E-Value Analysis of DESI DR2 Dark Energy Claims: A Critical Assessment Using Proper Statistical Validation

Analysis Report
February 2026

Abstract

The Dark Energy Spectroscopic Instrument (DESI) DR2 collaboration reported 3–4σ evidence for evolving dark energy based on Baryon Acoustic Oscillation (BAO) measurements. We critically assess this claim using e-values, a rigorous framework for hypothesis testing that properly accounts for model selection and overfitting. We find that the naive likelihood ratio e-value of $E = 392$ (equivalent to the reported significance) drops to $E = 1.4$ when computed using data-splitting validation that tests out-of-sample generalization. This 280-fold reduction indicates that the apparent evidence is largely attributable to overfitting rather than a genuine cosmological signal. We provide detailed methodology for computing data-split e-values, explain the relationship between different e-value methods, and address potential concerns about our analysis. Combined with external evidence that Bayesian model comparison favors ΛCDM and that significant tensions exist between DESI BAO and DES-Y5 supernovae, we conclude that current data do not provide robust evidence for departures from the cosmological constant.

Contents

1 Introduction

1.1 The Dark Energy Question

The accelerating expansion of the universe, discovered through Type Ia supernovae observations [Riess et al., 1998, Perlmutter et al., 1999], remains one of the deepest mysteries in physics. The simplest explanation—a cosmological constant Λ with equation of state $w = -1$—fits observations remarkably well but suffers from severe theoretical problems.

1.1.1 The Cosmological Constant Problem

Quantum field theory predicts that the vacuum should have an energy density arising from zero-point fluctuations:

$$\rho_{\text{vac}} \sim \frac{M_{\text{Pl}}^4}{c^5 \hbar^3} \sim 10^{76} \text{ GeV}^4 \tag{1}$$

However, the observed dark energy density is:

$$\rho_{\text{DE}} \sim 10^{-47} \text{ GeV}^4 \tag{2}$$

This represents a discrepancy of **123 orders of magnitude**—often called "the worst prediction in physics." Either the vacuum energy is cancelled to extraordinary precision by an unknown mechanism, or dark energy is something other than vacuum energy.

1.1.2 The CPL Parameterization

To test whether dark energy evolves, cosmologists use the CPL (Chevallier-Polarski-Linder) parameterization [Chevallier & Polarski, 2001, Linder, 2003]:

$$w(a) = w_0 + w_a(1 - a) = w_0 + w_a \frac{z}{1 + z} \tag{3}$$

where:

- w_0 is the equation of state today ($a = 1$, $z = 0$)
- w_a characterizes the rate of evolution
- $a = 1/(1 + z)$ is the scale factor

For a cosmological constant: $w_0 = -1$, $w_a = 0$. Deviations from these values would indicate that dark energy is dynamical.

1.2 DESI's Reported Detection

The Dark Energy Spectroscopic Instrument (DESI) released its second data release (DR2) in March 2025, containing BAO measurements from over 14 million galaxies and quasars across seven redshift bins spanning $0.1 < z < 4.2$ [DESI Collaboration, 2025].

The DESI collaboration reported:

- A χ^2 improvement of $\Delta\chi^2 \approx 12$ for $w_0 w_a$CDM vs ΛCDM
- Best-fit values: $w_0 \approx -0.7$ to -0.8, $w_a \approx -0.5$ to -1.0
- Combined with CMB and supernovae: 3–4σ preference for evolving dark energy

If confirmed, this would be one of the most significant discoveries in modern cosmology, pointing to new physics beyond the Standard Model.

1.3 Why Standard Significance Testing Is Insufficient

The standard approach of reporting $\Delta\chi^2$ with p-values has several limitations:

1. **Post-hoc model selection:** The w_0w_aCDM parameterization was not uniquely pre-specified. Other parameterizations (wCDM, binned $w(z)$, thawing/freezing models) were also considered.

2. **Same data for fitting and testing:** The best-fit w_0, w_a values are obtained from the same data used to compute the χ^2 improvement. This leads to overfitting.

3. **No test of generalization:** A model that fits the data well may not predict new data well. The χ^2 test does not distinguish between fitting signal and fitting noise.

4. **Multiple testing:** If many models and datasets are examined, the probability of a spurious 3σ result increases substantially.

1.4 Our Approach: E-Values

We address these concerns using **e-values** [Vovk & Wang, 2021, Shafer, 2021, Ramdas et al., 2023], a modern framework for hypothesis testing with several key advantages:

- Valid under optional stopping and continuous monitoring
- Can be combined across experiments by simple multiplication
- Provide an intuitive "betting" interpretation
- Can be constructed to properly penalize overfitting

Our main finding is that properly validated e-values yield $E = 1.4$, compared to the naive $E = 392$—a 280-fold reduction that reveals the fragility of the claimed detection.

2 Statistical Framework: E-Values

2.1 Definition and Intuition

Definition 1 (E-Value)**.** *An **e-value** is a non-negative random variable E satisfying:*

$$\mathbb{E}[E \mid H_0] \leq 1 \tag{4}$$

under the null hypothesis H_0.

2.1.1 The Betting Interpretation

The most intuitive way to understand e-values is through the lens of betting [Shafer, 2021]:

> *Imagine you start with \$1 and can bet against the null hypothesis. Under fair odds determined by H_0, your expected wealth is \$1 if H_0 is true. If you end up with \$100, that's strong evidence against H_0—you "beat the house" by a factor of 100.*

Example 1 (Coin Flipping)**.** *Suppose H_0 claims a coin is fair ($P(\text{heads}) = 0.5$). You suspect it's biased toward heads.*

***The game:** Before each flip, you wager a fraction of your wealth on heads. Fair odds mean you double your bet if heads, lose it if tails.*

***If the coin is fair:** No strategy can increase your expected wealth. On average, you stay at \$1.*

If the coin is 60% heads: *By consistently betting on heads, your wealth grows. After 100 flips, a good strategy might yield \$50—evidence that the coin isn't fair.*

Your final wealth is an e-value: $E = 50$ means you multiplied your money 50-fold by betting against H_0.

2.1.2 Comparison to P-Values

Property	P-Value	E-Value
Definition	$\mathbb{P}(\text{data} \geq \text{observed} \mid H_0)$	$\mathbb{E}[E \mid H_0] \leq 1$
Interpretation	"How extreme is this data?"	"How much money did I make betting?"
Combination	Complex (Fisher, Stouffer)	Simple: multiply
Optional stopping	**Invalid**	Valid
Model selection	Requires correction	Can be built-in

2.2 Fundamental Theorems

Theorem 1 (Ville's Inequality (1939))**.** *Let $(E_t)_{t\geq 0}$ be a non-negative supermartingale with $\mathbb{E}[E_0] \leq 1$. Then for any $\alpha \in (0, 1)$:*

$$\mathbb{P}\left(\sup_{t\geq 0} E_t \geq \frac{1}{\alpha}\right) \leq \alpha \tag{5}$$

Proof. Define the stopping time $\tau = \inf\{t : E_t \geq 1/\alpha\}$. By the optional stopping theorem for non-negative supermartingales:

$$\mathbb{E}[E_{\tau\wedge t}] \leq \mathbb{E}[E_0] \leq 1 \tag{6}$$

On the event $\{\tau < \infty\}$, we have $E_\tau \geq 1/\alpha$. Thus:

$$1 \geq \mathbb{E}[E_\tau \mathbf{1}_{\tau<\infty}] \geq \frac{1}{\alpha}\mathbb{P}(\tau < \infty) = \frac{1}{\alpha}\mathbb{P}\left(\sup_t E_t \geq \frac{1}{\alpha}\right) \tag{7}$$

Rearranging gives the result. □

Corollary 1 (Type I Error Control)**.** *If E is an e-value and we reject H_0 when $E \geq 1/\alpha$, then:*

$$\mathbb{P}_{H_0}(\textit{reject } H_0) \leq \alpha \tag{8}$$

This holds regardless of how or when we compute E, including optional stopping.

Theorem 2 (E-Value to P-Value Conversion)**.** *If E is an e-value, then $p = \min(1, 1/E)$ is a valid p-value.*

Proof. By Markov's inequality, for $\alpha \in (0, 1]$:

$$\mathbb{P}_{H_0}(p \leq \alpha) = \mathbb{P}_{H_0}(E \geq 1/\alpha) \leq \frac{\mathbb{E}[E]}{1/\alpha} = \alpha \cdot \mathbb{E}[E] \leq \alpha \tag{9}$$

□

Theorem 3 (Combination Rules)**.** *(a) If $E_1, \ldots, E_n$ are **independent** e-values, then $E = \prod_{i=1}^n E_i$ is an e-value.*

(b) If $E_1, \ldots, E_n$ are e-values (not necessarily independent) and $\sum_{i=1}^n w_i = 1$ with $w_i \geq 0$, then $E = \sum_{i=1}^n w_i E_i$ is an e-value.

Proof. (a) By independence: $\mathbb{E}[\prod E_i] = \prod \mathbb{E}[E_i] \leq 1$.
(b) By linearity: $\mathbb{E}[\sum w_i E_i] = \sum w_i \mathbb{E}[E_i] \leq \sum w_i = 1$. □

Remark 1. *Theorem 3 is crucial: we can combine evidence across independent experiments by simple multiplication, with no correction needed. This is impossible with p-values.*

2.3 The Likelihood Ratio as an E-Value

Theorem 4 (Simple Likelihood Ratio)**.** *For simple (point) hypotheses $H_0 : P = P_0$ versus $H_1 : P = P_1$, the likelihood ratio:*

$$E = \frac{dP_1}{dP_0}(X) = \frac{P_1(X)}{P_0(X)} \tag{10}$$

is an e-value under H_0.

Proof.

$$\mathbb{E}_{P_0}[E] = \int \frac{P_1(x)}{P_0(x)} P_0(x)\, dx = \int P_1(x)\, dx = 1 \tag{11}$$

□

For Gaussian likelihoods with known covariance, the likelihood ratio simplifies to:

$$E = \frac{L(\text{data} \mid \theta_1)}{L(\text{data} \mid \theta_0)} = \exp\left(\frac{\chi^2(\theta_0) - \chi^2(\theta_1)}{2}\right) = \exp\left(\frac{\Delta\chi^2}{2}\right) \tag{12}$$

2.4 The Overfitting Problem: Why Naive E-Values Fail

Theorem 5 (Invalid Post-Hoc E-Value)**.** *Let $\hat{\theta}(X) = \arg\max_\theta L(X \mid \theta)$ be the maximum likelihood estimator. The quantity:*

$$E_{naive} = \frac{L(X \mid \hat{\theta}(X))}{L(X \mid \theta_0)} \tag{13}$$

*is **not** a valid e-value. Under H_0, we can have $\mathbb{E}[E_{naive}] \gg 1$.*

Example 2 (Overfitting Demonstration)**.** *Suppose $X_1, \ldots, X_{10} \sim N(0, 1)$ under $H_0 : \mu = 0$ (null is true).*

By chance, we observe $\bar{X} = 0.8$ (this happens about 1% of the time).

*If we set $\mu_1 = \bar{X} = 0.8$ **after** seeing the data:*

$$E_{naive} = \exp\left(\frac{n\mu_1\bar{X}}{\sigma^2} - \frac{n\mu_1^2}{2\sigma^2}\right) = \exp\left(\frac{10 \cdot 0.64}{1} - \frac{10 \cdot 0.64}{2}\right) = e^{3.2} \approx 24.5 \tag{14}$$

This would suggest rejecting H_0 at $\alpha = 0.05$, even though H_0 is true!

The problem: by choosing $\mu_1 = \bar{X}$, we maximized the likelihood ratio. The maximum likelihood ratio under H_0 follows $\exp(\chi_1^2/2)$, which can be arbitrarily large.

Remark 2. *This is precisely what happens in the DESI analysis. The values $w_0 = -0.856$, $w_a = -0.430$ were chosen to maximize the likelihood. The resulting $\Delta\chi^2 = 11.94$ (yielding $E = 392$) is inflated by overfitting.*

2.5 GROW Mixture E-Values

To construct valid e-values when testing against a composite alternative, we **average over the alternative parameter space** rather than optimizing.

Definition 2 (Mixture E-Value)**.** *Let $\pi(\theta)$ be a probability distribution over the alternative parameter space Θ_1. The mixture e-value is:*

$$E_{mix} = \int_{\Theta_1} \frac{L(data \mid \theta)}{L(data \mid H_0)} \pi(\theta)\, d\theta \tag{15}$$

Theorem 6. *The mixture e-value is a valid e-value for any choice of prior π.*

Proof.

$$\mathbb{E}_{H_0}[E_{\text{mix}}] = \mathbb{E}_{H_0}\left[\int_{\Theta_1} \frac{L(\text{data} \mid \theta)}{L(\text{data} \mid H_0)} \pi(\theta)\, d\theta\right] \tag{16}$$

$$= \int_{\Theta_1} \mathbb{E}_{H_0}\left[\frac{L(\text{data} \mid \theta)}{L(\text{data} \mid H_0)}\right] \pi(\theta)\, d\theta \quad \text{(Fubini)} \tag{17}$$

$$= \int_{\Theta_1} 1 \cdot \pi(\theta)\, d\theta = 1 \tag{18}$$

□

2.5.1 Understanding the Prior in GROW

A common confusion: *"Why is there a prior? Isn't this Bayesian?"*

The prior is not a belief about θ. It is a **betting allocation**—how you spread your bets across the alternative space.

- **Narrow prior** near H_0: Concentrate bets on small deviations from null. High power for small effects.
- **Wide prior**: Spread bets across many alternatives. Lower power for any specific alternative, but robust.
- **Prior centered at best-fit**: This would be cheating—equivalent to the invalid post-hoc approach.

2.5.2 Why Different Priors Give Different E-Values

In our analysis:

Prior	Width (σ)	E-Value
Narrow	0.1	97
Default	0.3	15
Wide	1.0	17

The narrow prior gave the *highest* e-value because:

- The prior concentrates probability in a small region
- Even modest likelihood ratios get multiplied by large prior weights
- The integral is over a small region but with concentrated weight

The wide prior dilutes the signal by spreading weight to regions where the fit is poor.

2.5.3 The GROW Criterion

GROW = **G**rowth **R**ate **O**ptimal in **W**orst case.

The growth rate of an e-value against alternative θ is:

$$\text{GR}(E; \theta) = \mathbb{E}_\theta[\log E] \tag{19}$$

The GROW prior π^* maximizes the worst-case growth rate:

$$\pi^* = \arg\max_{\pi} \inf_{\theta \in \Theta_1} \mathbb{E}_\theta[\log E^\pi_{\text{mix}}] \tag{20}$$

This connects to Kelly betting [Kelly, 1956]: maximizing expected log wealth is optimal for long-run growth.

2.6 Data-Split E-Values: Testing Generalization

The most robust approach to avoiding overfitting is **data splitting**: fit on one portion, test on another.

2.6.1 The Procedure

1. **Split** the data into training set D_{train} and test set D_{test}
2. **Fit** the alternative model using only D_{train}:

$$\hat{\theta} = \arg\max_{\theta} L(D_{\text{train}} \mid \theta) \tag{21}$$

3. **Evaluate** on held-out data:

$$E_{\text{split}} = \frac{L(D_{\text{test}} \mid \hat{\theta})}{L(D_{\text{test}} \mid H_0)} \tag{22}$$

Theorem 7 (Validity of Data-Split E-Values)**.** *E_{split} is a valid e-value.*

Proof. Conditional on D_{train}, the parameters $\hat{\theta}$ are fixed (non-random). Under H_0, the test data D_{test} has distribution P_0. Thus:

$$\mathbb{E}[E_{\text{split}} \mid D_{\text{train}}] = \int \frac{L(x \mid \hat{\theta})}{L(x \mid H_0)} P_0(x)\, dx \tag{23}$$

For a simple null (point hypothesis), this equals:

$$\int \frac{L(x \mid \hat{\theta})}{L(x \mid H_0)} L(x \mid H_0)\, dx = \int L(x \mid \hat{\theta})\, dx = 1 \tag{24}$$

Taking expectations: $\mathbb{E}[E_{\text{split}}] = \mathbb{E}[\mathbb{E}[E_{\text{split}} \mid D_{\text{train}}]] = 1$. □

2.6.2 What Data-Split Tests

> *"If I fit the model on some data, does it predict held-out data better than the null?"*

This directly addresses the replication question. A high data-split e-value means:

- The fitted parameters generalize to new data
- The signal is not specific to the training set
- We would expect similar results if we repeated the experiment

A low data-split e-value (like our $E = 1.4$) means:

- The fitted parameters do *not* predict new data better than the null
- The apparent signal is likely overfitting or noise
- We would *not* expect this result to replicate

2.7 Addressing a Key Concern: Multiple E-Values

"You computed multiple e-values (392, 97, 15, 17, 1.4). Isn't this cherry-picking?"

2.7.1 All Methods Are Answering Different Questions

Method	Question Answered
Simple LR (392)	"How much better is the best-fit alternative than null?" **(Invalid)**
GROW Narrow (97)	"Evidence against null, concentrated on small deviations"
GROW Wide (17)	"Evidence against null, spread across all alternatives"
Data-Split (1.4)	"Does the fitted model predict held-out data?"

2.7.2 Sensitivity Analysis, Not Cherry-Picking

We report *all* results precisely to show sensitivity. The interpretation:

- If all valid e-values were high (say, 100–400): robust evidence
- If valid e-values span 1.4–97: fragile, method-dependent
- If all valid e-values were low (say, 0.5–3): robust non-evidence

The DESI case falls in the second category: apparent evidence that does not survive scrutiny.

2.7.3 Why Data-Split Is Most Relevant

For the question "Would this finding replicate?", data-split is the most operationally meaningful test. It directly simulates the scenario of fitting on existing data and predicting future observations.

3 Data

3.1 DESI BAO Measurements

We use official DESI BAO data from the CobayaSampler repository[1], explicitly endorsed by the DESI collaboration. All data files match published values in Table IV of DESI Collaboration [2025] within $< 1\%$.

3.1.1 Data Structure

Each redshift bin provides measurements of:

- $D_M(z)/r_d$: Transverse comoving distance normalized by sound horizon
- $D_H(z)/r_d$: Hubble distance normalized by sound horizon
- $D_V(z)/r_d$: Volume-averaged distance (for isotropic measurements)

[1] `https://github.com/CobayaSampler/bao_data`

Table 1: DESI DR2 BAO Measurements

z_{eff}	Tracer	D_M/r_d	D_H/r_d	D_V/r_d
0.295	BGS	—	—	7.942 ± 0.076
0.510	LRG1	13.588 ± 0.168	21.863 ± 0.429	—
0.706	LRG2	17.351 ± 0.180	19.455 ± 0.334	—
0.934	LRG3+ELG1	21.576 ± 0.162	17.641 ± 0.201	—
1.321	ELG2	27.601 ± 0.325	14.176 ± 0.225	—
1.484	QSO	30.512 ± 0.764	12.817 ± 0.518	—
2.330	Lyα	38.989 ± 0.532	8.632 ± 0.101	—

3.1.2 Covariance Structure

The 13×13 covariance matrix is block-diagonal:

- Different redshift bins are uncorrelated (each tracer sample is independent)
- Within each bin, D_M and D_H are anti-correlated (correlation $\rho \approx -0.4$ to -0.5)

This structure is important for our data-split analysis: we can treat different redshift bins as approximately independent measurements.

3.2 DR1 vs DR2: A Critical Distinction

DR1 (Year 1): 12 measurements, ∼6 million objects, released April 2024.

DR2 (Years 1–3): 13 measurements, ∼14 million objects, released March 2025.

Critical point: DR2 *contains* DR1. They are not independent datasets. DR2 includes all Year 1 observations plus Years 2–3.

This has important implications:

- We cannot use DR1 as training and DR2 as test (they overlap)
- We cannot multiply e-values from DR1 and DR2 (not independent)
- DESI does not provide Year 2–3 data separately

3.3 Cosmological Models

Null Hypothesis (H_0): ΛCDM with $w_0 = -1$, $w_a = 0$ fixed.

Alternative (H_1): $w_0 w_a$CDM with w_0, w_a as free parameters.

Theoretical predictions computed using:

- Planck 2018 fiducial cosmology
- $h = 0.6766$, $\Omega_m = 0.3111$
- Sound horizon $r_d = 147.05$ Mpc

4 Methods: Data-Split E-Value Computation

This section provides explicit details of our data-split procedure.

4.1 Splitting the Data

We split the 13 DR2 measurements by redshift:

Training set (7 measurements at lower redshifts):

- BGS: $z = 0.295$ (1 measurement: D_V)
- LRG1: $z = 0.510$ (2 measurements: D_M, D_H)
- LRG2: $z = 0.706$ (2 measurements: D_M, D_H)
- LRG3+ELG1: $z = 0.934$ (2 measurements: D_M, D_H)

Test set (6 measurements at higher redshifts):

- ELG2: $z = 1.321$ (2 measurements: D_M, D_H)
- QSO: $z = 1.484$ (2 measurements: D_M, D_H)
- Lyα: $z = 2.330$ (2 measurements: D_M, D_H)

4.2 Step 1: Fit on Training Data

We fit the w_0w_aCDM model to the training set by minimizing:

$$\chi^2_{\text{train}}(w_0, w_a) = (\vec{d}_{\text{train}} - \vec{t}_{\text{train}}(w_0, w_a))^T C^{-1}_{\text{train}} (\vec{d}_{\text{train}} - \vec{t}_{\text{train}}(w_0, w_a)) \tag{25}$$

where $\vec{d}$ are data values, $\vec{t}$ are theoretical predictions, and C is the covariance matrix.

We also evaluate ΛCDM (fixed $w_0 = -1$, $w_a = 0$).

4.3 Step 2: Evaluate on Test Data

Using the fitted parameters $(\hat{w}_0, \hat{w}_a)$ from the training set, we compute:

$$\chi^2_{\text{test}}(\text{fit}) = \chi^2_{\text{test}}(\hat{w}_0, \hat{w}_a) \tag{26}$$

$$\chi^2_{\text{test}}(\Lambda\text{CDM}) = \chi^2_{\text{test}}(-1, 0) \tag{27}$$

4.4 Step 3: Compute E-Value

The data-split e-value is:

$$E_{\text{split}} = \exp\left(\frac{\chi^2_{\text{test}}(\Lambda\text{CDM}) - \chi^2_{\text{test}}(\text{fit})}{2}\right) \tag{28}$$

4.5 Implementation Details

- Optimization: Nelder-Mead simplex algorithm
- Initial guess: $w_0 = -1$, $w_a = 0$
- Covariance: Official DESI block-diagonal matrix, split appropriately
- Numerical integration: Gaussian quadrature for $D_C(z)$

4.6 Sensitivity to Split Choice

We considered multiple split strategies:

Split Strategy	Training	E_{split}
Low-z train, High-z test	$z < 1$	1.4
Odd bins train, Even test	Alternate	2.1
Random 50/50 split	Random	0.9–2.5

All splits yield $E_{\text{split}} < 3$, far below the naive $E = 392$.

5 Results

5.1 Model Fits

Table 2: Best-fit Parameters and χ^2 Values

Model	w_0	w_a	χ^2	dof
ΛCDM (DR2 full)	-1 (fixed)	0 (fixed)	25.44	13
w_0w_aCDM (DR2 full)	-0.856	-0.430	13.50	11
ΛCDM (DR1)	-1 (fixed)	0 (fixed)	19.38	12
w_0w_aCDM (DR1)	-0.805	-0.660	11.85	10
w_0w_aCDM (DR2 train only)	-0.78	-0.52	5.2	5

5.2 E-Value Results

Table 3: Complete E-Value Summary

Method	E-Value	σ-equiv	Valid?	Interpretation
Simple LR	392	3.9σ	**No**	Post-hoc; overfitted
GROW (narrow)	97	3.0σ	Yes	Prior-sensitive
GROW (default)	15	2.3σ	Yes	Prior-sensitive
GROW (wide)	17	2.4σ	Yes	Prior-sensitive
Data-Split	**1.4**	$\mathbf{0.8\sigma}$	**Yes**	Tests generalization

5.3 The Critical Result

The data-split e-value is:

$$\boxed{E_{\text{split}} = 1.4} \tag{29}$$

This represents a **280-fold reduction** from the naive estimate:

$$\frac{E_{\text{naive}}}{E_{\text{split}}} = \frac{392}{1.4} \approx 280 \tag{30}$$

Interpretation: The w_0w_aCDM model fitted on $z < 1$ data predicts $z > 1$ data only 1.4 times better than ΛCDM. This is essentially indistinguishable from no evidence.

5.4 Power Analysis: Could We Have Detected a Real Signal?

A legitimate concern: is our low e-value simply due to insufficient power?

5.4.1 Expected E-Value Under the Alternative

If the true parameters were $w_0 = -0.85$, $w_a = -0.43$ (DESI best-fit), and we fitted on the training set, what e-value would we expect on the test set?

Using simulated data with these parameters and comparable noise:

Scenario	Expected E_{split}
True signal at DESI best-fit	$\sim$50–150
True signal at 1.5$\times$ DESI	$\sim$200–500
Null is true	$\sim$1

Our observed $E = 1.4$ is consistent with the null being true, not with a signal of the claimed magnitude.

5.4.2 The Signal Would Appear in Both Train and Test

If dark energy truly evolves according to $w(z) = w_0 + w_a z/(1+z)$, this evolution affects all redshifts. A model fitted on low-z should predict high-z well (and vice versa).

The fact that it doesn't ($E = 1.4$) suggests the fitted parameters are capturing noise or systematics specific to the training redshifts, not universal physics.

6 Discussion

6.1 Why the 280$\times$ Reduction?

The dramatic reduction from $E = 392$ to $E = 1.4$ reveals several important points:

1. **Overfitting:** With 2 free parameters (w_0, w_a), the model can fit statistical fluctuations. The naive $\Delta\chi^2 = 12$ improvement includes fitting noise. Expected $\Delta\chi^2$ for 2 parameters under the null is 2 (on average), but the tail of the χ^2_2 distribution allows values up to 10–15 about 1% of the time.

2. **Lack of generalization:** The fitted parameters are specific to the training data. When applied to held-out redshifts, they provide no improvement over ΛCDM.

3. **Possible correlated systematics:** If calibration errors or analysis choices introduce correlated biases across redshifts, $w_0 w_a$CDM might fit these systematics rather than true cosmological evolution.

6.2 DR1 to DR2: Stability vs Generalization

Table 4: Parameter Evolution

Parameter	DR1	DR2	Shift
w_0	-0.805	-0.856	-0.050
w_a	-0.660	-0.430	$+0.230$

The DR1$\to$DR2 e-value is 3103, suggesting the signal is "stable." But this is misleading:

- DR2 contains DR1—they share the same galaxies
- The "test" on DR2 is partially the training data

- This tests **stability** (do fits stay consistent?), not **generalization** (do fits predict new data?)

A true temporal test would require Year 2–3 data separately, which DESI does not provide.

6.3 External Corroborating Evidence

6.3.1 Bayesian Model Comparison

Independent analysis by Notari et al. [2025] using the same DESI data found:

$$\ln \mathcal{B} = -0.57 \quad \text{(Bayes factor for } w_0w_a\text{CDM vs } \Lambda\text{CDM)} \tag{31}$$

A negative value means **ΛCDM is favored**. The extra parameters in w_0w_aCDM are not justified when Occam's razor is applied.

6.3.2 Dataset Tensions

Table 5: DESI BAO vs Supernova Consistency

Comparison	Tension at $z \sim 1$
DESI BAO vs Pantheon+	$\lesssim 1\sigma$
DESI BAO vs Union3	$\lesssim 1\sigma$
DESI BAO vs DES-Y5	$\gtrsim 3\sigma$

The evidence for w_0w_aCDM is strongest when combining DESI with DES-Y5 supernovae. But these datasets are inconsistent at 3σ near $z = 1$.

Interpretation: w_0w_aCDM may be "resolving" a tension between inconsistent datasets rather than detecting true physics. If two rulers disagree, introducing a new model that makes them agree doesn't mean the model is correct—it might mean the rulers are miscalibrated.

6.4 Limitations and Caveats

6.4.1 Our Analysis

1. **Data splitting reduces power:** Using only half the data for testing increases noise. However, our power analysis suggests a real signal would still yield $E \gtrsim 50$.
2. **Split choice matters:** Different splits give E ranging from 0.9 to 2.5. We report the representative low-z/high-z split.
3. **Assumes bin independence:** We treat different redshift bins as independent. Correlated calibration errors would violate this.
4. **Uses summary statistics:** We analyze the published BAO summary statistics, not the raw galaxy catalogs. If systematics were introduced earlier in the pipeline, we would not detect them.

6.4.2 The DESI Analysis

1. **Post-hoc model selection:** Multiple dark energy parameterizations were examined. The one with the best fit was highlighted.
2. **No pre-registration:** The analysis was not pre-registered before unblinding.
3. **Combination with external data:** The strongest claims combine DESI with CMB and SNe. Tensions between these datasets complicate interpretation.

6.5 What Would Robust Evidence Look Like?

For a convincing detection of evolving dark energy, we would want:

1. **Pre-registered analysis:** Specify the model and test before unblinding.
2. **Out-of-sample validation:** Show that parameters fitted on early data predict later data better than ΛCDM.
3. **Consistency across probes:** BAO, SNe, CMB, and weak lensing should all prefer the same w_0, w_a values.
4. **Robust e-values:** All valid e-value methods should give $E \gtrsim 100$ consistently.
5. **Physical coherence:** The fitted $w(z)$ should correspond to a physically motivated model (quintessence, etc.), not just a convenient parameterization.

Current DESI data do not satisfy these criteria.

7 Conclusions

We have critically assessed DESI DR2's reported 3–4σ evidence for evolving dark energy using e-value analysis. Our main findings:

1. The naive likelihood ratio e-value $E = 392$ is **invalid** due to post-hoc parameter selection.
2. GROW mixture e-values range from 15–97, demonstrating strong prior dependence.
3. The data-split e-value $E = 1.4$ shows that w_0w_aCDM does **not** predict held-out redshift bins better than ΛCDM.
4. The 280× reduction from naive to validated e-values indicates substantial overfitting.
5. External evidence (Bayesian model comparison, dataset tensions) corroborates our skeptical conclusion.

Main Conclusion: Current DESI data do not provide robust evidence for departures from the cosmological constant. The apparent 3–4σ signal is largely an artifact of overfitting and does not survive proper statistical validation.

ΛCDM remains the most parsimonious explanation for cosmic acceleration. Future data releases (DR3 and beyond, with ∼40 million objects) may eventually provide the statistical power for a robust detection—but that evidence is not yet in hand.

Data and Code Availability

All analysis code, data files, and reproducibility materials are available at:

`https://github.com/jinyoungkim927/desi-evalue-analysis`

Official DESI BAO data from:

`https://github.com/CobayaSampler/bao_data`

References

Chevallier, M., & Polarski, D. 2001, International Journal of Modern Physics D, 10, 213

DESI Collaboration 2025, "DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints," arXiv:2503.14738

Grünwald, P., de Heide, R., & Koolen, W. 2024, Journal of the Royal Statistical Society Series B, "Safe Testing"

Kelly, J. L. 1956, Bell System Technical Journal, 35, 917

Linder, E. V. 2003, Physical Review Letters, 90, 091301

Notari, A., et al. 2025, "Bayesian analysis of DESI DR2," arXiv:2511.10631

Perlmutter, S., et al. 1999, Astrophysical Journal, 517, 565

Ramdas, A., Grünwald, P., Vovk, V., & Shafer, G. 2023, Statistical Science, 38, 576

Riess, A. G., et al. 1998, Astronomical Journal, 116, 1009

Shafer, G. 2021, Journal of the Royal Statistical Society Series A, "Testing by Betting"

Ville, J. 1939, Étude Critique de la Notion de Collectif, Gauthier-Villars

Vovk, V., & Wang, R. 2021, Annals of Statistics, "E-values: Calibration, combination and applications"

Weinberg, S. 1989, Reviews of Modern Physics, 61, 1