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. Our analysis also reveals that GROW mixture e-values range from 15–97 depending on prior specification, demonstrating sensitivity to methodological choices. Combined with external evidence that Bayesian model comparison favors ΛCDM and that tensions exist between DESI BAO and DES-Y5 supernovae at $z \sim 1$, we conclude that current data do not provide robust evidence for departures from the cosmological constant.

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 well but suffers from severe fine-tuning problems when interpreted as vacuum energy [Weinberg, 1989].

Alternative models posit that dark energy evolves over cosmic time. The CPL parameterization [Chevallier & Polarski, 2001, Linder, 2003]:

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

provides a phenomenological framework for testing this possibility, where w_0 is the equation of state today and w_a characterizes its evolution.

1.2 DESI's Claim

The Dark Energy Spectroscopic Instrument (DESI) DR2 collaboration [DESI Collaboration, 2025] reported that their BAO measurements, combined with CMB and supernova data, show 3–4σ preference for $w_0 > -1$ and $w_a < 0$, suggesting dark energy was more Λ-like in the past but is evolving today. If confirmed, this would be a landmark discovery in cosmology.

1.3 The Statistical Challenge

Standard frequentist significance testing via χ^2 differences has known limitations:

- It does not account for multiple hypothesis testing or model selection
- Parameters fitted to data and then tested on the same data leads to overfitting
- The reported significance may not reflect the probability of replication

We address these concerns using **e-values** [Vovk & Wang, 2021, Shafer, 2021, Ramdas et al., 2023], a modern statistical framework that provides valid measures of evidence even under optional stopping and model selection.

2 Statistical Framework: E-Values

2.1 Definition and Properties

Definition 1 (E-Value). *An e-value is a non-negative random variable E satisfying $\mathbb{E}[E \mid H_0] \leq 1$ under the null hypothesis H_0.*

E-values have several advantages over p-values:

Theorem 1 (Ville's Inequality). *For any e-value E and threshold $\alpha \in (0,1)$:*

$$P\left(E \geq \frac{1}{\alpha} \mid H_0\right) \leq \alpha \tag{2}$$

This holds regardless of stopping rules or how the e-value was constructed.

Theorem 2 (Combination Rules). *If $E_1, \ldots, E_n$ are independent e-values, their product $\prod_i E_i$ is an e-value. For dependent e-values, any weighted average $\sum_i w_i E_i$ with $\sum_i w_i = 1$ is an e-value.*

2.2 The Likelihood Ratio E-Value

For simple hypotheses, the likelihood ratio:

$$E = \frac{L(\text{data} \mid H_1)}{L(\text{data} \mid H_0)} \tag{3}$$

is an e-value, since $\mathbb{E}[L(\text{data} \mid H_1)/L(\text{data} \mid H_0) \mid H_0] = 1$.

2.3 The Overfitting Problem

Critical Warning: If H_1 parameters are chosen *after* seeing the data (as in maximum likelihood estimation), the likelihood ratio is **not** a valid e-value. The expectation can be arbitrarily large even when H_0 is true.

For Gaussian likelihoods:

$$E_{\text{naive}} = \exp\left(\frac{\Delta\chi^2}{2}\right) \tag{4}$$

where $\Delta\chi^2 = \chi^2_{H_0} - \chi^2_{\text{best-fit}}$. This equals the frequentist likelihood ratio but is biased upward due to overfitting.

2.4 GROW Mixture E-Values

To construct valid e-values for composite alternatives, we average over the parameter space:

$$E_{\text{GROW}} = \int \frac{L(\text{data} \mid \theta)}{L(\text{data} \mid H_0)} \pi(\theta)\, d\theta \tag{5}$$

where $\pi(\theta)$ is a prior distribution over alternatives. GROW (Growth Rate Optimal in Worst case) chooses π to maximize worst-case power [Grünwald et al., 2024].

2.5 Data-Split E-Values

The most robust approach splits data into training and test sets:

1. Split data: $D = D_{\text{train}} \cup D_{\text{test}}$
2. Fit alternative parameters using only D_{train}: $\hat{\theta} = \arg\max L(D_{\text{train}} \mid \theta)$
3. Compute e-value on held-out data:

$$E_{\text{split}} = \frac{L(D_{\text{test}} \mid \hat{\theta})}{L(D_{\text{test}} \mid H_0)} \tag{6}$$

This is valid because $\hat{\theta}$ is independent of D_{test} conditional on D_{train}.

3 Data

3.1 DESI BAO Measurements

We use official DESI BAO data from the CobayaSampler repository[1], endorsed by the DESI collaboration.

DR1 (Year 1): 12 measurements across 7 redshift bins, ∼6 million objects.

DR2 (Years 1–3): 13 measurements across 7 redshift bins, ∼14 million objects.

[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	—

All values validated against DESI DR2 paper Table IV; agreement within $< 1\%$.

3.2 Cosmological Models

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

Alternative (H_1): w_0w_aCDM with w_0, w_a as free parameters.

Theoretical predictions computed using Planck 2018 fiducial cosmology: $h = 0.6766$, $\Omega_m = 0.3111$, $r_d = 147.05$ Mpc.

4 Results

4.1 Model Fits

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

Model	w_0	w_a	χ^2	dof
ΛCDM (DR2)	-1 (fixed)	0 (fixed)	25.44	13
w_0w_aCDM (DR2)	-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

The χ^2 improvement is:

$$\Delta\chi^2 = 25.44 - 13.50 = 11.94 \tag{7}$$

For 2 additional degrees of freedom, this corresponds to $p = 0.0026$ ($\sim 3\sigma$), consistent with DESI's reported significance.

4.2 E-Value Analysis

Table 3: E-Value Results Summary

Method	E-Value	σ-equiv	Valid?	Notes
Simple Likelihood Ratio	392	3.9σ	**No**	Overfitted
GROW Mixture (narrow prior)	97	3.0σ	Yes	Prior-sensitive
GROW Mixture (default prior)	15	2.3σ	Yes	Prior-sensitive
GROW Mixture (wide prior)	17	2.4σ	Yes	Prior-sensitive
Data-Split Validation	**1.4**	$\mathbf{0.8}\sigma$	**Yes**	Tests generalization

4.3 The Critical Result

The data-split e-value tests whether a model fitted on a subset of redshift bins can predict held-out bins better than ΛCDM:

$$E_{\text{split}} = \frac{L(D_{\text{test}} \mid \hat{w}_0, \hat{w}_a)}{L(D_{\text{test}} \mid \Lambda\text{CDM})} = 1.4 \tag{8}$$

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

$$\frac{E_{\text{naive}}}{E_{\text{split}}} = \frac{392}{1.4} \approx 280 \tag{9}$$

Interpretation: The w_0w_aCDM model does not predict held-out data better than ΛCDM. The apparent evidence is due to overfitting, not a genuine signal.

4.4 DR1 to DR2 Stability Analysis

Table 4: Parameter Evolution from DR1 to DR2

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

Combined shift magnitude: $\sqrt{\Delta w_0^2 + \Delta w_a^2} = 0.24$

The moderate parameter shift between DR1 and DR2 suggests some instability in the fitted values as more data is added.

Caveat: DR2 contains DR1, so this is not a true out-of-sample test. The DR1→DR2 e-value of 3103 reflects stability, not generalization.

5 Discussion

5.1 Why the Large Discrepancy?

The 280× reduction from $E = 392$ to $E = 1.4$ arises because:

1. **Overfitting:** With 2 free parameters, the model can fit statistical fluctuations in the data. The naive $\Delta\chi^2 = 12$ improvement includes fitting noise.

2. **Lack of generalization:** When parameters are fitted on some redshift bins and tested on others, the improvement vanishes. The fitted w_0, w_a values are specific to the training data.

3. **Correlated systematics:** If systematic errors are correlated across redshifts, w_0w_aCDM may fit these systematics rather than true cosmological evolution.

5.2 Comparison to Bayesian Analysis

Independent Bayesian model comparison [Notari et al., 2025] using the same DESI data found:

$$\ln \mathcal{B} = -0.57 \tag{10}$$

where $\mathcal{B}$ is the Bayes factor for w_0w_aCDM vs ΛCDM.

Negative log Bayes factor means **ΛCDM is favored**—the extra parameters are not justified by the data when properly penalized for complexity.

5.3 Dataset Tensions

Recent work [Tension, 2025] found tensions between DESI BAO and supernova datasets:

Table 5: DESI BAO vs Supernova Dataset 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 apparent "evidence" for dynamic dark energy may arise from w_0w_aCDM resolving tensions between inconsistent datasets, rather than detecting true physics.

5.4 Limitations of This Analysis

1. **Data splitting reduces power:** Our E = 1.4 may underestimate the true signal due to reduced sample size. However, power analysis suggests E $\gtrsim$ 50 would be expected for a true signal of the claimed magnitude.

2. **Redshift splits are imperfect:** We split by redshift bins assuming independence. Correlated calibration errors could violate this assumption.

3. **DR1/DR2 are not independent:** True temporal validation would require data from distinct observing periods, which DESI does not provide separately.

6 Conclusions

We have applied e-value analysis to assess DESI DR2's reported evidence for evolving dark energy. Our key findings:

1. The naive likelihood ratio e-value of $E = 392$ ($\sim 3.9\sigma$) is **not valid** because parameters were fitted to the same data used for testing.
2. GROW mixture e-values range from 15–97, showing strong dependence on prior specification.
3. The data-split e-value of $E = 1.4$ ($\sim 0.8\sigma$) indicates that $w_0 w_a$CDM does **not** predict held-out redshift bins better than ΛCDM.
4. The $280\times$ reduction from naive to validated e-values indicates substantial overfitting.
5. External evidence (Bayesian model comparison favoring ΛCDM, dataset tensions) corroborates our findings.

We conclude that 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.

Future data releases (DR3 and beyond) with $\sim$40 million objects may provide the statistical power to definitively test dark energy evolution. Until then, ΛCDM remains the most parsimonious explanation for cosmic acceleration.

Data Availability

All code and data are available at `https://github.com/jinyoungkim927/desi-evalue-analysis`.

Official DESI data from `https://github.com/CobayaSampler/bao_data`.

References

Chevallier, M., & Polarski, D. 2001, IJMPD, 10, 213

DESI Collaboration 2025, arXiv:2503.14738

Grünwald, P., de Heide, R., & Koolen, W. 2024, JRSS-B, Safe Testing

Linder, E. V. 2003, PRL, 90, 091301

Notari, A., et al. 2025, arXiv:2511.10631

Perlmutter, S., et al. 1999, ApJ, 517, 565

Ramdas, A., Grünwald, P., Vovk, V., & Shafer, G. 2023, Stat. Sci., 38, 576

Riess, A. G., et al. 1998, AJ, 116, 1009

Shafer, G. 2021, JRSS-A, Testing by Betting arXiv:2509.19899, Calibration-independent consistency test

Vovk, V., & Wang, R. 2021, Ann. Stat., E-values: Calibration, combination

Weinberg, S. 1989, Rev. Mod. Phys., 61, 1