On Measuring and Reducing Selection Bias with a Quasi-Doubly Randomized Preference Trial

Ted Joyce*
Baruch College, CUNY and NBER

Dahlia Remler
Baruch College, CUNY and NBER

David A. Jaeger
CUNY Graduate Center, University of Cologne, IZA, and NBER

Sean Crockett
Baruch College, CUNY

Onur Altindag
Havard University

Stephen D. O'Connell
Massachusetts Institute of Technology

November, 2016

Acknowledgements: This research was supported, in part, by a CUNY Collaborative Incentive Research Grant 20 to Ted Joyce and David Jaeger. We thank John Choonoo and Paul Bachler of Baruch's Institutional Research and Program Assessment for help with data. We received helpful comments from seminar participants at Baruch College, Universidad Carlos III, Universitat Pompeu Fabra, the University of Michigan, the University of Reading, the CUNY Higher Education Policy Series, Eric Chiang, our discussant at the American Economic Association's Conference on Teaching and Research in Economic Education in Minneapolis, MN, May 2015 and anonymous referees. There are no conflicts of interest.

***Corresponding Author:** Baruch College, Department of Economics and Finance, 55 Lexington Avenue, New York, NY 10010. Email: Theodore.Joyce@baruch.cuny.edu

Randomized experiments are widely seen as the most convincing research design to determine how well policies work. In most settings, however, people choose whether to participate in a given program, often choosing the programs that work best for them. Therefore, the average treatment effect estimated by randomized experiments may be less relevant to situations with choice (Heckman et al. 1998, Heckman and Smith 1995). Research designs that both incorporate choice and produce unbiased program effects could produce more relevant estimates.

We demonstrate how to leverage a randomized experiment by running a subsequent study that is identical in every way, except that subjects choose their treatment. Our approach is modeled after the doubly randomized preference trial (DRPT), in which subjects are first randomly assigned to an experimental arm or a choice arm. Those in the experimental arm are then randomized again between treatments, while those in the choice arm select between the same treatments (Janevic et al., 2003; Long, Little and Lin, 2008; Little, Long and Lin, 2008; Shadish, Clark and Steiner 2008). Ideally, subjects' preferences over the experimental treatments are solicited prior to the first randomization.

DRPTs have two fundamental advantages. First, a DRPT is a rigorous "within-study" design in which the difference between non-experimental and experimental treatment effect estimates reveals the magnitude of selection bias (LaLonde 1986). A DRPT potentially identifies those control variables that can reduce or even eliminate this bias, if as many control variables as possible are measured (Cook, Shadish and Wong 2008; Shadish, Clark and Steiner 2008; Steiner et al 2010). Second, DRPTs can determine how unbiased estimates of treatment effects vary by preference for the treatments in the study (Long, Little and Lin, 2008; Little, Long and Lin, 2008; Wing and Clark, forthcoming) and are altered by subjects being allowed to choose a treatment (Marcus et al 2012). DRPTs are especially useful when preferences for particular treatments are

strong and when outcomes or treatment effects vary considerably by those preferences. Such situations likely include weight loss strategies, smoking cessation programs, substance abuse treatments, job training programs, or teen pregnancy prevention, because active participation and compliance are critical to treatment effectiveness.

True DRPTs are challenging. They require recruiting more subjects than a randomized experiment, with double the number required to power both experimental and choice arms and more than double to power the comparison between designs. In many contexts, it may be infeasible to recruit and randomize so many subjects at one time. Relative to a traditional randomized experiment, the logistical burden of conducting a DRPT is greatly increased.

To address these logistical and sample size issues, we develop the idea of a *quasi-DRPT*, consisting of an experimental arm and a choice arm with the same treatments and outcome measures, but offered before or after the experiment. Because the subjects in a quasi-DRPT come from the same setting in both arms, it is plausible that they are allocated "as if" they were randomly assigned between the experimental and choice arms (Cook, Shadish and Wong 2008). In both a true DRPT and quasi-DRPT subjects are randomized between treatments in the experimental arm of a study. By conducting the choice and experimental arms at two different points in time, a quasi-DRPT reduces the recruitment and logistical burden compared to a true DRPT. In addition, subjects in the choice arm of the quasi-DRPT may not need to provide explicit consent, because they are not randomly assigned but simply being observed, while subjects in the choice arm of a true DRPT must provide explicit consent to random assignment. We also discuss design features that increase the value of quasi-DRPTs, such as a pre-treatment survey that solicits subjects' preferences for the treatments and identical measures of control variables across designs.

To illustrate how a quasi-DRPT can be effectively accomplished in a real-world setting, we examine the effect of class time on student performance (including test scores and grades) in an undergraduate introductory microeconomics course at a large public university. The treatment is a compressed format that meets only once a week compared to a traditional lecture that meets twice a week. Students were randomly assigned to one of these two formats in Fall 2013 (Joyce, *et al*., 2015) and in the subsequent year, Fall 2014, we allowed students to choose their format but otherwise kept all other conditions identical (See Figure 1).[1] The classes in the choice arm of our quasi-DRPT were offered on the same days, at the same times, in the same classrooms, with the same instructors, and using the same materials as in the experimental arm. The exams differed in the specific questions that were asked, but care was taken to ensure their close similarity in tested material and difficulty. We demonstrate that students' measured characteristics in the two design/years are well-balanced. The allocation to design by year effectively matches intact equivalent groups, increasing the likelihood that subjects in our choice arm are "as if" randomly assigned (Cook, Shadish and Wong 2008).

Our primary analysis is a within-study comparison. Surprisingly, we find little evidence of selection bias in the choice arm. While in both designs we find moderate effects of class time on student performance, there was little difference in the magnitude of the estimates across the choice and experimental arms. Individual characteristics that determine selection, such as having a self-reported learning style suited to online versus in-person instruction, are almost all otherwise unrelated to performance outcomes. Variables that are highly predictive of performance outcomes, such as Math SAT scores and GPA, are almost uniformly unrelated to selection of format. The

[1] Figlio, Rush and Yin (2013) and Alpert, Couch and Harmon (2016) conducted similar randomized experiments of online formats with introductory economics at the undergraduate level.

only variable that appears to be related to both format choice and outcomes is a preference for quantitative courses. Not controlling for format selection reduces the magnitude of the effect of being in the compressed format by .03 standard deviations, a statistically insignificant difference of little practical consequence.

Given the time, effort, and resources required to implement randomized experiments, we demonstrate how their value can be substantially enhanced and show that it is possible to determine the extent of treatment selection bias in a particular context. We also show how it may be possible to reduce or eliminate such selection bias, offering support for future studies with similar settings and treatments to be accomplished without randomized experiments. If many such quasi-DRPTs are conducted, over time it may be possible to generalize about the extent and determinants of selection bias in many settings.

Literature

Designs that Reproduce Randomized Experiment Estimates

Cook, Shadish, and Wong (2008, henceforth CSW) review the within-study literature.[2] They describe three cases under which non-experimental studies reproduce the unbiased estimates of treatment effects from randomized experiments. Two cases are relevant to our quasi-DRPT. CSW's Case II to reproduce randomized experimental estimates is matching "intact comparison groups [which] are purposively selected to maximize overlap with the treatment group on at least pretest values of the outcome measure" (CSW, p. 732). The empirical success of this approach is

[2] The term "within-study" is shorthand for analyses that contrast experimental and non-experimental estimates. The most common approaches use the experiment treatment or control group and compare treatment effects using an untreated comparison group from observational data as the counterfactual. The seminal study in economics is Lalonde (1986).

surprising, because, as CSW note, such matching does not logically rule out differences in unobservables related to the outcome. In effect, by choosing the same or similar geographical settings, time periods, and other institutional features, the groups are also matched on unobservables that determine outcomes.

CSW's Case IIII comprises studies where treatment and comparison groups differ manifestly but "the selection process into treatment is known" (p. 737). If the problem with observational studies in which subjects choose their treatment is omitted variables bias, then the solution is to ensure that no needed controls are omitted. The keys are that the selection process into treatment is fully known and that all determinants of selection have valid and reliable measures. That is, of course, easier said than done.

One of CSW's illustrative studies where the selection process is known is Shadish, Clark, and Steiner's (2008, henceforth SCS) DRPT, in which undergraduate students choose or are randomized into short term math or vocabulary coaching and the outcomes are tests in those fields. In SCS's DRPT, the selection process into math or vocabulary coaching cannot be known for sure, but they strive to measure variables of all kinds, including pre-test outcomes, and preference for and fear of both fields that determine selection. They demonstrate that adjusting estimates of the effect of coaching on test outcomes in the experimental arm by the extensive set of covariates does not alter the unadjusted estimates, suggesting that randomization within the experimental arm provides unbiased treatment effects. In a follow-up study, Steiner *et al.* (2010) show that dislike of math was the key control variable needed to eliminate selection bias in the choice arm, but that a rich other set of correlated covariates, beyond demographic and administrative variables, could also remove omitted variables bias, provided there are sufficient indicators to remove noise.

Our study is inspired by SCS's DRPT. We also aim to address some of the limitations of their study noted by Hill (2008): its "short duration," "laboratory setting," "easy to recruit participants," and (implicitly) zero stakes (p. 1347). Because quasi-DRPTs are easier to implement in many settings, they may enable longer term and higher stakes studies relative to true DRPTs.

Preference Effects: How Unbiased Treatment Effects Vary by Choice

DRPTs, however, have uses beyond rigorously demonstrating the extent of selection bias and the variables needed to eliminate it in observational studies. As Little, Long, and Lin (2008) note in their comment on SCS, the original idea of a DRPT, as emphasized in its name, was to examine how preferences and choice alter estimated treatment effects in a randomized experiment. They use SCS's data "to actually compare the effectiveness of treatments in subgroups with different treatment preferences" (p. 1344). In real world circumstances in which people have a choice over which treatment they receive, the relevant treatment effects are among those choosing each treatment.

In our case, students whose learning style is more suited to in-person interaction may perform worse in the compressed format, leading to heterogeneity in the effect of format on outcomes. If students are aware of what works best for them and choose accordingly, then the average treatment effect estimated from a randomized experiment maybe less relevant to the real world of choice (Heckman et al. 1998, Heckman and Smith 1995). More useful estimates would be the unbiased causal effect of the compressed format *among* those who would choose the compressed format and the unbiased effect of the compressed format *among* those who would choose the traditional format. Little, Long and Lin's preference effect is the double difference of

the causal effect of treatment relative to control between those who would choose treatment and those who would choose control.

Marcus et al (2012) also use SCS's DRPT to estimate how outcomes differ between those randomized to a particular treatment and those choosing the same treatment, conditional on all other variables that affect outcomes. Their goal is to examine the pure psychological effects of choosing a treatment, compared to being randomized to that same treatment. The preference effect estimated by Marcus et al. (2012), however, depends crucially on strong ignorability and having removed all selection bias. Thus, further uses of DRPTs—and quasi-DRPTs—hinge on their quality as a within-study comparison.

DRPTs can also be analyzed with an instrumental variable (IV) framework (Wing and Clark, *forthcoming*). In a DRPT, three groups are randomly assigned: those in the treated group of the experimental arm, those in the control group of the experimental arm and those in choice group. The experimental arm provides the familiar estimate of the average treatment effect if both groups full comply with their assignment. Wing and Clark's insight is that the randomization indicator (between the choice and experimental arms) can be used as an instrument to obtain estimates of the average treatment effect on the treated (ATT) or the average treatment effect of the untreated (ATU). For example, contrast the outcomes of those in the choice arm who select the treatment option with those in the control group of the experimental arm and use the randomization indicator to instrument for those in the choice arm who select treatment. The IV estimand in this case is the ATT as long as those in the experimental arm's control group have no access to treatment (see Bloom 1984 and Angrist and Pischke 2009). If some members of the control group gain access to treatment, the IV estimand can be interpreted as a local average treatment effect (Angrist, Imbens and Rubin 1996). In the quasi-DRPT, subjects are not randomly

assigned to the choice group so an IV framework can only be applied if allocation between experimental and choice arms is "as if" random.

Methods

Our quasi-DRPT design is illustrated in Figure 1, adapted from SCS's Figure 1. Rather than randomizing students to experimental or choice arms, as in a true DRPT, cohorts of students were allocated based on whether they enrolled in the course in Fall 2013 or Fall 2014. Each cohort consists of students in Introductory Microeconomics (excluding honors and evening sections) at the same college, with the same instructors, in the same classrooms, and using the same materials. In the experimental arm of the study, we randomized 725 students in the fall of 2013 into a traditional lecture format that met twice a week for 150 minutes or a compressed format that met once a week for 75 minutes. In the fall of 2014, we offered the exact same course taught by the same professors at the same times and in the same classrooms. Instead of randomizing students between formats, however, 769 students enrolled in the format of their choice.[3]

In both 2013 and 2014 all sections used N. Gregory Mankiw's *Principles of Microeconomics* (6th Edition) as the textbook, along with Cengage Learning's Aplia web application to administer and grade online quizzes. Each week students in both the traditional and compressed format took a "pre-lecture quiz" due on Sundays covering material to be taught in the upcoming week and a "post-lecture quiz" due on Saturdays covering material that had been taught during the week. The pre-lecture quizzes were pass/fail (students who correctly answered at least

[3] Eighty-nine students in the choice arm of the study were not enrolled in their preferred format based on the pre-course survey because of scheduling or lack of availability. We report results dropping those students below. In both years, students were kept from selecting their course and section based on the professor, as both professors were listed as the course instructors on all sections at the time of registration.

half of the questions received full credit for the quiz) and were generally easier than the post-lecture quizzes which were graded as a percentage of 100. The midterm and final accounted for 35 and 45 percent of their course grade, respectively. Grades on the Aplia quizzes accounted for the 20 percent. Thus, the only difference in the classes between the compressed and traditional formats was 75 minutes of class time per week.

Table 1 shows the comparability of students in the two arms on all observables. We have no pre-test outcome measures, since most students had never taken economics. GPA and SAT math scores are highly predictive of performance in economics (Joyce, *et al*. 2015), however, and students are well matched across arm/year. While differences in SAT verbal scores between the two arms are statistically significant, the normalized difference suggests that balance is obtained (Imbens and Rubin 2015).

In principle, students in the two years could differ on unobservables related to outcomes, but the institutional setting makes that unlikely. Students had no non-honors or daytime alternatives to the sections used in the study during both years.[4] The course is required for applying to the college's business school from which approximately 80 percent of students graduate. For many students, postponing the course will delay their graduation. Consequently, students in the choice arm represent an "intact comparison group" that facilities overlap with subjects in the experimental arm providing justification for the assumption that those in the choice and experimental arms are "as if" randomly assigned. Thus, our quasi-DRPT meets almost all of CSW's criteria for a valid within-study design. The one exception is that midterm and final exams differed across the years, as described below.

[4] The one exception is that a small section-class of 40 students was opened just prior to the Fall 2013 semester to accommodate students with scheduling difficulties.

Both true DRPTs and quasi-DRPTs are designed for cases where each treatment is desired by some subset of subjects. If the number of available places for each of the treatments is constrained (e.g. because of the number of available seats in a classroom), then it is advantageous if the preference for each treatment is roughly equal in the subject pool. In our case the study's design would not have been possible if one format had proven vastly more popular and filled up, shutting off choice for many students. In addition, students could have been prevented from enrolling in their choice of format if it was not available at the day or time they needed to take the course, due to work obligations or scheduling constraints. In fact, 89 students in the choice arm (about 14%) were not enrolled in their preferred format, roughly equal between formats.[5]

Our choice arm is an example of CSW's Case III in which we try to replicate the choice arm of a true DRPT (Shadish, Clark and Steiner's 2008). In our choice arm, like most observational studies, students self-select their treatment—compressed or traditional format—and that choice is potentially endogenous. If determinants of format choice are also related to outcomes, omitting those determinants from the analysis means selection into treatment lacks strong ignorability. As a result, estimates of the treatment effect will be biased. To create strong ignorability and have unbiased estimates requires that selection into treatment be fully known and that all determinants of selection be measured and used as controls.

We used as many methods as possible to theorize about all possible determinants of format choice, following Freedman's (1991) exhortation to use "shoe leather" to investigate causes of treatment choice. Specifically, based on Shadish, Clark and Steiner (2008), informal interviews

[5] Of the 648 students who completed pre-course survey in the Fall of 2014, 284 (43.8%) chose the compressed format, 275 (42.4 %) chose the traditional format, 47 (7.3%) enrolled in the compressed format but would have preferred the traditional format, and 42 (6.5%) enrolled in the traditional format but would have preferred the compressed format.

with students, and our experiences as professors, we hypothesized that the following constructs could conceivably determine format choice: learning style (online vs. in-person); past experience with hybrid or online courses (any; if any, how well it worked); perceived orientation towards quantitative vs. writing courses; conscientiousness; importance of the course to a student's major; risk aversion (in general and in this context); time and burden of commuting; and time on paid employment. We developed a survey instrument to measure these constructs in the choice arm shown in the Appendix. We also asked the students whether they were able to enroll in their preferred format and if not, the reason. Table 2 shows the survey questions in the Appendix that were indicators for each construct. Some, but not all, latent constructs were measured with multiple indicators.

Ideally, the confounders would be measured prospectively before students choose their format. Unfortunately, registration for most students began five months before the fall semester and continued up until the first week of class. We therefore chose to standardize the distribution of the survey by sending an email with a link to the online survey to all registered students the week before classes began and through the first week of classes. We sent multiple reminders including suggestions to do the survey, if not already done, by end of the first day of class.[6] The online survey included a mobile option. Ninety-six percent of students completed the pre-course survey.

A much shorter and somewhat different survey was fielded in the experimental arm. The constructs asked were: past experience with online courses and time on paid employment (the

[6] With IRB approval, students received three percentage points added to their final course average if they completed both the pre-course and post-course surveys and 1.5 percentage points if they did only one. The IRB required that students have an alternative path to the extra credit: a short additional assignment taking equal time.

randomized experiment survey instrument is shown in the Appendix). In both semesters we obtained data on student characteristics prior to enrollment from the college's Office of Institutional Research and Program Assessment. These variables include current GPA, transfer GPA, SAT math/verbal scores, cumulative credits, age and indicators for part-time, underclass, female, white, Asian, black/Hispanic/other, and native English speaker. We refer to these as administrative variables.

We measure student performance outcomes with the midterm exam score, the final exam score, combined midterm and final, online quiz scores, and overall course grade. The midterm and final exams consisted of 30 and 40 multiple choice questions, respectively. We were conscientious of the need to keep the content and difficulty of exams as similar as possible between the two years. Nevertheless, we unit-standardize scores on all tests within year. We also analyze withdrawal rates, counting as withdrawals students who enrolled in the class but failed to finish.

We first investigate selection into format in the choice arm by comparing student characteristics in the compressed and traditional formats and by estimating linear probability models to predict format choice. We then estimate regression models that recover the effect of the compressed format on performance, first controlling only for student characteristics and indicators for day of the week and for class size.[7] We add the hypothesized potential confounders that significantly drove format choice, and observe the difference these controls make for coefficient magnitudes. This reveals any selection bias arising from limiting the vector of controls to the student characteristics from the administrative data, rather than the richer set of hypothesized

[7] As discussed in Joyce, et al. (2015), one instructor always taught in a smaller classroom, while the other instructor taught in a larger classroom. For this reason, it is not possible to separate instructor and classroom size effects. Both the choice and experimental studies are balanced across format in both professor and classroom size.

potential confounders obtained from the survey. We do all our analysis both including and dropping those students who were not able to get their preferred format, even if only due to lack of availability at their preferred day and time.

We pool data from the two arms to provide a direct test of the difference in treatment effects between the two arms. We estimate the following regression:

$$P_{idf} = \alpha_0 + \alpha_1 C_{if} + \alpha_2 D_{id} + \alpha_3 (C_{if} \times D_{id}) + \alpha_4 MW_{idf} + \alpha_5 SmallClass_{idf} + \sum \beta_k X_{kidf} + \epsilon_{idf} \quad (1)$$

where P_{idf} is the academic performance of student i, in design d, and format f, C_{if} is an indicator for the student being in the compressed (treatment) format, D_{id} is one if the student is in the experimental arm and zero if she is in the choice arm, MW_{idf} is an indicator for the Monday-Wednesday sections to control for any systematic difference between students that choose different days of the week, $SmallClass_{idf}$ is a fixed effect for the small classroom, and X_{kidf} is the vector of student characteristics and preferences available in both designs.[8] The main effect of being in the compressed format in either year is captured by α_1 while α_2 measures the main effect of the research design. Our focus is α_3, which shows how the treatment effect of being in the compressed format relative to the traditional format differs between the experimental and choice arms, testing the effect of being able to choose a compressed or traditional lecture format on student performance. The estimate for the "compressed" lecture format relative to the "traditional" lecture format in 2013 is $\alpha_1 + \alpha_3$.

Our overall study was not originally designed as a quasi-DRPT, having recognized the potential for and value of a quasi-DRPT after the randomized experiment was conducted. Measurements of potential confounders and preferences were therefore unfortunately measured

[8] We do not have consistent survey measures for subjects in the randomized experiment, unfortunately, and only the student characteristics from the administrative data are available for the regressions that pool the two years.

inconsistently between designs and not prospectively to treatment choice or assignment. Ideally, all potential confounder and moderator variables are measured prior to assignment of treatment or prior to choice of treatment. The psychology and behavioral economics literatures show that being "endowed" with a particular treatment alters attitudes towards the treatment (Kahneman, Knetch and Thaler 1991). The same could be true for characteristics related to treatment choice. For example, students who are assigned to or who choose the compressed format could then rationalize that situation by minimizing the value of interaction with the professor or peers. As it turns out, our estimated treatment effects in the choice study are relatively insensitive to the inclusion of the survey variables that predicted choice and the pooled estimates remain informative in this case.

Although our choice study was originally intended only as a within-study comparison, we also estimate preference effects by estimating the effect of the compressed format *among* those who would choose the compressed format and the effect of the compressed format *among* those who would choose the traditional format.[9] Students who prefer the compressed format are defined as those who would choose it when allowed choice; such choice is only observed in the choice study. There are several possible approaches to estimate preference for format in the experimental arm. If the richest possible set of predictors of format choice were measured, they could be used to predict choice in the randomized experiment and because the preference proportions are correctly known in the choice year, there is further information to improve the estimate. Long, Little, and Lin (2008) use preferences in the choice arm to estimate preference effects in both arms using a maximum likelihood estimator.

[9] Above and beyond whether students get their preferred format, their performance could be affected purely by being randomized to format, rather than choosing for themselves. Such motivational and other psychological effects may be substantial, especially in medical settings, but it seems unlikely to be substantial in this educational setting. Marcus et al. (2012) discuss and estimate this distinct preference effect.

Another approach to estimating preference effects is to ask directly about preference in the experimental arm. With sufficient sample size, the experimental arm could be stratified by preference and heterogeneous treatment effects, estimated similarly to randomized experiments designed to estimate heterogeneous treatment effects. If the within-study estimates suggest that the covariates measured in the choice arm are sufficiently rich to eliminate selection bias, then preference effects can be estimated in the choice arm as well and the two arms can be pooled for added precision. We take this last approach. Unfortunately, preferences, like potential confounders, were not measured perfectly consistently between designs or prospectively to treatment choice or assignment. In the experimental arm, however, the pre-experiment survey did ask about preferences during the first week of class after students were randomly assigned to a section, with a somewhat different question than in the choice study (see the Fall 2013 survey instrument in the Appendix).[10]

To estimate preference effects, we create a set of dummy variables for the mutually exclusive categories of preference and assignment: *PrefTT*, *PrefTC*, *PrefCT*, *PrefCC*, with *T* indicating traditional and *C* indicating compressed. The first letter of the suffix indicates the actual format to which the student was assigned or chose and the second letter indicates the student's expressed preference. For example, *PrefTT* is one for students that were in the traditional format and who stated they preferred that format whereas *PrefTC* is one for students in the traditional

[10] The distribution of preferences measured in the experimental arm differed between formats, showing endowment effects (Kahneman, Knetch and Thaler 1991). Those randomly assigned to the compressed arm, 77% said they preferred the compressed format, while of those randomly assigned to the traditional format, 43% said they preferred the compressed format (while 57% preferred the traditional). Our preference measure in the randomized experiment is therefore biased towards actual assignment. While cognizant of the problems with our preference measures, we present estimates of preference effects, demonstrating further value of the quasi-DRPT.

format who would have preferred the compressed format. The combined experimental and choice data are then analyzed with the regression equation:

$$P_{it} = \alpha_0 + \alpha_1 PrefCC_{it} + \alpha_2 PrefCT_{it} + \alpha_3 PrefTC_{it} + \\ + \alpha_4 MW_{it} + \alpha_5 SmallClass_{it} + \sum \beta_k X_{kit} + \epsilon_{it} \quad (2)$$

Except for the format/preference indicators, the other variables are the same as in equation (1). The reference category among the format/preference indicators are those who were in and who preferred the traditional format ($PrefTT_{ij}$). Thus, α_2, for example, contrasts the academic performance of those in the compressed format but who preferred the traditional relative to those who were in and who preferred the traditional format.

Results

Compressed Format Selection

We first analyze the choice arm as if it were a stand-alone observational study, examining the hypothesized predictors of format choice and then using significant ones as controls to try to best eliminate omitted variables bias. Table 3 shows that students did not differ between compressed and traditional format in any of the administrative data variables, the usual kinds of controls available in similar studies. We cannot tell, however, if the lack of differences arises because students do not differ in any meaningful way or if the administrative data simply lacks the relevant variables to show the differences.

We next turn to student preferences, which measure other determinants of format choice. Table 4 shows how these variables differ between formats.[11] Although all variables are categorical, for brevity we present differences in means and show the *p*-values for a χ^2 test based on all categories.[12] Three of the four determinants of selection that are significant at the 5% level relate to self-perceived online versus in-person learning style: "My learning style is well-suited to hybrid format", "Traditional lectures work well for me" and "Interaction with professor and other students helps me learn." Among those who had taken a prior hybrid or online course, how well it worked was also associated with choosing the compressed format. One of the questions intended to get at the same construct, "Prefer electronic devices to read than paper" had a *p*-value of 0.16, close to significance at the 10% level. We conclude that self-perceived learning style is a clear predictor of format choice.

The other statistically significant predictor is whether students "prefer quantitative courses to writing courses," even though "writing courses are my strength" is not significant. Those who agree or strongly agree with having a preference for quantitative courses are 10 percentage points more likely to choose the compressed format and the chi-square test has a *p*-value <0.001. Hours of paid work during the semester follows an expected pattern but was not statistically significant (*p*=0.15). For example, those who did not work at all were 5 percentage points less likely to choose the compressed format.

[11] These results include the 89 students who did not get their preferred format, but we also examined all results dropping them and there were no important differences, only some small gains in statistical significance in Table 4.

[12] Most of the variables are 5-category Likert scales, with the exception of risk attitudes, which measures subjects' willingness to take risks on a scale of 0 to 10. This risk aversion question has been widely used and experimentally validated (Dohmen et al. 2011).

Columns (1) and (2) of Table 5 shows the format selection results from a linear probability model with predictors using all administrative variables and those survey, format determinant variables that differed significantly by format at the 5% level in the descriptive statistics. The standard errors in this and all subsequent regressions are adjusted for a general form of heteroscedasticity. The survey choice variables were collapsed into 3 categories: agree (combining strongly agree and agree); neutral (i.e., neither agree nor disagree); and disagree (combining strongly disagree and disagree), which is the omitted category in the regressions.[13] Echoing the descriptive statistics, the most significant predictors are learning style and preference for quantitative courses. Agreement with or neutrality towards the statement "my learning style is well-suited to a hybrid format" increases the probability of choosing the compressed format by 32 and 17 percentage points, respectively, relative to those who disagree. Agreeing or neither agreeing nor disagreeing with "I prefer quantitative courses" increases the probability of choosing compressed by about 12 and 16 percentage points, respectively. No other survey variables and no administrative variables are significant predictors of format choice.

Estimates and Sources of Selection Bias in a Choice Setting

To what extent does controlling for those few variables that do determine format choice affect estimates relative to estimates with only the administrative controls? In columns 3 through 6 of Table 5 we show that the estimated effect of compressed format on the combined midterm and final exam is –0.149 standard deviations with only administrative controls (columns 3 and 4) and -0.180 adding the significant survey controls (columns 5 and 6). The 0.031 increase in the

[13] Missing values were not dropped but treated as a separate response category.

magnitude of the estimate when adding the survey controls is not statistically significant nor of a meaningful magnitude.

We can also learn which determinants of format selection are related to outcomes and which lead to the apparent reduction in selection bias of the estimated treated effect. The only statistically significant survey control variable is preference for quantitative courses. Those who agree with preferring quantitative courses scored 0.30 standard deviations higher, relative to those who disagree. Our finding is therefore similar to those of SCS and Steiner et al. (2010) that disliking math was the essential factor in removing selection bias in their study.

Learning style does indeed affect format choice, but it is not related to performance. Meanwhile, those variables that determine performance, such as the SAT math score and GPA, seem to have no impact on format choice. We therefore find no detectible selection bias and, at most, very limited selection bias due only to preference for quantitative courses.

Comparing the Experimental and non-Experimental Estimates

If this were a purely observational study, we would not know the degree to which our estimates still suffer from selection bias due to unobserved confounders. A valid quasi-DRPT provides direct estimates of the selection bias in the choice setting with different sets of controls. Table 6 shows the results from the estimation of equation (1) for all four outcome measures: midterm, final, midterm and final combined, and course grade, estimated both without the set of administrative control variables $\boldsymbol{X}$ (odd-numbered columns) and with $\boldsymbol{X}$ (even-numbered columns). For all four outcomes, the effect of the compressed format does not differ between the two years/designs. The compressed effect estimate is lower in magnitude in the choice arm, by between 0.01 standard deviations for the final exam (column 4) and 0.08 standard deviations for

the midterm (column 2). In the middle is an estimated 0.05 standard deviation difference for the combined midterm and final (column 6), which we discussed in the previous section. The effect of classroom size (or professor) is quite strong and discussed in Joyce et al. (2015). The last column of Table 6 shows there are no differences in the effect of format on course withdrawal in either year.

Students in the compressed format of the choice arm (2014) scored 0.14 standard deviations less on the combined midterm and final than those in the traditional format, based on the results in Table 6 (column 6). This estimate is without the extensive set of format determinant covariates obtained from the survey that were only available in 2014. Controlling for those survey variables yields an estimate of the compressed of -0.18 standard deviations (Table 5, column 5). Although this estimate differs neither statistically or substantively from the compressed estimate for 2014 in Table 6 (-0.14), it is closer to the estimates obtained in the experimental arm (-0.19).[14] The quasi-DRPT has allowed us to estimate the selection bias in the choice arm and determine how much the survey data reduced that selection bias, relative to as estimate with only the administrative data.

The lack of statistically significant differences in the estimated treatment effects between the experimental and choice arms in Table 6 does not prove the two designs provide equivalent results. The strength of the evidence depends on the sample size and power. In our case, the minimal detectable effect size of the difference between the experiment and choice arms is 0.22

[14] Returning to Table 5, we dropped those who did not get their preference which reduced the estimated impact of being in the compressed format both with and without controls by 0.03 standard deviations. This is likely because those who do not get their preference were systematically different than those who did: an average of 0.3 standard deviations lower test scores, differences that are not fully captured by the controls variables. The amount of bias reduction from adding the survey controls is not affected by dropping those who don't get their choice, however.

standard deviations and thus we could not reject educationally meaningful differences. We would have needed a total sample of 1,572 students to detect an effect size of 0.20 standard deviations and 6,280 students to detect a difference of 0.10 standard deviations with 80 percent power and an alpha of 0.05.[15]

An alternative approach to demonstrating the statistical power needed to draw policy-relevant conclusions is to define a region of equivalency based on differences that would be meaningful to the particular setting. This avoids accepting the null of no differences and assumed equivalence because of small samples or tests of low power (Walker and Nowacki 2010). In our case we have defined the region of equivalency as differences between plus or minus 0.2 standard deviations based on the average contrast from a meta-analysis of online teaching formats by the U.S. Department of Education (2010). Accordingly, we reject the null of equivalence if the 90 percent confidence interval for the estimated difference between the traditional and the compressed formats by design is not completely contained within the region of equivalency. In our case, we can establish equivalency for the combined midterm and final of -0.05 because the 90 percent confidence interval (-0.198 to 0.098) in Table 6 lies just within the region of equivalency (-0.20 to 0.20).[16]

[15] As noted previously we were well-powered to test for differences in student performance within the experimental arm and choice arm. Testing differences in treatment effects between the two arms, a difference-in-differences (DD), requires much larger samples sizes because the standard error of the DD is essentially the square root of the sum of the variances of each arm. In addition, the minimal detectable effect size of the DD likely will be smaller than the minimal detectable effect size within each arm of a DRPT and quasi-DRPT given sampling error and modest bias (Steiner et al. 2010).

[16] A region of plus or minus 0.2 standard deviations has appeal because it represents approximately 2.5 percentage points on a scale of zero to 100 in our sample. This is just enough to change a student's grade from B- to B or B+ to A-. We thank an anonymous referee for suggesting a test of equivalency.

Preference Effects

We estimate preference effects with equation (2). We find that relative to the omitted category of "got traditional and wants traditional," the coefficient α_1, "got compressed and wanted compressed" is -0.17 standard deviations with a robust standard error of 0.055 ($p<.05$); α_2, "got compressed and wanted traditional," is -0.20 with a standard error of 0.10 ($p<.05$); and α_3, "got traditional and wanted compressed," is -0.04 with a standard error of 0.071 ($p<.55$). These results imply a compressed effect among those who prefer compressed (α_1-α_3) of -0.12 (with a standard error of 0.072) and a compressed effect among those who prefer traditional (α_2) of -0.20. Following Little, Long, and Lin (2008) we can estimate a preference effect, the difference in the compressed effect between those who prefer compressed and those who prefer traditional [(α_1-α_3) – α_2], of 0.074 (with a standard error of 0.115).

The effect of being in a compressed class was therefore less negative for those who preferred the compressed format than students in the compressed format who would have preferred to be in a traditional class, but the difference was not statistically significant. We do not rely heavily on these estimates, however, given the problems of measuring preference after assignment or choice and the lack of statistical power. They nonetheless show the potential value of a quasi-DRPT and the prospective measurement of preferences to uncover possibly meaningful heterogeneous treatment effects.

Conclusions

Our study illustrates the value and practicality of implementing a quasi-DRPT by following or preceding a randomized experiment with a study that is identical, except that the usual processes

of treatment selection, including self-selection, replaces random assignment. By creating overlap in the setting, subjects, and treatment between the experimental and choice arms, quasi-DRPTs can accomplish the same strong within-study comparisons that gauge selection bias and estimates of the variation in treatment effects by preference that are found in true DRPTs. While a quasi-DRPT is logistically easier than a true DRPT, both are feasible whenever some subjects would choose both treatments.

We also illustrate several potential difficulties and problems in conducting quasi- DRPTs, providing lessons in how to conduct them most effectively. The first lesson is to measure as many potential determinants of treatment choice as is feasible prospectively, prior to treatment choice or assignment (Rubin 2008a). If estimates of preference effects are desired, the second lesson is to measure subjects' preferred treatment prospectively and consistently in the choice and experimental arms. The third lesson is to ensure a sufficient number of subjects to power both the comparison of treatment effects between designs and the comparison of treatment effects between preferences.

We found little or no selection bias associated with subjects' choice of a compressed or traditional class format. One reason was that preference for format was not relevant to success in this class. The determinants of format selection, such as learning style, were unrelated to performance, while the variables related to performance, such as SAT math scores, were unrelated to format selection. A second reason was the substantial similarities between those choosing compressed and traditional formats. As noted, this was a large, required class that most students took within their first two years at the college in order to apply to the business school. Results might differ by format for elective classes in which preferences are more likely to affect course selection and where there may be less overlap between students by format.

Both true DRPT and quasi-DRPTs are likely to be most valuable when subjects have strong preferences for particular treatments and when outcomes and/or treatment effects vary considerably by those preferences. DRPTs were created for clinical settings with those features, when subjects are not blinded to their treatment and preferences are likely to affect the outcomes through compliance and motivation.

DRPTs could also enhance the value of large randomized field experiments. Consider the well-known Moving to Opportunity study of economic mobility (Kling, Liebman and Katz 2007). In this study, a fourth arm could have been created in which subjects would have had a choice of being in either the experimental group with housing vouchers only usable in a low-poverty census track and other support or the Section 8 group that would have received a conventional housing voucher. While this would have required a substantial increase in participant recruitment, it could have provided insights as to subjects' choices, information about controls needed to reduce selection bias, and the average treatment effect among the untreated (Wing and Clark, forthcoming). DRPTs also have value in other settings in which subjects are likely to have strong preferences for treatment, including substance abuse rehabilitation, job training, and birth control.

As an example, several states are considering teen pregnancy prevention programs modeled after Colorado's Family Planning Initiative (Lindo and Packham, forthcoming). Before launching a statewide program, a DRPT could be mounted at family planning clinics in which teens would receive either the contraceptive pill or Long-Acting Reversible Contraception (LARC). In a DRPT, after obtaining consent, half the teens recruited would be randomly allocated to the experimental arm and half to the choice arm. Those in the experimental arm would be randomly assigned again to either the pill or LARC and those in the choice arm would choose between the two. While a traditional randomized experiment would estimate the average effect of

LARC vs. the pill, a DRPT could estimate two important effects of interest. First, whether young women who receive their preferred method are more likely to avoid pregnancy because they stay with the method or use it more consistently than those receive their non-preferred method. Second, whether those who choose a given method are more likely to avoid pregnancy than those randomly assigned to the method. In this example, a quasi-DRPT would be potentially even more practical with teens in the choice arm enrolled either before or after the experimental arm was completed. Using the same clinics and collecting the same data as temporally close to the experimental arm as possible would increase overlap between the two arms, meeting CSW's suggestion for an intact comparison group. Prior to any assignment or choice, teens should be surveyed about preferences regarding contraception or their indifference based on past contraceptive use and other salient features. Data for the choice arm of a quasi-DRPT could be collected by survey for those who attend the same clinics and who choose the pill or LARC as part of their normal visit.

There has been tremendous growth in use of randomized designs for policy evaluation. This has prompted a major debate as to their primacy in determining policy (Deaton and Cartwright 2016; Imbens 2010). Quasi-DRPTs offer a practical means of improving the external validity of evaluations based on randomized designs.

References

Alpert, William T., Kenneth A. Couch, and Oskar R. Harmon (2016) "A Randomized Assessment of Online Learning," *American Economic Review* 106(5):378-82.

Angrist, Joshua and Jorn-Steffen Pischke (2009) *Most Harmless Econometrics* (Princeton,New Jersey: Princeton University Press).

Angrist, Joshua, Guido Imbens, and Donald Rubin. (1996) "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91(434):444-454.

Bloom, Howard S. (1984) "Estimating the Effect of Job-Training Programs, Using Longitudinal Data: Ashefnfelter's Findings Reconsidered," *Journal of Human Resources* 19:544-556.

Cook, Thomas D. (2003) "Why Have Educational Evaluators Chosen Not to Do Randomized Experiments?" *The Annals of the American Academy of Political and Social Science* 589(1):114-140.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong (2008) "Three Conditions under which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-study Comparisons," *Journal of Policy Analysis and Management* 27(4):724–750.

Deaton, Angus and Nancy Cartwright (2016) "Understanding and Misunderstanding Randomized Controlled Trials," National Bureau of Economic Research Working Paper 22595.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner (2011) "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association* 9(3):522–550

Figlio, David, Rush, Mark, and Yin, Lu (2013) "Is It Live or is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning," *Journal of Labor Economics* 31(4):763–784.

Freedman, David A. (1991) Statistical models and shoe leather. *Sociological Methodology* 21: 291-313.

Heckman, James J. and Jeffrey A. Smith (1995) "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9(2):85-110.

Heckman, James J., H. Ichimura, Jeffrey A. Smith, and P.E. Todd (1998) "Characterizing Selection Bias Using Experimental Data" *Econometrica* 66, 1017-1098.

Hill, Jennifer (2008) "Comment on: 'Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments'," *Journal of the American Statistical Association* 76(4):1346-1350.

Imbens, G. (2010) "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature* 48(2):399-423.

Janevic, Mary R., Nancy K. Janz, Julia A. Dodge, Xihong Lin, Wenqin Pan, Brandy R. Sinco, and Noreen M. Clark (2003) "The Role of Choice in Health Education Intervention Trials: A Review and Case Study," *Social Science and Medicine* 56(7):1581–1594.

Joyce, Theodore J., Sean Crockett, David A. Jaeger, Onur Altindag, and Stephen D. O'Connell (2015) "Does Classroom Time Matter?" *Economics of Education Review* 46:64–77.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler (1991) "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives* 5(1):193–206.

Kling, Jeffrey R., Jeffrey B. Leibman and Lawrence F. Katz (2007) "Experimental Analysis of Neighborhood Effects," *Econometrica* 75(1):83-120.

Lalonde, Robert J. (1986) "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review* 76(4):604–620.

Lindo, Jason M. and Analisa Packham (forthcoming) "How Much can Expanding Access to Long-Acting Reversible Contraceptives Reduce Teen Birth Rates?" *American Economic Journal: Economic Policy*.

Little, Roderick J., Qi Long, and Xihong Lin (2008) "Comment on: 'Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments'," *Journal of the American Statistical Association* 103(484):1344–1346.

Long, Qi, Roderick J. Little, and Xihong Lin (2008) "Causal Inference in Hybrid Intervention Trials Involving Treatment Choice," *Journal of the American Statistical Association* 103(482):474–484.

Marcus, Sue M., Elizabeth A. Stuart, Pei Wang, William R. Shadish, and Peter M. Steiner (2012) "Estimating the Causal Effect of Randomization versus Treatment Preference in a Doubly Randomized Preference Trial," *Psychological Methods* 17(2):244–254.

Rubin, Donald B. (2001) "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation," *Health Services and Outcomes Research Methodology* 2:169-188.

Rubin, Donald B. (2007) "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials," *Statistics in Medicine* 26(1):20-36.

Rubin, Donald B. (2008a) "Comment on: 'Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments'," *Journal of the American Statistical Association* 76(4):1350-1353.

Rubin, Donald B. (2008b) "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics* 2(3):808–840.

Shadish, William R., M. H. Clark, and Peter M. Steiner (2008) "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments," *Journal of the American Statistical Association* 103(484):1334–1350.

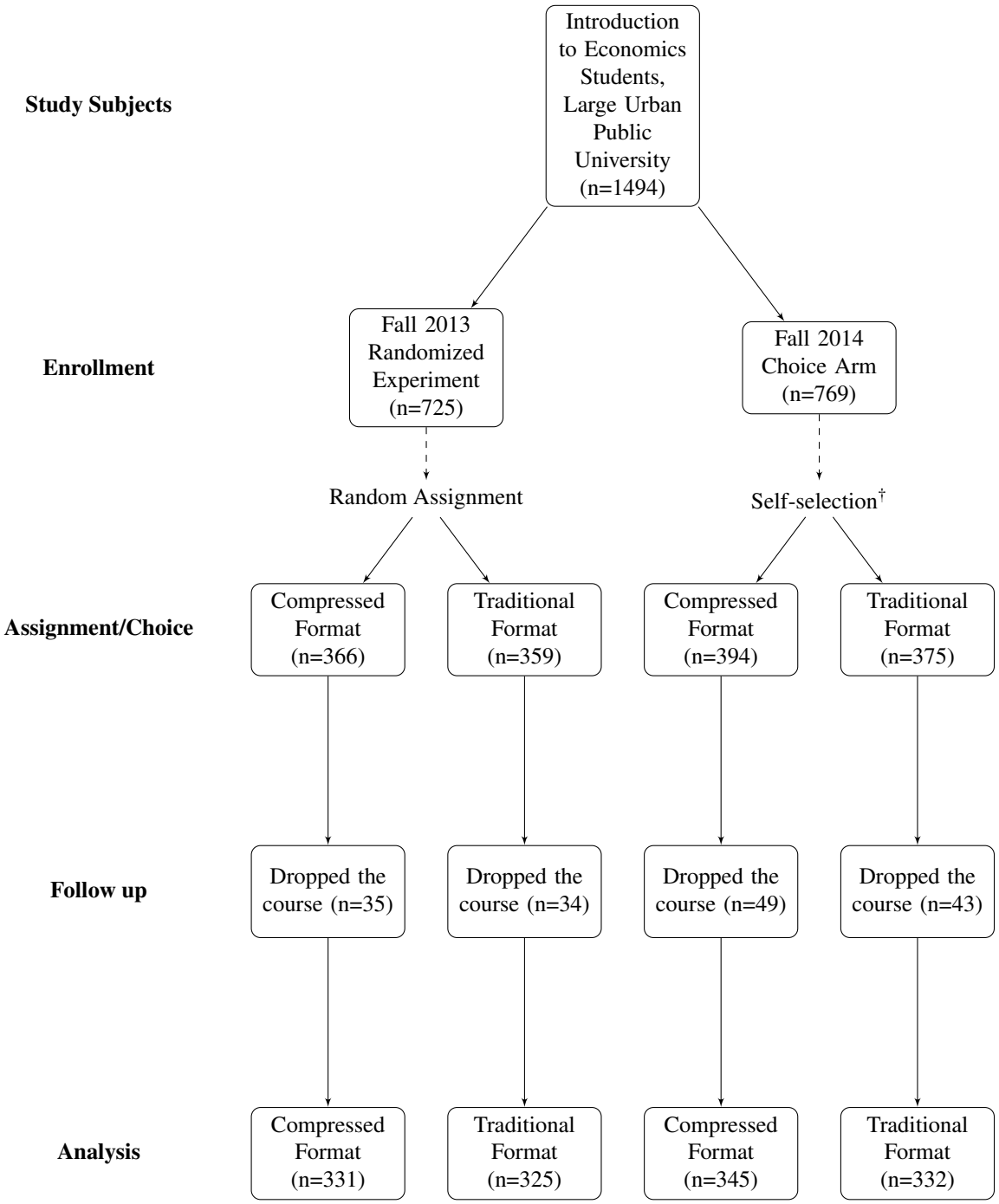
Steiner, Peter M., Thomas D. Cook, William R. Shadish and M. H. Clark (2010) "The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies," *Psychological Methods* 15(3): 250-267.

Department of Education, Office of Planning, Evaluation, and Policy Development (2010). Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. (Washington, D.C.: U.S Department of Labor). Available at https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf (last accessed November 7, 2016)

Walker, Esteban and Amy Nowacki. (2010) "Understanding Equivalence and Noninferiority Testing," *Journal of General Internal Medicine* 26(2): 192-6.

Wing, Coady and M. H. Clark (forthcoming) "What Can We Learn from a Doubly Randomized Preference Trial?" *Journal of Policy Analysis and Management*.

Figure 1. Quasi DRPT



Note: Adapted from Shadish, Clark and Steiner (2008).
†48 students who wanted the compressed instead took the traditional format due to lack of availability of their preferred timeslot, and 42 of them completed the course. 54 students who wanted the traditional format instead took the compresses for the same reasons, and 47 of them completed the course.

Table 1. Baseline Characteristics of Participants by Experimental (2013) and Choice (2014) Arms

	2013	2014	Diff.	Norm. Diff.	Log Ratio SD	*N*
Prior Academic Performance						
Baruch GPA	3.03	3.04	-0.01	-0.01	0.03	1102
Transfer GPA	3.31	3.30	0.00	0.01	-0.18	541
SAT Verbal	540.74	556.51	-15.77***	-0.13	0.19	918
SAT Math	604.03	608.73	-4.70	-0.04	0.03	1055
Prior Academic Experience						
Cumulative Credits	44.59	44.88	-0.29	-0.01	-0.01	1333
Part time	0.07	0.08	-0.01	-0.01	-0.03	1333
Underclass	0.76	0.75	0.01	0.02	-0.02	1333
Demographic Characteristics						
Age	20.96	21.08	-0.12	-0.02	-0.20	1333
Female	0.45	0.47	-0.02	-0.02	0.00	1333
White	0.27	0.29	-0.02	-0.03	-0.02	1163
Asian	0.45	0.46	-0.01	-0.02	0.00	1163
Black, Hispanic, Other	0.28	0.25	0.03	0.05	0.04	1163
Native English Speaker	0.53	0.55	-0.03	-0.04	0.00	991
Withdrawal rate	0.10	0.12	-0.02	-0.06	-0.10	1494

Note: This table reports the average background characteristics of students in randomized field experiment (RFE) in Fall 2013 and contrast them with students who enrolled to the same course in the Fall 2014. Sample includes students who completed the course. The column "Diff." shows the difference in means for the indicated variable. Statistical significance means between 2013 and 2014 tested using two sample t-tests assuming unequal variances. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$. The column "Norm. Diff" shows the normalized differences, and equals the difference in average covariate values, normalized by the standard deviation of these covariates, i.e. $(\bar{X}_{2013} - \bar{X}_{2014})/\sqrt{s^2_{X,2013} + s^2_{X,2014}}$. The column "Log Ratio SD" shows the logarithm of the ratio of standard deviations and measures of dispersion in the distributions of two covariates. The sample analog of this is calculated as the difference in the logarithms of the two sample standard deviations, i.e. $\log(s_{X,2013}) - \log(s_{X,2014})$. The column "*N*" shows the number of non-missing observations that are used in the comparison.

Table 2. Constructs of Survey Instrument

Construct of potential confounder or preference	Questions corresponding to construct indicator in choice study survey	Questions corresponding to construct indicator in randomuzed experiment survey
Learning Style (online vs. in person)	Q1.1, Q3.5, Q3.7, Q3.10, Q3.14	
Past experience with hybrid or online course (if any, how well it worked)	Q1.2	
Orientation towards quantitative vs. writing courses	Q3.1, Q3.9	
Conscientiousness	Q3.2, Q3.4, Q3.11	
Stakes of this course	Q.3.3	
Risk aversion (in general and in this context)	Q3.6, Q3.8, Q3.13, Q3.15	
Time and burden of commuting	Q3.12, Q3.16	
Time on paid employment	Q3.17	Q5
Lecture Format Preference	Q2.1, Q2.2, Q2.3	Q4

Note: Table 2 describes the indicators in the pre-course survey for each construct. Please see the Appendix for a copy of the pre-course surveys.

Table 3. Baseline Characteristics of Participants by Lecture Format in the Choice Arm (2014)

	Compressed	Traditional	Diff.	Norm. Diff.	Log Ratio SD	*N*
Prior Academic Performance						
Baruch GPA	3.05	3.03	0.02	0.03	0.06	584
Transfer GPA	3.29	3.32	-0.03	-0.05	0.03	311
SAT Verbal	555.20	557.75	-2.55	-0.02	0.03	407
SAT Math	604.92	612.37	-7.45	-0.06	0.00	544
Prior Academic Experience						
Cumulative Credits	44.76	44.99	-0.24	-0.01	0.09	677
Part time	0.07	0.09	-0.01	-0.04	-0.08	677
Underclass	0.73	0.76	-0.02	-0.04	0.03	677
Demographic Characteristics						
Age	21.02	21.14	-0.12	-0.02	0.08	677
Female	0.50	0.44	0.06	0.09	0.01	677
White	0.28	0.30	-0.02	-0.03	-0.02	617
Asian	0.46	0.47	-0.01	-0.01	0.00	617
Black, Hispanic, Other	0.26	0.23	0.03	0.05	0.04	617
Native English Speaker	0.59	0.53	0.06	0.08	-0.01	430
			p-value, joint χ^2-test = 0.615			

Note: This table reports the average background characteristics of students in "compressed" format (lectures once per week) and contrast them with students in "traditional" format (lectures twice per week) for Fall 2014. Sample includes students who completed the course. The column "Diff." shows the difference in means for the indicated variable using two sample t-tests assuming unequal variances. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$. The column "Norm. Diff" shows the normalized differences and equals the difference in average covariate values, normalized by the standard deviation of these covariates, i.e. $(\bar{X}_{\text{compressed}} - \bar{X}_{\text{compressed}})/\sqrt{s^2_{X,\text{compressed}} + s^2_{X,\text{traditional}}}$ The column "Log Ratio SD" shows the logarithm of the ratio of standard deviations and measures of dispersion in the distributions of two covariates. The sample analog of this is calculated as the difference in the logarithms of the standard deviations, i.e. $\log(s_{X,\text{compressed}}) - \log(s_{X,\text{traditional}})$. The column "*N*" shows the number of non-missing observations that are used in the comparison.

Table 4. Student Responses to the Pre-class Survey in the Choice Arm (2014)

	Traditional	Compressed	Scale	χ^2-test (p)	N
	(1)	(2)	(3)	(4)	(5)
My learning style is well-suited to a hybrid format	3.29	2.69	[1-5]	0.00***	648
No previous hybrid or fully online course before	0.34	0.40	[0-1]	0.09*	647
Hybrid worked well for me	2.93	2.58	[1-5]	0.06*	240
Writing focus courses are my strenght	2.94	2.92	[1-5]	0.79	647
I use every available course supplement	2.00	2.05	[1-5]	0.86	648
Economics is not very relevant to my major	3.88	3.94	[1-5]	0.49	646
I typically do not finish my classwork	4.15	4.14	[1-5]	0.50	648
Traditional lectures work well for me	2.30	2.51	[1-5]	0.02**	647
I avoid hard grader professors	2.48	2.47	[1-5]	0.30	648
I need structure to get my class work done	2.02	2.20	[1-5]	0.06*	647
Getting at least A- is a high priority for this class	1.50	1.50	[1-5]	0.19	647
Prefer quantitative courses to writing-focused	2.57	2.54	[1-5]	0.00***	645
Prefer electronic devices to read than paper	3.38	3.23	[1-5]	0.16	647
I am a disciplined person, no need deadlines	2.73	2.75	[1-5]	0.57	646
Commute to campus on weekdays is difficult	3.16	2.95	[1-5]	0.23	646
Economics is a challenging course	2.43	2.45	[1-5]	0.26	646
Interaction with professor and other students helps	1.81	2.12	[1-5]	0.00***	647
Risk preference	6.64	6.78	[0-10]	0.33	648
Commute time to school	2.42	2.52	[1-4]	0.30	647
Paid work during the semester	2.51	2.35	[1-4]	0.15	648

Note: This table reports the differences in pre-class survey responses between the students who chose the "compressed" format (lectures once per week) and the students who chose the"traditional" format (lectures twice per week) during the Fall 2014 semester. Figures in column (1) and (2) are the average score for each question. Column (3) reports the survey question scale. All [1-5] questions used a 5-point Likert scale from 1 to 5 with strongly agree 1, agree 2, neither agree or disagree 3, disagree 4 and strongly disagree 5. The possible answers to the first question is binary and equals one if the answer is "yes". Among the last three questions, the risk preference question has a continuous scale from 1 to 10 and increases in risk-seeking. The commute question has 4 categories: [1] "less than 30 minutes", [2] "between 30 minutes and 60 minutes", [3] "between 60 minutes and 90 minutes" , and [4] "more than 90 minutes". The last question has 4 categories: [1] "No paid work" , [2] "Working less than 15 hours per week", [3] "Working between 15-30 hours per week", and [4] "Working more than 30 hours per week". Column (4) show the p-values from the χ^2-test of independence among responses by format. Column 5 reports the number of non-missing observations for the indicated survey question. Significance levels are indicated by $* < .1$, ** $< .05$, *** $< .01$..

Table 5. Lecture Format Choice and Student Performance in the Choice Arm (2014)

Covariate	"Compressed" Coef.	Std. Err.	Midterm+Final Coef.	Std. Err.	Midterm+Final Coef.	Std. Err.
	(1)	(2)	(3)	(4)	(5)	(6)
Compressed			-0.149	0.070*	-0.180	0.073*
Baruch GPA	0.003	0.037	0.642	0.065***	0.654	0.065***
Transfer GPA	0.081	0.063	0.243	0.102*	0.230	0.102
SAT Verbal	-0.000	0.000	0.001	0.001	0.001	0.001
SAT Math	0.000	0.000	0.004	0.001***	0.003	0.001***
Cumulative Credits	0.001	0.002	-0.001	0.003	-0.001	0.003
Part time	0.024	0.076	-0.035	0.129	-0.029	0.133
Underclass	0.082	0.087	0.009	0.152	0.036	0.153
Age	0.001	0.007	0.003	0.012	0.001	0.012
Female	-0.072	0.039	-0.362	0.066***	-0.346	0.066***
Asian	0.007	0.049	-0.051	0.079	-0.061	0.080
Black, Hispanic, Other	0.013	0.055	-0.066	0.096	-0.050	0.096
Native English Speaker	-0.048	0.048	-0.088	0.082	-0.089	0.081
Mon.-Wed.			0.011	0.074	0.014	0.074
Small Classroom			0.389	0.071***	0.385	0.070***
Well-suited to hybrid						
Neutral	0.172	0.049***			0.098	0.083
Agreed	0.318	0.051***			0.042	0.090
Traditional lectures work						
Neutral	-0.057	0.067			0.113	0.111
Agreed	-0.082	0.063			0.045	0.108
Prefer quantitative						
Neutral	0.157	0.060**			0.054	0.089
Agreed	0.123	0.058*			0.298	0.092**
Interaction helps						
Neutral	0.017	0.094			0.156	0.133
Agreed	-0.163	0.084			0.016	0.109
N	677		676		676	
R^2	0.131		0.364		0.393	

Note: The dependent variable in column (1) is a dichotomous indicator that is 1 if the student chose the "compressed" format and 0 if she chose the "traditional format". The dependent variable in columns (3)-(6) is the score on the combined midterm and final standardized with mean zero and standard deviation of one. Estimated with OLS. The covariates are Baruch GPA, Transfer, GPA, Verbal SAT, Math SAT, Cumulative Credits, Age, indicator variables for Part-Time Student, Underclassman, Female, Asian, Black/Hispanic/Other, and Native Speaker plus indicator variables for missing Baruch GPA, Transfer GPA, SAT scores, Race, and Native English Speaker. *Mon.-Wed* is a dichotomous indicator of whether students attended classes on Mon and/or Wednesday versus Tuesday and/or Thursday. Small classroom is a dichotomous indicator of whether students attended class in the room that held 114 students versus the room that could accommodate 272 students. Heteroskedasticity-consistent standard errors are in the adjacent column with the significance levels, indicated by $* < .1$, ** $< .05$, *** $< .01$.

Table 6. Student Perfomance

	Midterm		Final		Midterm + Final		Course Grade		Withdraw	
Covariate	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Compressed (2013)	-0.21***	-0.20***	-0.18**	-0.14**	-0.22***	-0.19***	-0.22***	-0.20***	0.01	0.00
	(0.08)	(0.07)	(0.08)	(0.07)	(0.08)	(0.07)	(0.08)	(0.06)	(0.02)	(0.02)
Compressed (2014)	-0.15*	-0.12*	-0.14*	-0.13*	-0.16**	-0.14**	-0.16**	-0.14**	0.02	0.02
	(0.08)	(0.07)	(0.08)	(0.07)	(0.08)	(0.07)	(0.08)	(0.06)	(0.02)	(0.02)
Diff. (2013 – 2014)	-0.07	-0.08	-0.04	-0.01	-0.06	-0.05	-0.06	-0.07	-0.01	-0.01
	(0.11)	(0.09)	(0.11)	(0.09)	(0.11)	(0.09)	(0.11)	(0.08)	(0.03)	(0.03)
Mon.-Wed.	0.04	-0.05	0.07	-0.01	0.07	-0.03	0.05	-0.05	-0.01	-0.01
	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)	(0.02)	(0.02)
Small Classroom	0.25***	0.24***	0.31***	0.30***	0.32***	0.30***	0.30***	0.28***	-0.01	-0.01
	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)	(0.02)	(0.02)
Other Covariates		X		X		X		X		X
R^2	0.021	0.342	0.025	0.297	0.029	0.388	0.027	0.426	0.002	0.085
N	1332		1333		1332		1333		1492	

Note: This table reports the differences between student performance in "compressed" format (lectures once a week) and in "traditional" format (lectures twice a week) for the Fall 2013 and Fall 2014 semesters. Coefficients are from the estimation of equation (1) in the text which for convenience we show here. $P_{idf} = \alpha_0 + \alpha_1 C_{if} + \alpha_2 D_{id} + \alpha_3 (C_{id} \times D_{if}) + \sum \beta_k X_{ikdf} + e_{ifd}$. The estimate for the "compressed" lecture format relative to the "traditional" lecture format in 2013 is $\hat{\alpha}_1 + \hat{\alpha}_3$. All outcomes are based on a standardized normal scale with a mean of zero and a standard deviation of 1 within each semester. Estimated with OLS. Heteroskedasticity-consistent standard errors in parentheses. *Mon.-Wed* is a dichotomous indicator of whether students attended classes on Mon and/or Wednesday versus Tuesday and/or Thursday. Small classroom is a dichotomous indicator of whether students attended class in the room that held 114 students versus the room that could accommodate 272 students. Other covariates are Baruch GPA, Transfer, GPA, Verbal SAT, Math SAT, Cumulative Credits, Age, indicator variables for Part-Time Student, Underclassman, Female, Asian, Black/Hispanic/Other, and Native Speaker plus indicator variables for missing Baruch GPA, Transfer GPA, SAT scores, Race, and Native English Speaker. Course Grade includes curved midterm and final grades, penalties for missed classes, and the participation bonus. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.