

Dynamics of Behavioral Risk-Taking: A Dynamic-Hazard BART and the DOSPERT-30

An Honors Seminar Report

The dynamic-hazard task, all behavioral metric definitions, the Monte Carlo optimization study, the full-stack web platform, and the complete analytical pipeline described below were designed and implemented independently by the author. The live platform is available at bart.aselimyilmaz.com. All empirical findings are reported as exploratory; the single confirmatory test in this report is the pre-registered external replication described in Section 4.6, and its outcome is reported in full, including the hypotheses that did not replicate.

Abstract

For decades, a stubborn puzzle has sat at the center of the psychology of risk: people's stated attitudes toward risk barely predict how they actually behave when something is on the line. Ask someone how much financial risk they will tolerate, then watch them play a financial risk-taking game, and the two numbers hardly move together. This report argues that part of this famous attitude-behavior gap is not a fact about people at all, but an artifact of the games we use to measure them. The canonical Balloon Analogue Risk Task (BART; Lejuez et al., 2002) sets its optimal stopping point so high that almost no one reaches it, which collapses the task into a measure of raw exposure rather than skilled, calibrated risk-taking. To address this, a modified BART was engineered in which the probability of bursting rises linearly with every pump, governed by a sequential Bernoulli hazard model. This single change lowers the expected-value-optimal stopping point from an unreachable 64 pumps to a reachable handful (approximately the square root of each balloon's capacity), restoring the task's ability to separate the cautious, the calibrated, and the reckless. The platform was deployed as an open-source web application and administered to laboratory and online cohorts (final $N = 70$). In exploratory analysis, a counterintuitive pattern emerged and survived a multi-estimator gauntlet of canonical correlation analysis, robust Bayesian estimation, exploratory factor analysis, and machine-learning feature importance: the self-report subscale that best tracked successful, calibrated behavioral risk-taking was not financial-gambling, the domain-matched candidate, but the DOSPERT *Social* subscale, which indexes willingness to challenge authority and voice unpopular views. A pre-registered test on a large independent dataset ($N = 1,505$; Frey et al., 2017) confirmed that the basic social-behavior link generalizes, but showed that social's specific dominance is task-dependent: it appears only when the task actually rewards calibration. Read together, these results suggest that the psychometric signal linking stated and behavioral risk has been there all along, but only becomes visible through task architectures built to reveal it.

1. Introduction

1.1 The psychometric structure of risk and the attitude-behavior gap

"Risk" names a family of properties of the natural and social world, from probability and outcome variance to danger and unquantifiable uncertainty. Human responses to those properties, unified conceptually as *risk preference*, form a basic building block of psychological and economic theories of choice (Weber, Blais, & Betz, 2002; Frey et al., 2017). Yet despite decades of interdisciplinary effort, a single reliable metric of risk preference has proven elusive.

A large body of empirical work has documented a persistent divide between *stated* (self-reported) risk preference and *elicited* (behavioral) risk-taking. In the most comprehensive study to date, Frey et al. (2017) administered 39 risk-taking measures to more than 1,500 adults and found that self-report propensity measures and reports of real-world risky activity cohere into a highly reliable general factor of risk preference, labeled R, that accounts for more than half of the systematic variance in those measures, much as g organizes intelligence. Behavioral tasks, however, including the canonical BART (Lejuez et al., 2002), the Columbia Card Task (Figner, Mackinlay, Wilkening, & Weber, 2009), and lottery paradigms (Holt & Laury, 2002), barely load onto this general factor. Each behavioral task instead captures idiosyncratic variance that is largely unrelated to stated preference, and often unrelated even to other behavioral tasks in the same battery (Pedroni et al., 2017).

This attitude-behavior gap poses a serious problem. Behavioral tasks have long been preferred by economists for their incentivized structure and grounding in expected-utility theory, yet they do not appear to measure a unified risk preference across situations. When researchers attempt to establish the convergent validity of a behavioral measure, they typically apply a *domain-matched* assumption: that stated willingness to take financial or gambling risks should predict actual risk-taking in a financial gambling task such as the BART. In practice, this domain-matched validation routinely fails, returning weak, non-credible, or entirely absent associations (Szrek, Chao, Ramlagan, & Peltzer, 2012; Frey et al., 2017). The growing use of game-based and gamified assessments in personnel selection, where such tasks are increasingly deployed precisely because they are thought to resist the social-desirability and faking biases of self-report (Armstrong, Landers, & Collmus, 2016; Landers & Sanchez, 2022; Paulhus, 1984), makes resolving this gap a practical as well as a theoretical concern.

1.2 The failure of canonical task architectures

The usual interpretation of this failure is that the behavioral task fails to capture the underlying trait, which fuels the long-running debate over whether risk preference is domain-general or domain-specific (Weber et al., 2002). A structural reading suggests a different culprit: the rigid architecture of the canonical task itself.

In a traditional static-hazard BART, participants inflate a single type of balloon whose explosion point is drawn from a uniform distribution, and the maximum capacity is

typically set high, often 128 pumps. Under that scheme the expected-value (EV) optimal stopping point is also high, near 64 pumps. The empirical record shows that human participants are uniformly risk-averse in this paradigm, routinely terminating trials well short of the optimum, often around 40 pumps on average across large cohorts (Frey et al., 2017; Pleskac, 2008).

Because virtually no one reaches the mathematically optimal threshold, the relationship between pump count and money earned becomes almost perfectly monotonic. In the Frey et al. (2017) data, the two correlate at roughly 0.93. The task therefore stops measuring *calibrated* risk-taking, in which a participant must balance greed against rapidly escalating probability, and degrades into a one-dimensional index of raw exposure or luck. A task that measures only gross, undifferentiated exposure cannot reasonably be expected to correlate with a sophisticated, domain-specific psychological trait. This motivates a concrete design goal: to decouple successful, calibrated risk-taking from indiscriminate high-volume exposure, the digital environment must be re-engineered so that the optimum is reachable and overshooting it is penalized.

1.3 An exploratory origin

The recognition that domain-matched financial attitudes fail to track calibrated behavioral risk did not arrive as a top-down hypothesis. It surfaced from an unsupervised, exploratory data-mining phase. The first analyses of the new dynamic-hazard task mapped its behavioral telemetry against the six subscales of the DOSPERT-30.

The pattern appeared first in pairwise k-means clustering grids, which searched for structure in two-dimensional spaces crossing each DOSPERT subscale against single BART metrics. Financial-gambling against earnings produced diffuse, unstable clusters. Social against earnings, by contrast, partitioned cleanly into joint clusters, separating high-social and high-earning participants from low-social and low-earning ones. This unexpected visualization became the hypothesis-generating origin of the present work. Because unsupervised clustering offers exploratory vision rather than confirmatory proof, a large analytical battery was then built to stress-test the observation across canonical correlation analysis, robust Bayesian estimation, exploratory factor analysis, and machine-learning feature importance, with the explicit question of whether the social subscale is genuinely the self-report dimension that tracks successful risk execution.

1.4 The present contribution

This report makes three connected contributions. First, it describes a methodological one: a dynamic-hazard BART, grounded in a sequential Bernoulli model and validated by Monte Carlo simulation, that restores the measurement of calibration and is openly deployed as a live web platform. Second, it reports an exploratory empirical finding: across multiple estimators with different failure modes, the DOSPERT Social subscale, not the domain-matched financial-gambling subscale, best tracks successful behavioral risk-taking in this sample. Third, it reports a pre-registered external replication that delineates the boundary of that finding, showing that the basic link generalizes to a large independent sample while its specific strength depends on whether the task rewards calibration. The remainder of the

report follows a conventional structure: methods, implementation, results, discussion, and future work, with all technical detail beyond what is needed for interpretation deferred to the supplementary materials.

2. Methods

2.1 The dynamic-hazard BART and the sequential Bernoulli model

In the canonical BART, the explosion probability on a given pump is effectively static, governed by a single uniform draw over the balloon's capacity. The task engineered here replaces that with a sequence of independent Bernoulli trials in which the probability of bursting rises linearly with each successive pump. For pump attempt k , the probability of explosion is

$$P(\text{explode at pump } k) = k / N,$$

where N is the maximum pump limit for a given balloon profile. The probability of surviving s sequential pumps is the product of the per-pump survival probabilities,

$$P(\text{survive } s) = \prod_{k=1}^s (1 - k / N).$$

By convention this returns 1.0 for $s \leq 0$ and is floored at 0.0 once the product reaches or would cross zero, which caps the distribution at N .

Because the hazard accelerates linearly, the expected reward of stopping at s does not grow monotonically. With a reward of one unit per pump, the expected value of intending to stop at s is

$$EV(s) = s \times \prod_{k=1}^s (1 - k / N).$$

This nonlinear EV curve produces a distinctive optimization landscape. In the static uniform-hazard task, where the burst point is uniform on the capacity, the optimum sits at half the capacity, $N / 2$, which is why a 128-pump balloon has an unreachable optimum near 64. Under the linear-hazard model the optimum is far lower. A short derivation makes this precise. Taking logs of the survival product and using the leading term of the expansion of the logarithm,

$$\ln P(\text{survive } s) \approx - \sum_{k=1}^s (k / N) \approx - s^2 / (2N),$$

so that survival is approximately Gaussian in s , $P(\text{survive } s) \approx \exp(-s^2 / 2N)$, and

$$EV(s) \approx s \cdot \exp(-s^2 / 2N).$$

Differentiating and setting the result to zero gives the closed-form optimum

$$s^* = \sqrt{N}.$$

The square-root law is the central mathematical property of the design. It guarantees that EV-optimal play is psychologically reachable for ordinary participants without inducing the fatigue of clicking dozens of times, and by making the optimum reachable it restores the

task's ability to distinguish overly conservative players, mathematically optimal players, and over-exposed reckless players, the very separations the static task cannot make.

Two refinements matter for interpretation. First, because the continuous approximation drops higher-order terms that make survival decay slightly faster, the realized *integer* optima are not the rounded square root but its floor. For the three profiles used here (capacities 128, 32, and 8), the closed-form optima are $\sqrt{128} \approx 11.3$, $\sqrt{32} \approx 5.7$, and $\sqrt{8} \approx 2.8$, and the exact discrete optima computed from the full EV product are 11, 5, and 2 pumps respectively. Second, to confirm these integer optima without relying on the approximation, a Monte Carlo simulation (developed initially as a computational simulation course project) was implemented: for each candidate stopping threshold on each balloon profile, many sequential-Bernoulli trials were simulated and the realized payoffs averaged, with the maximizing threshold recorded. The simulation reproduced the integer optima of 11, 5, and 2 and confirmed the $\sqrt{N}$ approximation as an accurate description of the continuous landscape. This pairing of a closed-form solution with a Monte Carlo verification provides both an analytic and an empirical foundation for the task's scoring.

2.2 Multi-risk color profiles

To assess behavioral adaptability and risk diversification within a single session, the task uses three balloon profiles distinguished by neutral colors. Neutral colors were chosen deliberately to avoid the framing biases carried by red, which signals danger, or green, which signals safety.

- **Purple (low risk, high capacity).** Maximum $N = 128$, optimal stop ≈ 11 pumps. Targets patience, prolonged engagement, and willingness to exploit a forgiving hazard environment.
- **Teal (medium risk, benchmark).** Maximum $N = 32$, optimal stop ≈ 5 pumps. Serves as the baseline risk benchmark bridging the extremes.
- **Orange (high risk, low capacity).** Maximum $N = 8$, optimal stop ≈ 2 pumps. Tests acute impulse control and sensitivity to rapid hazard escalation.

By rotating these profiles unpredictably across a compressed 30-trial session, the task forces continuous cognitive flexibility and penalizes participants who apply a single uniform strategy across shifting conditions (Di Plinio, Pettorruso, & Ebisch, 2022, who demonstrated in simulation that hazard structure materially changes the risk profiles a BART can recover).

2.3 Learning and adaptation metrics

A central analytical advantage of the multi-schedule design is the ability to measure within-task learning. Learning is operationalized through a linear regression of pump count on trial number, fit separately for each color profile (pumps $\sim$ trial number).

To preserve measurement fidelity, the scoring engine applies a mandatory RNG-truncation correction. On a balloon that bursts, the recorded final pump count reflects only where the

random number generator terminated the trial, not the participant's intended stopping threshold. Including those truncated values introduces a severe downward bias, so the learning regression is computed on collected (non-exploded) balloons. This concern with the difference between intended and realized stopping behavior follows the sequential-sampling tradition in BART modeling (Pleskac, 2008).

Because the three profiles demand different responses for optimal play, the directional meaning of a learning slope is profile-dependent. For orange ($N = 8$), optimal learning requires reducing pumps over time, so adaptation is scored as the negated weighted slope. For purple ($N = 128$), optimal learning requires overcoming baseline aversion to exploit the large capacity, so adaptation is scored as the positive weighted slope. Teal ($N = 32$) is excluded from the composite learning metric because its directionality is ambiguous, depending on whether a participant's initial baseline fell above or below the six-pump optimum. To suppress noise from erratic single trials, each raw slope is weighted by its coefficient of determination (R^2), producing a weighted-slope metric that rewards consistent, deliberate adaptation.

2.4 Behavioral metric families and the Lejuez-3 composite

From the raw telemetry, a large feature matrix was extracted and then subjected to a systematic data audit to remove multicollinearity and structural redundancy before any modeling. The audit empirically confirmed and dropped six alias or composite metrics (for example, a money-efficiency index numerically identical to money collected, and a financial-mean score that was the exact average of its two subfacets), leaving 23 vetted features grouped into five conceptual families: magnitude, EV-calibration, learning, post-explosion adjustment (PES), and consistency or timing.

For primary cross-estimator comparison, analyses centered on a Lejuez-3 composite, modeled on the canonical triad of the original task (Lejuez et al., 2002):

- **RNG-Normalized Pumps.** Average intended risk, corrected for RNG truncation.
- **Explosion Rate.** Gross proportion of trials ending in a burst, a proxy for uncalibrated recklessness.
- **Money Collected.** Total financial yield. Because the dynamic-hazard design penalizes overshooting the optimum, this metric indexes EV-calibration and efficiency rather than serving as a redundant proxy for exposure.

The primary composite behavioral outcome is the mean of the z-scores of these three metrics, providing a stabilized index of active, profitable risk engagement. Variables with near-zero variance, such as a binary flat-strategy flag on which 65 of 70 participants scored zero, were excluded from correlation modeling to prevent degenerate estimates (a deliberate parallel to the exclusion of a poorly distributed measure in Frey et al., 2017).

2.5 The DOSPERT-30 (Turkish adaptation)

Self-reported domain-specific risk propensity was measured with the 30-item Domain-Specific Risk-Taking scale (DOSPERT; Weber et al., 2002; revised short form, Blais & Weber,

2006), in a validated Turkish adaptation suited to the linguistic and cultural context of the sample [**Turkish DOSPERT-30 adaptation: please confirm the exact reference for the validation you administered; candidates located include a 2019 Turkish DOSPERT-30 adaptation study and Dinç & Tez, 2019**]. Participants rate their willingness to engage in risky activities on a 1-to-7 Likert scale.

For high-resolution mapping, the financial domain is split, yielding six functional subscales: Financial-Gambling (for example, betting a day's income at high-stakes poker), Financial-Investment (for example, investing a share of income in a speculative stock), Health and Safety (for example, driving without a seatbelt), Recreational (for example, skydiving), Ethical (for example, passing off another's work as one's own), and Social (for example, speaking up to challenge an unpopular opinion or an authority figure). The default assumption in behavioral economics and psychology is that the Financial-Gambling subscale should anchor convergent validity for a financial risk task such as the BART. Testing that domain-matched assumption is the core objective of the analytic pipeline. One psychometric caveat is worth noting in advance: in a large reliability meta-analysis of the DOSPERT, the social subscale showed the lowest internal consistency of all six domains (mean coefficient alpha of 0.68 versus up to 0.80 for recreational; Highhouse, Nye, Zhang, & Rada, 2017), a point revisited in the limitations.

2.6 Analytic strategy

The analysis advanced deliberately from exploratory structure discovery to confirmatory probability statements, so that no single method carried the full evidential weight. The pipeline comprised dimensionality reduction (principal component analysis and manifold embedding), unsupervised behavioral phenotyping (k-means with bootstrap stability assessment), supervised feature importance (random forests with permutation importance), canonical correlation analysis between the behavioral and self-report blocks, robust Bayesian parameter estimation following the reporting framework of Frey et al. (2017), and exploratory factor analysis with a Schmid-Leiman general-factor decomposition. Confirmatory weight rests on a single pre-registered external replication. The specific methods, priors, and diagnostics are introduced alongside their results in Section 4 and detailed in the supplementary materials.

3. Implementation

3.1 Open-source web platform

The experiment was built as a full-stack, open-source web application, designed with modern interface principles to bridge the divide between tightly controlled laboratory settings and scalable online collection. The platform remains live at bart.aselimyilmaz.com and renders the task, administers the DOSPERT-30, captures pump-level behavioral telemetry with timestamps, and runs the scoring and validation logic described below. Releasing the platform openly is intended to let other researchers deploy, score, and analyze dynamic-hazard behavioral data without rebuilding the infrastructure.

3.2 Laboratory and online administration

Data were collected through two channels using the same application. A laboratory channel administered the task under proctored conditions, and an online channel allowed remote participation through the hosted platform. Running the identical instrument across both channels permits a direct comparison of behavioral structure across modalities while holding the measurement instrument constant. Because modality was confounded with demographic differences between the recruitment pools, such as age and employment profiles, modality is treated analytically as a comparison variable rather than as a covariate to be adjusted away, a decision examined empirically in Section 4.

3.3 Automated session-validation pipeline

Remote behavioral data are vulnerable to bot automation, distracted participation, and superficial engagement. A strict validation pipeline was therefore embedded directly in the scoring engine, flagging or voiding sessions on the following automated criteria:

- **Completeness and attrition.** Sessions with fewer than 15 of 30 balloons are voided; sessions with 15 to 29 are retained but flagged as incomplete.
- **Color distribution.** Fewer than 5 balloons of a profile triggers a distribution warning; fewer than 2 collected balloons of a profile causes the engine to fall back to including exploded balloons for that profile's metrics, preventing division-by-zero in feature extraction.
- **Data integrity and monotonicity.** Event timestamps logged out of sequence, which indicate browser lag, network desynchronization, or tampering, void the session.
- **Speed and automation defenses.** Sessions completed in under 30 seconds are flagged; any session with more than 10 balloons whose pump-count standard deviation is near zero (below 0.5) is flagged as non-genuine, catching bots and mindless button-mashing.

3.4 Sample and cohort characteristics

After applying the exclusion criteria, the final dataset comprised 70 complete cases, drawn from proctored laboratory ($n = 17$) and remote online ($n = 53$) cohorts. The unsupervised discovery analyses (Sections 4.1 to 4.3) were run on the 69 complete cases available at an earlier data export; a subsequent refresh added one online participant, yielding the 70 used in the Bayesian convergent-validity analyses. Shared-IP flags in the laboratory environment were reviewed and retained after sensitivity checks, as they reflected local network routing rather than duplicate participation. Demographics collected included age, gender, faculty affiliation, degree, employment status, and prior task experience.

Table 1 compares the two cohorts on core behavioral metrics. Variance is marginally higher online, as expected in unproctored settings, but mean performance is tightly aligned across channels. As Section 4.1 reports, modality intermixed continuously in every projection, indicating that the delivery channel did not fundamentally distort behavioral phenotypes.

Metric	Lab mean (n = 17)	Lab SD	Online mean (n = 52)	Online SD	Δ mean
Total time (min)	2.46	0.60	2.42	1.64	0.04
Total pumps	129.00	23.61	131.92	37.27	-2.92
Explosions (count)	14.71	2.80	15.06	4.23	-0.35
Explosion rate	0.49	0.09	0.50	0.14	-0.01
Risk-taking (normalized)	0.98	0.20	0.93	0.24	0.05
Money collected	18.10	2.76	17.72	4.10	0.39
Impulsivity	0.29	0.19	0.37	0.23	-0.08
Average pumps (adjusted)	4.95	1.48	5.23	2.10	-0.28
Mean latency (ms)	579.32	166.69	516.27	256.51	63.06

Note. Values from the vetted descriptive matrix; one incomplete online session affects the n for specific metrics.

4. Results

All analyses in Sections 4.1 to 4.5 and 4.7 are exploratory and were conducted on the same sample in which the central pattern was discovered, so they characterize and triangulate that pattern rather than confirm it. Section 4.6 reports the one confirmatory test, pre-registered and run on an independent sample.

4.1 Dimensionality and unsupervised phenotyping

Mapping the 23 vetted features with principal component analysis (PCA) immediately surfaced a structural anomaly: on the first principal component, the DOSPERT Social subscale loaded in the opposite direction to every other DOSPERT subscale, an early signal of its distinct psychometric behavior. To visualize the broader structure, a three-dimensional Uniform Manifold Approximation and Projection embedding was applied (UMAP; McInnes, Healy, & Melville, 2018), which preserves fuzzy topological structure better than linear projection or t-SNE (van der Maaten & Hinton, 2008); for this dataset its utility was largely confirmatory of the PCA picture rather than additive. Embedding

coordinates are treated strictly as visualization aids, and no cluster statistics were computed on them, since manifold embeddings do not preserve metric distances.

Moving from visualization to modeling, k-means clustering (MacQueen, 1967) was applied to the BART feature matrix alone, with no input from the DOSPERT self-reports, and the stability of every solution was assessed by nonparametric bootstrap (Hennig, 2007). This established behavioral phenotypes from task performance only. Restricting to the behavioral subspace produced clusters that, for the first time in the pipeline, crossed the conventional bootstrap-stability threshold, in contrast to attempts to recover hand-coded rule-based archetypes, which agreed only weakly with the data-driven solution (Adjusted Rand Index ≈ 0.20 ; Hubert & Arabie, 1985). Critically, because these clusters were generated without any self-report input, they constitute a non-circular target for the predictive test that follows.

4.2 Random Forest feature importance

To run an independent, non-circular test of convergent validity, a random forest classifier (Breiman, 2001) was trained to predict the unsupervised behavioral clusters from the six DOSPERT subscales, with permutation importance used to rank each subscale's contribution to predictive accuracy. The ranking placed the Social subscale first, identifying it as the most informative self-report predictor of behavioral phenotype, while the domain-matched Financial-Gambling subscale ranked last. At this sample size the confidence intervals around permutation importance are wide, so the result is read as ordering evidence that corroborates the linear models that follow rather than as a precise effect estimate.

4.3 Canonical correlation analysis: the two-axis story

To map the multidimensional relationship between attitudinal subscales and behavioral execution, canonical correlation analysis was applied (CCA; Hotelling, 1936). On the well-powered Lejuez-3 feature set (ratio of observations to variables ≈ 7.8), the first canonical axis (CC1, canonical correlation $\rho \approx 0.53$) described a pathway of successful, calibrated risk-taking. On that axis the Social subscale loaded at +0.85, far above any other subscale, with recreational and financial-gambling both near +0.25.

Stress-testing against the full 14-feature behavioral matrix (underpowered, with a ratio of 3.5) left the structure intact (CC1 $\rho \approx 0.72$): Social loaded at +0.64, recreational at +0.40, and financial-gambling collapsed to a negligible -0.01 . The CCA also extracted a second, orthogonal axis on which all DOSPERT subscales loaded together in a diffuse pattern resembling the weak general propensity factor described in prior literature, and on which Social loaded near zero. Because canonical axes are orthogonal by construction, this pattern indicates that social risk willingness and financial-gambling willingness occupy different attitude-behavior bridges: the familiar weak general gap lives on the second axis, while successful risk execution is governed largely by the social pathway on the first.

To probe generalizability across modality, an out-of-sample transfer test was run: CCA weights were trained on the online cohort alone ($n = 53$) and projected onto the unseen

laboratory cohort ($n = 17$). Despite the strict penalty of projection, the first canonical correlation remained positive at 0.32, an honest cross-environmental estimate of the structure's stability.

4.4 Robust Bayesian estimation

Given a complex feature space and a sample of 70, frequentist p-values are vulnerable to outlier distortion, false-positive volatility, and the garden of forking paths (Gelman & Loken, 2014). To quantify the social-behavior link probabilistically, an outlier-tolerant Bayesian Markov Chain Monte Carlo framework was implemented, mirroring the reporting approach of Frey et al. (2017) and the robust estimation framework of Kruschke (2013), using the PyMC probabilistic programming library (Salvatier, Wiecki, & Fonnesbeck, 2016) with the No-U-Turn Sampler (Hoffman & Gelman, 2014).

Each bivariate model used a Student-t likelihood whose degrees-of-freedom parameter was itself estimated ($\nu = 2 + \text{Exponential}(1/29)$), allowing the model to down-weight extreme observations in the heavy tails. Broad, mildly informative priors were placed on the location and scale parameters and a uniform prior on the correlation ρ , with four chains of 1,000 draws each following 1,000 tuning iterations. All reported fits converged ($\hat{R} \leq 1.01$); the poorly distributed flat-strategy variable failed to converge and was excluded, consistent with its exclusion as a near-degenerate measure.

The posterior modes and 95% highest-density intervals sharply separated the social and financial-gambling domains. Of 90 behavioral-attitudinal correlations, exactly five had 95% intervals excluding zero, and four of those five belonged to the Social subscale:

Pair	Posterior mode	95% HDI	$P(\rho > 0)$
Social $\times$ Money Collected	+0.44	[+0.24, +0.63]	1.000
Social $\times$ Lejuez Composite	+0.43	[+0.23, +0.62]	1.000
Social $\times$ RNG-Normalized Pumps	+0.41	[+0.17, +0.58]	1.000
Social $\times$ EV Ratio Score	+0.37	[+0.17, +0.57]	0.999

The domain-matched Financial-Gambling subscale did not cross the threshold of credibility for any behavioral metric. Its association with the composite peaked at a mode of +0.19 with an interval, $[-0.08, +0.38]$, that comfortably spans zero. Within this sample the posterior placed effectively all of its mass above zero for the social correlations, but the more informative detail is the *profile* of those correlations. Social's link to the raw explosion rate was notably weaker (mode +0.18) than its link to money and calibration, which isolates the behavioral mechanism: high-social participants were not indiscriminately pumping until destruction, but extracting financial yield while maintaining EV-calibration. They resemble the successful, calibrated risk-taker rather than the reckless gambler.

To test directional superiority formally, a trivariate Student-t model with an LKJ prior on the correlation matrix (Lewandowski, Kurowicka, & Joe, 2009) produced a direct posterior contrast, $\Delta\rho = \rho_{\text{social}} - \rho_{\text{financial-gambling}}$. For Money Collected the difference was decisively positive (mode +0.38, HDI [+0.11, +0.67], $P(\Delta\rho > 0) = 0.995$). On the broader composite the contrast strongly favored Social ($P = 0.960$), though the lower bound of the interval grazed zero, [-0.04, +0.54]. The contrast vanished on explosion rate ($P = 0.671$), reinforcing that the social advantage governs the variance of *profitable* risk specifically rather than gross exposure.

A multiple Bayesian regression (standardized coefficients, $N(0,3)$ priors, the prior used by Frey et al., 2017) cemented this pattern. Regressing the behavioral composite on all six subscales simultaneously, Social emerged as the only credible predictor ($\beta = +0.45$, HDI [+0.22, +0.69]), and adding age, gender, and modality as covariates left it essentially unchanged ($\beta = +0.42$, HDI [+0.21, +0.68]), indicating the signal is robust to demographic and environmental confounding within this sample. Financial-gambling's partial coefficient was weak and non-credible under every specification ($\beta = +0.16$, HDI [-0.08, +0.41]).

4.5 Exploratory factor analysis and Schmid-Leiman decomposition

To examine the structure at the level of latent factors, an exploratory factor analysis was conducted on the combined DOSPERT and BART matrix. Diagnostics supported factorability (Kaiser-Meyer-Olkin = 0.608, Kaiser, 1974; Bartlett's test of sphericity $\chi^2 = 205.5$, $p = 1.2 \times 10^{-25}$, Bartlett, 1950). Horn's parallel analysis with 1,000 simulations and a 95th-percentile criterion (Horn, 1965) unanimously indicated two factors. A maximum-likelihood extraction with Promax rotation cleanly recovered the classic divide:

- **Factor 1 (stated risk):** dominated by the DOSPERT subscales, with health and safety at 0.78, recreational at 0.68, ethical at 0.65, financial-investment at 0.64, and financial-gambling at 0.48.
- **Factor 2 (behavioral risk):** anchored by the dynamic-hazard BART metrics, with normalized pumps at 1.02, explosion rate at 0.64, and money collected at 0.61.

The two factors correlated at only $\phi = 0.25$, confirming that stated willingness and behavioral execution operate in largely separate dimensions. Within that divided structure, one anomaly stood out. The Social subscale showed the weakest loading among the self-reports on the stated-risk factor (0.29), yet it was the only DOSPERT subscale to cross-load on the behavioral factor, at +0.36 in the primary set and up to +0.41 in the 14-feature robustness set.

A Schmid-Leiman orthogonalization (Schmid & Leiman, 1957) approximated a general-factor decomposition. The transformation revealed an exceptionally weak general risk factor accounting for under a quarter of the common variance (explained common variance = 0.234; omega-hierarchical = 0.326; McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005). Once that weak general factor was isolated, Social was the only self-report measure carrying appreciable specific variance onto behavior (specific loading 0.31). A confirmatory

factor analysis was deliberately omitted, since estimating roughly 55 free parameters at $n = 70$ would invite non-convergence or Heywood cases; at this sample size the exploratory decomposition is the defensible choice (Frey et al., 2017, had more than 1,500 participants and still dropped a measure for convergence).

4.6 Pre-registered external replication and the boundary condition

A sample of 70 cannot establish generalizability on its own. To test the social-behavior link externally, a pre-registered confirmatory analysis was run against the open-access Basel-Berlin Risk Study ($N = 1,505$; Frey et al., 2017), with the analysis plan frozen before any DOSPERT-by-BART statistic was computed on that dataset, ensuring hypothesis isolation.

The basic link replicated. In the large independent sample, the Social subscale correlated with behavioral risk at $\rho = 0.115$ (95% HDI [0.062, 0.166]), credibly above zero. The hypothesized ordering, that Social out-predicts Financial-Gambling, was strikingly consistent across the two studies: the posterior probability of the social advantage was 0.957, essentially identical to the 0.960 obtained in the dynamic task.

Social's *specific dominance*, however, did not replicate. In the Basel-Berlin data, which uses a classic static-hazard BART, the propensity-behavior relationship was diffuse across all subscales. Social was not the sole cross-loading factor, and in regression no single subscale carried a credible unique partial effect. This partial non-replication is itself the key theoretical result, and the task architecture explains it. The Basel-Berlin balloon has a 128-pump capacity with an EV-optimal threshold of 64, yet the sample mean is only 40.3 pumps. Because the entire cohort severely under-pumps, EV-efficiency correlates with raw mean pumps at roughly 0.93. In that regime calibration cannot be measured, because meaningful hazard variation does not exist, and the task captures gross exposure alone. Stated preferences from all domains therefore correlate weakly and diffusely with that generalized exposure metric, and the distinctive social signal dissolves into the background of the general attitude-behavior gap. The social specificity observed in the present study appears to require a task that actively rewards calibration, adaptation, and efficiency.

One further result emerged from the external data and is flagged as exploratory. In the fully adjusted regression on the large sample, the Financial-Gambling subscale's partial coefficient was not merely weak but credibly negative ($\beta = -0.068$, HDI [-0.124, -0.016]), a pattern that held across several scoring schemes. Because the zero-order gambling correlation in that sample is near zero rather than negative, this negative *partial* is consistent with a suppression relationship among correlated predictors rather than a simple negative association, and it should be confirmed in its own right before being interpreted substantively. Read conservatively, it reinforces the most robust cross-dataset fact: the domain-matched financial-gambling subscale is the weakest behavioral predictor in both samples.

4.7 Quadrant analysis: the mechanics of success

To crystallize the findings, a cohort-referenced re-scoring mapped High-Risk against Low-Risk and High-Success against Low-Success, forming four quadrants. In the dynamic-

hazard task, membership in the High-Risk and High-Success quadrant, the successful calibrated risk-taker, behaved as a social phenomenon. The success residual, that is, earning beyond what one's risk level alone would predict, credibly tracked the Social subscale (+0.28, HDI [+0.031, +0.459]) while every other subscale straddled zero, and a robust group comparison placed the social effect on quadrant membership near one standard deviation (Cohen's $d \approx 0.97$). In the static Basel-Berlin task, by contrast, the success-residual association was null for all six subscales, implying that reaching the top quadrant in a single-capacity static task is largely a matter of luck rather than skill. The same metric thus tracks an attitude where the task makes success a matter of calibration, and tracks nothing where the task makes success a matter of chance.

5. Discussion

5.1 Deconstructing the social risk construct

If successful, calibrated behavioral risk-taking is best predicted by the DOSPERT Social subscale in dynamic environments, the meaning of "social risk" deserves a second look. The subscale's items assess willingness to take actions such as speaking up against an unpopular opinion or disagreeing with an authority figure. These are not measures of physical thrill-seeking, nor of financial calculation. They plausibly index a readiness to act under ambiguity, where immediate feedback may be socially punishing but the longer-run expected value of asserting oneself is positive.

Navigating the dynamic-hazard task may call on an overlapping profile. The participant must tolerate the anxiety of an unpredictable burst, absorb the negative feedback of occasional immediate failures, and hold to a steady, rational strategy to accumulate earnings over time. One interpretation, offered tentatively, is that the psychological resources used to speak out in a contentious setting are related to those used to navigate algorithmic volatility efficiently, such that the Social subscale acts as a proxy for active, calibrated engagement under uncertainty. This reading is consistent with the present data but remains speculative, and the construct-validity studies proposed in Section 6 are needed to test it.

A cross-cultural note follows from the sample. The social items, which concern challenging authority and voicing unpopular views, may carry different connotations in a predominantly Turkish cohort, where their relationship to broader dispositions such as non-conformity could differ from Western validation samples. This is a question worth examining directly rather than assuming, and the comparatively low internal consistency of the social subscale across studies (Highhouse et al., 2017) makes such scrutiny more, not less, important.

5.2 The illusion of financial-gambling traits

Why does the Financial-Gambling subscale fail so comprehensively, even trending negative as a partial predictor in the large sample? One plausible reading is that self-report

items about betting a day's income capture an idealized self-image of financial boldness rather than a behavioral readiness to face risk. When a participant is placed in an active, incentivized task that demands experiential learning, impulse control, and rapid state evaluation, that idealized self-image offers little operational purchase, and high self-reported gambling propensity may even co-occur with impulsivity or weak calibration. This interpretation should be held loosely given the suppression caveat noted above, but it is consistent with the broader literature in which domain-matched behavioral validation of risk attitudes routinely underperforms (Pedroni et al., 2017; Szrek et al., 2012).

5.3 Assumptions

Several assumptions underlie the analyses and should be made explicit. First, the behavioral composite weights its three constituent metrics equally; because explosion rate correlates only weakly with money, the composite leans toward the pumps-and-money axis, and the per-metric tables are provided so a reader can reweight. Second, modality is treated as a comparison variable rather than a covariate, on the grounds that it is confounded with demographics and that no behavioral divergence between channels was observed; an analysis that adjusted for modality directly would make a different assumption. Third, the linear-hazard model assumes independent Bernoulli trials with a per-pump explosion probability of k/N , and the optimal-stopping results follow from that specification. Fourth, the DOSPERT subscales are treated as continuous, which is standard for subscale means but is an approximation of ordinal item data.

5.4 Limitations

The primary limitation is sample size. At $N = 70$ the more complex procedures, particularly the 14-feature CCA and the factor models, are underpowered and rely on heavily regularized priors; the exploratory findings should be read accordingly. Relatedly, while the directional claim that Social correlates positively with behavior is interval-solid, the superiority contrast over Financial-Gambling, though highly probable at roughly 0.96, has a 95% interval that touches zero on several composite metrics, which precludes a conclusive superiority claim without a larger sample. The Social and Recreational subscales are moderately correlated ($r \approx 0.34$), producing some credit-sharing in bivariate analyses, although the partial regressions address this directly. The social subscale also carries the lowest internal consistency of the six domains (Highhouse et al., 2017), which adds measurement noise to precisely the construct of interest. Finally, the analysis rests on measure-level aggregate data; trial-level Markov-chain behavioral modeling and item-level DOSPERT factor analysis were beyond the present scope and represent natural avenues for greater granularity. The single confirmatory test, while genuine in being frozen before contact with the external data, is pre-registration in spirit rather than registry-grade, and the next study should carry an independent timestamp.

6. Future work

The convergence of factor structure, Bayesian posteriors, machine-learning importance,

and external re-scoring points to a methodological reframing. For decades the attitude-behavior gap has been read as evidence that behavioral tasks are merely noisy or that the general propensity factor is too weak to capture variance. The present analysis suggests instead that the psychometric signal is present, but becomes visible only under task architectures that reward calibration rather than gross exposure.

Several concrete steps follow. The most immediate is a pre-registered, registry-timestamped replication on a larger sample, with the primary confirmatory test defined in advance as the Bayesian Δp contrast between the Social and Financial-Gambling subscales on the behavioral composite, and ideally with the success-residual dissociation registered as a secondary confirmatory target. A construct-validity battery, integrating broad taxonomies such as the NEO Five-Factor framework (Costa & McCrae, 1992) along with focused measures of extraversion and approach motivation, would help identify what latent trait the social subscale is indexing in non-social environments. Cross-task generalization using the Iowa Gambling Task (Bechara, Damasio, Damasio, & Anderson, 1994) and the Cambridge Gambling Task (Rogers et al., 1999) would test whether the social tracking is specific to sequential hazard accumulation or extends to calibrated risk more broadly. Finally, continued development and open release of the dynamic-hazard platform, with an interface designed for researchers across disciplines, is intended to lower the barrier to deploying, scoring, and analyzing sophisticated behavioral risk telemetry, so that the theoretical argument and the software infrastructure advance together.

References

- Armstrong, M. B., Landers, R. N., & Collmus, A. B. (2016). Gamifying recruitment, selection, training, and performance management: Game-thinking in human resource management. In D. Davis & H. Gangadharbatla (Eds.), *Emerging research and trends in gamification* (pp. 140-165). IGI Global.
- Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, 3(2), 77-85.
- Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1-3), 7-15.
- Blais, A.-R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1), 33-47.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.
- Di Plinio, S., Pettorruso, M., & Ebisch, S. J. H. (2022). Appropriately tuning stochastic-psychometric properties of the Balloon Analog Risk Task. *Frontiers in Psychology*, 13, 881179. <https://doi.org/10.3389/fpsyg.2022.881179>

- Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 709–730.
- Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10), e1701381. <https://doi.org/10.1126/sciadv.1701381>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465.
- Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1), 258–271.
- Highhouse, S., Nye, C. D., Zhang, D. C., & Rada, T. B. (2017). Structure of the Dospert: Is there evidence for a general risk factor? *Journal of Behavioral Decision Making*, 30(2), 400–406. [reliability/structure of the DOSPERT; please verify against the specific meta-analytic source you cite for the social subscale's coefficient alpha]
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1593–1623.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321–377.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218.
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603.
- Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection and Assessment*, 30(1), 1–13.
- Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, 8(2), 75–84.
- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9), 1989–2001.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46(3), 598–609.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, 1(11), 803–809.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 167–185.

Rogers, R. D., Everitt, B. J., Baldacchino, A., Blackshaw, A. J., Swainson, R., Wynne, K., Baker, N. B., Hunter, J., Carthy, T., Booker, E., London, M., Deakin, J. F. W., Sahakian, B. J., & Robbins, T. W. (1999). Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers. *Neuropsychopharmacology*, 20(4), 322–339.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2, e55.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1), 53–61.

Szrek, H., Chao, L.-W., Ramlagan, S., & Peltzer, K. (2012). Predicting (un)healthy behavior: A comparison of risk-taking propensity measures. *Judgment and Decision Making*, 7(6), 716–727.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263–290. <https://doi.org/10.1002/bdm.414>

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123–133.

Turkish DOSPERT-30 adaptation (to be confirmed by the author): the specific Turkish validation administered should be cited here. Candidates identified include a 2019 Turkish adaptation of the DOSPERT-30 short form and Dinç & Tez (2019); please insert the exact reference for the version you used.