Sensitivity Analysis

Despite the excellent agreement of the different mathematical methods of estimating ATE from the scores and the covariates there are reasons to worry about what all the models leave out, i.e. unmeasured or inadequately measured confounders. The covariates used are necessarily imperfect measurements of the relevant student traits, and imperfect measurement typically leads to under-adjustment bias. Such bias can arise simply because measurements of the covariates are imprecise or because the covariates don’t fully capture the relevant traits (Draper and Smith 1998). An informal quantitative sensitivity argument(Liu, Kuramoto, and Stuart 2013; VanderWeele, Tchetgen, and Halloran 2014) helps quantify an estimate of how much completing the potentially inadequate set of covariates would affect ATE. Although the factors used in estimating further adjustments will sound troublingly subjective, it helps to keep in mind that the range of plausible corrections will turn out to be smaller than the statistical error bars.

We have used covariates that give good indicators of students’ general strengths but none that are particularly sensitive to an important trait: how much effort they would be likely to put into this particular course. Furthermore, the correlation between the difference in ATEs between semesters and the differences in the adjustments to the raw scores looks suspicious. F17 is the only semester for which the overall model adjustment to the raw scores using both categorical and quantitative covariates is near zero. For the other semesters the adjustments are substantial and negative. We think the true ATE should be almost constant over time, although perhaps a bit lower in F17. Does this mean that a more complete adjustment with a more complete set of covariates would bring all the semesters close to F17, i.e. with the point estimate of the ATE around -1.5? Although not large, that would be worth noting as a negative effect.

Although it is fairly likely that the OL system did not work quite as well in F17 as in later semesters, it turns out that that differences in student traits account for most of the difference between the F17 ATE and those of the three other semesters. Table 6 shows the student strength covariates in each semester. In the Fall semesters the IP group was stronger on almost every pure predictor (not counting HW, an outcome) but in the Spring semesters the OL group was stronger. Although our adjusted ATE results are therefore not nearly as spread out as the raw score differences are, they remain anti-correlated with the adjustment for quantitative covariates, suggesting that the adjustment was under-estimated due to attenuation bias (Draper and Smith 1998). It is convenient that we have variation in the student strength balance between the semesters, since those differences will serve as an instrumental variable (Pokropek 2016) allowing more reliable causal inferences than could be made without it.

Before presenting several fairly complicated calculations of the sensitivity to inadequately measured confounders, it helps to have a very simple preliminary estimate of what a full correction would give if a fuller set of confounders were available. Once the categorical variables are controlled for via simple MLR, the ATE estimate is 1.17 points, corresponding to a Cohen d of ~0.12. There are a collection of student strength covariate measurements with Cohen d’s running from -0.01 to 0.11, once controlled for the categorical variables, as shown in Table 6. A reasonable guess would be that if there were no OL effect the OL group would score higher by d somewhere in the -0.01 to 0.11 range, i.e. -0.1 to 1.1 points. The actual score difference after controlling for categorical variables was a little higher than that, by 0.1 to 1.2 points, which should be attributed to an actual OL effect plus random statistical noise. Since the most important covariate, ACTmath, is also the most imbalanced after conditioning on the categorical variables the best initial guess would be a true effect toward the lower end of the 0.1 to 1.2 interval. That agrees fairly well with the range we shall estimate from more elaborate methods. This estimate does not, unfortunately, include any covariate sensitive to the specific commitment to this course, the confounder which we initially suspected based on related data (Joyce et al. 2014) would bias the results to favor IP.

Our next method starts with the model based on the measured covariates and then uses extrapolation based on potential Cohen d’s of unmeasured confounders to extend to a hypothetical maximally predictive model. In a simple model including only categorical variables as covariates, the ATEMLR is 1.17 with a remaining standard deviation of the errors (SDerr )of 9.53. Adding the three ACT variables reduced the ATE by 0.54 and reduced SDerr to 8.22, i.e. the ACTs “explained” an SDerr of (9.532 - 8.222)0.5 = 4.83. The ATE adjustment was -0.11 times the SE explained, with the factor 0.11 not coincidentally typical for the Cohen d’s of the explanatory variables, of which ACTmath was the most important. A corresponding calculation for the HSGPA group but including HSGPA in the added variables gave a factor of 0.095. Thus the adjustments can be described as the SDerr explained times a weighted average Cohen d' of the explanatory variables.

Only a portion of the SDerr could be attributable to unmeasured confounders. For a trait to be correlated with lecture mode it must exist at the time of registration. Unpredictable individual contingencies involved in taking a few tests with particular questions at particular times would not be among such traits. We can set a lower bound to such inexplicable contributions to the variance by looking at the variations between the three exam scores (each corrected for class average) available for each student for Sp19 and Fa19. This short-term random component would account for 11% of the between-student variance in the three-exam average and an estimated 33% of the variance of the single-exam values used for the first two semesters. Thus we should not expect there to be any set of student traits that exist at the time of registration that would predict scores with better than R2 =0.78 (100% minus the average of exam variation across the semesters, ie the average of 11% and 33%). This is an approximate estimate, but we are only using it to set approximate limits on a small correction term. The model including ACTs and HSGPA has R2= 0.37 leaving an unexplained R2 of 0.63 of which ~0.41 is potentially explicable by factors present at registration. That remaining explicable SDerr  for this model is 6.6 (square root of 0.41/0.63 multiplied by SDerr).

We can look at a case in which the remaining explicable SDerr of 6.6 is explained by variables as unbalanced as the collection of measured variables. (Although we have not used HW in the estimate due to its causal ambiguity, as discussed below, its d’ of 0.10 is similar to that of the other covariates, although some of that difference could be a treatment effect. ) This assumption would imply that our ATE estimate should be reduced by 6.6\*0.10= 0.66 giving ATE = 0.08. The key assumptions in reaching this were that all the unexplained variance that was not expected based on the between-exam variance is due to factors that exist at registration time and are as unbalanced as the factors measured, in the same direction. We think this is likely to be an over-correction.

A third estimate makes use of the differences between the adjustments for the different semesters to interpolate to a hypothetical balanced sample. The differences between the OL and IP students for the three ACT variables are given in Table 7 for each semester, after adjustment by MLR for the other categorical variables. Both Fall semesters have higher scores in the IP groups while both Spring semesters have higher scores in the OL groups. Thus both Fall semesters’ ATEs are adjusted up when ACTs are included while both Spring semesters’ ATEs are adjusted down. Weighting these ACT score differences by their MLR coefficients from Table 5 gives effective net ACT differences for the semesters, as shown in Table 7. The regression of the ATEs obtained for each semester vs. the ACT imbalance has a positive slope consistent with the starting expectation that the corrections used were somewhat attenuated versions of the true corrections. The intercept corresponding to a balanced sample gives ATE=0.36. The same procedure may be followed for the HSGPA subset, including HSGPA as a predictor in the overall model and remembering to subtract 0.21 from the balanced intercept to correct for the original ATE difference between this subset and the full set. The result is ATE=0.85. We think that these similar point estimates of the true ATE make reasonable allowance for the incompletely measured covariates.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Trait | Cohen d | d’ | d’ F17 | d’ Sp18 | d’ Sp19 | d’ F19 |
| ACT | 0.11 | 0.06 | -0.17 | 0.14 | 0.35 | -0.07 |
| ACTmath | 0.11 | 0.11 | -0.14 | 0.16 | 0.56 | -0.06 |
| ACTmajor | 0.18 | 0.07 | -0.06 | 0.26 | 0.18 | -0.02 |
| HSGPA | 0.10 | -0.01 | 0.02 | 0.10 | 0.14 | -0.33 |
| HW\* | 0.13\* | 0.10\* | -0.31\* | 0.31\* | 0.28\* | 0.15\* |

**Table 6.**  Cohen d’s are shown for variables including homework scores (HW). Here d’ describes the residual after controlling for the categorical variables in an MLR model. The asterisk for HW is a reminder that its d and d’ may be affected by the treatment rather than being purely a pre-existing condition.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| score | weight | F17 | Sp18 | Sp19 | Fa19 |
| ACT | 0.224 | -0.545 | 0.466 | 1.125 | -0.212 |
| ACTmath | 1.065 | -0.467 | 0.526 | 1.872 | -0.213 |
| ACTmajor | 0.518 | -0.142 | 0.634 | 0.442 | -0.049 |
| Weighted sum | NA | -0.69 | 0.99 | 2.47 | -0.30 |
| ATEMLR (HSGPA adjusted) | NA | -1.63 | 1.86 | 2.18 | 1.30 |

**Table 7.** The mean imbalance between the OL and IP students’ scores for the three ACT measures left after conditioning on the categorical variables is shown for each semester, along with their weighted sum and the semester’s ATEMLR.

The key assumption shared by these estimates is that the unmeasured confounders are about as unbalanced as the measured ones. With the exception of HSGPA, which was almost perfectly balanced in the overall sample after conditioning on categorical variables, all the differences in pre-treatment covariates had opposite signs in Fall semesters and Spring semesters. Unless the other traits follow a qualitatively different pattern we have set plausible limits on their effects.

The main unmeasured confounders with which we were concerned were not ones that would be partially picked up by our measures of broadly relevant student traits but rather ones that were more course-specific, perhaps just applying in the semester the course was taken. These were statistics-specific preparation, as opposed to general facility with math, and course-specific commitment, as opposed to general conscientiousness. Given the very close balance in self-reported prior statistics courses taken, statistics preparation would be an implausible major confounder. Students’ unmeasured commitment to the course seems the most obvious variable likely to differ between groups choosing different versions, to affect learning outcomes, and not to show up adequately in our more general covariates. Thus we looked for a way to see if there were signs of such unmeasured motivational confounding.

Homework scores (HW) are the one variable we have that would be strongly sensitive to effort. Course-specific commitment should especially affect how much effort students put into doing the homework. In fact HW was less predictable than the objective exam score from the covariates to which we had access (R2 = 0.11 for predicting HW in the HSGPA group, in contrast to 0.37 for exam scores). That indicates that HW reflects causes not much picked up by our standard covariates. In contrast to exam scores, HW was much better predicted by HSGPA than by ACT scores, consistent with our intuition that HSGPA and HW might show relatively large effort-dependent contributions. Unlike HSGPA, however, HW is specific to this course.

The problem with including HW as a covariate is that it is not only a measure of prior student traits but also an outcome of the treatment. Fig. 1 shows a causal directed acyclic graph to help picture this issue. We leave out the other covariates because they are only weakly linked to HW. Conditioning on HW could bias estimates of ATE toward zero because it blocks the path OL🡪HW🡪ObjEx, because it is a collider between UC and OL, and because it can simply serve as a marker for intermediate stages of the treatment effects, not included in the diagram (Greenland 2003).

For example, one way that HW could appear on the path from OL to ObjEx would be that OL students could spend more time doing HW in conjunction with watching the lectures online rather than with more time separation. That could tend to raise the HW scores of OL students even if their commitment to the course were the same as IP students. Whether that difference in style of linking lectures and homework would tend to increase or decrease the retained knowledge measured by ObjEx is hard to guess. Thus using HW as a covariate is useful only as a way to pick up signs of possible large imbalances in commitment, since it can create some bias in the ATE estimate.

Chart

Description automatically generated

**Fig. 1.** A directed acyclic graph picturing the role of HW and unmeasured confounders (UC) in changing the association between the treatment (OL/IP) and the outcome (ObjEx). The figure is drawn using the online Dagitty tool.

As we see in Table 6, the Cohen d for HW, after conditioning on the categorical variables, is 0.10, so that we would not expect that including it as a covariate would strongly affect the overall ATE estimate. That is supported by the more detailed analysis below.

When HW is included as a covariate R2 = 0.47 in the whole sample leaving SDerr = 7.41. The same R2 was found in a separate analyses of the HSGPA subsample. In the HSGPA subsample inclusion of HSGPA increased ATE by 0.12 with HW included, almost identical to the 0.11 without HW in the model, so we will continue to add 0.11 to estimates for the whole sample obtained via models without HSGPA.

In the whole sample, inclusion of HW reduced (HSGPA-adjusted) ATEMLR from 0.74 to 0.44 ±1.0 (statistical). The statistical uncertainty is slightly reduced due to the better model fit. The HW adjustment to the ATE estimate of -0.30 is consistent either with losing some of the ATE by conditioning on an outcome or with properly adjusting for a confounder, or combinations of those effects . Thus the true correction is probably in the interval (-0.30,0), although it could in principle be outside that interval. Such ambiguity arises from including a variable downstream from the treatment.

We now take a closer look at HW effects for the individual semesters. For the four semesters including HW in the model for the full group gives the mean between our two limits for the estimated ATEs of (-0.95±0.98, 1.02±1.05, 1.18±1.07, 0.68±0.86), each adjusted +0.11 for HSGPA, with the error bars given just being the statistical ones and although the systematic uncertainty of roughly ±0.15 from the inclusion of HW is still present. The semester results when HW is added are more tightly clustered than the values obtained without HW, indicating that HW may pick up an important confounder that accounts for much of the between-semester ATE variation.

We also have anonymous survey data for three semesters on one relevant affective trait. Students were asked near the start of the course whether they “like math”, with a 3-level response choice: positive, neutral, and negative. On average, the in-person students responded more positively, but to a statistically non-significant degree. In F17 the difference was statistically significant and in Sp19 it non-significantly had the opposite sign. It is probably no coincidence that the semester in which the OL students had worse HW scores than the IP students is also the semester in which they said they liked math less. It is the semester before the reputation of the OL version was established. Given the correspondence of the survey response differences to the different ATEs in this semesters, it might have been useful to have an anonymized version of this response linked to the student records to include in the model, but the surveys are always strictly unlinked to any identifiers. Most importantly for our current argument, there was no indication that the online students initially liked math more overall.

One cannot rigorously exclude the possibility of some other important confounder, but if one existed it would have to meet some stringent requirements. Readers can judge for themselves the plausibility of finding important covariates that meet three conditions:

* + 1. Exist at registration time.
    2. Are orthogonal to the variety of covariates already in the model, including HW.
    3. Are substantially more imbalanced than the covariates we have.

The different dropout rates between the online and in-person groups can also create a bias in estimating ATE. If the extra 4.7% of those who would have finished in person dropped out from online were concentrated near the pass-fail border with exam scores ~25 points lower than the mean, then that would shift the raw score advantage for the online group up by 3.9% of 25, i.e. 1.2 points. We think that is a seriously high estimate because many students who drop out have personal reasons unrelated to their expected grade, e.g. illness or having taken on an overly ambitious schedule. If we had a perfect model, with R=1, the resulting ATE would be unaffected by dropouts since the predictive coefficients on the remainder would be unchanged. We have an imperfect model with R2=~0.4, so in this extreme case the ATE would be biased by ~0.7 points in favor of the online treatment. If, however, a more predictive model using better-measured covariates were to lower ATE that would reduce any such selection bias. The plausible bias due to dropouts is nearly the same as the plausible bias due to missing confounders, which we estimated above. Since inclusion of missing confounders would remove potential dropout selection bias these two potential biases cannot be combined. We note also that most of the gap in dropout rate occurred in F17, so that the semesters in which the estimated ATE was positive had smaller potential selection bias, under 0.27 points. In an extrapolated model with R2 =0.78, i.e. the sort for which we found the lowest ATE estimate, the further downward correction for dropout selection bias in these semesters would be less than 0.11.

As discussed, there could also be a small additional adjustment, probably downward, for motivational variables linked to HW. The approximate overall balance in the outcome variable HW indicates that course-specific motivational differences are unlikely to be much different from the other imbalances found. Our sensitivity analysis, however, does nothing to remove the statistical error bars but only indicates that large systematic adjustments are unlikely to be needed.

To summarize the sensitivity analysis, barring some unlikely coincidence there will be unmeasured confounding affecting our ATE. Given that including the available covariates slightly reduced the ATE estimate, it is reasonable to guess that a more complete set would reduce it a little more. Plausible estimates of such effects are constrained to be small. Since the upwardly adjusted Fall ATEs and the downwardly adjusted Spring ATEs all lie within a narrow interval, the additional systematic uncertainty should be small even if the adjustment magnitudes are underestimated. One fragment of evidence on a suspected confounder (“like math”) even points toward increasing the true ATE. Dropout selection bias could conceivably be about as big as unmeasured confounding, but eliminating unmeasured confounding would almost eliminate selection bias, so these two potential corrections cannot both be added.

Draper, N. R., and Smith, H. (1998), "Bias in Regression Estimates," in Applied Regression Analysis, New York: Wiley, pp. 235-240.

Greenland, S. (2003), "Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias," *Epidemiology*, 14 (3), 300-306.

Joyce, T. J., Crockett, S., Jaeger, D. A., Altindag, O., and O'Connell, S. D. (2014), "Does Classroom Time Matter? A Randomized Field Experiment of Hybrid and Traditional Lecture Formats in Economics," National Bureau of Economic Research.

Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013), "An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research," *Prevention science : the official journal of the Society for Prevention Research*, 14 (6), 570-580. DOI: 10.1007/s11121-012-0339-5.

Pokropek, A. (2016), "Introduction to instrumental variables and their application to large-scale assessment data," *Large-scale Assessments in Education*, 4 (1), 4. DOI: 10.1186/s40536-016-0018-2.

VanderWeele, T. J., Tchetgen, E. J. T., and Halloran, M. E. (2014), "Interference and Sensitivity Analysis," *Statistical Science*, 29 (4), 687–706. DOI: 10.1214/14-STS479.