MAE4054 Study Designs for Causal Inference

Grading guide, school exam 9 October 2026

Section A, Concepts (30 points)

A1 (6 points)

(i) 2 points. The difference between a pupil's test score if the pupil attends the summer school and the same pupil's test score if the pupil does not attend. An average of this difference over pupils, for all pupils or for those who attended, also gets full credit.

(ii) 2 points. Each pupil either attends or does not, so only one of the two potential outcomes is observed. The other is missing, so the difference cannot be computed for any pupil.

(iii) 2 points. 1 point for naming selection bias (confounding or omitted variable bias is also accepted) and explaining it: the pupils who attended would have scored differently from the pupils who did not, even without the summer school, so the comparison between the groups mixes the causal effect with this difference. 1 point for a concrete example, such as pupils with motivated parents being more likely to attend and scoring higher regardless.

A2 (5 points)

(i) 3 points. With a lottery, who gets a place does not depend on the pupils' characteristics. Winners and losers are therefore comparable in expectation on all characteristics, observed and unobserved, and would have scored the same, on average, without the summer school (2 points). The difference in mean test scores between winners and losers therefore estimates the causal effect of winning a place (1 point).

(ii) 2 points. Any reasonable discussion of the setting, for example: a lottery requires more applicants than places; refusing places by lottery may be seen as unfair; pupils who lose may find other help, which reduces the contrast between the groups. Two sensible points give 2 points, one gives 1.

A3 (9 points)

(i) 4 points. The open backdoor paths are T ← P → Y and T ← N → S → Y (1 point each). A set that closes both must contain P and either N or S, and must not contain E or A: {P, N}, {P, S}, or {P, N, S} (2 points). A set that contains E or A gets 0 points on the set. A list that includes T → E → Y or T → A ← Y as a backdoor path gets at most 1 of the 2 points for the paths.

(ii) 3 points. E (pupil effort) should not be controlled for: it lies on the causal path T → E → Y, so controlling for it removes part of the effect (1.5 points). A (advanced-class placement) should not be controlled for: it is a collider, T → A ← Y, and controlling for it creates a non-causal association between T and Y (1.5 points).

(iii) 2 points. If P is not observed, the path T ← P → Y cannot be closed, and the effect is not identified by controlling for covariates (1 point). The strategies are the quasi-experimental designs: instrumental variables, regression discontinuity, and difference-in-differences. They have in common that they use variation in T that is as good as random, from a lottery, a threshold, or the timing of a change, instead of trying to measure all confounders (1 point). Naming the family, or two of the designs, is enough; a complete list is not required, and including fixed effects is not an error.

A4 (5 points)

The assumption (2.5 points): conditional on the covariates in the regression, attendance is independent of the potential outcomes, so no variable that affects both attendance and the test score is left out. **Why it is hard to defend (2.5 points):** the variables that determine who attends, such as motivation, ability, and parental engagement, are usually unobserved or poorly measured, and the assumption cannot be tested with the data. A statement in words or in notation both get full credit.

A5 (5 points)

What the fixed effects remove (2.5 points): everything about a pupil that does not change over the years, observed or unobserved, such as ability and family background. The effect is estimated from changes within pupils. **Two limitations (2.5 points, any two):** confounders that change over time remain; reverse causality is not solved, for example if a bad year leads to attendance; only pupils who attend in some years and not in others contribute, so precision is lower and the estimate applies to these pupils; variables that do not change within pupils cannot be studied.

Section B, Evaluating studies (45 points)

B1, The app lottery (15 points)

(i) 3 points. Pupils who used the app, or their families, chose to use it. Users and non-users therefore differ in motivation, resources, and need, and the difference in test scores mixes the effect of the app with these differences. Full credit requires both the self-selection and its consequence for the comparison.

(ii) 7 points. 1 point for the setup: winning is random and raises app use, so the effect of winning on test scores, scaled by the effect of winning on app use, gives the effect of app use. 4 points for the assumptions, each stated and assessed for this setting (1 point each). Relevance: winning raises use from 8% to 62%. Independence: the lottery is random among the families that applied, so winning is unrelated to the pupils' potential outcomes. Exclusion: winning affects test scores only through app use; this is the least certain assumption here, since families that win may also change their behavior in other ways. Monotonicity: winning makes no pupil less likely to use the app, which is plausible. An answer that lists five assumptions, including no interference between pupils, is also correct. 2 points for evidence: relevance is checked by the difference in use; independence by comparing winners and losers on characteristics measured before the lottery, and on participation in the test; exclusion and monotonicity cannot be checked with the data. A list of assumptions that is not related to the setting gets at most 4 of the 7 points.

(iii) 5 points. The compliers: pupils who used the app because their family won a license and would not have used it otherwise (2 points). The estimate does not cover always-takers, who use the app anyway, or never-takers, who do not use it even with a license, and the effect for them may differ (1.5 points). If the city buys licenses for all pupils, the app reaches mainly never-takers and pupils in families that did not apply, and the estimate need not apply to them (1.5 points). "The compliers are all pupils who used the app" gets at most 1 of the 2 points for the definition.

B2, The reading threshold (15 points)

(i) 5 points. A regression discontinuity design (1 point). Pupils just below the threshold, who took the course, score about 0.25 standard deviations higher than the line fitted on the other side predicts at the threshold; the sign is positive (1 point), and a size between 0.2 and 0.3 is accepted (1 point). The jump at the threshold carries the causal claim. The slope of the lines does not (2 points). Reading the slope as the effect gives 0 points on the last part.

(ii) 6 points. 2.5 points for each of two threats, discussed for this setting. The two strongest threats: manipulation of the score, since teachers score the test and know the threshold, and may

place pupils just below it to give them the course, or just above it; and something else changing at the same threshold, for example other support triggered by the same score. Also accepted: the fitted lines may have the wrong form, so the jump depends on the bandwidth and the functional form; few pupils are close to the threshold, so the estimate is imprecise. 1 point for a concrete check tied to a threat: the distribution of screening scores should not bunch on one side of the threshold; pupil characteristics measured before the screening should not jump at the threshold; the estimate should not change much with the bandwidth.

(iii) 4 points. The design estimates the effect for pupils close to the threshold (2 points). Offering the course to all pupils extends it to pupils far above the threshold, for whom the effect is unknown and may be different (2 points). An answer that only discusses precision, or says that more research is needed, gets at most 2 points.

B3, The extended school day (15 points)

(i) 5 points. Parallel trends, stated for this setting: without the extended school day, test scores in the treatment municipalities would have followed the same trend as test scores in the control municipalities. The two groups may differ in level. A correct but general statement that does not mention the municipalities and their test scores gets 3 points. “The two groups must be similar” gets at most 2 points.

(ii) 5 points. Before the introduction, the estimates are close to zero and the confidence intervals include zero: the two groups moved in parallel before the change. This supports the assumption but does not prove it, since the assumption concerns the years after the introduction (2.5 points). From the year of introduction, the difference is positive, grows over the first years to about 0.05 standard deviations, and then levels off. The confidence intervals exclude zero but widen in the later years (2.5 points).

(iii) 5 points. Differences in the level of test scores between the two groups are removed when scores are compared before and after the introduction. The assumption concerns the trend, so the municipalities’ choice is a problem only if it is related to how their scores would have developed anyway (2.5 points). An example (2.5 points): municipalities introduced the extended day because their results were falling, so their scores would have developed differently, or recovered, without it; or they introduced it together with other measures. “Richer municipalities introduced the extended day and score higher” describes a difference in level and gets 0 points on the example.

Section C, Design task (25 points)

There is no single correct design. Grade the reasoning: whether the design uses a feature of the plan, whether the comparison is specified, whether the assumptions are stated for this setting, whether the student says who the estimate applies to, and whether the threats are the ones that matter here. Two designs are defensible as the main proposal, and a combination of the two is also acceptable.

Regression discontinuity at the cutoff. Compare outcomes after fall 2027 in municipalities just above and just below the cutoff. Assumption: municipalities close to the cutoff are comparable, and nothing else changes at the cutoff. The cutoff is based on 2025 statistics fixed before the decision, so manipulation is not a concern. The estimate applies to municipalities near the cutoff. Threats: few municipalities are close to the cutoff, so the estimate is imprecise; other measures may be tied to the same cutoff; the estimate depends on the functional form.

Difference-in-differences. Compare the change in outcomes from before to after fall 2027 in municipalities at or above the cutoff with the change in municipalities below it. Assumption: without free school meals, outcomes in the two groups would have followed the same trend. The estimate applies to the municipalities that get free school meals. Threats: trends may differ between

municipalities with high and low shares of low-income families; other measures aimed at the same municipalities in the same years. The yearly data before 2027 allow a check of the trends before the start.

(i) 12 points: a design that uses a feature of the plan (3); the comparison that identifies the effect, clearly specified (3); assumptions stated for this setting (2); who the estimate applies to (2); the most important threats (2). A comparison of municipalities with and without free school meals in 2028 with no identification strategy gets at most 4 points. Naming a design without specifying the comparison gets little credit.

(ii) 7 points: a coherent alternative (3); its advantages and disadvantages compared with the main design (4). The expected trade-off: the regression discontinuity design is credible but local and imprecise; the difference-in-differences design covers all municipalities that get free school meals but rests on the parallel trends assumption. If the student proposes a combination in (i), grade (i) and (ii) together

(iii) 6 points: 3 points for each of two changes with a justification. Expected changes: introduce free school meals in the municipalities below the cutoff in random order over the following years; hold a lottery among municipalities close to the cutoff; apply the cutoff strictly, with no exceptions; avoid other measures aimed at the same municipalities in the same year; keep the survey unchanged; decide the outcomes and the analysis before the start. The justification must say how the change improves identification or precision.