

Final Examination — Blueprint

Econ 6376 — Applied Time Series Econometrics — Fall 2026

The conditions

This examination is in person, in the university final examination period, on 17 December 2026. It is worth 45% of your grade — the largest single component of this course, and the only one written under supervision.

- **Two and a half hours.** One sitting.
- **Closed book. No notes, no textbook, no devices of any kind** — including a calculator. Every piece of arithmetic on the paper is built to be done with a pen.
- **Cumulative.** Every module from the first week to the last is represented.
- **Everything you need is printed in the exhibit booklet:** every plot, every regression, every test statistic, every critical value. Nothing to look up, and nothing left out on the assumption that you remember it.
- **You write in sentences, by hand.** A formula with no statement of what it means earns part marks. A claim with no evidence behind it earns none.

There is no take-home component, no resubmission, no grade replacement.

What is *not* on this exam

Read this before you build a revision plan. It exists to stop you studying things that will not be examined.

No R. You will not write code, and no question's answer is a script. Output appears as R prints it, so *reading* a printout is assumed; *producing* one is not tested.

No memorizing critical values. Verified across the whole booklet: every test that reports a statistic arrives with either its critical values printed beside it or its p -value. Sample Exhibit 2 in this document is built to the same standard as the booklet's, and prints its Augmented Dickey-Fuller thresholds like this —

tau2 critical values 1%: -3.46 5%: -2.88 10%: -2.57

— with the exact values depending, as they must, on the test specification. The booklet does the same for KPSS, for Engle-Granger and for the Johansen trace table. You will never lose a mark for forgetting a number from a table.

But you must know the null hypothesis of every test this course taught. Do not read the paragraph above as permission to skip this. Where an exhibit *reports* a test it states that test's null in words so you can read the exhibit — a reading aid, not an exemption. Six of the seven questions open by asking you to define something or state a null yourself, and those opening parts alone are worth **at least 26 of the 100 points**. Definitional recall is the most reliably earnable block on the paper and the most commonly under-prepared.

No derivations. Where mathematics is asked for it is a statement — a condition, a statistic, a formula — with what it means. Nothing runs longer than a few lines.

No ARCH, no GARCH, no volatility modelling. The course ends at Module 13, FEVD and VECMs, lectured 3 December. Module 14 was absorbed when the travel week was; it is on neither the syllabus nor the exam.

Nothing from your own data. The midterm dataset and the problem sets built on it are not examined here. Every series on this paper is supplied.

No cascading failures. The seven questions are independent, and where a later one would naturally use an earlier result, that result is supplied in the question. An unfinished Question 3 costs you Question 3 and nothing more.

What is on it

Seven questions, thirteen modules, one hundred points. This is the exam's own rubric, reproduced without alteration.

Question	Covers	Format	Points
1 — Stationarity	Modules 1–2	Define, apply, decide	15
2 — Identification and information criteria	Modules 3–4	Reconciliation	16
3 — Residual diagnostics and seasonality	Module 5	Error-finding	14
4 — Forecast evaluation	Modules 6–7	Error-finding, interpretation	14
5 — Dynamic multipliers and Granger causality	Modules 8–9	Define, apply, decide	14
6 — Cointegration and system specification	Modules 10–11	Reconciliation	14
7 — Impulse responses and variance decomposition	Modules 12–13	Define, apply, decide; interpretation	13

Question	Covers	Format	Points
Total			100

Two things follow from that table.

It is genuinely cumulative, Modules 1 through 6 included — a consequence of the redesign, not an oversight. Your problem sets and midterm are take-home; this is the only assessment written under supervision, so it is the only one that certifies anything, the early material included. Stationarity is examined in December, for fifteen points.

Module 13 has no problem set behind it. FEVD and VECMs are lectured 3 December, Problem Set 5 stops at Module 12, and the exam is two weeks later. That block is therefore examined *definitionally and interpretively* — reading a supplied table and saying what it says, naming a specification and the sign a coefficient must take — rather than computationally. Nothing on this paper asks you to derive or estimate a VECM. But definitional means you have to know the definitions: this redirects your preparation, it does not reduce it.

Every question ends with a blockquote headed **What earns the points here**, in the same voice as the ones on your midterm. Read it before you start writing.

The four question formats

Every question is one of four shapes, and the last two are trained by nothing else in this course.

1. Define, apply, decide. State a definition precisely, use it on an exhibit, make a determination. The definition is not a warm-up — it is the instrument you are about to use, and if it is wrong the rest cannot be right. *To prepare:* write the definitions out from a blank page, not one you are looking at. The exam wants the mathematics and the plain-English statement to visibly say the same thing, so write the condition, write the sentence, and check that they match.

2. Interpretation. Here is output; what does it tell you. *To prepare:* read output you did not produce. For the rest of the semester, when you run something, cover the code before you look at the result and ask what you would conclude from the output alone.

3. Reconciliation. Two bodies of evidence point different ways. Which governs, and why? *To prepare:* learn what each procedure is *for*, not just how to run it. A memorized workflow assumes the evidence agrees, which is exactly the assumption these questions remove. Most reconciliations resolve because one instrument is being asked something outside its remit — so for every test you know, be able to say what it establishes and what it cannot.

4. Error-finding. *A student concluded X from this output. They are wrong. Explain why, and what they missed.* *To prepare:* **knowing a procedure is not enough — you have to know what wrong looks like**, and those are different skills. Running a test correctly a hundred times does not teach you to recognise it run incorrectly, because you never saw the incorrect version. So produce them on purpose: take a finished answer and run the test at the wrong horizon, drop the degrees-of-freedom correction, difference one time too many. Then look at what the output does.

Note also that the errors are usually **not equal**, and ranking them is part of the question. Find the one that actually overturns the conclusion, and say so.

Three worked samples

Each is a question in the shape and wording the exam uses, with a real exhibit, an answer that would earn full marks, and a note on what half marks looks like.

The answers matter more than the questions. What you cannot calibrate from a syllabus is *how much writing earns full marks*, and one worked answer teaches more than five bare questions. They are written as you would write them — in sentences, on paper, in the time available. Sample A is also a working calibration of what a minute a point buys: look at how much is written there, and at how little of it is throat-clearing.

None of these three questions is on the exam. Publishing a worked answer spends the exhibit it is worked from, so these were built fresh, on determinations the exam does not ask for. Do not go looking for Series Z in December.

Sample A — Error-finding (*Modules 1–3, 15 points*)

Series Z is a series of 180 observations. A student writes:

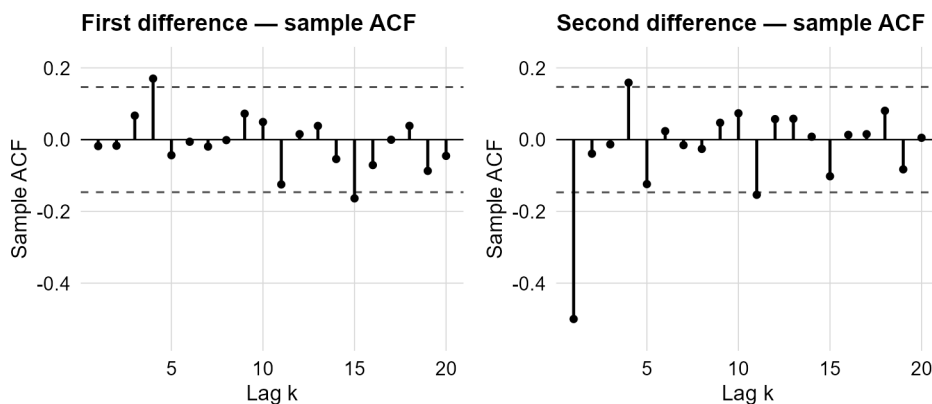
“The ADF test on the level of Series Z gives -1.29 , well above the 5% critical value of -2.88 , so I cannot reject a unit root. I took a first difference and the ADF statistic is -9.39 , which rejects. To be safe I differenced once more: the statistic on the second difference is -15.08 , which rejects even more strongly. Since the second difference is the most stationary of the three, I will set $d = 2$ and model $\Delta^2 y_t$.”

The student is wrong. Sample Exhibits 1 and 2 contain everything needed to see why.

(a) [4 points] State the null hypothesis of the Augmented Dickey-Fuller test, and say what rejecting it does and does not establish.

(b) [7 points] Identify what the student got wrong. There is more than one error. For each, say what it is and what difference it makes — and be clear about which one actually overturns the conclusion.

(c) [4 points] State what you would do instead, and what you would check to confirm it had worked.



SAMPLE EXHIBIT 2 - Series Z: unit root tests and summary statistics
 Series Z: T = 180 observations.

 AUGMENTED DICKEY-FULLER (drift specification: constant, no trend)

H0: the series contains a unit root.

Lag length chosen by BIC from a maximum of 4.

	tau2 statistic
Series Z, in levels	-1.29
Series Z, first difference	-9.39
Series Z, second difference	-15.08
tau2 critical values	1%: -3.46 5%: -2.88 10%: -2.57

 KPSS (mu specification: stationarity around a constant)

H0: the series is stationary.

Series Z, first difference	0.070
critical values	10%: 0.347 5%: 0.463 2.5%: 0.574 1%: 0.739

 SUMMARY STATISTICS

	1st diff	2nd diff
standard deviation	1.026	1.468
rho-hat_1	-0.018	-0.500
rho-hat_2	-0.017	-0.039
two standard errors, 2/sqrt(T)	0.149	0.150

 AN MA(1) FITTED TO THE SECOND DIFFERENCE

Arima(diff(diff(Z)), order = c(0, 0, 1), include.mean = FALSE)

	ma1
coefficient	-1.000
s.e.	0.018

A full-credit answer

(a) The ADF null is that the series contains a unit root — that it is integrated and needs differencing. The alternative is that it is stationary around whatever deterministic terms the test regression includes, here a constant, since the drift specification was used. Rejecting says the data are inconsistent with a unit root at the stated level. It does not establish stationarity in any absolute sense, and it says nothing about *how many* differences are appropriate: each run asks one question about one series.

(b) Two errors, and they are not equal.

The first is real but not decisive. The student treats a more negative statistic as stronger evidence for differencing, as though the ADF statistic were a stationarity score to be maximised. It is not. It is a test statistic referred to a critical value, and it returns a verdict — reject or fail to reject — not a ranking. Once the first difference rejects at -9.39 against a 5% value of -2.88 , the question has been answered. Running the test again on the second difference asks a *new* question about a *new* series, and the answer to that one — that $\Delta^2 y_t$ has no unit root — is perfectly true and entirely

beside the point. Note that the student's premise is correct: -15.08 really is more negative than -9.39 . It is the inference drawn from it that fails.

The second overturns the conclusion. The second difference is **over-differenced**, and Sample Exhibit 1 shows it. The correlogram of Δy_t is what white noise looks like: $\hat{\rho}_1 = -0.018$, nothing systematic after. The correlogram of $\Delta^2 y_t$ has a single large negative spike at lag 1, $\hat{\rho}_1 = -0.500$, more than three times the two-standard-error band of 0.150, with everything past lag 1 back inside the band. That is not a feature of the data. It is the signature of a unit root in the *moving-average* polynomial: differencing an already-stationary series multiplies it by $(1-L)$, contributing an MA(1) factor with $\theta = -1$, whose theoretical lag-1 autocorrelation is exactly -0.5 .

Sample Exhibit 2 confirms it twice more. The standard deviation *rises* from 1.026 to 1.468 — differencing has added variance rather than removed it, by a factor close to the $\sqrt{2}$ the algebra predicts — and an MA(1) fitted to $\Delta^2 y_t$ returns $\hat{\theta} = -1.000$, exactly on the invertibility boundary. A non-invertible MA has no convergent AR(∞) representation, so the innovations cannot be recovered from the observed past, and ordinary inference at that boundary is not trustworthy.

The evidence that one difference sufficed is on the same exhibit: KPSS on Δy_t is 0.070 against a 5% critical value of 0.463, nowhere near rejecting. ADF and KPSS have opposite nulls, and on the first difference they agree.

(c) Set $d = 1$ and model Δy_t . Confirm it three ways rather than one: the residual correlogram should show no systematic structure; a Ljung-Box test at the course horizons, with degrees of freedom adjusted for the estimated ARMA parameters, should not reject; and no fitted MA coefficient should sit near -1 . That last check is the one that would have caught this. An MA estimate piling up at the invertibility boundary after differencing is the over-differencing warning light.

Half marks. Spotting that $\Delta^2 y_t$ is over-differenced and stopping there, without engaging the argument the student actually made — the claim that a more negative statistic is better evidence *is* the reasoning error, and leaving it standing corrects the answer without correcting the mistake. The reverse earns half too: a paragraph on why ADF statistics are not scores, with no use made of Exhibit 1, which is where the proof is. Naming the lag-1 spike and calling it “autocorrelation the model should capture” is worse than either — it is not structure to be modelled, it is structure that was manufactured, and the fix is to stop differencing, not to add an MA term. Nothing is earned by noting that one or two of the twenty lags in the *first* difference's correlogram sit just outside the band; twenty draws from white noise put about one outside it every time. The lag-1 spike in the second difference is -0.500 against a band of 0.150. Telling a pattern from a coin flip is the whole of this question.

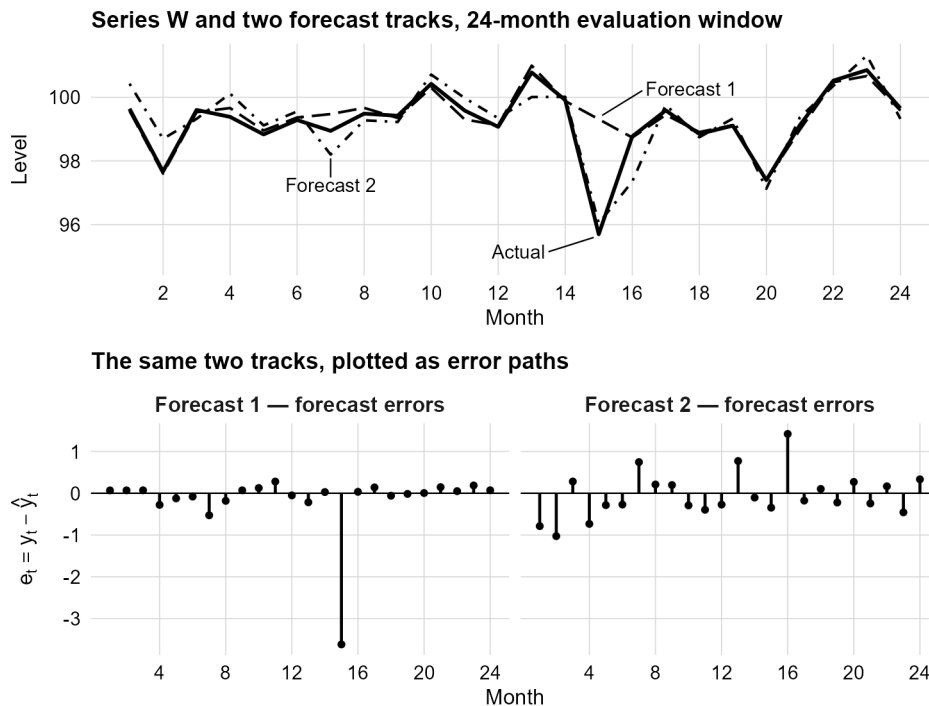
Sample B — Reconciliation (Modules 6–7, 14 points)

Two forecast tracks for Series W were produced over the same 24-month evaluation window and scored against the same actuals. Sample Exhibit 3 plots them, and plots their forecast errors. Sample Exhibit 4 is the scorecard.

(a) [4 points] Define the root mean squared error and the mean absolute error. Say what each measures, and what differs between them beyond the arithmetic.

(b) [7 points] The two metrics rank these tracks in **opposite** orders. Set out why, using both exhibits, and say which summary is telling you the truth about these forecasts.

(c) [3 points] State what you would need in order to choose between the two tracks, and when that choice should have been made.



SAMPLE EXHIBIT 4 - Out-of-sample scorecard, Series W

Both tracks are scored on the same 24 months, against the same actuals.

Neither model was re-estimated inside the window.

	RMSE	MAE	MAPE (%)
Forecast 1	0.757	0.270	0.278
Forecast 2	0.528	0.421	0.424

SHAPE OF THE TWO ERROR PATHS (plotted in Sample Exhibit 3)

	Forecast 1	Forecast 2
largest single $ e_t $	3.616	1.419
median $ e_t $	0.075	0.282
months with $ e_t > 1.0$	1	2
months this track is the closer one	23	1

A full-credit answer

(a) With forecast errors $e_t = y_t - \hat{y}_t$ over n periods, $RMSE = \sqrt{n^{-1} \sum_t e_t^2}$ and $MAE = n^{-1} \sum_t |e_t|$. Both summarise the size of the errors in the units of the series. What differs is how each weighs a large error against a small one. MAE weights every error in proportion to its size, so it reports

the typical miss. RMSE squares first, so an error twice as large counts four times as much; the square root at the end returns the answer to the original units but does not undo that weighting. RMSE is a statement about the whole error distribution, tail included; MAE is a statement about its middle.

(b) They disagree because the two error paths have different *shapes*, and each metric responds to the feature it was built to notice.

Forecast 1 is the closer track in 23 of the 24 months, with a median absolute error of 0.075 against Forecast 2's 0.282: in an ordinary month it is almost exactly right. In month 15 it misses by 3.616. The lower panel of Sample Exhibit 3 makes this unmissable — Forecast 1's error path lies flat against zero with one enormous spike, while Forecast 2's is a scatter of moderate errors with no spike at all, its largest being 1.419.

MAE averages absolute errors, so twenty-three near-zero months outweigh one large one and it ranks Forecast 1 ahead, 0.270 to 0.421. RMSE squares, so the single miss enters as roughly 13 while a typical Forecast 2 error of 0.4 enters as 0.16 — one month is enough to reverse the ranking, 0.757 against 0.528. MAPE agrees with MAE, and that is not a third opinion: MAPE is also linear in the absolute error, so it was never going to disagree with MAE about how to weight a tail.

Neither summary is telling the truth about these forecasts, because neither can. A scorecard is a lossy compression of an error path, and these two numbers disagree precisely because they compress different features of it. What is true about these forecasts is the picture: one is excellent except when something unusual happens, the other mediocre but never badly wrong. Both metrics report that accurately. They report different halves of it.

(c) What is needed is not more data. It is a statement of what being wrong *costs* — whether one month at -3.6 hurts the decision this forecast feeds more than twenty-four months at ± 0.4 . If the forecast supports something with a threshold or a buffer that one bad month breaks, the quadratic weighting is right and Forecast 2 wins. If it feeds a decision remade every month that absorbs errors additively, Forecast 1 wins. That question is prior to the data and must be answered before the scorecard is computed. Reporting both metrics and then adopting the one that favours your preferred track is not a comparison. It is a search.

Half marks. Reporting that RMSE and MAE disagree and stopping — the disagreement is the question, not the answer. Declaring one metric “better”, usually RMSE, on no ground except that it is the one you have seen most often. Explaining the difference from the two formulas without ever looking at the lower panel of Exhibit 3, where the shape driving the whole result actually is: an argument about squaring that never identifies the month doing the squaring is arithmetic, not evidence. And treating MAPE as a tie-breaking third vote — count your instruments, not your columns; there are two here.

A general point, since this format produces more confident nonsense than any other: **check that the evidence disagrees before you reconcile anything.** Where two instruments agree there is nothing to resolve, and manufacturing a conflict so you have something clever to say about it costs you twice. You lose the marks for reading the exhibit correctly, and you spend time you needed elsewhere.

Sample C — Define, apply, decide (Module 3, 12 points)

Sample Exhibit 5 reports two fitted AR(2) specifications, for two different series. Not every exhibit on this paper is a picture.

(a) [4 points] State the stationarity condition for an AR(p) process. Say what the condition is on, and say in plain language what fails when it fails.

(b) [5 points] Apply it to both specifications and classify each. Show the arithmetic.

(c) [3 points] Say what you would do with the specification that fails, and note anything in the exhibit that makes your call less than clean.

SAMPLE EXHIBIT 5 - Two fitted AR(2) specifications
Series U and Series V are two different series. Each was fitted with

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

Series U: ordinary least squares, T = 200, 198 usable observations.

Series V: ordinary least squares, T = 200, 198 usable observations.

	Series U	s.e.	Series V	s.e.
phi1-hat	0.50	0.067	0.70	0.068
phi2-hat	0.36	0.067	0.30	0.069

REMINDER OF THE ARITHMETIC, NOT OF THE CONDITION

For an AR(2) the characteristic equation is

$$\lambda^2 - \phi_1 \lambda - \phi_2 = 0$$

$$\text{with roots } \lambda = (\phi_1 \pm \sqrt{\phi_1^2 + 4 \phi_2}) / 2 .$$

Both discriminants here are perfect squares. You will not need a calculator.

A full-credit answer

(a) An AR(p) process $y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \epsilon_t$ is stationary if and only if every root of its lag polynomial $\Phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$ lies **outside** the unit circle — equivalently, if and only if every root of the characteristic equation $\lambda^p - \phi_1 \lambda^{p-1} - \dots - \phi_p = 0$ has modulus **less than one**. The two are the same condition written twice, related by $\lambda = 1/L$.

The condition is on the **roots**, not on the coefficients. It is not enough that each $\hat{\phi}_j$ is below one in absolute value, and checking coefficients one at a time is not a rough version of the right test — it is a different test, and it returns wrong answers. What fails when the condition fails is mean reversion: a shock entering the system is no longer damped geometrically, so there is no fixed mean or variance to return to, and the moments this course is built on do not exist.

(b) For an AR(2) the characteristic roots are $\lambda = (\hat{\phi}_1 \pm \sqrt{\hat{\phi}_1^2 + 4\hat{\phi}_2})/2$.

Series U, $\hat{\phi}_1 = 0.50$, $\hat{\phi}_2 = 0.36$. Discriminant $0.25 + 1.44 = 1.69$, square root 1.30. Roots $(0.50 + 1.30)/2 = 0.90$ and $(0.50 - 1.30)/2 = -0.40$. Both have modulus below one, so **Series U's specification is stationary**.

Series V, $\hat{\phi}_1 = 0.70$, $\hat{\phi}_2 = 0.30$. Discriminant $0.49 + 1.20 = 1.69$ again, square root 1.30. Roots $(0.70 + 1.30)/2 = 1.00$ and $(0.70 - 1.30)/2 = -0.30$. One root has modulus exactly one, so **Series V's specification is not stationary** — it has a unit root.

Note what separates them. Every coefficient in both columns is comfortably below one in absolute value, and Series V's are individually no larger than Series U's. Read one coefficient at a time and both look harmless; the discriminants are even identical. The only thing that differs is $\hat{\phi}_1$, and it is enough.

(c) Series V should not be modelled in levels as a stationary AR(2). That $\hat{\phi}_1 + \hat{\phi}_2 = 1.00$ exactly is the fingerprint of an integrated series fitted with a stationary specification: the autoregressive weights of an I(1) process sum to one, and least squares has found that as closely as it can. The next step is not another AR fit but a unit root test on the level, and if it fails to reject, model the first difference.

What makes the call less than clean is that the estimates sit exactly on the boundary, and with standard errors near 0.07 a sum of 1.00 is not distinguishable from, say, 0.97. That is a reason to test, not a reason to call Series V stationary — the unit-root boundary is precisely where ordinary t -ratios do not have their usual distribution, which is why there is a Dickey-Fuller table at all. Deciding from the estimated roots alone would be asking the standard errors a question they cannot answer.

Half marks. Reading down the coefficient column, observing that everything is below one, and calling both specifications stationary. That is the error this exhibit is built around and the most common one on this material. Also: stating the condition on the roots correctly and then never computing a root; or computing Series U's roots, finding them comfortable, and assuming Series V is similar because its numbers look similar. They do look similar. That is the point. And a version of (c) that says “difference it” with nothing about what would confirm that, or that treats a root of exactly 1.00 as settling the matter with no acknowledgement that these are estimates.

Pacing

One hundred points, one hundred minutes of writing — **a minute a point**. The exam prints a suggested allocation for each question on its front page, and because that allocation is proportional, the simple rule reproduces it. Add roughly **twenty minutes for reading the exhibits**, which is real work and is counted inside the time the paper is built to take. That is 120 minutes.

The slot is 150. **The remaining thirty minutes are deliberate slack**, there so you can go back through your answers at the end, not because there is more work to do. Do not pace against 150: a student who spreads the work across the whole slot over-invests in Questions 1 and 2 and reaches Questions 5, 6 and 7 — forty-one points between them — with nothing left.

If you are well past the suggested time on a question, leave it and come back. Nothing depends on anything else.

Sample A's answer runs about 550 words, against the 18 minutes a 15-point question buys. That is achievable — but not if you spend the first five minutes deciding where to begin, which is what the **What earns the points here** note is there to prevent.

How to prepare

Write the definitions out from a blank page. Twenty-six points open the questions, and the exam wants the mathematics and the plain-English sentence to visibly say the same thing. Produce both, then check they match.

Rework old problem set questions without opening your code. If you cannot reach the answer with a pen, you did not learn it — you retrieved it. This is the highest-return item here, and the reason the problem sets are worth doing yourself.

Break things on purpose. For error-finding there is no substitute. Take a finished answer, run the test at the wrong horizon, drop the degrees-of-freedom correction, difference one time too many, and watch what the output does.

Use the AI packs. They are the best practice tool available for an exam an AI cannot sit for you. The packs are released with their modules and are built for exactly this. Have the assistant simulate a series and **hide the data generating process from you.** Then read the correlogram, or the test output, cold: make a determination, write down your reasoning, and only afterwards ask what it generated. That is the December situation reproduced exactly, and there is no other way to reproduce it. When it gets comfortable, ask for two exhibits that point in different directions, or for a confident wrong answer you have to find the error in.

This is not in tension with the disclosure policy on take-home work. That policy governs work you submit. This is training against a tool that will not be in the room with you, which is precisely the right use of it.

Nothing in December is meant to be a surprise. If anything about the exam changes, this document changes first.