golden" referents instead. There may still be value in calculating the MVL of a model, to support using it as a lower-level referent which could generate representative data for situations that require extremely large referent data volumes, or data at specific points in the factor space that are not represented in the "golden" referent set. However, even in these cases, users of a direct-comparison referent scale would encounter a new problem. While this method of establishing authority would be inherently numeric, and not require an additional weighting scale, the lower authority levels would no longer be tied to a labeled scale, making interpretation of the authority of the lower-level referents much less clear.

## Discussion

The strengths and weaknesses mentioned above for the three possible methods of building a referent scale are summarized in Table 2. Before developing these methods, we explored an example that established three core requirements for shared trust. Those were the chain of authority, the necessity of trust, and the necessity of standardization. The chain of authority means that authority is inherited from a central source of authority. All of the proposed scale options would support the chain of models' authority being inherited from referents. The key concerns in selecting the best scale are ensuring that the scale is suitably rigorous to satisfy the necessity to invest trust in it, and ensuring that it can be standardized in order to support shared trust. Additionally, the scale selected must satisfy these constraints while being practical to construct and use.

**Table 2**

*Referent Scale Method Comparison*

| Referent Scale Method | Strengths | Weaknesses |
|---|---|---|
| **TRL-Based** | • Simple & familiar<br>• Easy to construct<br>• Easy to standardize<br>• Authority levels are easily interpretable | • Vague<br>• Requires additional weights<br>• May not apply completely to all cases<br>• May not be complete |
| **Community-Informed** | • Immediate buy-in<br>• Authority levels are easily interpretable | • Difficult to get input from all stakeholders<br>• May be difficult to get agreement<br>• Potentially very complicated<br>• Difficult to standardize<br>• Requires additional weights<br>• Difficult to ensure completeness |
| **Direct Referent Comparison** | • Standard, objective rating for every referent<br>• No additional weights required | • Requires large body of high-authority "golden" referent data<br>• Lower-level referents redundant<br>• Requires all but "golden" referents to be validated before use<br>• "Golden" referent availability will limit completeness of referent scope coverage<br>• Authority of lower-level referents difficult to compare or interpret |

The method of constructing a scale by direct referent comparison is attractive due to its objectivity, but it also suffers from some key weaknesses, with the chief among them being the need for high authority "golden" referent data against which to validate all other referents. The drive to use lower-level referents comes from the frequent unavailability of higher authority data in real scenarios. Unfortunately, this results in a paradox, because it is unlikely that "golden" referent data would be available in most practical use cases, meaning that the authority of lower-level referents could not be established. This practical limitation makes the direct comparison method useless for assessing the validity of modeled performance for all but the most mature systems, and therefore unsuitable for development of a standard scale for referent authority in model validation.

While the community-informed scale may seem like the next best option, and a logical choice to ensure shared buy-in, the difficulty of securing community engagement encountered in practice has actually made it more difficult to construct. Furthermore, as noted before, any scale developed by this method would probably require multiple revisions to secure buy-in from any stakeholders who do not participate in its initial construction. The likely need for revision of a community-informed scale also increases the likelihood of producing multiple disconnected scales, as opposed to a single standard scale. Because of these difficulties, the community-informed scale is also an impractical option for developing a standardized referent authority scale.

Of the three options discussed, the TRL-based scale has the most benefits, due to its ease of construction, built-in standardization, simplicity, interpretability, and connection to a familiar framework, all of which should promote easy adoption. By default, it would generate a single cohesive scale and provide an interpretable ranking structure for lower-level referents. The most concerning weakness of the TRL-based scale is the need for a secondary weighting scale (a weakness shared with the community-informed scale), which may be questionable from a perspective of rigor. Several authors have advanced algorithms for constructing simple weighting structures to appropriately reflect non-linear scaling between levels in ranked scales, like TRLs (Barron & Barrett, 1996; Conrow, 2011; Kunsch, 2019). There is disagreement between these authors as to which method is the most effective, both in the case of TRLs and as a general-purpose method. However, all share a structure which gives increasing weight to higher levels, reflecting the large increases in investment and technical maturity that are expected at each successive TRL relative to the previous one. Given the consensus in the literature on the general trend of non-linear growth across levels, but the disagreement on a single best algorithm, it is reasonable to construct a referent weighting scale that is tailored in support of MVLs, using a method that incorporates increasing growth across levels but is calibrated to work with the other mathematical elements of the MVL framework. While the potential for incompleteness of the scale may at first also be concerning, the openness of the referent type definitions and traceback to the TRL framework should facilitate integration of additional referent types as they are identified. TRL-based referent types, their uses, and their interpretations will be discussed in the upcoming STAT COE best practice, “Model Validation Levels: Methods and Implementation” (Weeks, Provost, & Jones, 2023). Because of its numerous strengths and the ready availability of mitigations for its greatest weaknesses, the TRL-based scale is the most practical option for a standardized scale of referent authority.

**Conclusion**

A standardized referent scale is critical to establishing a basis for shared trust in MVLs that will support objective model validation and model reuse in a DE environment. While several options might be considered to produce such a scale, the TRL-based method is the most practically achievable option which fulfils all three requirements for establishing shared trust. The TRL-based method will produce a scale that is simple, interpretable, standardized, familiar enough to promote quick adoption, and supports rigorous quantification of an MVL. The STAT COE has pursued this option to develop a referent authority scale with calibrated weights for use in MVLs as described in “Elements of a Mathematical Framework for Model Validation Levels” (Provost, Weeks, Jones, & Sieck, 2022). Additional information on the development and use of the scale, the associated weighting scheme, practical considerations for the collection of referent data, and clarification on other potential points of ambiguity or confusion with the use of a TRL-based referent scale will be given in the upcoming STAT COE best practice, Model Validation Levels: Methods and Implementation” (Weeks, Provost, & Jones, 2023).

## References


Ahner, D. K., Jones, N., Key, M., Adams, W., Burke, S., & Weeks, C. (2021). *A Conceptual Framework for the Establishment of Model Readiness Levels.* Scientific Test & Analysis Techniques Center of Excellence (STAT COE). Retrieved from https://www.afit.edu/images/pics/file/STAT%20COE%20MRL%20Whitepaper_A%20Conceptual%20Framework%20for%20the%20Establishment%20of%20Model%20Readiness%20Levels.pdf

Barron, F. H., & Barrett, B. E. (1996). Decision Quality Using Ranked Attribute Weights. *Management Science, 42*(11), 1515-1523.

Conrow, E. H. (2011). Estimating Technology Readiness Level Coefficients. *Journal of Spacecraft and Rockets, 48*(1), 146-152.

DoD Instruction 5000.61. (2018). *DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A), Change 1.* United States Department of Defense.

Government Accountability Office. (2020). *Technology Readiness Assessment Guide: Best Practices for Evaluating the Readiness of Technology for Use in Acquisition Programs and Projects.* Government Accountability Office.

Hall, K. (2013). *Standards and Industry Regulations Applicable toCertification Authorities.* Trend Micro, Inc. Certification Authority Security Council (CASC). Retrieved July 20, 2023, from https://casecurity.org/wp-content/uploads/2013/04/Standards-and-Industry-Regulations-Applicable-to-Certification-Authorities.pdf

Kunsch, P. L. (2019). *A critical analysis on Rank-Order-Centroid (ROC) and Rank-Sum (RS) weights in Multicriteria-Decision Analysis.* Vrije Universiteit Brussel. Retrieved from https://researchportal.vub.be/en/publications/a-critical-analysis-on-rank-order-centroid-roc-and-rank-sum-rs-we

Mankins, J. C. (1995). *Technology Readiness Levels.* White Paper, National Aeronautics and Space Administration (NASA), Advanced Concepts Office, Office of Space Access and Technology.

Mankins, J. C. (2009). Technology readiness assessments: A retrospective. *Acta Astronautica, 65*, 1216-1223.

Provost, K., Weeks, C., Jones, N., & Sieck, V. (2022). *Elements of a Mathematical Framework for Model Validation Levels.* Scientific Test & Analysis Techniques Center of Excellence (STAT COE). Retrieved from https://www.afit.edu/images/pics/file/0930_ProvostMVLBP_2_2.pdf

Stewart, J. M., Chapple, M., & Gibson, D. (2012). *CISSP Study Guide* (Sixth Edition ed.). Indianapolis: John Wiley & Sons, Inc.

Weeks, C., Provost, K. J., & Jones, N. (2023). *Model Validation Levels: Methods and Implementation.*