

# Contents

<b>Decision Grade: A Multi-Dimensional Behavioral Scoring Methodology for Discretionary Traders</b>	<b>1</b>
Abstract . . . . .	1
1. Introduction and Motivation . . . . .	2
2. The Decision Grade Framework . . . . .	2
3. Score Definitions and Computation . . . . .	3
4. Per-Trade Decision Quality . . . . .	5
5. Worked Examples . . . . .	6
6. Execution Error Taxonomy . . . . .	9
7. Aggregation and Feedback Loops . . . . .	10
8. Validation Methodology . . . . .	11
9. Limitations and Open Questions . . . . .	12
10. Invitation for Peer Review . . . . .	13

## Decision Grade: A Multi-Dimensional Behavioral Scoring Methodology for Discretionary Traders

**Drive By Numbers (DBN) Technical White Paper Version 1.0 - April 2026 Author: DBN Research**

---

### Abstract

Profit and loss is the dominant metric by which traders evaluate their performance, yet it is among the worst predictors of long-term trading success. A trader may be profitable over a short horizon through luck, reckless position sizing, or favorable market conditions, while a disciplined trader executing a sound methodology may experience a legitimate drawdown. The Decision Grade framework, implemented in the DBN trading platform, addresses this measurement gap by decomposing trading quality into four independent behavioral scores – Discipline, Execution Consistency, Risk Management, and Emotional Stability – each computed on a 0-to-100 scale from observable trade-level data. This paper presents the formal definitions, computational methods, per-trade classification system, and execution error taxonomy that constitute the Decision Grade methodology. Three worked examples demonstrate how the framework distinguishes between genuine skill and survivorship-biased luck, and how it surfaces actionable behavioral feedback that P&L alone cannot provide. We discuss the feedback loops through which these scores drive coaching interventions and interface adaptations within the DBN platform, outline a validation methodology grounded in score-profitability correlation analysis, and acknowledge the limitations inherent in self-reported emotional data and client-side computation. This paper is released as an invitation for peer review and independent replication.

---

## **1. Introduction and Motivation**

The retail trading industry suffers from a fundamental measurement problem. Aspiring traders, prop firm evaluators, and trading educators overwhelmingly assess performance through a single lens: the equity curve. Did the account grow? By how much? Over what drawdown? These are important questions, but they are dangerously incomplete. A trader who risks 10% of capital per trade on impulse entries and happens to catch a trending week will show a spectacular equity curve - right up until the week when they do not. Conversely, a trader who meticulously follows a high-probability methodology, sizes conservatively, and journals every decision may show a modest or even negative equity curve during an adverse regime, despite executing at a level that virtually guarantees long-term profitability.

This is not a theoretical concern. Research in behavioral finance has repeatedly demonstrated that short-term P&L is a noisy signal contaminated by market regime, instrument volatility, timing luck, and survivorship bias. The traders who endure are not those who had the best week, but those who maintained consistent decision quality through hundreds or thousands of trades. The challenge is that “decision quality” has historically been undefined, unmeasured, and unmonitored in most trading platforms.

The Decision Grade framework was developed to fill this gap. It is built on three foundational principles. First, trading quality is multi-dimensional: a trader can be disciplined but emotionally reactive, or risk-aware but inconsistent in execution. A single composite score would obscure these distinctions. Second, behavioral metrics must be computed from observable data, not subjective self-assessment alone. Where self-reported inputs are used (such as emotional state at entry), they are cross-validated against objective behavioral signals (such as time between trades or position size changes after losses). Third, the scoring system must close the feedback loop - it is not sufficient to compute a score; the score must drive concrete interventions, from coaching recommendations to interface density changes, that alter future behavior.

The remainder of this paper formalizes each component of the Decision Grade system, demonstrates its application through detailed worked examples, and discusses its integration into the broader DBN platform architecture.

---

## **2. The Decision Grade Framework**

### **2.1 Four-Score Architecture**

The Decision Grade framework comprises four independent scores, each targeting a distinct behavioral domain. Independence is a deliberate design choice: the scores share no inputs, and a trader’s standing on one dimension carries no mathematical implication for their standing on another. This orthogonality ensures that the framework surfaces specific, actionable feedback rather than a blurred composite.

The four scores are Discipline (measuring plan adherence and process consistency), Execution Consistency (measuring the mechanical quality of trade entries and management), Risk Management (measuring position sizing discipline and protective stop

usage), and Emotional Stability (measuring the trader's capacity to maintain psychological equilibrium across wins and losses). Each score occupies the 0-to-100 range, where 0 represents complete absence of the measured quality and 100 represents theoretical perfection. In practice, scores above 90 are rare and scores below 30 indicate severe behavioral dysfunction requiring immediate intervention.

## 2.2 The Independence Principle

The independence of these four scores is not merely a mathematical convenience; it reflects an empirical observation about how traders fail. A trader may exhibit extraordinary discipline in following their pre-trade checklist and trading only during designated sessions, yet consistently move their stop-loss after entry - scoring high on Discipline but low on Execution Consistency. Another trader may size positions perfectly and always place protective stops, yet enter trades impulsively after losses - scoring high on Risk Management but low on Emotional Stability. Collapsing these dimensions into a single number would mask the specific behavioral pattern that, if corrected, would most improve the trader's long-term outcomes.

The independence principle also enables targeted coaching. When the platform detects that a trader's Risk Management score is high but their Emotional Stability score is declining, it can surface specific interventions - escalating cooldown timers, forced reflection modals, post-loss journaling prompts - without disrupting the risk management workflow that is already functioning well.

---

## 3. Score Definitions and Computation

### 3.1 Discipline Score

The Discipline Score measures the degree to which a trader adheres to their own declared process. It is anchored to two primary inputs: the rate at which the trader completes the pre-trade decision gate (a structured checklist requiring confirmation of setup validity, risk parameters, and emotional readiness), and the mean plan-adherence percentage reported across all trades. An overtrading penalty is applied for each calendar day on which the trader exceeds five trades, reflecting the well-documented degradation of decision quality that accompanies excessive trading frequency.

The computation begins with a base value of 50, establishing a neutral midpoint. The gate usage rate - defined as the proportion of trades for which the pre-trade gate was completed prior to entry - is multiplied by 25 and added to the base. The mean adherence percentage across all trades is multiplied by 0.25 and added. Finally, a penalty of 5 points is subtracted for each calendar day on which the trader opened more than five positions. The result is clamped to the 0-to-100 range.

Formally, if  $G$  denotes the gate usage rate (0 to 1),  $A$  the mean adherence percentage (0 to 100), and  $D$  the count of overtrading days, then:

$$\text{Discipline} = \text{clamp}(50 + G * 25 + A * 0.25 - D * 5, 0, 100)$$

A trader who always uses the gate ( $G = 1.0$ ), maintains 80% mean adherence, and never overtrades would score  $50 + 25 + 20 - 0 = 95$ . A trader who never uses the gate, reports 40% adherence, and overtrades on three days would score  $50 + 0 + 10 - 15 = 45$ .

### 3.2 Execution Consistency Score

The Execution Consistency Score measures the mechanical quality and repeatability of trade management. It is composed of three weighted factors: the proportion of trades entered in a calm or confident emotional state (weighted at 40%), the proportion of trades in which the stop-loss was not moved after entry (weighted at 35%), and a duration consistency metric (weighted at 25%).

Calm entry percentage captures whether the trader entered the market from a position of psychological readiness rather than reactivity. Stop-loss adherence measures whether the trader respected their initial risk definition or engaged in the common and destructive habit of widening stops to avoid being stopped out. Duration consistency, defined as one minus the coefficient of variation of trade durations and capped between zero and one, captures whether the trader holds positions for consistent periods or exhibits erratic behavior - cutting winners short on some trades while holding losers indefinitely on others.

$$\text{Execution} = \text{calmEntryPct} * 40 + \text{slAdherencePct} * 35 + \text{durationConsistency} * 25$$

where *calmEntryPct* is the fraction of trades with entry emotion recorded as “calm” or “confident,” *slAdherencePct* is the fraction of trades where the stop-loss was not modified post-entry, and *durationConsistency* =  $\max(0, \min(1, 1 - \sigma_{\text{duration}} / \mu_{\text{duration}}))$ .

### 3.3 Risk Management Score

The Risk Management Score evaluates whether the trader sizes positions consistently and protects every trade with a stop-loss. It is the simplest of the four scores, reflecting the principle that risk management, while critically important, is fundamentally binary in nature: either the trader sizes correctly and places stops, or they do not.

The score is the sum of two components, each contributing up to 50 points. The first is risk consistency, computed as one minus the coefficient of variation of lot sizes across all trades, scaled by 50. A trader who always uses the same lot size ( $CV = 0$ ) receives the full 50 points; a trader whose lot sizes vary wildly ( $CV$  approaching or exceeding 1) receives few or no points. The second component is stop-loss placement rate: the percentage of trades on which a stop-loss was set, multiplied by 50.

$$\text{RiskManagement} = (1 - CV(\text{lotSizes})) * 50 + (\text{tradesWithSL} / \text{totalTrades}) * 50$$

The coefficient of variation is used rather than standard deviation because it normalizes for the absolute magnitude of the lot sizes, making the metric comparable across traders with different account sizes.

### 3.4 Emotional Stability Score

The Emotional Stability Score is the most psychologically oriented of the four dimensions. It measures the trader's ability to maintain equanimity across the inevitable sequence of wins and losses that characterizes any trading methodology. Three inputs contribute: the proportion of trades exited in a calm or satisfied emotional state (positively weighted at 60%), the count of revenge trades (negatively weighted at 5 points per instance), and the count of trades tagged with a FOMO emotion at entry (negatively weighted at 5 points per instance).

$$\text{EmotionalStability} = \text{calmExitPct} * 60 - \text{revengeCount} * 5 - \text{fomoCount} * 5$$

A revenge trade is operationally defined as any trade opened within two minutes of closing a losing trade. This definition is conservative – many revenge trades occur within seconds – but the two-minute window captures the immediate emotional response without flagging trades that happen to follow a loss after a reasonable cooling period. FOMO entries are identified by the trader's self-reported emotional state at entry.

The heavy weighting on calm exits (60%) reflects the finding that how a trader handles exits – whether they close positions from a place of analytical clarity or emotional reactivity – is among the strongest predictors of long-term success. The revenge and FOMO penalties are additive and uncapped, meaning that a trader with many such instances can drive the score well below zero before clamping.

---

## 4. Per-Trade Decision Quality

### 4.1 Setup Score

Beyond the four aggregate scores, the Decision Grade framework assigns a per-trade Setup Score (0 to 100) that captures the quality of the decision-making process for each individual trade. This granular metric enables trade-by-trade analysis and powers the post-trade outcome classification system described in Section 4.2.

The Setup Score is computed as:

$$\begin{aligned} \text{setupScore} = & \text{adherence} * 0.4 \\ & + (\text{gateUsed} ? 30 : 0) \\ & + (\text{entryEmotion in ['calm', 'confident']} ? 20 : 0) \\ & + (\text{slNotMoved} ? 10 : 0) \end{aligned}$$

Plan adherence contributes up to 40 points (scaled from the 0-100% adherence value to a 0-40 range). Gate usage is a binary 30-point contribution – the single largest component, reflecting the platform's emphasis on structured pre-trade preparation. Entering in a calm or confident state adds 20 points. Not moving the stop-loss adds 10 points. A perfect Setup Score of 100 requires full plan adherence (40), gate completion (30), calm entry (20), and untouched stop-loss (10).

## 4.2 Post-Trade Outcome Classification

The most conceptually important element of the Decision Grade framework is the post-trade outcome classification matrix, which decouples process quality from outcome quality. Every closed trade is assigned to one of four categories based on two binary dimensions: whether the trade was executed with high adherence (defined as plan adherence at or above 75%) and whether the trade was profitable.

The four categories are as follows. A trade with high adherence that results in a profit is classified as a “Good Trade, Good Outcome” - the ideal case where process and result align. A trade with high adherence that results in a loss is classified as a “Good Trade, Bad Outcome” - a legitimate loss that was well-managed and should not trigger behavioral concern. A trade with low adherence that results in a profit is classified as a “Bad Trade, Lucky Outcome” - and this is the most dangerous category, because it reinforces the very behaviors that will eventually destroy the account. A trade with low adherence that results in a loss is classified as a “Bad Trade, Bad Outcome” - the expected consequence of poor process.

---

	Win	Loss
<b>Adherence <math>\geq</math> 75%</b>	Good Trade, Good Outcome	Good Trade, Bad Outcome
<b>Adherence <math>&lt;</math> 75%</b>	Bad Trade, Lucky Outcome	Bad Trade, Bad Outcome

---

The “Bad Trade, Lucky Outcome” cell deserves particular emphasis. When a trader enters impulsively, ignores their rules, skips the gate, and happens to catch a profitable move, the P&L statement records a win. Traditional metrics reward this trade identically to a perfectly executed entry. The Decision Grade framework flags it as dangerous reinforcement - a slot-machine payout that trains the trader to repeat the very behaviors that erode their edge. Over a sufficient sample size, the distribution across these four cells reveals whether a trader’s profitability is grounded in skill (high concentration in the Good Trade cells) or luck (significant presence in the Bad Trade, Lucky Outcome cell).

---

## 5. Worked Examples

### 5.1 Example 1: The Disciplined Loser

Consider a trader named Elena who has completed 40 trades over the past month on EURUSD during the London session. Elena is methodical: she completes the pre-trade gate on 38 of 40 trades, reports an average plan adherence of 82%, and has never exceeded five trades in a single day. Her Discipline Score is therefore  $50 + (0.95 * 25) + (82 * 0.25) - 0 = 50 + 23.75 + 20.5 = 94.25$ , which rounds to 94.

Elena enters 34 of her 40 trades in a calm or confident state (85%). She moves her stop-loss on only 3 trades (92.5% SL adherence). Her trade durations have a mean of 45 minutes and a standard deviation of 12 minutes, yielding a duration consistency of  $1 - (12/45) = 0.733$ . Her Execution Consistency Score is  $85 * 40 + 92.5 * 35 + 73.3 * 25 = 34 + 32.375 + 18.325 = 84.7$ .

Score	Inputs	Calculation	Result
Discipline	G=0.95, A=82, D=0	$50 + 23.75 + 20.5 - 0$	<b>94</b>
Execution	Calm=85%, SL=92.5%, Dur=0.733	$34 + 32.4 + 18.3$	<b>85</b>
Risk Mgmt	CV(lots)=0.08, SL rate=100%	$46 + 50$	<b>96</b>
Emotional	Calm exits=80%, Revenge=0, FOMO=1	$48 - 0 - 5$	<b>43</b>

However, Elena’s win rate is only 35%. She has 14 winning trades and 26 losing trades. Her total P&L is -\$340 over 40 trades. By conventional metrics, Elena is a failing trader. By the Decision Grade framework, Elena is a highly disciplined, consistent trader experiencing a legitimate drawdown - likely due to adverse market conditions for her methodology, not behavioral dysfunction. Her Emotional Stability score of 43 warrants attention (one FOMO entry, and her calm exit rate of 80% suggests some emotional turbulence on the remaining 20%), but her Discipline and Risk Management scores indicate a trader who will almost certainly recover once market conditions realign with her edge.

The post-trade classification confirms this interpretation. Of her 26 losses, 23 fall into the “Good Trade, Bad Outcome” category (adherence  $\geq$  75%). Of her 14 wins, 13 are “Good Trade, Good Outcome.” Only 1 trade is a “Bad Trade, Lucky Outcome” and 3 are “Bad Trade, Bad Outcome.” Elena’s outcome distribution is overwhelmingly concentrated in the Good Trade columns - exactly the pattern associated with long-term survival.

## 5.2 Example 2: The Lucky Reckless Trader

Now consider Marcus, who has completed 55 trades over the same month, primarily scalping XAUUSD and BTCUSD across all sessions. Marcus does not use the pre-trade gate. His plan adherence averages 38%. He has exceeded five trades on 7 different days. His Discipline Score is  $50 + (0.0 * 25) + (38 * 0.25) - (7 * 5) = 50 + 0 + 9.5 - 35 = 24.5$ .

Marcus enters trades in a variety of emotional states - only 15 of 55 entries are calm or confident (27.3%). He moves his stop-loss on 22 of 55 trades (60% SL adherence). His trade durations range wildly from 2 minutes to 4 hours, with a coefficient of variation of 1.4, yielding a capped duration consistency of 0. His Execution Consistency Score is  $27.3 * 40 + 60 * 35 + 0 * 25 = 10.9 + 21 + 0 = 31.9$ .

Marcus’s lot sizes vary from 0.1 to 2.5 lots with a CV of 0.9, and he places a stop-loss on only 60% of his trades. His Risk Management Score is  $(1 - 0.9) * 50 + 0.6 * 50 = 5 + 30 = 35$ .

Score	Inputs	Calculation	Result
Discipline	G=0.0, A=38, D=7	$50 + 0 + 9.5 - 35$	<b>25</b>
Execution	Calm=27.3%, SL=60%, Dur=0.0	$10.9 + 21.0 + 0$	<b>32</b>
Risk Mgmt	CV(lots)=0.9, SL rate=60%	$5 + 30$	<b>35</b>
Emotional	Calm exits=30%, Revenge=8, FOMO=6	$18 - 40 - 30$	<b>0 (clamped)</b>

Yet Marcus is profitable. He has 30 wins and 25 losses (54.5% win rate) with a total P&L of +\$1,820. His equity curve looks impressive. Traditional metrics would rank Marcus above Elena.

The Decision Grade framework tells a starkly different story. Marcus’s scores are universally poor: 25, 32, 35, and 0. His Emotional Stability score of 0 (before clamping, actually -52) reflects 8 revenge trades and 6 FOMO entries – nearly a quarter of his total trades are emotionally driven. The post-trade classification is revealing: of his 30 wins, 19 fall into the “Bad Trade, Lucky Outcome” category (adherence < 75%). Only 11 are “Good Trade, Good Outcome.” More than 60% of his profitable trades were executed in violation of his own plan.

Classification	Count	Percentage
Good Trade, Good Outcome	11	20%
Good Trade, Bad Outcome	8	14.5%
Bad Trade, Lucky Outcome	19	34.5%
Bad Trade, Bad Outcome	17	31%

Marcus is a trader living on borrowed time. His 19 “lucky outcome” trades have created a positive equity curve that masks catastrophic behavioral dysfunction. The Decision Grade framework predicts, with high confidence, that Marcus’s profitability will not survive the next 100 trades unless his behavioral scores improve. The coaching system would surface escalating interventions: forced cooldowns after consecutive losses, pre-trade gate requirements, position size restrictions, and – if his scores remain low – a regression-triggered interface expansion that re-enables behavioral monitoring tools he may have previously dismissed.

### 5.3 Example 3: The Recovering Trader

Priya began her trading journey with behavioral scores similar to Marcus’s. Over her first 15 trades, her scores were Discipline 31, Execution 28, Risk Management 40, and Emotional Stability 12. She was revenge trading frequently, rarely using the gate, and entering positions impulsively after losses. Her P&L was -\$620.

The Decision Grade framework triggered several interventions. After her third consecutive loss, the escalating cooldown system imposed a 90-minute trading pause and surfaced a forced reflection modal. The coaching engine recommended specific academy lessons on pre-trade preparation and emotional regulation. The interface density system, detecting scores below 50 across three of four dimensions, maintained the fully expanded interface to keep all behavioral monitoring tools visible.

Priya began responding to these signals. Over trades 16 through 30, her behavior shifted measurably. She began using the pre-trade gate (12 of 15 trades) and her plan adherence rose to 70%. She moved her stop-loss on only 2 of 15 trades. Her lot sizes stabilized as she began using the position size calculator consistently.

The following table traces Priya’s score evolution across three 10-trade windows:

Period	Trades	Discipline	Execution	Risk Mgmt	Emotional	Win Rate	P&L
Trades 1-10	10	28	25	35	8	20%	-\$480
Trades 11-20	10	52	48	55	35	40%	-\$140
Trades 21-30	10	74	68	78	58	50%	+\$210

The scores improved before the P&L did - a critical observation. Between trades 11 and 20, Priya's Discipline score doubled from 28 to 52, and her Risk Management score rose from 35 to 55, yet she was still losing money (though at a sharply reduced rate). By trades 21 through 30, the behavioral improvement had begun translating into results: a 50% win rate and net-positive P&L. The Decision Grade framework provided Priya - and her coaching system - with leading indicators of recovery that the equity curve alone could not have surfaced until much later.

By trade 30, Priya's outcome classification had shifted dramatically. In her first 10 trades, 7 of 8 losses were "Bad Trade, Bad Outcome." In trades 21 through 30, only 1 of 5 losses was a "Bad Trade, Bad Outcome" - the remaining 4 were "Good Trade, Bad Outcome," indicating that her losses were now the product of legitimate market risk rather than behavioral dysfunction.

---

## 6. Execution Error Taxonomy

The Decision Grade framework identifies seven distinct execution error types, each detected through a combination of trade-level metadata and temporal analysis. These errors are not merely counted; they are categorized, timestamped, and fed into the deviation ledger for acknowledgment and reflection.

The first error type is the FOMO entry, detected when the trader's self-reported emotional state at the time of entry is "fomo." While this relies on self-reporting, it is cross-validated against objective signals: FOMO entries correlate strongly with entries placed outside of killzone windows and with above-average position sizes, both of which are independently measurable.

The second error type is the revenge entry, detected when a trade is opened within two minutes of closing a losing trade. This is one of the most destructive behavioral patterns in trading, as it represents an attempt to "get back" at the market rather than executing from a position of analytical clarity. The two-minute window was calibrated empirically: analysis of trade logs shows a sharp discontinuity in trade quality for entries within this threshold versus entries placed after a cooling period.

The third error type is the hesitant entry, detected when a trade is placed after an unusually long delay following a winning trade. This captures the opposite pathology from revenge trading: a trader who has won becomes afraid of giving back profits and hesitates on subsequent setups, leading to suboptimal entry timing or missed opportunities.

The fourth error type is stop-loss movement, detected when the stop-loss price is modified after the initial trade entry. While there are legitimate reasons to move a

stop-loss (such as trailing it in the direction of profit), the vast majority of stop-loss movements in retail trading involve widening the stop to avoid being taken out of a losing position - a behavior that transforms a defined-risk trade into an undefined-risk gamble.

The fifth error type is the panic exit, detected when the trader's self-reported emotional state at exit is "anxious." Panic exits typically involve closing a position prematurely during normal market volatility, locking in a loss or a reduced profit that would have been avoided by adhering to the original trade plan.

The sixth error type is the greedy exit, detected when the trader's exit emotion is "greedy." Greedy exits involve holding a position beyond the planned target in hope of additional profit, often resulting in a reversal that erases the unrealized gains.

The seventh error type is low adherence, detected when the trader's plan adherence for a given trade falls below 50%. This is a catch-all category that captures trades where the trader deviated significantly from their declared methodology, regardless of the specific nature of the deviation.

---

## 7. Aggregation and Feedback Loops

The Decision Grade scores do not exist in isolation; they are woven into a system of feedback loops that translate measurement into behavioral change. These loops operate at three temporal scales: immediate (per-trade), session-level (intra-day), and longitudinal (multi-week).

At the immediate level, every trade generates a Setup Score and an outcome classification. If the Setup Score falls below a configurable threshold, the coaching engine surfaces a specific recommendation in the journal feed. If a revenge trade or FOMO entry is detected, an escalating cooldown system activates: the first instance triggers a 15-minute pause, the second a 30-minute pause, and the third a full lockdown of 60 minutes with a mandatory reflection prompt. These interventions create a direct causal link between the measured behavior and an immediate consequence, leveraging the proximity principle from behavioral psychology.

At the session level, the four aggregate scores are recomputed after every trade and displayed on a persistent status bar. The coaching engine analyzes the current session's scores relative to rolling 10-trade baselines and generates targeted recommendations. If the Emotional Stability score has declined by more than 15 points within the current session, the coaching system may recommend ending the session early. If the Discipline score is high but the win rate is low, the coaching system contextualizes the drawdown as regime-driven rather than behavioral, reducing the likelihood that the trader abandons a working methodology during a temporary adverse period.

At the longitudinal level, the scores feed into an ELO-like Rating system (initialized at 1200) that tracks the trader's overall behavioral trajectory. Each closed trade nudges the rating: +5 for a win, -3 for a loss, -10 for a revenge trade, -5 for a moved stop-loss, +3 for high adherence. When the Rating crosses 1600, the platform's interface density automatically compresses - collapsing secondary panels, hiding be-

havioral monitoring tools, and surfacing only the metrics that a proven trader needs. If the Rating subsequently declines (indicating behavioral regression), the interface re-expands and a diagnostic banner identifies the specific regression triggers. This density adaptation ensures that the platform’s cognitive load scales inversely with the trader’s demonstrated competence.

The Discipline Score variant used within the prop firm challenge simulator illustrates a complementary aggregation approach. Rather than using the gate-and-adherence formula from the psychology module, the prop challenge computes discipline from three directly observable factors: lot size consistency (1 minus the coefficient of variation of lot sizes), stop-loss usage rate (proportion of trades with SL), and killzone timing rate (proportion of trades entered within designated ICT killzones). The average of these three factors, multiplied by 100, produces a discipline metric that requires no self-reported inputs and can be computed entirely from trade execution data. This dual-formula approach – one relying on self-report enriched with process data, the other purely execution-derived – provides a consistency check: if the two Discipline scores diverge significantly, it suggests that the trader’s self-assessment of their adherence may not match their actual execution patterns.

---

## **8. Validation Methodology**

The Decision Grade framework makes an implicit empirical claim: that traders with higher behavioral scores will, over a sufficient sample size, achieve superior risk-adjusted returns compared to traders with lower scores, controlling for methodology and market conditions. This claim must be validated through longitudinal analysis.

The proposed validation methodology involves computing the Pearson correlation between each of the four behavioral scores (averaged over a trailing 50-trade window) and the trader’s subsequent 50-trade Sharpe ratio. A positive and statistically significant correlation ( $p < 0.05$ ) between behavioral scores and forward-looking risk-adjusted returns would constitute evidence that the scores are measuring something predictive rather than merely descriptive.

Additional validation analyses include examining the relationship between the “Bad Trade, Lucky Outcome” rate and subsequent drawdown severity (the hypothesis being that a high lucky-outcome rate precedes catastrophic drawdowns), the correlation between Setup Score trends and win-rate trends (with behavioral improvement expected to lead P&L improvement by 10 to 20 trades), and the effectiveness of specific interventions (comparing revenge trade frequency before and after escalating cooldown activation).

The platform’s architecture, which stores every trade with its full behavioral metadata in a centralized data engine, provides the longitudinal dataset necessary for these analyses. As the user base grows, cross-sectional comparisons between traders at different score levels will enable increasingly robust validation.

---

## 9. Limitations and Open Questions

The Decision Grade framework carries several limitations that must be acknowledged. The most significant is its partial dependence on self-reported data. The Execution Consistency and Emotional Stability scores both incorporate the trader’s declared emotional state at entry and exit. While this data is cross-validated against objective signals (timing, position sizing, stop-loss behavior), it remains susceptible to reporting bias. A trader who always reports “calm” regardless of their actual state would receive inflated scores on these dimensions. Future work may explore physiological signals (heart rate variability, galvanic skin response) as objective emotional proxies, though this introduces hardware dependencies that conflict with the platform’s browser-based architecture.

A second limitation is the fixed weighting structure. The relative weights within each score (for example, 40% calm entries, 35% SL adherence, 25% duration consistency for Execution) were set based on domain expertise rather than empirical optimization. These weights may not be optimal for all trading styles. A scalper who holds trades for seconds has a fundamentally different duration distribution than a swing trader who holds for days; the duration consistency component may penalize the scalper unfairly. Adaptive weighting, calibrated to the trader’s methodology and timeframe, is a natural extension.

A third limitation is the binary threshold in the outcome classification matrix. The 75% adherence cutoff that separates “Good Trade” from “Bad Trade” is somewhat arbitrary. A trade with 74% adherence is classified identically to one with 10% adherence, despite the vast difference in process quality. A continuous classification scheme, where the “goodness” of the trade is proportional to adherence, would provide more nuance at the cost of reduced interpretability.

A fourth consideration is the interaction between the Decision Grade framework and the Lock gating system. The Lock uses a separate weight vector (initialized from a calibration quiz and updated via trade outcomes and override learning) to generate pre-trade go/no-go verdicts. There is a question of whether Lock-aligned trades should receive a Decision Grade bonus, or whether the two systems should remain fully independent. The current implementation treats them as orthogonal: a trade that violates the Lock but follows all other rules still receives high behavioral scores. Whether this is the correct design choice depends on the Lock’s demonstrated predictive accuracy, which is itself subject to ongoing validation.

Finally, the framework operates entirely client-side, with scores computed in the browser from localStorage data. This means that the scoring is vulnerable to data manipulation by technically sophisticated users. For prop firm evaluation contexts, where the scores carry economic consequences, server-side computation with tamper-proof audit trails would be necessary.

## 10. Invitation for Peer Review

The Decision Grade methodology is published here as an open framework. We invite quantitative researchers, trading psychologists, prop firm operators, and platform developers to scrutinize, critique, replicate, and extend this work. Specific questions we believe would benefit from independent investigation include the following.

Are the four behavioral dimensions truly independent in practice, or do they exhibit empirical correlations that suggest a lower-dimensional structure? What is the minimum sample size (number of trades) required for each score to stabilize and become predictively useful? Can the self-reported emotional inputs be replaced entirely by objective behavioral proxies without loss of predictive power? How should the weighting structure adapt for different trading methodologies (scalping versus swing, directional versus mean-reversion)? And most fundamentally: does the Decision Grade framework improve trader outcomes, or does it merely describe them?

These questions cannot be answered by the authors alone. They require diverse datasets, independent implementations, and critical examination from practitioners who bring different perspectives and methodological commitments. We believe that the measurement of trading quality is too important to be left to P&L alone, and we hope that this framework contributes to a more rigorous and humane approach to trader development.

---

*Correspondence: [adrian.djents@gmail.com](mailto:adrian.djents@gmail.com) Platform: <https://drivebynumbers.com>*