

Quantifying Case Strength Through Evidence, Judicial Response, and Legal Consequence

Christopher R. Harrison

Analyzed Discovery Ltd., Nevada, United States

chris@analyzeddiscovery.com

Methods research article October 2026

Abstract

Criminal defense assessment requires judgments about what the record supports, how the forum is likely to respond, and what obtaining relief would accomplish. This article presents a formal additive scoring model that makes those judgments explicit and combines them under a public numerical specification. Source-linked component ratings produce an issue index, logistic scaling supplies a bounded display, and consequence weighting summarizes the developed defense position. The design is defended through traceability, separation of distinct judgments, and mathematically interpretable aggregation. Symbolic derivation and exhaustive analysis of 5,508 permitted profiles establish the model's ordering, sensitivity, and aggregation properties. Across 11,376,774 cross-type profile comparisons, removing the structural adjustment produces strict reversals in 5.27% of pairs, while removing all type adjustments produces strict reversals in 20.19%. These results quantify the influence of declared assumptions over the design space. The analysis also derives score-sensitivity bounds and the conditions under which issue subdivision preserves the aggregate. The model supplies a reproducible framework for examining how factual judgments, institutional assumptions, and consequence priorities determine assessed defense strength. Rating reliability and judicial outcome validation remain empirical tasks. The contribution is a fully specified scoring rule, an account of its design choices, and reproducible mathematical results available for independent examination.

Keywords. criminal defense; legal judgment; structured assessment; composite indices; sensitivity analysis; formal models

Contents

1	Introduction	3
2	The construct and its boundaries	5
3	Constructing the assessment	6
4	Analytical study	9
5	Analytical results	11
6	Substantive choices in specification	16
7	Evaluation and validation	19
8	Discussion	21
9	Conclusion	23
A	The disclosed numerical specification	24
B	Derivations supporting the sensitivity analysis	25
	Acknowledgments and declarations	25
	References	26
C	A Trial-Court Benchmark for Defense Issue Assessment	28

1 Introduction

Criminal defense counsel routinely assess three different questions. What proposition does the available record support? How is the relevant court likely to treat that proposition under the governing law and procedural conditions? What would success on the issue accomplish for the defense? Those judgments can point in different directions. A well-supported objection may confront an unfavorable institutional setting. A difficult suppression argument may concern the prosecution's central evidence. A highly persuasive impeachment point may have little effect on an independently corroborated charge.

A scoring method must preserve those distinctions while explaining how they contribute to an overall assessment. The problem has two levels. At the issue level, evidence and conditions of relief must be combined into an assessment of a specified defense proposition. At the case level, issues with different consequences must be summarized under an explicit priority rule. Combining everything into a single intuitive conclusion leaves the relationship between those judgments unstated. A formal model makes that relationship an object of inspection.

This article presents such a model. Each developed defense issue receives a type-specific reference position, an evidentiary rating, a forum and authority rating, and a bounded contextual adjustment. Addition produces an issue index. A logistic transformation places the index on a common 0–100 display. A consequence-weighted mean of the issue indices summarizes the developed defense position. The model therefore formalizes a sequence of professional judgments and the rule for combining them. It specifies what each number represents, what supports its assignment, and how changing it changes the result.

The methodological claim is that defense assessment can be made explicit through a compact scoring architecture whose assumptions and consequences are open to examination. The design places factual support, institutional reception, and the consequence of relief in identifiable roles. Source-linked reasons expose the basis of a rating. A fixed combination rule defines the numerical tradeoffs. Analytical results establish which conclusions follow from the rule and which depend on its chosen settings. These features provide the basis for defending the method as a research instrument and evaluating competing specifications.

Research on judgment supports investigating this approach. Dawes (1979) examined linear judgment models whose weights are assigned without optimizing their fit to observed outcomes. Grove et al. (2000) found that mechanical combination often matched or exceeded clinical judgment in the domains studied. The relevant methodological precedent is the use of explicit combination rules even when judgment supplies the inputs. Whether the present legal specification improves ranking, forecasting, or professional decisions requires its own evidence.

The paper makes three contributions. It specifies an issue-scoring and case-aggregation model with an explicit account of the intended construct. It derives the mathematical properties needed to interpret the scores, including sensitivity bounds and conditions for aggregation invariance. It exhaustively evaluates the disclosed rating grid to show how type assumptions, component spacing,

and issue formulation affect the output. The reference settings are held fixed throughout the analysis. Alternative specifications isolate the effects of particular choices.

Public disclosure is part of the method’s research value. It makes the parameters, source requirements, equations, and reported results available for independent implementation and criticism. The present contribution is the specified assessment architecture and the analysis reported here. Addition, logistic scaling, and weighted averaging are established mathematical operations. Judicial outcome validation is a subsequent empirical test of the disclosed model, while this paper supplies the model and its mathematical examination.

1.1 Relationship to existing methods

Multiattribute decision methods provide a relevant foundation for decomposing judgments. Edwards and Barron (1994) describe practical additive models with explicit assessment of attribute scales, ranges, and tradeoffs. Luce and Tukey (1964) show, at the level of measurement theory, that additive interval representations require substantive conditions on comparisons. These traditions support examining an explicit combination rule while directing attention to what its numerical increments mean. They do not establish that legal rating categories have equal intervals or that a unit of evidentiary merit should exchange exactly for a unit of forum assessment. (Edwards & Barron, 1994; Luce & Tukey, 1964)

Legal evidence models preserve structures that a scalar necessarily compresses. Fenton, Neil, and Lagnado (2013) use Bayesian networks to represent hypotheses, evidence accuracy, and dependencies. Bex et al. (2010) connect evidential arguments with explanatory narratives, including support, contradiction, and gaps. The present framework records a judgment about a developed issue rather than reconstructing the full inferential network. Its source-linked reasons are therefore essential to interpretation. A numerical merit rating cannot substitute for examining the supporting and defeating evidence. (Fenton et al., 2013; Bex et al., 2010)

Litigation decision analysis already combines explicit issue assessment with consequences. Victor (2015) describes conditional decision paths, case valuation, and the costs and uncertainties relevant to settlement. A consequence-weighted mean of issue indices has a different target. It describes the assessed issue set without specifying action alternatives, joint outcome probabilities, or a utility function. It therefore does not inherit the decision-theoretic interpretation of a litigation decision tree. (Victor, 2015)

Structured professional judgment must also be distinguished from mechanical aggregation. In one forensic risk study, structured final judgments supplied predictive information beyond a numerical instrument score. That result concerns another domain, but it illustrates why the relative merits of professional synthesis and a fixed combination rule require evaluation. Neither the clinical literature nor research on simple linear models establishes which approach improves defense assessment. (Douglas et al., 2003)

Composite-indicator research supplies the closest precedent for the present analytical design. Saisana, Saltelli, and Tarantola (2005) study how modeling choices affect indices and rankings.

Paruolo, Saisana, and Saltelli (2013) distinguish nominal coefficients from actual influence across a distribution of observations. Accordingly, this article examines explicit changes to assumptions and reports the conditions under which conclusions remain stable. Stability within a specified analysis is evidence about that specification’s behavior, rather than evidence that its ratings or predicted outcomes are correct. (Saisana et al., 2005; Paruolo et al., 2013)

2 The construct and its boundaries

2.1 Defense issue strength

Defense issue strength is the degree to which a specified defense proposition, developed from an identified record, has evidentiary support and is positioned to obtain a specified legal or tactical benefit in a specified forum and procedural posture. An issue may concern suppression, exclusion of evidence, an element of an offense, or impeachment. Its description must identify both the proposition and the intended benefit. “Witness credibility,” without a more specific claim and purpose, is not a sufficiently defined unit.

The assessment is conditional on the available record and the formulation of the issue. It does not describe every fact that exists or every defense that might be developed. A higher index expresses a more favorable assessment under the declared conventions. Whether the ordering agrees with independent professional assessments or observed outcomes remains an empirical question.

Strength and consequence require separate treatment. A minor impeachment point may be well supported while offering little practical benefit. A difficult motion may have substantial consequences if granted. Combining those dimensions too early makes it unclear whether an issue is described as strong because it is well supported or because success would matter. The framework therefore constructs an issue-strength index first and introduces consequence weights only when summarizing a set of issues.

Neither dimension alone determines the best next use of counsel’s time. Deadlines, preservation requirements, investigation costs, and the potential value of obtaining additional evidence also affect strategy. Similarly, issue strength is not synonymous with acquittal probability, sentencing exposure, plea value, or the expected benefit of filing a motion. Those quantities require additional definitions and models of the processes that produce them.

2.2 A formative index

The proposed assessment is a formative composite. Its components constitute the operational index. They are not interchangeable manifestations of a single latent trait. Strong factual support can coexist with an unfavorable procedural setting. Consequently, a high correlation among every component is not a prerequisite for a useful assessment, and low internal consistency would not by itself establish failure. Measurement theory distinguishes indicators that constitute an assessment from indicators that reflect an underlying variable. (Bollen & Lennox, 1991)

This interpretation does not establish validity by definition. It directs attention to whether the components cover the intended construct, whether their ratings have defensible meanings, and whether their combination supports the intended use. Construct validation also requires relationships with observations beyond the index’s own formula. (Cronbach & Meehl, 1955)

Several relationships are testable. Within comparable issue types and procedural settings, stronger independently assessed factual support should generally accompany higher indices. Blinded counsel should tend to rank higher-scored issues more favorably. Among issues sharing a defined adjudicative endpoint, higher scores should tend to precede favorable rulings. Changes to legally irrelevant features of otherwise equivalent records should not change the assessment. These hypotheses concern the meaning of the index rather than its arithmetic alone.

2.3 The object of a case summary

A summary across issues needs its own construct definition. This article considers the consequence-weighted central tendency of the developed issues’ indices. It describes the center of a distribution under stated priorities. It does not describe the probability that at least one issue will succeed or the combined strategic value of every available defense.

The distinction has an immediate implication. Developing an additional viable issue can increase counsel’s options while lowering an average if the new issue lies below the existing mean. A summary must therefore be interpreted with the included issue set, the record boundary, and the consequence weights. Comparisons across cases or over time require attention to differences in those conditions.

3 Constructing the assessment

3.1 Unit, components, and sources

The unit of analysis is an issue with a stated proposition, requested benefit, decision maker, and procedural stage. An assessment should establish inclusion rules before calculating scores and should account for excluded or unassessable issues. A set of issues developed from one record is an enumeration under those rules, not a probability sample of cases or defendants. Issues may share a witness, document, event, or legal premise; separate entries do not make them independent observations.

For issue i , let $t(i)$ denote its assigned type. Table 1 separates the components by their intended function. The notation distinguishes two possible type adjustments so that their relationship can be examined. An implementation may instead use their combined value directly.

Table 1. Components of a structured issue assessment.

Symbol	Component	Intended role	Basis
$\beta_{0,t}$	Baseline type adjustment	Reference position associated with a route to relief	Declared convention, elicitation, or an identified estimation procedure
δ_t	Structural adjustment	Additional allowance for specified procedural or remedial conditions	Declared convention or independently supported estimate
m_i	Evidentiary merit	Support for the defense proposition in the available record	Source-linked factual assessment
g_i	Forum and authority assessment	Conditions affecting recognition or use of the proposition in the target setting	Applicable authority and documented procedural or court information
c_i	Contextual adjustment	Relevant circumstances not already represented	Identified additional record-supported consideration
w_i	Consequence weight	Relative importance of obtaining the specified benefit	Stated priority tied to the case and requested relief
L_i	Issue index	Additive combination of the strength components	Computed
S_i	Displayed score	Bounded transformation of the issue index	Computed

Forum assessment concerns more than whether a motion will be granted. For a trial proposition, it can concern the conditions for admitting, using, or establishing the proposition. Different intended benefits will require different outcome definitions in empirical evaluation.

The source of support must match the kind of claim. Factual premises should be linked to identifiable passages in the record. Legal propositions require legal authority. Declared numerical constants require an explicit methodological rationale; a document citation cannot establish their empirical magnitude. Keeping these sources distinct prevents the appearance of evidence support from substituting for support of the particular numerical choice.

3.2 Assignment and traceability

The implementation motivating this study places assessment at the end of an automated discovery-analysis report. Its scoring instructions apply the fixed rule to major defense issues already developed in the preceding analysis. They prohibit new facts or arguments in the scoring stage, require source-linked reasons, and require recomputation of displayed values. These are requirements of the disclosed specification. This article does not measure how consistently software executions satisfy them or how completely the preceding analysis identifies issues.

Issue development should precede numerical assessment. The analyst first states the proposition, its factual and legal basis, and the benefit sought. The scoring stage then expresses the assessment of that developed issue. New facts or arguments should trigger revision of the underlying issue

analysis rather than enter silently as a numerical adjustment. This sequence also makes it possible to distinguish a change in judgment from a change in the material being judged.

Each rated component should have a reason and a source reference. The issue description, type assignment, component values, consequence weight, and explicit calculation should be available together. These materials permit a reader to locate disagreement. A source may have been misread, contrary evidence omitted, a consideration assigned to the wrong component, or a rating chosen inconsistently. Traceability facilitates review; it does not establish that the underlying interpretation is correct.

Merit and forum assessment should have separate assigned subject matter. Merit concerns the evidentiary support for the proposition. Forum assessment concerns the legal and institutional conditions for obtaining the stated benefit. A rule requiring that separation cannot establish statistical independence. The same circumstances may influence both assessments, and evidentiary strength may affect how a court receives an issue. The practical requirement is to state where a consideration enters and to justify any repeated contribution.

Within the forum rationale, governing authority and observed institutional practice should remain separately traceable. Authority supplies the applicable legal conditions. Historical decisions may supply evidence about how comparable issues are treated collectively in a jurisdiction, procedural setting, and period. An observed ruling does not itself establish the correctness of the legal analysis. The present specification retains one forum rating, but its rationale should identify these different sources and any tension between them. A claimed practice pattern requires an identified evidentiary basis rather than an assumption about individual judges.

A residual contextual component requires similar discipline. It should identify a relevant consideration that the other components miss and explain its direction. An explicit zero is appropriate when an adequate review identifies no additional adjustment. A zero assigned because the necessary information is unavailable has a different meaning and should not be presented as an evidence-supported neutral judgment.

Calculations should be recomputed independently of the narrative assessment. The same inputs must produce the same numerical output, regardless of whether ratings are supplied by counsel, another analyst, or an automated system. Arithmetic consistency is a property that can be enforced deterministically. Agreement on the ratings themselves requires a separate reliability study.

3.3 Classification, missingness, and scale construction

Classification matters whenever issue types receive different adjustments. A taxonomy that combines procedural categories, such as suppression, with categories describing evidence quality can produce overlap. The same corroborated proposition may support impeachment, an element dispute, or a motion. A classification manual should explain which characteristic controls, link distinct legal uses of shared evidence, and state how unresolved classifications are handled. Where two assignments remain defensible, their effects can be reported through sensitivity analysis.

Rating scales also require substantive anchors. Adjacent numerical values should correspond to distinctions the record and applicable authority can support. Regularly spaced codes make calculation straightforward but assume that the represented steps can be treated as equal increments. Coarse categories can reduce the number of distinctions requested of a rater; they do not establish that those distinctions are reliable.

When an input lacks a defensible basis, the assessment should identify it as unresolved. Conditional scenarios or a range may be more informative than an unconditional point score. A summary based only on assessable issues must identify that subset and the omissions. An empty set has no defined mean. These reporting choices make the information boundary visible rather than allowing missingness to enter as an undisclosed rating convention.

Reproducibility requires the record version, inclusion rules, classification and rating manuals, and numerical specification. Automated rating additionally requires the model and software versions, instructions, relevant settings, and source-retrieval procedures. Reproducible arithmetic conditional on those inputs remains distinct from repeatable judgments and complete issue identification.

Supplementary Material 1 supplies a proposed operational codebook with issue boundaries, availability rules, rating anchors, classification precedence, and treatment of missing information. Those conventions have not been tested with raters and are not claimed to have governed prior reports. Separating the disclosed numerical specification from proposed operational refinements prevents a revised manual from being mistaken for evidence about an existing implementation.

4 Analytical study

4.1 Design and reference specification

The study combines symbolic derivation with complete enumeration of the reference specification's permitted component codes. A local analysis plan was written before the calculations were executed and is included with the reproducibility materials. It was not externally preregistered. There is no sampled case series, simulated judicial outcome, or claim of empirical validation.

Merit and forum assessment each have nine permitted codes, from -2.0 to $+2.0$ in half-unit increments. Context has seventeen codes, from -0.8 to $+0.8$ in tenth-unit increments. Their Cartesian product contains $9 \times 9 \times 17 = 1,377$ profiles for each of four issue types, or 5,508 typed profiles. The type adjustments, their component split, and consequence weights are disclosed in Appendix A. Enumeration retains every permitted numerical combination, including combinations that may be uncommon, legally incoherent, or excluded by a future operational codebook. It does not impose a probability distribution on actual issues.

The cross-type ordering analysis compares every profile of one type with every profile of each other type. Each of the six unordered type pairs contributes $1,377^2 = 1,896,129$ comparisons, producing 11,376,774 comparisons overall. Profiles with the same summed ratings retain their multiplicity because they are distinct permitted assignments. The resulting fractions depend on the grid and the equal enumeration of type pairs. They are design-space summaries, not population estimates.

4.2 Comparators and sensitivity definitions

Two comparators isolate the role of the fixed settings. The first removes the structural adjustment δ_t while retaining the baseline $\beta_{0,t}$. The second removes both type adjustments while retaining the rated components. Neither comparator is selected or assessed using court outcomes. A strict reversal requires a positive difference between two issues under one rule and a negative difference under the other. Ties in either rule are reported separately. Within-type comparisons are unchanged by these operations and do not enter the cross-type denominator.

A second analysis considers componentwise dominance. For each pair of types, the lower-offset issue must be at least as favorable on every rated component and strictly more favorable on at least one. The analysis asks whether that advantage survives the difference in type adjustments. Cross-type dominance is a diagnostic of what the offsets can outweigh; it is not assumed to settle substantive comparisons between different remedies.

Rating sensitivity allows deviations of up to one permitted step from a reference profile in each component, namely 0.5 in merit, 0.5 in forum assessment, and 0.1 in context. Each deviation is clipped at its component’s endpoints. The resulting lower and upper scores form a conditional sensitivity envelope. Opposite endpoints can differ by two steps per component. The envelope is not a confidence interval, and its construction does not assert that these disagreements occur with any particular frequency.

The aggregation analysis derives the effects of priority weights, issue addition, and subdivision. A separate ordinal-recoding counterexample examines whether the framework depends only on category order or also on the assigned numerical spacing. Recoding changes the specification and is not treated as another allowed rating under the frozen reference rule.

4.3 Computation and verification

All profile-index calculations and ordering comparisons use integer tenths. Logistic values are computed at full floating-point precision and rounded only for display. The main analysis uses Python 3.14.7 and the standard library. A direct enumeration checks the frequency representation used in the ordering calculations. A separately written verifier uses sorted totals and binary-search bounds to reproduce the reversal counts. Algebraic identities are also checked over a deterministic grid of 343 triples. These checks address implementation errors; the mathematical arguments supply the general results.

The code, analysis plan, full profile grid, comparison counts, and proposed codebook accompany the article as supplementary materials. No significance tests, outcome-performance statistics, or sampling confidence intervals are calculated. AI assistance in preparing the text, code, and technical checks is disclosed in the acknowledgments.

5 Analytical results

5.1 Additive construction and its assumptions

Let $\alpha_t = \beta_{0,t} + \delta_t$ denote the combined type adjustment. Consider the additive index

$$L_i = \beta_{0,t(i)} + m_i + g_i + \delta_{t(i)} + c_i = \alpha_{t(i)} + m_i + g_i + c_i. \quad (1)$$

This specification is compensatory. A favorable contribution can offset an unfavorable one. It assigns the same index change to one numerical unit of merit, forum assessment, or context, and it contains no interaction terms. The component scales therefore determine the tradeoffs. Equal coefficients do not imply equal substantive influence when the components have different permitted ranges or distributions.

Statistical independence is not required to compute Equation 1. Dependence instead raises questions about redundancy, double counting, and whether important combinations of conditions require interaction terms. Calling contributions additive does not establish that they are independent pieces of evidence.

The compensation assumption also has a legal boundary. Favorable evidence cannot create a remedy that governing law makes unavailable. Legal availability and necessary prerequisites must be addressed in the issue analysis. If they are contested, the score must remain conditional on the stated legal position. A numerical total cannot resolve that dispute or substitute for identifying the relief sought.

For two issues of the same type, subtracting their indices cancels the fixed adjustment.

$$L_i - L_j = (m_i - m_j) + (g_i - g_j) + (c_i - c_j), \quad t(i) = t(j).$$

Changing a type adjustment therefore leaves within-type rankings unchanged when the ratings are held fixed. If one issue is at least as favorable on every rated component and more favorable on at least one, it receives a higher index. Cross-type comparisons additionally depend on the differences between type adjustments. These properties identify which comparisons depend on the declared reference positions and which do not.

5.2 Logistic scaling

A bounded display can be obtained through

$$S_i = 100 \sigma(L_i), \quad \sigma(x) = \frac{1}{1 + \exp(-x)}. \quad (2)$$

The transformation is strictly increasing, so it preserves the ordering of the indices. Finite indices map strictly between 0 and 100; zero maps to 50; and $\sigma(-x) = 1 - \sigma(x)$ gives symmetry around

that midpoint. The slope declines toward the extremes. These are properties of the chosen display, not empirical findings about how legal strength or judicial behavior varies.

Writing $p_i = S_i/100$ gives the algebraic identity

$$\log\left(\frac{p_i}{1-p_i}\right) = L_i. \quad (3)$$

Thus the index is the logit of the displayed fraction. This does not establish that the fraction is a probability of a legal event. The quantity $\exp(L_i)$ is an implied-odds representation of the score. An increment d multiplies it by $\exp(d)$, but the rated contributions have not thereby become log-likelihood ratios. A resemblance to logistic regression or Bayesian updating does not supply an outcome distribution or a probabilistic derivation.

Other increasing transformations would preserve the same issue ordering. A common positive rescaling of the index would also preserve ordering while changing the numerical display. The logistic function is therefore a presentational choice with explicit mathematical consequences. It does not improve ranking discrimination by itself, make 50 an even chance of success, or establish that a score of 80 represents twice the strength of 40.

The derivative provides a sensitivity bound.

$$\frac{dS_i}{dL_i} = 100 \sigma(L_i) \{1 - \sigma(L_i)\} \leq 25. \quad (4)$$

An index change of magnitude $|d|$ can consequently change the displayed score by no more than $25|d|$ points. The exact continuous maximum is $100 \tanh(|d|/4)$, obtained when the two indices are symmetric about zero. Appendix B gives the derivation. A half-unit change therefore has a sharp bound of approximately 12.44 points. These bounds describe the transformation's responsiveness rather than rating error or statistical uncertainty.

5.3 Consequence-weighted aggregation

For $n \geq 1$ issues with positive weights, let $W = \sum_i w_i$. Define

$$\bar{L}_w = \frac{\sum_{i=1}^n w_i L_i}{W}, \quad O = 100 \sigma(\bar{L}_w). \quad (5)$$

The weighted mean is the unique value a minimizing $\sum_i w_i (L_i - a)^2$. The derivative of that expression is $2W(a - \bar{L}_w)$ and its second derivative is $2W > 0$. The weighted mean is therefore the center minimizing consequence-weighted squared deviations on the additive scale. This characterization does not require independent issues or a normal distribution. It does require a decision to use that scale, those weights, and squared-deviation loss.

If $o_i = \exp(L_i)$, then

$$\frac{O}{100 - O} = \exp(\bar{L}_w) = \prod_{i=1}^n o_i^{w_i/W}. \quad (6)$$

The transformed aggregate thus corresponds to a weighted geometric mean of the scores' implied odds. An arithmetic mean of the displayed scores would summarize a different scale. Neither approach is universally preferable. In particular, averaging displayed scores does not always produce a value nearer 50 than transforming the average index; the logistic curve changes curvature at zero. The summary scale must follow the intended interpretation rather than a general claim that one averaging rule is inherently superior.

The aggregate lies between the smallest and largest issue scores. Holding the weights fixed, increasing any issue index increases the aggregate, with

$$\frac{\partial O}{\partial L_i} = 100 \sigma(\bar{L}_w) \{1 - \sigma(\bar{L}_w)\} \frac{w_i}{W}. \quad (7)$$

An issue's influence consequently depends on its relative weight as well as the location of the aggregate on the logistic curve. Multiplying all weights by the same positive constant leaves the result unchanged. These properties describe the declared priorities without establishing that the priorities correspond to measured utilities or effects on case outcomes.

The weights' effect can be stated without a case sample. With one issue in each consequence category, the difference between the 3–2–1 weighted mean and the equal-weight mean is $(L_3 - L_1)/6$, where the subscript indicates consequence category. The major issue's index cancels from that difference. Whether weighting raises or lowers the summary therefore depends on the relative strength of the dispositive and tactical issues. Ordered consequence labels alone do not determine the 3–2–1 numerical tradeoff.

5.4 Dependence on issue formulation

Duplicating the entire issue set leaves the weighted mean unchanged. Duplicating one issue generally changes it. Subdividing an issue preserves the aggregate for arbitrary other issues when the subdivisions conserve both the parent's total weight and its weighted index contribution. Assigning each subdivision a new full weight does not provide that guarantee. Coincidental cancellation may preserve a particular mean without conservation, but it is not a general subdivision rule.

This dependence makes issue formulation part of the measurement procedure. Multiple evidentiary propositions supporting one legal issue should not acquire additional aggregate influence merely because they are described separately. Conversely, genuinely distinct routes to relief should not disappear into a single broad heading without explanation. A declared unit definition and related-issue coding are needed to make either practice reviewable.

There is also a compatibility limit. Arbitrary subdivision invariance cannot be guaranteed while assigning every resulting issue a full positive integer weight from 1 to 3. A weight-1 parent divided

into two identical children would require the children’s weights to sum to 1 for invariance in arbitrary surrounding issue sets. Two permitted positive integer weights cannot meet that condition. The implementation must therefore control the issue unit, restrict subdivisions, or adopt a different rule for allocating group influence. This is a mathematical constraint on the specification, not an observed defect rate.

If an additional issue has weight v and index x , then

$$\bar{L}_{\text{new}} - \bar{L}_w = \frac{v(x - \bar{L}_w)}{W + v}. \quad (8)$$

An issue below the existing mean lowers the summary. That behavior is consistent with a central-tendency construct but inconsistent with treating the summary as accumulated defensive opportunity. Even calibrated issue probabilities would not make Equation 5 the probability of any success, all successes, or acquittal. Those targets require separate models and attention to the dependence among legal and factual routes to an outcome.

5.5 Exhaustive results for the declared settings

Every type has 1,377 component profiles but only 97 distinct summed index values. Different distributions of merit, forum assessment, and context can therefore produce the same score. The components remain necessary to distinguish those assessments.

The permitted displayed ranges differ by type. Suppression spans approximately 0.06 to 90.02, evidentiary or procedural issues 0.37 to 98.20, element or contradiction issues 0.67 to 99.00, and corroborated facts 1.34 to 99.50. These floors and ceilings are consequences of the declared offsets and finite rating ranges. They are not discovered limits on legal success.

Table 2 reports strict ordering reversals. Across all cross-type comparisons, omitting the structural adjustment reverses 599,784 of 11,376,774 comparisons, or 5.27%. Omitting all type adjustments reverses 2,297,316 comparisons, or 20.19%. These fractions use all cross-type comparisons as the denominator, including pairs tied in one or both specifications.

Table 2. Strict reversals of the reference ordering under two comparator rules.

Compared types	Structural adjustment omitted	All type adjustments omitted
Suppression and evidentiary or procedural	221,118 (11.66%)	435,831 (22.99%)
Suppression and element or contradiction	196,275 (10.35%)	561,960 (29.64%)
Suppression and corroborated fact	182,391 (9.62%)	684,066 (36.08%)
Evidentiary or procedural and element or contradiction	0 (0.00%)	135,795 (7.16%)
Evidentiary or procedural and corroborated fact	0 (0.00%)	317,061 (16.72%)
Element or contradiction and corroborated fact	0 (0.00%)	162,603 (8.58%)

Each row contains 1,896,129 profile pairs. Percentages describe the enumerated grid. A strict reversal excludes ties in either specification. With the structural adjustment omitted, 106,863 pairs are tied only in the reference rule, 123,471 only in the comparator, and 26,808 in both. With all type adjustments omitted, the corresponding counts are 133,671, 167,382, and zero. The supplementary comparison file reports agreement and all tie categories for every row.

The absence of strict reversals in some rows does not mean the structural terms have no effect. The evidentiary and element categories share the same structural adjustment, so removing it leaves their differences unchanged. The 0.1 differences involving the corroborated-fact category can create or remove ties on the tenth-unit grid without moving a strictly positive difference to a strictly negative one.

The dominance analysis further characterizes the type assumptions. Each ordered lower-offset versus higher-offset comparison contains 308,448 profile pairs in which the lower-offset issue is strictly componentwise dominant. For suppression versus corroborated fact, the lower-offset issue nevertheless receives a lower index in 150,918 such pairs, or 48.93%, and ties in 7,810. The offset difference is 3.1 units. Thus the type adjustment can outweigh a favorable rating on every assessed dimension. That outcome is coherent with a model intentionally assigning different remedial conditions to different types, but it demonstrates how much substantive work the type assumptions perform.

Classification sensitivity is also substantial when ratings are held fixed. Substitution of the corroborated-fact adjustment for the suppression adjustment shifts the index by 3.1 and can change the displayed score by approximately 64.96 points on the permitted grid. This is a diagnostic of alternative coding, not a claim that these legal classifications are interchangeable. It makes unresolved overlap between quality-based and function-based categories material to interpretation.

5.6 Rating envelopes and numerical spacing

At an interior profile with index zero, the simultaneous one-step sensitivity analysis permits indices from -1.1 to $+1.1$ and displayed scores from approximately 24.97 to 75.03. Each type has a permitted profile attaining this full envelope. Its 50.05-point width spans opposite deviations from the reference assessment. A single simultaneous upward step from the central profile instead moves the score from 50.00 to 75.03. The distinction prevents the envelope from being mistaken for the effect of one rater making one adjacent reassignment.

For two fixed-type issues whose permitted index envelopes do not overlap, their order survives every rating change within those envelopes. When envelopes overlap, the sensitivity analysis cannot establish a strict order for every allowed combination. This provides a practical reporting rule. A point ranking can be accompanied by a statement of whether it survives the declared disagreements, without assigning probabilities to those disagreements.

The ordinal-recoding analysis shows why category order alone is insufficient. Consider two same-type profiles with context zero and core ratings $(2, -0.5)$ and $(1, 1)$. Their sums are 1.5 and 2.0. Recode each core scale through the strictly increasing function $f(x) = x|x|/2$, which preserves zero and the endpoints -2 and $+2$. The sums become 1.875 and 1.0, reversing their order. This

counterexample changes the coding convention. It demonstrates dependence on interval spacing rather than instability under another rating allowed by the original half-step rule.

6 Substantive choices in specification

6.1 Distinguishing evidentiary support from conditions of relief

Legal doctrine supplies reasons to separate evidentiary support from the conditions governing a remedy. In *Franks v. Delaware*, the Supreme Court required a substantial preliminary showing of a knowing or reckless material falsehood to obtain a hearing challenging an affidavit; ultimate relief requires additional findings. In *United States v. Leon*, the Court recognized circumstances in which objectively reasonable reliance on a warrant prevents exclusion despite a warrant defect. These examples support assessing factual merit and remedial conditions separately. They do not establish a numerical penalty for suppression or a common difficulty across all suppression questions. (*Franks*, 438 U.S. 154, 155–156, 171–172 (1978); *Leon*, 468 U.S. 897, 922–923 (1984))

An element dispute has a different structure. The prosecution’s obligation to establish the facts necessary for criminal liability beyond a reasonable doubt distinguishes it from a defense request for an exclusionary remedy. That allocation of the burden does not imply a 50% starting probability that an element challenge succeeds. (*In re Winship*, 397 U.S. 358, 364 (1970))

Evidentiary issues also vary. Federal Rule of Evidence 403 provides a balancing framework, whereas other rules impose different conditions and exclusions. A broad evidentiary category is therefore a practical grouping rather than a homogeneous class with a known rate of success. Federal doctrine illustrates these distinctions; state rules, constitutional protections, and procedural settings require their own analysis. (Federal Rules of Evidence, rules 403–404)

These distinctions support differentiated assessment. They do not determine the numerical ordering or spacing of every issue type. Suppression should not be treated generally as unfettered judicial discretion. An entitlement to relief under governing law cannot be displaced by an unfavorable numerical convention. A type adjustment is a stated modeling choice, not an additional legal burden.

6.2 Identification and overlapping contributions

If both $\beta_{0,t}$ and δ_t are fixed functions of the same type, only their sum affects the score. Outcome data cannot separately identify those contributions under this specification without additional restrictions or independent information. The split may express a conceptual distinction between baseline difficulty and structural conditions, but separate columns do not create separately estimable effects. A fitted model should ordinarily use a single type adjustment unless its design supports the distinction.

A structural adjustment also requires a clear reference setting. Its numerical value is not, by itself, evidence of judicial bias or measured institutional behavior. If a general procedural obstacle is

already represented in the type adjustment, a forum rating should not charge it again without explaining the additional information. One approach is to define the forum rating as a supported departure from the declared type reference. That approach requires the reference to be specified sufficiently well for raters to apply it.

Classification can create another form of overlap. If a category is favorable because a defense fact is independently corroborated, and corroboration also increases merit, the same consideration may contribute twice. A coding rule must allocate the category-defining characteristic and the remaining evidence explicitly. Alternatively, classification can follow legal function while all evidence-quality considerations enter merit. These are competing specifications to compare, not distinctions resolved by a category label alone.

6.3 Declared settings, estimation, and consequence weights

Declared settings and estimated parameters answer different questions. A declared setting expresses a scoring convention or priority. An estimated parameter describes a relationship in a sampled population under an estimation model. Estimation can inform a convention, but does not automatically settle which consequences should receive greater emphasis in professional deliberation.

The numerical ranges of rated inputs determine their possible contributions and the extent of compensation. A smaller contextual range limits its maximum contribution relative to core components; it does not make context unimportant. Context may determine the ordering of otherwise similar issues. Type adjustments combined with finite input ranges also create type-specific score floors and ceilings, as illustrated in Appendix A. These are properties of the specification, not discovered limits on legal strength.

Consequence weights introduce an additional judgment. Categories such as dispositive, major, and tactical have an ordered interpretation. Entering them as 3, 2, and 1 in a weighted mean imposes numerical ratios. For an equal change in index, a weight-3 issue contributes three times as much to the weighted mean as a weight-1 issue. Those ratios should be presented as declared priorities unless a separate study supplies another interpretation. Research on simple predictor weights does not establish that this particular ratio is optimal for aggregating distinct legal issues.

Weights must follow the consequence in the actual case. Impeachment can be tactical when it offers limited leverage and major when it substantially affects the prosecution's central proof. The label of the legal technique alone cannot establish the importance of its successful use. Sensitivity to reasonable alternative weights should be examined whenever the aggregate is used to support a material comparison.

6.4 Rationale for the reference design

The reference model adopts traceability, separation of distinct judgments, and interpretable comparison as design criteria. Its small set of components gives each issue a common assessment structure while retaining the source-linked reasons needed to evaluate the assignment. The index is

intended to be read with those reasons. That pairing lets a reader identify whether a disagreement concerns the record, the legal setting, the rating, or the numerical tradeoff.

Separating evidentiary support from forum conditions addresses a recurring feature of legal assessment. The strength of a factual proposition and the prospect that a court will afford a requested benefit are different judgments. Governing authority and observed institutional practice can also pull in different directions. Recording both in the forum rationale preserves that tension for review. A history of adverse rulings is relevant to an expectation about reception; it does not resolve whether the evidence supports the defense proposition. This separation gives the model a way to represent a well-supported issue with poor practical prospects.

Type adjustments provide a reference position from which issue-specific assessments proceed. They express an assumption that routes to relief differ in their starting conditions. The rated terms then express the contribution of the particular record and forum. The combined offset is the operative numerical assumption. Keeping its proposed baseline and structural explanations visible permits their substantive rationales to be examined, while the identification result establishes that the two fixed terms cannot be estimated separately from the present specification.

The magnitudes have an operational interpretation through their relationship to the rating ranges. A suppression offset of -2.6 requires a combined rated contribution greater than 2.6 to produce a positive index. The corresponding requirement for an element issue is greater than 0.2 . Two core components, each permitted to reach $+2.0$, allow a sufficiently favorable assessment to overcome either reference position. The contextual maximum of $+0.8$ is smaller than the maximum of either core component. These relationships state the balance chosen by the specification. They are the quantities a reviewer can challenge or replace when proposing a different balance. The numerical origin is a reference point on the index, not a decision threshold for legal success.

Additivity makes the contribution of each component explicit. Holding the others fixed, a one-unit change contributes one unit to the index, and its effect can be recomputed directly. The choice supplies a compact reference model with no hidden interactions. It also makes compensation an express assumption. Whether evidence and forum conditions instead require an interaction, a threshold, or a noncompensatory rule can be investigated against this disclosed reference. Any legal bar to the requested relief must remain explicit in the underlying issue analysis; the numerical index cannot alter that rule.

The logistic display serves a defined presentation function. It is smooth, bounded, symmetric about the index origin, and strictly increasing. It expresses every permissible index on the same display while preserving issue order. Its smaller display changes near the endpoints follow directly from the chosen curve. This provides a mathematically specified account of how the display behaves. The probability interpretation belongs to a separately evaluated outcome model. The scoring procedure itself uses the logistic function as a display transformation.

Consequence-weighted aggregation answers a different question from issue scoring. It summarizes the assessed position while giving greater influence to issues whose success would have greater consequences. The 3–2–1 rule makes that priority concrete and easily recomputable. For the same

index change, a weight-3 issue contributes three times as much to the weighted numerator as a weight-1 issue. Averaging on the additive scale retains the scale on which the issue judgments were combined. The resulting summary is useful for describing the center of the developed position under those priorities. Its composition and weights must accompany it because they determine what the summary represents.

Together, these choices specify an assessment procedure that can be implemented, checked, varied, and compared. The analytical study evaluates the consequences of the disclosed choices. Empirical studies can then address the additional questions of whether raters apply them consistently and whether the resulting assessments support their intended professional uses.

7 Evaluation and validation

7.1 Claims requiring evidence

Validity concerns the support for an interpretation and use of scores. It is not an intrinsic property conferred on an index by its mathematical form. An adequate account connects the proposed meaning to evidence and to the consequences of acting on that meaning. (Messick, 1995; Kane, 2013)

The *Standards for Educational and Psychological Testing* identify five sources of validity evidence. These concern content, response processes, internal structure, relationships with other variables, and consequences. Here they direct attention to coverage of relevant legal considerations, fidelity of ratings to their stated reasons, overlap among components, agreement with independent assessments, and effects on professional decisions. Applying this framework does not establish compliance with testing standards. (AERA, APA, & NCME, 2014, pp. 14–21)

A transparent specification establishes what the assessment purports to express and permits its mathematical behavior to be checked. Accurate rankings, calibrated probabilities, and improved decisions are further claims. They should be evaluated separately because evidence supporting one may provide little support for another.

7.2 Rating reliability and independent judgment

Reliability research should distinguish rating a fixed list of issues from developing and rating issues from an entire record. The first isolates variation in component assignment. The second also examines omissions, classification, and subdivision. Source audits should assess whether cited passages support the factual and legal propositions independently of whether raters agree numerically.

Analyses should report absolute disagreement and uncertainty as well as relative agreement. A high overall association can coexist with systematic score differences or poor agreement within issue types. An intraclass correlation must match the sampling design and intended generalization; a generic coefficient or threshold cannot establish adequacy for legal use. (Shrout & Fleiss, 1979; Koo & Li, 2016)

Independent counsel should assess the same materials without seeing the index or its component ratings. Agreement among attorneys should be described alongside agreement with the structured assessment. Professional judgment is a relevant comparator whose own variability must be measured. Consequence judgments should be evaluated separately from strength judgments, preserving the distinction that motivates weighting.

7.3 Outcomes, blinding, and calibration

Outcome studies require issue-specific endpoints. A suppression ruling, admission of evidence, recognition of an element challenge, and useful impeachment are different events. An issue never filed or decided cannot automatically be treated as a loss. The endpoint must also distinguish a favorable legal finding from the requested relief, and a prediction about what a court will do from an assessment of what it should do.

Retrospective reconstruction from published decisions introduces selection and hindsight. Removing names does not remove reasoning that reveals the outcome, and the opinion’s factual account was itself selected after adjudication. Forecasting should therefore be distinguished from recognition of an outcome already reflected in the input. Contemporaneous pre-ruling records provide a stronger basis for a prediction study. (Medvedeva & McBride, 2023)

A blinded evaluation should specify who can see which information. Source preparation may require access to the outcome, while scorers receive only reviewed predecision material. Outcome coding should be completed without access to predicted scores. Ratings and outcome labels should be locked before comparison, and legal authorities should respect the declared prediction time. Separate models or analysts are useful only insofar as the information barriers are implemented. For automated systems, possible prior exposure to public opinions remains a limitation even when retrieval is disabled.

Public appellate opinions, including the Nevada Supreme Court and Court of Appeals materials identified for future work, can support development of a retrospective evaluation. Normalization and identity removal alone do not create predecision records. Preparation requires issue-level eligibility decisions, preservation of the relevant procedural history, removal of target-outcome reasoning, separately coded outcomes, and review for information leakage. A preparation model and a scoring model can implement separate roles; calling the study double-blind additionally requires specifying and enforcing what scorers and outcome adjudicators cannot see. No such outcome study is reported in this article.

Where an appellate source sufficiently describes an earlier trial-court issue and ruling, those observations can be extracted as a separate decision event. The governing standard, the lower court’s stated standard, its applied analysis, and subsequent appellate treatment should retain separate source references and missing-information flags. An appellate disposition must not replace the underlying trial outcome. Patterns among reviewed cases require an explicit denominator and cannot establish statewide trial practice without an appropriate sampling frame. Historical information

used for a forum rating must precede the target decision. Evaluation on later cases can then examine whether the earlier collective pattern remains informative.

Evaluation should compare the complete rule with informative alternatives, such as type-only, merit-only, and forum-only indices, using equivalent source information. Otherwise, apparent performance may be explained by the issue-type mix or a single component. Related issues and proceedings require appropriate grouping when dividing development and evaluation data and when estimating uncertainty. A frozen evaluation set should not become a repeated source of feedback for revising the rule.

Outcome concentration also matters. If nearly every observed ruling permits prosecution to continue, a constant prediction can agree with the rulings while providing little discrimination among defense issues. Evaluation should report outcome frequencies and compare any prespecified binary decision rule with the corresponding constant-decision baseline. A subset containing only one outcome category cannot establish discrimination between favorable and unfavorable rulings. Independent review of the evidence-to-element reasoning supplies a distinct assessment and should remain separate from the recorded judicial result.

If the index is later mapped to outcome probabilities, discrimination and calibration require separate assessment. Discrimination concerns ordering favorable and unfavorable outcomes; calibration concerns agreement between predicted probabilities and observed frequencies in the intended setting. An effective ranking can have poor probability calibration. (Van Calster et al., 2019) The Brier score measures overall probabilistic prediction error, rather than calibration alone. (Brier, 1950) A calibration mapping should be developed and evaluated on distinct data, with uncertainty and population limits reported. Calibrating an issue endpoint would still not convert the weighted aggregate into an acquittal probability.

7.4 Consequences of professional use

A decision-use study should examine whether structured assessment improves identification and discussion of material issues, recognition of unsupported premises, and the quality of selected next steps. It should also examine anchoring on numerical displays and mistaken probability interpretations. Relevant comparisons include the same source materials with and without numerical aggregation, so that any contribution of the combination rule can be distinguished from the contribution of organizing the evidence.

These questions concern practical effects. Transparency makes reasons inspectable, but inspection may be incomplete and visible numbers may receive undue weight. Professional review is part of the proposed use; its presence alone does not establish a beneficial consequence.

8 Discussion

This study establishes the behavior of a formal scoring model for criminal defense analysis. It defines an issue index, a display transformation, and an aggregate, and derives the relationship

among them. The reference implementation is fully enumerated over its permitted input grid. The resulting findings identify how the model orders issues, how strongly its assumptions can affect those orders, and what must remain constant for an aggregate to retain its interpretation.

The model's principal methodological value is the organization of legal judgment into inspectable contributions. Evidence supports a proposition. Legal and institutional conditions affect its reception. The requested benefit determines its consequence. Assigning those considerations identifiable roles makes the basis of a favorable or unfavorable assessment visible. The combination rule then states how the numerical assignments affect the conclusion. This is a concrete form of accountability in the representation of judgment. A critic can challenge an evidentiary premise, a forum inference, a type assignment, a scale convention, or the aggregation rule and show the resulting difference.

The distinction between evidentiary support and institutional reception is especially important when observed outcomes are highly uniform. Consistent rejection of a class of defense arguments may be informative about practical prospects, but it cannot alone establish that every underlying objection lacks merit. Treating the observed ruling as the sole definition of strength would erase that distinction. The present architecture retains separate components and source rationales so that empirical work can examine agreement with actual rulings alongside independent assessment of the evidence and legal reasoning.

The exhaustive results demonstrate why the constants belong in public methodological disclosure. Removing the structural adjustment produces strict reversals in 5.27% of cross-type pairs in the enumerated design space. Removing all type adjustments produces strict reversals in 20.19%. The rated evidence and forum judgments remain fixed in those comparisons. The reversals therefore identify the contribution of the type assumptions to the resulting order. They give reviewers a quantitative basis for evaluating those assumptions and for specifying comparisons with an alternative model. They do not estimate which ordering would better predict rulings.

The sensitivity results also provide practical conditions for interpreting an assessment. When permitted rating envelopes do not overlap, the issue order survives every reassignment within the declared envelopes. When they overlap, that same assurance is unavailable. Reporting these conditions makes the dependence of a conclusion on disputed ratings explicit. The analysis of ordinal recoding further shows that preserving category order is insufficient to preserve additive results. Numerical spacing is part of the model and must be defended along with the category descriptions.

The aggregation analysis identifies the precise object of an overall case score. The consequence-weighted mean describes the center of the developed issue set under declared priorities. It places greater influence on the issues whose success would matter more according to those priorities. The subdivision result establishes when reformulating one issue into several preserves the summary. Conservation of total weight and weighted index contribution is necessary and sufficient for invariance across surrounding issue sets. This result makes issue organization a substantive requirement of aggregation. A case score should therefore be presented with the included issues, their relationships, and their assigned consequences.

The declared specification is a reference model for criticism and comparison. Its compact form makes the arithmetic and tradeoffs visible, and the supplied analysis makes its behavior reproducible. A challenger can propose a different type structure, rating scale, interaction, or aggregation rule and compare the consequences under the same inputs. Whether a change improves legal assessment should then be examined against independent judgments or defined outcomes. An empirical comparison must also preserve equivalent source information so that the contribution of the numerical rule can be distinguished from differences in the underlying analysis.

Four limits define the reach of the reported results. The numerical grid exhausts the declared codes rather than the population of actual legal issues; some permitted combinations may be substantively inadmissible. The alternative specifications isolate assumptions rather than estimate an optimal rule. The proposed operational codebook has not undergone rater evaluation, particularly for its reference conditions and contextual anchors. Finally, no judicial outcomes, independent counsel ratings, or user decisions were analyzed. These limits locate the remaining empirical questions without altering the mathematical findings.

Implementation fidelity is a further object of study. The motivating software contract requires developed issues, source-linked reasons, separate component responsibilities, and correct calculation. Verifying those requirements in automated executions requires source audits and execution evidence. The present paper discloses the scoring rule and supplies its analytical evaluation. The author's relationship to its software implementation is disclosed, and independent replication is invited. Publication creates a defined object for criticism, testing, and improvement.

9 Conclusion

This paper presents a formal model for scoring criminal defense issues and summarizing overall case strength. Its components distinguish evidentiary support, conditions of relief, context, and consequence. Its equations specify the numerical tradeoffs, and its source requirements make the grounds of an assessment inspectable. The completed analysis establishes the model's ordering and aggregation properties, quantifies sensitivity to its declared assumptions, and derives conditions under which comparisons remain stable.

The contribution is a public, reproducible scoring specification with a substantive design rationale and mathematical results. Rating reliability, judicial outcome performance, and effects on professional decisions are the next empirical questions. The published model and accompanying materials give independent researchers the means to reproduce the analysis, contest the assumptions, and evaluate stronger alternatives.

Appendix A The disclosed numerical specification

The following settings define the reference specification evaluated in this article. They are declared conventions rather than estimates from an outcome dataset. Their provenance is identified in the declarations. The general mathematical results do not establish their magnitudes as preferable to alternative settings.

Table A1. Declared type adjustments and combined reference scores.

Assigned issue type	$\beta_{0,t}$	δ_t	α_t	$100\sigma(\alpha_t)$
Suppression, including affidavit and warrant challenges	−1.4	−1.2	−2.6	6.9
Ordinary evidentiary or procedural issue	−0.6	−0.2	−0.8	31.0
Factual contradiction or element failure	0.0	−0.2	−0.2	45.0
Strong corroborated defense fact	+0.6	−0.1	+0.5	62.2

The final column is the displayed value when the rated inputs are zero. It is a reference score, not an assumed or observed success rate. The last category describes evidentiary quality, whereas other categories describe legal function; the overlap and allocation problem discussed in Section 6.2 therefore applies directly to this example.

Merit and forum ratings each range from −2.0 to +2.0 in half-unit steps. Context ranges from −0.8 to +0.8 in tenth-unit steps, with zero used when adequate review identifies no residual adjustment. A contribution of +2 multiplies implied odds by approximately 7.39, while −2 divides them by that factor. The contextual endpoints correspond to a multiplier of approximately 2.23 or its reciprocal. These are consequences of the scale choices, not measured effects on legal outcomes.

The ranges create type-specific limits. For suppression, $L \in [-7.4, 2.2]$, giving displayed scores from approximately 0.06 to 90.02. The other types have different limits because their offsets differ. This example shows why review should consider the full interaction of offsets and rating ranges, rather than evaluating the ordering of type labels alone.

The consequence scale assigns 3 to a dispositive issue capable of suppressing core evidence, defeating an element, or ending the case; 2 to a major issue substantially affecting trial posture, admissibility of key evidence, or sentencing exposure; and 1 to a tactical benefit such as limited impeachment, leverage, or procedural advantage. Assignment follows the specified consequence in the case. The ratios are priorities for aggregation, not measured utilities or empirically established effect sizes.

Each issue’s components, reasons, citations, weight, and computed index should be reported together. The aggregate should display its weighted numerator and denominator and identify its issue set. Calculations should retain full precision until display; one decimal place can facilitate inspection without implying that uncertainty is one tenth of a score point. Changing these declared settings produces a different specification and should be documented before evaluating its performance.

Appendix B Derivations supporting the sensitivity analysis

For a positive index shift d , the largest logistic display difference is obtained by maximizing $100\{\sigma(x+d) - \sigma(x)\}$. The derivative vanishes uniquely at $x = -d/2$. The difference tends to zero at either extreme and is positive between them, so this stationary point is the maximum. The zero-shift case is immediate. Symmetry gives

$$\max_x 100|\sigma(x+d) - \sigma(x)| = 100\{2\sigma(|d|/2) - 1\} = 100 \tanh(|d|/4).$$

For a symmetric index envelope $[-h, h]$, the full score width is consequently $100 \tanh(h/2)$. With $h = 1.1$, this is approximately 50.0520 points. This width compares the two extreme reassessments, not one extreme with the center.

For subdivision, let a parent have weight w and index x . Let R and Q be the total weight and weighted sum of the other issues. Replace the parent by children with total weight v and weighted sum z . Requiring $(Q + wx)/(R + w) = (Q + z)/(R + v)$ for arbitrary surrounding (R, Q) first equates the coefficients of Q , giving $v = w$, and then requires $z = wx$. Conversely, those conditions preserve the mean immediately. Thus conservation of weight and weighted index contribution is necessary and sufficient for invariance in arbitrary surrounding issue sets. A particular case can exhibit accidental cancellation without satisfying the general condition.

The number of strict componentwise-dominance comparisons between a designated pair of types follows from the component grids. Each nine-level core scale supplies $9 \times 10/2 = 45$ weakly ordered pairs. The seventeen-level contextual scale supplies $17 \times 18/2 = 153$. Their product includes 1,377 identical profiles. Excluding those gives $45^2 \times 153 - 1,377 = 308,448$ strict dominance pairs. Integer enumeration then partitions them according to whether the summed rating advantage is smaller than, equal to, or larger than the type-offset gap.

Acknowledgments and declarations

Author and interests. Christopher R. Harrison is a trial attorney, an AI systems engineer, and the founder of Analyzed Discovery Ltd., a Nevada software provider. He also has an ownership and management role in Battlefront Legal LLC, a Nevada law firm, which holds a 0% ownership interest in Analyzed Discovery Ltd. The numerical specification originated in the Analyzed Discovery software project.

Materials and data availability. Supplementary Material 1 contains the proposed operational codebook. Supplementary Material 2 contains the analysis plan, executable source and independent verifier, complete enumerated profile grid, and machine-readable results. All numerical results can be regenerated using Python 3.10 or later without external packages. No case records, participant data, or judicial outcomes were used in the reported analytical study.

Funding. This work was supported by the author.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. Publisher's open-access edition.
- Bex, F. J., van Koppen, P. J., Prakken, H., & Verheij, B. (2010). A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law*, 18(2), 123–152. DOI 10.1007/s10506-010-9092-x.
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement. A structural equation perspective. *Psychological Bulletin*, 110(2), 305–314. DOI 10.1037/0033-2909.110.2.305.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. Publisher's article.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. DOI 10.1037/h0040957.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. DOI 10.1037/0003-066X.34.7.571.
- Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services*, 54(10), 1372–1379. DOI 10.1176/appi.ps.54.10.1372.
- Edwards, W., & Barron, F. H. (1994). SMARTS and SMARTER. Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes*, 60(3), 306–325. DOI 10.1006/obhd.1994.1087.
- Federal Rules of Evidence, rules 403–404. Official compilation.
- Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science*, 37(1), 61–102. DOI 10.1111/cogs.12004.
- Franks v. Delaware*, 438 U.S. 154 (1978). United States Reports.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction. A meta-analysis. *Psychological Assessment*, 12(1), 19–30. DOI 10.1037/1040-3590.12.1.19.
- In re Winship*, 397 U.S. 358 (1970). United States Reports.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. DOI 10.1111/jedm.12000.

- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. DOI 10.1016/j.jcm.2016.02.012.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement. A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1), 1–27. DOI 10.1016/0022-2496(64)90015-X.
- Medvedeva, M., & McBride, P. (2023). Legal judgment prediction. If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023* (pp. 73–84). Association for Computational Linguistics. DOI 10.18653/v1/2023.nllp-1.9.
- Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. DOI 10.1037/0003-066X.50.9.741.
- Paruolo, P., Saisana, M., & Saltelli, A. (2013). Ratings and rankings. Voodoo or science? *Journal of the Royal Statistical Society, Series A*, 176(3), 609–634. DOI 10.1111/j.1467-985X.2012.01059.x.
- Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society, Series A*, 168(2), 307–323. DOI 10.1111/j.1467-985X.2005.00350.x.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations. Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. DOI 10.1037/0033-2909.86.2.420.
- United States v. Leon*, 468 U.S. 897 (1984). United States Reports.
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W., on behalf of Topic Group “Evaluating diagnostic tests and prediction models” of the STRATOS initiative. (2019). Calibration. The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230. DOI 10.1186/s12916-019-1466-7.
- Victor, M. B. (2015). Decision tree analysis. A means of reducing litigation uncertainty and facilitating good settlements. *Georgia State University Law Review*, 31(4), 715–739. Journal article.

Appendix C A Trial-Court Benchmark for Defense Issue Assessment

Proposed design · October 2026

This document develops a benchmark proposal for the methods paper. It reports no collected national sample, blinded predictions, or outcome-validation results. Its allocations are planning choices, not an externally registered protocol or a completed sample-size calculation.

The supplied source collection has now received a local corpus assessment¹. It contains a substantial Nevada appellate and U.S. Supreme Court opinion pool, including 2,675 Nevada TXT records in criminal-labeled folders. The earlier Nevada extraction can be reused. Adequate predecision criminal trial-court packets and all-state coverage have not yet been established. The immediate study should therefore begin with a separately reported, retrospective opinion-based cohort. The national trial-court design below remains an expansion target, and its source requirements remain applicable.

C.1 What the benchmark would test

The primary question is whether the existing issue-strength index assigns higher values to defense issues that subsequently obtain a specified trial-court benefit than to comparable issues that do not. That is evidence about predictive or criterion validity. Actual rulings do not, by themselves, establish that a legal argument was normatively correct or that the full construct of defense issue strength has been validated.

A second component would compare the index with experienced defense attorneys' assessments of the same record, made without seeing the index or the outcome. That supplies a different source of evidence about the intended construct. The two components should be reported separately.

The unit is one identified issue from one case family, assessed at a specified point before a target ruling. A benchmark question is a legally grounded case task, not a general legal-knowledge question. It identifies the proposition, requested benefit, available evidence, governing law, procedural posture, and target decision.

The primary cohort should consist of criminal-defense issues with a discernible trial-court ruling. Suppression and admissibility questions are practical starting points. Factual support and impeachment are not always resolved through a separate grant-or-denial ruling. They can enter the attorney-assessment component, but a conviction or acquittal must not be used as a substitute label for every underlying issue.

C.2 A concrete first allocation

The proposed 5,000–10,000 cases form a candidate inventory. They are not presumed to be complete, eligible case files. One case may contain several potentially eligible issues, and one selected issue may require several documents.

¹Companion material supplied with the manuscript, `benchmark_corpus/corpus_assessment.md`.

Table C1. Candidate inventory and proposed benchmark allocation.

Collection	Proposed size	Purpose
Candidate inventory	5,000–10,000 case families	Establish document availability, jurisdiction, dates, issue coverage, and eligibility
Development pilot	50 case families outside the final test	Resolve packet preparation, label definitions, and scoring instructions
State trial-court test	250 case families	Five qualifying cases from each of the 50 states
Federal district-court test	250 case families	Deliberate coverage across districts, regions, issue families, and time periods
Locked primary benchmark	500 case tasks	One prespecified issue from each selected case family

The 50-state quota provides geographic coverage. Five cases per state cannot establish state-specific accuracy. Equal state and federal counts also do not reproduce the distribution of criminal cases or defense issues in practice. This is a deliberately balanced benchmark, and its pooled result describes that designed mixture.

A nationally representative estimate would require a defined national sampling frame, known selection probabilities, attention to unavailable records, and appropriate weighting. Post hoc weights cannot recover jurisdictions or outcome groups that were never observable.

Choose a fixed period and issue scope before sampling. Build a jurisdiction roster containing actual trial-court names and case types rather than relying on the words district or supreme in a court’s name. Randomly select eligible case families within declared strata and retain the selection seed, replacement rules, and reasons for every exclusion. Do not choose cases to obtain a favorable model result.

Use the 50-case pilot to estimate how often a candidate yields an adequate packet and usable outcome. The final sample size should then be assessed against the expected number of favorable and unfavorable rulings and the desired precision. Five hundred is a sensible initial benchmark target, not a guarantee of adequate precision for every issue family or jurisdiction. Prediction-model sample-size research supports tailoring validation size to the intended estimates rather than relying on a universal count. Its clinical examples are methodological analogies, not sample-size estimates for this legal study. Riley et al., 2024.

C.3 The records required

The strongest retrospective packet is assembled from documents available before the target decision, such as the motion, response, exhibits, relevant discovery, and a hearing transcript that does not disclose the ruling being predicted. A written order entered after an oral ruling does not move the prediction cutoff forward. Information revealed by the oral ruling is already outcome information.

Every included case needs both an adequate predecision packet and a separately held outcome source. A docket entry is not a complete evidentiary record. A judicial opinion’s account of facts

is also not equivalent to the material available before the ruling. Its selection of facts and framing of disputes can reveal the result even when the disposition is removed. Research on legal judgment prediction identifies this distinction between forecasting and recognizing outcomes from decision texts. [Medvedeva and McBride, 2023](#).

Federal RECAP and PACER materials are candidate sources, subject to document-level coverage. RECAP search can return case metadata without filings, and PACER excludes some categories of material. A searchable case therefore is not automatically a usable scoring packet. [Free Law Project’s coverage explanation](#) and [PACER availability guidance](#).

State trial-court access requires a separate jurisdiction audit. For example, Maryland’s judiciary reports that its online Case Search omits information about criminal charges ending in dismissal, acquittal, or nolle prosequi, although some records remain accessible at the courthouse. This creates a possible outcome-dependent sampling gap. It does not establish that the underlying records are unavailable through every route. [Maryland court-record guidance](#).

The candidate register must distinguish a located case, available docket metadata, available pre-decision documents, an adequate scoring packet, and an adjudicated target outcome. Missingness and exclusions should be reported by jurisdiction, issue family, time period, and outcome category where the outcome can be lawfully established. Availability may depend on how a case ended, so the final claims must address that selection.

The existing Nevada appellate collection can help develop extraction and masking procedures. It is not a substitute for the national trial-court cohort. Supreme court and other appellate decisions belong in a separate appellate test, with their own prediction time, standard of review, and requested relief.

C.4 Normalize structure and mask identity without changing the legal problem

Normalization should standardize document structure, encoding, fields, and source locators. Identity masking should replace irrelevant names, captions, docket numbers, filenames, and target-decision identifiers with stable aliases.

Do not change everything. Preserve the relationships among people, disputed status of assertions, quantities, ages, timing, jurisdiction, and other facts that affect the legal analysis. Changing those facts may create a different case for which the original ruling is no longer an appropriate comparison.

Remove the target ruling, later appellate history, headnotes, outcome-revealing summaries, and reasoning that resolves the target issue. Retain earlier rulings only when they were available at the prediction time and concern a different decision. Assemble governing authority as of that time and exclude the target opinion and its subsequent treatment.

Each proposition must retain a private source locator and a neutral evidence identifier in the scoring packet. A reviewer should verify that preparation has preserved both favorable and adverse evidence. A normalization model must not silently transform an allegation into an established fact.

Renaming does not prove that a model cannot recognize a public case. Distinctive facts or remembered text can remain identifying. An anonymization experiment on the CaseHOLD legal benchmark expressly reported an inconclusive memorization diagnostic; it did not establish that masking removed training contamination. Its holding-completion task also differs from prediction of a trial ruling. Arvin, 2025. Record known training and version information, reported recognition, and exposure concerns. Masked retrospective performance should be described with that limitation. A later prospective cohort, scored before the actual rulings occur, would provide stronger evidence against answer exposure.

C.5 Information separation and execution

The preparation and scoring roles should be separate. The preparation process may access the complete source record to construct packets and an outcome key. The scoring process should receive only the approved packet and permitted historical authority.

Table C2. Information access and blinding by study role.

Role	Permitted information	Information withheld until lock
Packet preparation and review	Original documents, source mappings, and preparation instructions	Scoring outputs
Outcome coding and adjudication	Target ruling and material needed to interpret it	Model scores and attorney strength ratings
Model scoring	Approved predecision packet, authority packet, frozen scoring specification	Outcome key, original-case lookup, later rulings
Attorney assessment	The same approved predecision materials	Model scores and actual outcome
Analysis after lock	Frozen scores, labels, and approved sampling metadata	No further separation needed for the comparison

Use genuinely separate access boundaries. Different filenames in a shared accessible directory are insufficient. A fresh model conversation is useful but does not remove information from the model’s pretraining. Two models do not automatically make a study double-blind. The precise intended procedure is outcome-blinded scoring and prediction-blinded outcome coding.

Freeze the numerical constants, rating instructions, model version, retrieval access, execution settings, and error-handling rules before the test. The scoring model must produce the existing method’s component ratings and source-linked reasons. A generic prediction of the ruling would test a different procedure. Recompute the index and logistic display deterministically from those components.

Lock predictions before joining them to outcomes. Keep related proceedings and duplicated records in the same partition. Do not revise prompts or constants in response to the held-out results and then describe the revised performance on the same cases as independent validation.

C.6 Outcome labels and independent judgment

Define the requested benefit from the predecision materials before inspecting the disposition used as the label. The primary binary endpoint is whether the trial court granted that prespecified benefit. The label manual must state how partial grants, alternative grounds, procedural denials, withdrawal, mootness, and unresolved issues are handled.

Keep those original categories even when a binary label is derived. A finding of error is not necessarily the relief sought. An unreached or withdrawn issue is not automatically a defense loss. Such issues belong in the benchmark’s accounting and may fall outside the primary adjudicated-issue analysis.

Two independent legally qualified coders should label outcomes without seeing model predictions, with disagreements adjudicated before the join. Model-assisted extraction can reduce preparation work, but the final reference label must have a reviewed source basis.

For the construct component, obtain independent attorney assessments on a prespecified subset, provisionally 100 of the test packets. Attorneys should judge strength without the scorecard formula or output and assess consequence separately. Record their disagreement as well as their agreement with the index. The developer’s own ratings are not an independent validation panel.

An issue-ruling study will not validate the overall case mean or the 3–2–1 consequence weights. Those require their own criterion. It also will not, by itself, test whether the full automated system identifies every important issue in raw discovery. Supplying an already identified issue evaluates the scoring stage.

C.6.1 Governing law, court practice, and the decision sequence

The study should distinguish the standard supplied by governing authority from the standard a court states and the analysis it actually performs. It should also distinguish an assessment of legal correctness from an estimate of how comparable issues are decided. A historical ruling is an observation of court behavior. It does not establish, by itself, that the legal analysis was correct.

An appellate opinion can identify an earlier trial-court issue and ruling where it describes them sufficiently clearly. Extract those observations with their source basis, retaining the appellate proceeding as a separate event. The earlier issue, the issue preserved for review, and the issue ultimately reached may differ. Do not assume that a complete predecision record can be reconstructed merely because the opinion states what happened below.

Table C3. Decision sequence and required observations.

Record field	Required content
Exact issue and requested benefit	The challenged element, evidence, procedure, or legal proposition and the relief sought at the identified stage
Target decision	Court, court level, procedural stage, decision date, and information cutoff
Governing standard	Applicable authority as of the cutoff and the substantive decision rule it supplies

Record field	Required content
Stated standard	The target court’s express formulation, with a source locator, or an explicit indication that the available source does not disclose it
Applied analysis	The evidence, elements, inferences, and legal conditions the court actually addresses, to the extent the record permits observation
Issue-specific outcome	Benefit granted, denied, partly granted, or not reached, with the supporting source
Subsequent review	Reviewing court, issue actually reviewed, standard of review, and disposition, recorded separately from the earlier ruling
Observation quality	Original transcript or order, quoted lower-court passage, appellate paraphrase, or unavailable information

The target court’s reasoning and disposition belong on the outcome side of the information barrier. They cannot be used to construct that same decision’s forum rating. An earlier ruling can be a legitimate input when the target is a later appeal, subject to the declared cutoff. These uses must not be interchanged.

For the present specification, historical practice may inform the forum component through its stated rationale. Evidentiary merit remains a separate assessment. Keep the account of governing law and the account of institutional practice distinguishable in the rationale without introducing an undisclosed extra score or changing the frozen constants. When the sources do not support a practice estimate, record that limitation rather than assign an invented jurisdictional rate.

The Nevada probable-cause research note² supplies a concrete example. The statutory requirement, judicial descriptions of the low evidentiary threshold, and the actual application to a disputed element should be recorded separately. The Supreme Court’s own abbreviated formulations make terminology alone an inadequate indicator of doctrinal departure. This example motivates an empirical question about application and does not establish a statewide practice finding.

C.6.2 The reported any-evidence practice

The practitioner observation motivating this part of the pilot concerns the operative decision rule. The hypothesis is that the presence of some inculpatory evidence is treated as sufficient to continue a prosecution even when the defense challenges the inferential connection between that evidence and a required element. This is a reported practice to investigate, rather than an established frequency or a premise that every arrest results in trial.

For an eligible probable-cause issue, preparation should identify the evidence relied on, the intermediate factual propositions needed to connect it to the disputed element, the basis for each inferential step, the defense’s objection, and the court’s recorded response. Distinguish an inference the court expressly adopts from one reconstructed by the preparer. A source that omits the court’s reasoning should be coded as insufficiently informative about that reasoning. Omission from an appellate summary does not establish that the lower court performed no analysis.

²Companion material supplied with the manuscript, `research/nevada_probable_cause_standard.md`.

Reviewers should assess the support for each inferential step separately from the observed ruling and without seeing model scores. Record supported, contested, unsupported, or indeterminate links with reasons and source references. A longer inference chain is not automatically defective. The relevant question is whether the necessary connections have a defensible basis under the applicable standard. This review supplies an additional observation about evidentiary reasoning; it introduces no automatic chain-length penalty or new numerical input into the frozen score.

The record can therefore preserve a well-supported defense objection together with a ruling that allows the prosecution to continue. Observed reception informs the study of forum behavior. It must not retroactively determine whether the underlying evidence supports the defense proposition.

C.7 Comparisons and reporting

The primary analysis should report how well the frozen index orders favorable and unfavorable outcomes, including area under the ROC curve and uncertainty. Report federal and state results separately and show results by issue family where the data support estimation. Court-level or case-family dependence must be addressed in the uncertainty procedure selected before the test.

Historical practice should be summarized collectively by legally comparable issue, jurisdiction, procedural stage, and period. Individual judge rankings are outside this design. Record both the numerator and the population of observable decisions behind a frequency. A collection selected through appellate review estimates patterns among the observed reviewed cases; it does not establish how often all trial-court judges rule that way across all eligible matters. Using a familiar phrase is also insufficient to establish a departure from governing law. That claim requires examination of the analysis applied to the particular evidentiary and legal requirements.

Any historical information used to assess a target decision must predate that decision and come from the permitted development or reference material. Keep related proceedings together when separating development and evaluation. A later-period evaluation can test whether an earlier collective pattern remains informative, with changes in governing law and the composition of cases reported. Historical frequencies alone do not establish future performance.

Compare the full score with a type-only score, merit-only and forum-only scores, and the same underlying model making a judgment from the same packet without the scorecard. The comparison with the type-only score is particularly important because different issue families may have different outcome frequencies. Report paired performance differences and their uncertainty rather than only a headline accuracy number.

For the probable-cause subset, report how often prosecution continues and how often the specified defense relief is granted. If a binary decision rule is evaluated using a threshold fixed before testing, include an always-continue comparator. A high agreement rate in a nearly uniform outcome sample may reveal little about assessment of issue strength. If only one outcome category is observed, discrimination between favorable and unfavorable outcomes cannot be estimated in that subset. Report that limitation directly and retain the separate review of evidentiary reasoning. Do not force a 50-point threshold onto the uncalibrated score to manufacture this comparison.

The 0–100 display is currently an uncalibrated index. Do not automatically treat 50 as a win-loss threshold or 70 as a 70% probability. If a probability mapping is developed, estimate it on development cases and evaluate the frozen mapping on the untouched test. Report discrimination and calibration separately, with the Brier score as overall probabilistic prediction error. Do not equate good ranking with good probability calibration.

Do not artificially force equal numbers of wins and losses in the primary sample merely to make a balanced exam. Jurisdiction balance and outcome balance are different choices. A deliberately outcome-enriched auxiliary set can test rare situations, but it must be labeled and analyzed separately.

Report every attempted packet, exclusion, missing document, coding disagreement, failed execution, unscorable issue, and suspected exposure. A system’s refusal to score an eligible packet should remain visible as a coverage result rather than silently disappear from its accuracy denominator.

C.8 Recommended sequence

1. Audit candidate-source coverage and identify predecision trial records, beginning with a limited set of federal districts and state jurisdictions.
2. Prepare and review 50 development packets, finalize the issue and label manuals, and estimate record yield and outcome frequency.
3. Set the final allocations and precision targets, then preregister the protocol and freeze the scoring system.
4. Prepare the independently sampled 500-case test, keeping all related proceedings outside development.
5. Lock model scores, attorney assessments, and adjudicated outcome labels before comparison.
6. Publish the protocol, sampling flow, permitted benchmark materials, evaluation code, and all prespecified results. Retain or refresh a separate private test set because public release can expose future models to the answers.

The immediate recommendation, given the supplied corpus, is to begin the 50-case development stage with screened Nevada appellate opinions and evaluate whether a separate 400–500-case opinion-based test can be supported. That study must report its reconstructed information conditions and potential outcome leakage. The balanced 500-case trial-court design should follow when suitable predecision records and jurisdiction coverage are available. Results from the two information settings should be reported separately.