Psychology 434 Test 2 Study Sheet

Weeks 8–10

1. From ATE to CATE

Test 2 asks you to apply the workflow to short scenarios. Good answers name the causal question, state the relevant rule, and explain why that rule changes the conclusion.

The average treatment effect (ATE) causal estimand is

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)].$$

The conditional average treatment effect (CATE) causal estimand is

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

The ATE summarises a population. The CATE summarises subgroups defined by covariate values. The difference matters when the ATE hides meaningful variation across people.

2. Causal forests

A causal forest learns where in the covariate space $\tau(x)$ varies, by recursively splitting the data on whichever covariates and thresholds reduce within-node heterogeneity in treatment effect. Two design choices matter for inference.

Honest splitting. The tree's structure is learned on one subsample; leaf-level treatment effects are estimated on a different subsample. Honesty prevents the same data from being used to both choose splits and estimate effects, which would bias the estimates and the confidence intervals.

Doubly-robust scores. Each unit's contribution to the estimator is built from both an outcome model and a propensity model. The estimator is consistent if either model is approximately correct. This protects against misspecification of any single component.

The forest is an estimator, not an identification strategy. Consistency, exchangeability, and positivity still need to be argued from the design and the directed acyclic graph (DAG).

3. Heterogeneity as a hypothesis: calibration, RATE, Qini

A causal forest should not be treated as useful for targeting just because it returns different $\hat{\tau}(x)$ values. First ask whether the forest has evidence of real heterogeneity rather than noise.

In the Lab 9/10 workflow, the quick check is the causal-forest calibration test. Its "Differential prediction" coefficient tests the null hypothesis that the forest's predicted $\hat{\tau}(x)$ carries no information about real treatment-effect variation. A coefficient reliably above zero rejects that null and supports a targeting interpretation. An interval that includes zero leaves the null standing: the data cannot distinguish the forest's heterogeneity from noise, which still leaves open whether effects are homogeneous. Read the test as a pre-screen — it clarifies whether targeting could plausibly help, so a weak result calls for reading any policy tree cautiously.

If a forest's predicted $\hat{\tau}(x)$ ranks units by their estimated benefit, the **rank-weighted average treatment effect** (RATE) measures whether the high-ranked units actually have larger treatment effects than the low-ranked units. RATE was introduced in Week 8 as another way to assess ranking quality. The Lab 9/10 policy-tree workflow uses the calibration test as its screening step.

- **AUTOC** weights by ranks across the whole population (good for detecting heterogeneity).
- **Qini** weights by treated fraction (good for evaluating targeting under a budget constraint).

A small or zero RATE means the forest's ranking is not picking up treatment-effect heterogeneity that is large enough to act on. A wide CI on RATE means the data cannot distinguish heterogeneity from noise.

A **Qini curve** plots cumulative treatment effect against cumulative treated fraction. A curve well above the diagonal indicates targeting helps. A curve that hugs the diagonal indicates targeting and uniform treatment perform similarly.

Keep the roles separate: calibration/RATE/Qini assess heterogeneity or ranking quality; the policy tree summarises a simple allocation rule.

4. Policy trees

A policy tree converts $\hat{\tau}(x)$ into a deployable allocation rule by partitioning the covariate space into a small number of leaves and assigning each leaf an action (treat / do not treat). It chooses splits that maximise expected policy value under the objective supplied to the algorithm.

In the course workflow, the objective is outcome-only. A "do not treat" leaf means the rule assigns the no-treatment action for that covariate profile. It does not mean the analysis has computed money saved, staff time saved, or any other resource saving.

Reading a tree. Each non-leaf node names a covariate and a threshold. Each leaf says "treat" or "do not treat". A depth-2 tree asks at most three yes/no questions before committing.

Choosing depth (parsimony rule). Fit depth-1 and depth-2 as candidates, then prefer the simpler depth-1 tree unless the depth-2 tree improves held-out policy value, using the point estimate, by at least the prespecified gain threshold. In the course workflow, the default threshold is `min_gain_for_depth_switch = 0.01` outcome units. Uncertainty and stability guide how cautiously to interpret a threshold-clearing depth-2 rule; interval overlap is not the selection rule. Simpler rules are more interpretable, more stable across resamples, and easier to defend to non-technical decision-makers.

Off-policy evaluation. Policy values are estimated on held-out data not used to construct the rule, so they reflect out-of-sample performance.

Coverage. Coverage is the share of people the selected rule recommends for treatment. In the default workflow it is an output of the fitted rule, not a budget chosen in advance. If a programme can treat only 20% of people, that budget must be added explicitly, or the analyst should compare with a ranking rule that treats the top 20% by estimated benefit.

Split variables. A split variable is a targeting descriptor under the fitted objective. It is not an identified cause of differential response.

5. Equity, governance, democratic judgement

A policy tree maximises expected benefit under its objective. It does not decide what justice requires. Three live threats:

Proxy variables. A split on a deprivation, income, postcode, age, or baseline-wellbeing variable can affect social groups differently even when those groups are not named in the tree. Removing an explicit variable does not remove its proxies.

Most-versus-somewhat trade-off. Targeting those who benefit *most* may withhold treatment from those who benefit *somewhat*. Whether that is acceptable depends on values, law, institutional purpose, and democratic accountability.

Science and public judgement. Statisticians and psychologists can estimate benefits, harms, uncertainty, and subgroup patterns. They do not get to legislate justice. In a democracy, citizens and institutions debate and vote about the values that govern public allocation. The analyst's job is to make the trade-offs visible and to avoid presenting an objective function as a public decision.

A defensible deployment names the decision-maker, the public objective, the override conditions, the audit plan, and the limits of the model.

6. Outcome-wide design and multiple-testing correction

When a single exposure is tested against several outcomes, the chance that at least one CI excludes zero by chance alone exceeds the nominal per-test rate.

Bonferroni correction. With K outcomes at family-wise error rate $\alpha_{FW} = 0.05$, test each outcome at $\alpha = 0.05/K$. For $K = 4$, that is $\alpha = 0.0125$, equivalent to a 98.75% CI per outcome. Report multiplicity-adjusted intervals alongside (or in place of) the unadjusted ones.

The aim of an outcome-wide design is meta-analytic interpretation, not cherry-picking. Read the full pattern of estimates, not the one outcome that survives the cut.

State the causal estimands separately for each outcome, then interpret the vector of results together.

7. Sensitivity to unmeasured confounding: E-values

The **E-value** for an estimated effect is the minimum association strength, on the risk-ratio scale, that an unmeasured confounder would need to have with both the exposure and the outcome — over and above measured covariates — to fully explain away the effect. Larger E-values indicate greater robustness.

There is no universal threshold at which an E-value becomes "safe". Interpret it against the study design, the covariates already measured, and plausible omitted causes in the setting. Report the E-value for the point estimate and for the confidence-limit closest to the null, then explain what kind of unmeasured confounder would be needed for the result to disappear.

8. Measurement and causal inference

Reflective vs formative models. A reflective model treats indicators as effects of an underlying latent construct (the construct causes the items). A formative model treats indicators as constituents of the construct (the items together compose it). Each is awkward for causal inference: reflective models posit a single latent cause that may not exist as a real-world quantity; formative models lack a clear potential-outcomes interpretation, because intervening on a construct is not the same as intervening on each item.

Multiple versions of treatment. When one treatment label bundles substantively different versions, $Y(1)$ is not well-defined. Consistency is threatened. The recorded ATE is then a population-weighted average across versions rather than the effect of any single intervention.

Measurement invariance. Invariance asks whether items relate to the construct the same way across groups. Configural invariance fails if items load on different latent structures; metric invariance fails if loadings differ; scalar invariance fails if intercepts differ. Without scalar invariance, mean differences across countries or languages confound true differences with measurement differences.

Measurement error as a structural threat. Let Y^* be the true outcome and Y the recorded outcome. If exposure A also affects how Y is recorded (e.g., the intervention teaches participants to label distress differently), the path $A \to U_Y \to Y$ is causal but does not run through Y^*. Identification of the effect on Y^* then requires extra assumptions or design responses.

Confounding control by baseline measures. In a three-wave panel, including the baseline measures of both the exposure (A_0) and the outcomes ($Y_{k,0}$) in the adjustment set provides strong confounding control. Any unmeasured confounder would need to act on later exposure conditional on baseline exposure and baseline outcome — a demanding requirement.

9. How to study

For each scenario you encounter:

1. State the causal question and the causal estimand. Distinguish the causal estimand from the statistical one.
2. Draw the DAG. Name the adjustment set the DAG implies.
3. State the identification assumptions: consistency, exchangeability, positivity. Add the design responses (lagged-self adjustment, IPCW for attrition).
4. Choose an estimator. For heterogeneity, name the calibration or ranking-quality check and the targeting tool (policy tree).
5. Apply the parsimony rule for depth, the Bonferroni correction for multiple outcomes, and the E-value for sensitivity.
6. Conduct an equity audit before recommending deployment. State one value judgement the model cannot settle.
7. Comment on measurement: which version of the treatment, which scale, which invariance level holds.

Practise with new examples rather than only rereading definitions. For any scenario, ask what could go wrong at each step, and what evidence would change your answer.

10. Common answer frames

Policy-tree depth. "I would report the simpler depth-1 rule unless the depth-2 rule improves held-out policy value, using the point estimate, by at least the prespecified threshold. Fitting depth-2 is not enough; the extra complexity has to buy the prespecified gain. Wide uncertainty would make me interpret a threshold-clearing depth-2 rule cautiously."

Fixed budget. "The default policy tree reports the treated share implied by the selected rule. If the programme can treat only a fixed share, that capacity constraint must be added explicitly or compared with a ranking rule."

Equity. "Before acting, I would audit treatment recommendations across relevant social groups, inspect proxy variables, name the public objective, and state who may override the rule."

Measurement. "If the treatment label or outcome scale means different things across groups, the causal estimand is less clear. Consistency, measurement invariance, and differential measurement error have to be addressed before comparing effects."