Using instrumental variables properly to account for selection effects

Stephen R. Porter [*]

keywords: college students, instrumental variables, causal effects

April 2012

[*] Address: Department of Leadership, Policy, and Adult and Higher Education, Box 7801, North Carolina State University, Raleigh, NC 27695. Direct correspondence to srporter@ncsu.edu. I would like to thank Brad Curs, Nick Hillman, Ozan Jacquette, Tatiana Melguizo, Claire Porter, Paul Umbach, and Ryan Wells for comments.

Abstract

Selection bias is problematic when evaluating the effects of postsecondary interventions on college students, and can lead to biased estimates of program effects. While instrumental variables can be used to account for endogeneity due to self-selection, current practice requires that all five assumptions of instrumental variables be met in order to credibly estimate the causal effect of a program. Using the Pike et al. (2011) study of selection bias and learning communities as an example, the paper reviews these assumptions in the postsecondary context, and offers advice to researchers seeking to use instrumental variables for research on college students.

In a recent award-winning paper[1] published in *Research in Higher Education*, Gary Pike, Michele Hansen, and Ching-hui Lin (Pike et al., 2011, henceforth, PHL) argue that selection bias is problematic when evaluating the effects of first-year programs on college students. They propose instrumental variables (IV) as an alternative approach to multiple regression, to account for student self-selection into first-year programs. Specifically, they are interested in understanding the impact of participating in a learning community on academic performance, as measured by first-semester grade-point average. They argue for the use of participation in a summer bridge program prior to college entry, and declaration of an academic major at entry, as sources of exogenous variation that can be used to identify the causal effect of learning community participation on academic performance.

While I applaud their attention to issues of self-selection in postsecondary research, the paper itself is severely flawed and may mislead other scholars and assessment practitioners interested in identifying the effects of first-year programs on student outcomes. The purpose of this paper is to describe these flaws and, more generally, to illustrate how the underlying assumptions of IV designs should be evaluated for any postsecondary research endeavor.

Perhaps the biggest weakness of the paper is its incorrect statement that IV analyses rest on two assumptions (p. 198). This was indeed the standard view in the econometrics literature twenty years ago. But PHL wish to interpret their estimates as causal effects, grounded within the counterfactual approach to causality (pp. 196-197). I very much agree with PHL that this is the only way to use IV when studying program effects. But using IV within a counterfactual framework requires interpreting IV coefficients as Local Average Treatment Effects (LATE). Besides grounding IV within a clear causal framework, LATE assumes heterogeneous treatment effects. However, interpreting IV coefficients as LATE requires *five* assumptions to hold; these subsume the former two-assumption approach. If any fail, then the IV coefficients cannot be interpreted as causal effects.

These five assumptions were laid out by Angrist, Imbens and Rubin (1996) in their seminal

[1] The paper received the 2010 Association for Institutional Research Charles F. Elton Best Paper award.

paper, and further discussed in Angrist and Pischke (2009). While PHL cite both works in their paper, they neglect to discuss whether four of these five assumptions hold for their analysis. The assumptions are:

1. Stable Unit Treatment Value Assumption (SUTVA)
2. Random assignment
3. Exclusion restriction
4. Nonzero average causal effect of instrument on treatment
5. Monotonicity

I review these assumptions as they apply to PHL's analysis, while also discussing PHL's misinterpretation of some of their empirical results. Most notably, PHL assert that they can empirically determine whether their instruments are valid.

While an emphasis on reviewing statistical assumptions may appear overly academic, as applied researchers we often forget that our substantive conclusions depend crucially on whether the underlying assumptions of our methods hold. If they do not, then depending on the nature of the violation, our conclusions, and thus our policy recommendations, could be gravely mistaken. This is especially the case with IV: the bias and inefficiency of a faulty IV analysis could yield a more misleading estimate of a program effect than simply using OLS with an endogenous measure of treatment.

Background

Any discussion of IV within a causal framework relies on Rubin's Causal Model (Holland, 1986). With this approach, individuals have a set of "potential" outcomes for both treatment status and dependent variable, regardless of what actually occurs. The actual treatment assignment and observed value of the dependent variable are the factual outcomes; what would have occurred if the treatment assignment had been different (and thus the dependent variable possibly taking on a different value) are the counterfactual outcomes.

One way to think about potential outcomes is to picture an individual's thoughts when faced with treatment choices. PHL wish to estimate the effect of learning communities on academic performance. Endogeneity arises because individuals faced with the choice of participation in a learning community make their choices based on their potential outcomes; here, the potential outcomes are 1) their academic performance, after participating in a learning community, and 2) their academic performance, after not participating in a learning community. A student who is very motivated to obtain a college degree may believe that learning community participation will increase their academic success; such a student is basing participation on their potential outcomes. As a result, treatment participation becomes correlated with the error term in a multiple regression model.

In other words, we wish to estimate:

$$\textit{academic performance}_i = B_0 + B_1 \textit{learning community}_i + u_i \tag{1}$$

but instead we estimate

$$\textit{academic performance}_i = B_0 + B_1 \textit{learning community}_i + \overbrace{\textit{motivation}_i + v_i}^{u_i} \tag{2}$$

Motivation and many other variables affect academic performance, but because we typically lack measures of these variables, they end up in the error term. Because they also determine learning community participation, the learning community dummy variable is correlated with the error term u_i, and the coefficient B_1 is biased.

The IV solution is to estimate

$$\textit{learning community}_i = \pi_0 + \pi_1 \textit{exogenous instrument}_i + \overbrace{\textit{motivation}_i + r_i}^{q_i} \tag{3}$$

Like Equation 2, Equation 3 is estimated with motivation in the error term q_i, which illustrates why the coefficient B_1 in Equation 2 is biased: part of the variation in learning community participation is due to motivation. Suppose that we use the predicted values from Equation 3 as a new estimate of learning community participation:

$$\widehat{\textit{learning community}}_i = \pi_0 + \pi_1 \textit{exogenous instrument}_i \tag{4}$$

Variation in this new measure has been "purged" of the correlation with motivation, and variation here is driven by variation in the exogenous instrument; that is, a variable that is not correlated with either of the error terms u_i or q_i.

We can then re-estimate Equation 2, using the new measure of learning community participation:

$$\textit{academic performance}_i = B_0 + B_1 \widehat{\textit{learning community}}_i + \overbrace{\textit{motivation}_i + v_i}^{u_i} \tag{5}$$

Motivation still remains in the error term, but this no longer matters, because our new measure of learning community participation is now uncorrelated with motivation.

In their paper, PHL estimate the models in Equations 3 and 5 to understand whether learning communities affect grade-point average. They argue that participation in a summer bridge program, as well as whether a student declared a major at entry, are suitable instruments for an IV analysis of learning community participation and academic performance. The remainder of this paper demonstrates why this is not the case. I conclude with recommendations for using IV to estimate causal effects when studying college students.

Assumption 1: Stable Unit Treatment Value Assumption

A key assumption in experimental research is no interference between units. Here, Angrist, Imbens and Rubin assert than an individual's treatment status and outcome (i.e., their potential outcomes) cannot be affected by the treatment status of other individuals. As Imbens and Wooldridge note, "This lack-of-interaction assumption is very plausible in many biomedical applications. Whether one individual receives or does not receive a new treatment for a stroke or not is unlikely to have a substantial impact on health outcomes for any other individual" (Imbens and Wooldridge, 2009, p.13) This assumption is less plausible for postsecondary interventions, especially the assumption that person i's treatment status does not affect person j's outcome.

As described by PHL (p. 195), one of the courses that students in the learning community take is a first-year seminar. These courses usually aim to provide students with study skills and knowledge about the institution that are thought to enhance student success. Students in the treatment condition likely gain other useful information via interactions with advisors and faculty associated with the learning community. It is easy to imagine this information being shared by students in the treatment condition (learning community participation) with their roommates and friends in the control condition (no learning community participation). Such spillover effects can reduce the estimated treatment effect, as outcomes for students in the control condition are affected by the treatment, similar to students in the treatment.

I note that such spillover effects can work in the opposite direction. A recent paper, for example, finds that lottery incentives in surveys of college students increase response rates (Laguilles et al., 2011). In this study, members of several different treated groups received survey invitations mentioning the possibility of winning prizes such as iPods, while members of the control groups received a survey invitation without the possibility of winning a prize. The survey experiments were conducted within a single university, so that communication between members of the treatment and control groups was likely. Thus, the study's positive findings have two explanations. First, lottery incentives boost response rates, despite the fact that the expected value of the response incentive (cost of iPod$\times$probability of winning) was very low. Second, spillover effects resulted in members of the control group feeling resentful that they were not eligible for a prize. Instead of increasing response on the part of the treated, the treatment instead suppressed response on the part of the controls, resulting in a difference in response rates between the two groups.

Given the dearth of postsecondary research with strong internal validity, I would gladly accept an otherwise strong IV analysis that might perhaps violate some aspects of SUTVA. But a discussion of SUTVA, whether it holds for a given application, and what violations might mean for the estimated treatment effect are necessary for any research application using IV. More importantly, deep thinking about SUTVA can lead to secondary analyses testing whether the assumption holds for a particular project. For example, one could conduct a small fidelity of implementation analysis to determine the extent to which students in the treatment condition discuss the treatment with students in the control condition.

In sum, the particular treatment studied by PHL could violate SUTVA, and they provide no evidence (or discussion) as to whether SUTVA holds for their study. Given the nature of spillover effects, the positive impact of learning communities at their study institution could have been reduced because of interference between units. A violation of SUTVA is one possible explanation for their null findings.

Correlation Between the Instrument and Error Term

Traditional approaches to IV assert that for IV to provide an unbiased estimate of an effect, the instrument must be valid; that is, it must be uncorrelated with the error term in the second-stage equation (i.e., our main OLS equation of interest). More formally, if we are interested in knowing the effect of D on Y,

$$Y_i = B_0 + B_1 D_i + u_i \tag{6}$$

where D_i is a dummy variable indicating some kind of treatment (such as learning community participation) and Z_i is the instrument, then

$$E[Z_i \cdot u_i] = 0 \tag{7}$$

This is one of the most vexing issues with IV, because defending the validity of an instrument depends on theoretical arguments as to the relationship between the instrument Z_i, the dependent variable Y_i, and the error term u_i in the second-stage equation. Such arguments can become complicated very quickly, and it is not always clear if all possible arguments against Equation 7 have been considered.[2]

[2] Technically, a second assumption is also required, that Z_i not be correlated with the error term in the first-stage

One of the central insights of Angrist, Imbens and Rubin is that Equation 7 should actually be viewed as two separate assumptions, one focusing on the treatment assignment, and the other on alternate causal paths between the instrument and Y_i. I find such an approach helps one think more clearly about possible violations, because rather than consider the opaque question "Is my instrument uncorrelated with my error term?", researchers must instead consider two more specific questions:

1. Is my instrument similar to random assignment (or at least ignorably so)? (Assumption 2)
2. Are there any plausible causal paths between my instrument and the dependent variable other than through my treatment? (Assumption 3)

Only when these questions can be answered with an unequivocal "Yes!" for Question 1, and "No!" for Question 2, can one argue that Equation 7 holds for a given analysis.

Before discussing the next two assumptions in detail, it is useful to review exactly what the u_i in Equation 6 represents. From a causal effects perspective, it is most certainly *not* an error term as most people picture it. By calling u_i an error term, we are implicitly stating that Equation 7 holds (because "error" implies "random error" for most people), when in most applications this is an open question that can only be answered after lengthy investigation. Instead, it is important to understand that u_i contains the hundreds, perhaps thousands, of independent variables that affect Y_i, but that are not in our regression equation. I tell my students that u_i is best thought of as an "unmeasured factors" term rather than an error term, to help keep in mind the main problem with using standard OLS or HLM to estimate program effects: it is only when we can state that these thousands of variables are uncorrelated with our treatment D_i that we can estimate the causal effect of a treatment on an outcome. And, contrary to statements by PHL on p. 198 of their paper, self-selection cannot be considered a single variable. In the case of college students and postsecondary interventions, self-selection is usually due to many different variables, *each* of which must be uncorrelated with the treatment in order to estimate a causal effect.[3]

equation (Angrist et al., 1996).

[3] In keeping with the field, and to avoid confusion, I will use "error term" for the remainder of the paper.

Assumption 2: Random assignment

Assumption 2 states that "…the instrument is as good as randomly assigned: it is independent of the vector of potential outcomes and potential treatment assignments" (Angrist et al., 1996, p. 152). Here, "as good as" means that once a set of covariates has been taken into account, the instrument should be unrelated to potential outcomes (both treatment and dependent variables). Assumption 2 implies the causal diagram in panel (a) of Figure 1, which is assumed by PHL in their analysis. Note that there are no additional variables in the diagram, so that their instruments and dependent variable are connected, other than through learning community participation.

For Assumption 2 to hold for PHL's analysis, we must ask ourselves a simple question: can we consider summer bridge program participation and major declaration at entry as being randomly distributed across students? At most institutions, the answer would be a resounding no. Unfortunately, PHL do not describe how students enroll in the summer bridge program at IUPUI. From the Summer Bridge website (http://bridge.uc.iupui.edu) and Hansen et al. (2008), it seems as if many of the participants are first-generation college students who have received a scholarship, and that enrollment is open to students in colleges and majors that have participation agreements with the Summer Bridge program.

Thus, summer bridge enrollment is likely voluntary, and driven by a myriad of unobserved student characteristics, such as extrinsic motivation (desire to get a degree for financial purposes), intrinsic motivation (desire to excel in college and actively learn), personality (the Big Five personality constructs have been linked to a wide variety of academic and life behaviors and outcomes), a desire for faculty interaction, and the social and cultural capital that a student brings to college.

Indeed, PHL note that students who select first-year programs may "…feel relatively comfortable in settings where interaction with other students and faculty is paramount; they may not fear taking active steps to participate in co-curricular activities and experiential learning opportunities, and look forward to the self-reflection that is necessary for full participation in most first-year academic support programs" (pp. 207-208). This description could just as easily apply to students who seek summer bridge program participation. Clearly, the variables that make learning community participation endogenous in an academic performance equation also make summer bridge participation endogenous. A more accurate description of PHL's causal configuration is panel (b) of Figure 1, where unobserved student characteristics drive both the choice to participate in the summer bridge program, as well as academic performance in college.

A similar argument can be made for declaring a major upon entry. Students who know what they want to major in at entry are different psychologically from students who do not (Gordon, 2007). Many of these characteristics are related to academic performance, suggesting that panel (c) also describes the relationship between major declaration, academic performance, and unobserved student characteristics.

The solution to this situation is to include measures of these unobserved student characteristics in the second-stage equation of the instrumental variables analysis (Equation 5). Doing so pulls these variables out of the error term, such that one can argue that the instrument is now "as good as" randomly assigned, conditional on the covariates in the second stage equation.[4] In other words, the instrument can be considered randomly assigned if it is no longer correlated with the unmeasured factors in the the error term, once the right set of covariates has been included in the second-stage equation.

In their second-stage model PHL include as covariates gender, a dummy variable measuring if a student was both first-generation and low-income, a combination of high school GPA and SAT score, and application date. This small set of commonly available observables can only partially capture the differences between summer bridge participants and non-participants, and declared and undeclared majors. While PHL argue that college application date measures student motivation, no evidence is offered for this assertion. In addition, their first-stage equation presented in the bottom panel of their Table 3 shows no statistically significant relationship between application date (i.e., motivation) and learning community participation, contrary to theoretical expectations. Moreover, such proxies as application date can only *partially* control for factors such as motivation and social and cultural capital, which, in turn, means that both of their instruments will still be endogenous,

[4] Technically, these covariates must also be included in the first-stage equation as well; see Angrist and Pischke (2009).

conditional on their limited set of covariates.

By using as their instruments two measures of student behavior that are determined by voluntary student choice, and by not controlling for the many unobserved characteristics that drive both these choices and academic performance, PHL use instruments that are similar to their endogenous variable of interest, learning community participation. Their instruments are thus invalid, and the estimates presented in Table 3 of their paper tell us nothing about the causal effect of learning community participation on academic performance.

Assumption 3: Exclusion restriction

Let us assume for the sake of argument that SUTVA is not violated in this context, and that the limited covariates in PHL's IV model perfectly capture all possible differences between summer bridge participants and non-participants, as well as decided and undecided majors. We would still be forced to reject their analysis, because their instruments violate the third assumption of IV analyses. Only one causal path can exist between an instrument and the dependent variable of interest, and that path must pass through the endogenous regressor.

PHL argue that summer bridge participation affects student behavior, because "some students who participated in the bridge program were expected to participate in themed learning communities, whereas other bridge participants were strongly encouraged to participate in TLCs" (p. 203). In other words, summer bridge programs altered student behavior through the expectations and recommendations of faculty and summer bridge personnel, resulting in some summer bridge participants deciding to participate in a learning community.

Yet PHL overlook the primary function of summer bridge programs, which is to prepare students for college. Such preparation usually focuses on bolstering basic mathematics and language arts knowledge, as well as studying skills, so that students will be more successful academically than they would have been otherwise. Chism et al. (2008) assert that summer bridge program participants at IUPUI

> . . . establish networks for success with faculty, advisors, student mentors and librari-

> ans; make friends with other freshmen; learn to handle college-level expectations for reading and writing; receive individualized support for math; begin connecting with a school and major; become acquainted with the campus; and gain experience in using instructional technology" (p. 10).

A typical summer bridge causal chain of events is depicted in panel (a) of Figure 2. Here, summer bridge participation alters students in *two* different ways, both of which ultimately affect academic performance. However, only one causal path passes through their endogenous regressor. In addition, most postsecondary scholars and practitioners would argue that the second path is non-trivial: the only possibility that this other causal path does not occur is if summer bridge participation has absolutely no effect on the attitudes, knowledge and behavior that affect academic performance, constructs that summer bridge programs specifically aim to change.

A similar argument can be made for their second instrument, declaration of an academic major upon college entry. PHL argue that this will lead to greater participation in learning communities, because many of the learning communities are associated with specific majors (see panel (b)). However, declaration of an academic major at entry sets off another causal chain of events, as declared majors are by definition situated within specific disciplines, colleges, and departments. They receive different advising, associate with different groups of peers, and take different sets of courses than undeclared majors. It is difficult to argue that all of these variables will not affect academic performance, outside of learning community participation.

Can PHL empirically verify the validity of their instruments?

From the previous discussion, we can see that both of PHL's instruments are invalid; that is, they are correlated with the error term in the second stage equation. PHL offer two pieces of empirical evidence supporting the validity of their instruments, and that would appear to contradict my arguments regarding Assumptions 2 and 3. First, they show that their instruments have very low correlations with their dependent variable. The correlations with grade-point average are $.05$ for summer bridge participation and $-.01$ for declaration of major (p. 203). Second, they conduct an overidentification test "...to verify that the instruments were not correlated with the error term ..." (p. 204).

Unfortunately, we cannot definitively and empirically verify the validity of an instrument. If we could, IV analyses would be much more common than they are. At first blush, it may seem easy to do so; we can just estimate the correlation between our instrument and dependent variable, as PHL have done. But a low correlation between an instrument and the dependent variable does not tell us if the instrument is uncorrelated with unmeasured variables in the error term of the second-stage equation; it is the correlation with the latter with which we are concerned, not the former.

Moreover, consider a situation in which there is a strong, positive causal effect of a treatment. As such, we would expect the instrument to be correlated with the endogenous treatment variable, and the endogenous treatment variable to be correlated with the dependent variable; low correlations between the instrument and dependent variable would signify that something is amiss. This is precisely why methodologists recommend estimation of a reduced form equation using the instrument(s) as predictors of the dependent variable as an initial part of any IV analysis (Angrist and Pischke, 2009; Murray, 2010). If we followed PHL's implicit strategy of rejecting instruments that have high correlations with the dependent variable, we could easily end up rejecting the use of strong and valid instruments.

Their second piece of evidence is the failure to reject the null hypothesis of an overidentification test, which is possible to conduct if we have more than one instrument for an endogenous regressor. PHL report that Wooldridge's test of overidentifying restrictions was not statistically significant, "...indicating that the model was appropriately specified and the instruments were not correlated with the error term" (p. 206). This statement is false, because overidentification tests cannot conclusively decide if instruments are uncorrelated with the error term.

Conceptually, an overidentification test for two instruments creates two separate IV estimates, and tests whether they differ, given sampling error (Stock and Watson, 2007). If the instruments are valid, then we would expect both instruments to provide similar estimates of the effect of the endogenous regressor. This is essentially the null hypothesis that is being tested by an overidentification test. If the estimates significantly differ, then that would call into question the validity of one of the instruments. The previous sentence reveals one of the main issues with overidentification tests: they implicitly assume that one of the instruments is valid, such that the other instrument can be compared against it. Simply running the test, as PHL have done, is not sufficient to establish validity:

> Tests of over-identifying restrictions are most compelling when there are some instruments (enough to identify the equation) whose validity seems sure. It is then that we can be confident the test is itself valid. Tests of over-identifying restrictions are particularly suspect when all of the instruments share a common vulnerability to being invalid (Murray, 2010, p. 14).

As I have described above, one can make a strong argument that PHL's instruments share two common vulnerabilities: unobserved student characteristics and multiple causal paths between the instruments and dependent variable.

Further, while overidentification tests have been popular in the past, they now make little sense when estimating LATE causal effects. As Angrist and Pischke (2009) note,

> . . . each instrumental variable identifies a unique causal parameter, one specific to the subpopulation of compliers for that instrument. Different valid instruments for the same causal relation therefore estimate different things, at least in principle . . . [O]ver-identification testing of the sort discussed in section 4.2.2, where multiple instruments are validated according to whether or not they estimate the same thing, is out the window in a fully heterogeneous world (p. 166).

Finally, I am skeptical of using statistical tests to drive the assumptions of any statistical analysis, particularly in the absence of strong theory. Defenses of the underlying assumptions of IV should rely first and foremost on theory, the use of which is almost entirely absent in PHL's discussion of their instruments. The relevance of theory over statistical tests can be seen most clearly with tests of endogeneity. The idea behind these statistical tests is to see whether an OLS estimate is biased due to the endogeneity of an independent variable, in which case the use of IV is seen as justified.

Consider an archetype of IV analysis in the field of economics, Card's (1993) study of the effect of education on earnings. This analysis is widely cited and is often used to illustrate IV in econometric texts, particularly the use of distance as an instrument. He argues that years of education will be correlated with the error term in a model with earnings as the dependent variable, because the omitted variable ability will affect both educational attainment and earnings. Using distance to the nearest college as an instrument for years of education, he finds that the return to education is 25%-60% higher when estimated with IV. Yet my reanalysis of his data indicates that a test for endogeneity fails to reject the null hypothesis that years of education is exogenous.[5] If Card had blindly followed the result of this statistical test, he would have concluded that IV is not necessary to study the returns to education, and he never would have published his now-classic paper.

In sum, PHL offer very little discussion about possible correlation between their instruments and error term. In other words, they do not defend the validity of their instruments. Given what we know about student choice and the workings of summer bridge programs and academic majors, it is highly unlikely that their instruments are uncorrelated with the error term, and the limited covariates in their model do not solve this problem. While overidentfication tests can play an important role in any discussion of the validity of multiple instruments, they are useless when strong theoretical arguments against validity exist, and when all of the instruments being tested could be invalid for the same reason.

Assumption 4: Nonzero average causal effect of instrument on treatment

The fourth assumption of IV is that the instrument must be correlated with the endogenous regressor, and hopefully highly correlated. Here, PHL are on much stronger ground. First, they offer

[5] Using the specification of column 5 in his Table 3, neither the Durbin nor Wu-Hausman tests are statistically significant ($p < .28$).

theoretical explanations as to why their instruments should be correlated with learning community participation. Second, they show that one of their instruments, summer bridge participation, is highly correlated with their endogenous regressor ($r = .30$). Third, and most importantly, they report two common statistics used to determine the strength of instruments. The first is Shea's partial R-square ($R^2 = .09$) (Shea, 1997), calculated as the variance in the treatment variable explained by the instruments only (thus differing from the R^2 of the first-stage equation, which is the proportion of variance explained by the instruments and covariates combined). The second is an F statistic from the first-stage regression ($F = 81$), a joint hypothesis test that the coefficients of both instruments are equal to 0. As with the partial R-square, this F statistic is different from the F statistic reported for the entire first-stage equation. While invalid, PHL's instruments definitely demonstrate a nonzero effect on the treatment.

Unfortunately, PHL incorrectly state how to determine whether instrument(s) are weak as based on the F statistic joint hypothesis test:

> A F statistic representing the joint significance of the relationship between the two instruments and TLC participation was also calculated. Following the recommendation of Stock and Yogo (2005), a minimum F value of 10.00 was set as the standard for strong and reliable instruments (p. 204).

> …the robust F statistic was 81.01 ($df = 2, 2186; p < 0.05$). This F value was well above the threshold of 10.00 recommended by Stock and Yugo (2005) (p. 206).

These two statements directly contradict Stock and Yugo's conclusion in their 2005 paper:

> When there is the case of a single included endogenous variable [the case of PHL's analysis], this procedure provides a refinement and improvement to the Staiger-Stock (1997) rule of thumb that instruments be deemed "weak" if the first-stage F is less than 10. The difference between that rule of thumb and the procedure of this paper is that,

> instead of comparing the first-stage F to 10, it should be compared to the appropriate entry in Table 5.1 (TSLS bias) ... (Stock and Yogo, 2005, p. 106).

Stock and Yogo's paper is often cited but rarely read, which is why researchers still talk about the supposed minimal F statistic value of 10. Stock and Yogo define weak instruments in two specific and different ways. First, weak instruments can result in a biased IV estimator. Second, weak instruments can result in null hypothesis rejection rates that are greater than 5%. They provide two sets of critical values for these two definitions of weakness. The first set of critical values is undefined for the case of a single endonegous regressor with less than three instruments, so it cannot be used with PHL's analysis. The second set of critical values are provided for four sets of rejection rates, ranging from 10% to 25%. With two instruments and one endogenous regressor, the critical values are 19.9, 11.6, 8.8 and 7.3, for rates of 10%, 15%, 20%, and 25%, respectively. The size of the critical value varies greatly, depending on what actual rejection rate we are willing to tolerate, and illustrates the dangers of using simple rules of thumb to determine strength.

PHL also make much of their claim that OLS shows a positive effect of learning communities (over one-quarter of a grade-point), while their IV estimate is not statistically different from zero, and indeed is almost zero $(.01)$. They neglect to mention, however, that the standard error for their IV coefficient is quite large, so much so that the OLS and IV confidence intervals overlap.

One drawback to IV is that the standard errors of the coefficients are larger than OLS; researchers face a trade-off between bias and efficiency. This is one reason such emphasis is placed on having a strong instrument. A good equation to keep in mind is that with a single endogenous regressor and instrument (and homoskedastic errors), the standard error of the IV regression coefficient is equal to the standard error of the OLS coefficient, divided by the correlation between the instrument and endogenous regressor (Cameron and Trivedi, 2005):

$$V[\hat{\beta}_{IV}] = \frac{V[\hat{\beta}_{OLS}]}{{r^2}_{DZ}} \tag{8}$$

Given that the correlation between summer bridge participation and learning community participation is .30, Equation 8 implies that PHL's IV standard error will be 3.33 times larger than the OLS standard error, and that the confidence intervals for their IV estimate will be quite large.[6]

Figure 3 shows these confidence intervals. As can be seen, the 95% confidence interval for the IV estimate almost overlaps the OLS point estimate. Given that PHL have over 2,000 observations in their dataset, one could argue that $\alpha = .01$ is more appropriate than $\alpha = .05$. The 99% confidence intervals are shown on the right-hand side of the figure, and almost completely overlap the OLS confidence interval. Leaving aside the issue of instrument validity, it is not clear how PHL can claim that their IV results qualitatively differ from their OLS results.

Assumption 5: Monotonicity

The last assumption of IV is monotonicity, which gets at the heart of how IV works. Here, we must be able to assume that as the instrument changes, it either a) does not affect whether a unit is treated or b) affects all units the same way, for those that are affected. The concept is often illustrated with four behavioral groups: always-takers, never-takers, compliers, and defiers (Angrist et al., 1996). Understanding these four groups is important not only to evaluate whether monotonicity holds, but also for the policy relevance of any IV analysis.

Suppose that in order to randomize individuals for an experiment, we flip a coin to assign each individual to the treatment or control condition. If we flip the coin, make the assignment, then go back in time and repeat the coin flip, we can think of the coin flip as switching an individual back and forth between the treatment and control conditions, as the coin takes on the values of heads and tails. Given our conceptualization of instruments as "as good as" randomly assigned, we can also imagine our instrument changing in value, flipping back and forth, and as it changes, individuals are "assigned" to the treatment and control conditions. In PHL's application, as summer bridge participation (or major declaration) switches "on" and "off", we can imagine individual students being assigned to learning community participation and non-participation.

[6] Equation 8 also illustrates why it is important to have a strong instrument. If PHL used major declaration as a single instrument in their analysis, their IV standard error would be almost 17 times larger than their OLS standard error. Using the IV coefficient of .01, the corresponding estimated 95% CI would range from -1.2 to 1.2, a range of 2.4 grade points, or most of the grade distribution at a typical university.

One objection to this conceptualization is that not every student will obey their assignment, and this is where the four groups come in. Regardless of summer bridge participation, some students (always-takers) will always choose to participate in a learning community, for a variety of reasons. These students may be very anxious and avail themselves of all possible programs that could assist them academically, or they could simply have strong tastes for learning opportunities. Similarly, some students (never-takers) will always avoid learning community participation, regardless of summer bridge participation. For example, they may be extremely antisocial, or they may wish to exert the minimum effort possible to obtain a college degree. It is important to understand that *an IV analysis tells us nothing about the causal effect of a program for these groups*. The reason is simple: the instrument cannot affect their treatment participation. Treatment participation cannot change for these groups given the instrument, in turn meaning that we cannot estimate a treatment effect for them.

Instead, IV estimates the effect of the treatment on the third group, compliers. These are individuals who only join a learning community if they attend the summer bridge program (declare a major); they do not join if they do not participate in the bridge program (do not declare a major). The size of this group has important implications for the substantive relevance of any IV analysis; I return to this point below.

The monotonicity assumption stipulates that a fourth group, defiers, does not exist. Defiers are individuals who do the opposite of their treatment assignment. Like small children, when told to do A, they do B; but when told to do B, they insist on doing A. In PHL's analysis, these would be students who participate in a learning community if they do not take part in the summer bridge program (do not declare a major), but if time was rewound and they instead took part in the summer bridge program (declared a major), they would then refuse to join a learning community. The existence of defiers generally seems improbable, but they do exist in some applications (see Hoxby (2000b) for one example of a monotonicity violation due to organizational rules).

Here, defiers are probable for the summer bridge instrument. One can imagine students with little experience of programmatic efforts such as learning communities or summer bridge programs eagerly signing up for them, but being disappointed with the results. They might find the pace of the bridge program too slow, or the material too elementary. When given the opportunity to join a learning community, they may decline, due to their experience with the bridge program. Conversely, if these students had not participated in the bridge program, they might now join the learning community, as this would be their first exposure to these types of programs. The possibility of defiers for the summer bridge instrument is particularly problematic, because this is PHL's strongest instrument.

The possible presence of defiers makes a causal interpretation of PHL's IV coefficient difficult. As Angrist and Pischke (2009) note, "We might therefore have a scenario where treatment effects are positive for everyone yet the reduced form is zero because effects on compliers are canceled out by effects on defiers" (p. 156). The extent that this will happen depends on the proportion of defiers: if this proportion equals zero, then monotonicity is not violated. But as this proportion increases, the estimated treatment effect will decrease in size. With a large estimated treatment effect, one could argue that the estimate represents a lower bound in the presence of defiers. Null findings, such as those reported by PHL, could be due to either no treatment effect or to the presence of defiers.

Understanding the implications of monotonicity is important for policy reasons, as well as econometric reasons. PHL discuss their results as if they apply to all students, but as the preceding discussion has hopefully made clear, IV only yields the causal effect for a subgroup of a population, the compliers. This group might be small, as well as unrepresentative of the population, in turn limiting what we can say about the effect of a particular program. If D is the treatment variable, taking a value of 1 if treated, 0 otherwise, and Z is the binary instrument, and if D_{1i} indicates treatment when $Z_i = 1$ and D_{0i} indicates treatment when $Z_i = 0$, then

$$\mathrm{P}[D_{1i} > D_{0i}] = \mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0] \tag{9}$$

provides the probability of complying: it is the difference in the probability of treatment for those assigned to treatment by the instrument and those not assigned by the instrument (Angrist and Pischke, 2009). PHL's first-stage equation in the bottom panel of Table 3 on p. 206 is a linear probability model, so we can interpret these coefficients such that participating in a summer bridge program increases the probability of joining a learning community by 36 percentage points, while the same effect for declaring a major is 6 percentage points. Because these are probabilities, they can also be considered proportions, and here they tell us the proportion of students who are compliers: 36% and 6%.

We can take this one step further, and ask what proportion of learning community participants participated because the summer bridge instrument assigned them to participation. The compliance probability for the treated is

$$\mathrm{P}[D_{1i} > D_{0i}|D_i = 1] = \frac{\mathrm{P}[Z_i = 1](\mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0])}{\mathrm{P}[D_i = 1]} \tag{10}$$

The probability of being a complier for someone in the treated group is equal to the probability of being assigned by the instrument to treatment $(\mathrm{P}[Z_i = 1])$ times the difference in probability of treatment for those assigned and those not assigned $(\mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0])$, divided by the probability of being treated $(\mathrm{P}[D_i = 1])$ (Angrist and Pischke, 2009). A similar probability for the control group can also be calculated, using $1 - \mathrm{P}[Z_i = 1]$ and $1 - \mathrm{P}[D_i = 1]$.

In PHL's analysis, the probabilities of being assigned to treatment by the instruments are simply the proportion participating in summer bridge and declaring a major, .17 and .92 respectively, while the probability of treatment is the proportion participating in a learning community, .25 (p. 200). Together, these yield the proportions in Table 1.

Should we be concerned about these small proportions? Unfortunately, there is no clear answer, leading some to declare, "I find it hard to make any sense of the LATE" (Deaton, 2010). Assuming all five assumptions hold, we can first see that the effect estimated by PHL only applies to a small subset of their population (6%-36%). One could argue that students whose behavior can be changed by the instruments are students at the margin, and in terms of understanding the effects of learning communities, it is precisely these students who we most want to target with an intervention like a learning community. However, consider the never-takers. Many of these students are probably students who are disengaged with the institution, and who would be considered among the group of marginal students that we wish to target with our interventions. The analysis tells us nothing about the effect for these students, and the compliers in this context may not be all the students whom we wish to target.

Second, the small proportion of compliers raises the question of what would happen if learning community participation was expanded to all students. It is easy to forget that the effects of an intervention may not occur when the intervention is rolled out to a group of individuals who differ from those who took part in the initial evaluation. The small proportion of compliers in Table 1 certainly make this an important issue to consider before expanding learning communities. Third, the proportions in the last two columns of the table are less relevant for this context, than for others. The relatively small proportions of compliers among the treated indicates that a minority of learning community participants were summer bridge participants or major declarers. This is less important that in other contexts, e.g. the use of draft eligibility to estimate the effect of military service (Angrist and Pischke, 2009).

Fourth, if we assume that no defiers exist, then we can also calculate the proportion of always-takers and never-takers (Imbens and Wooldridge, 2009):

$$\mathrm{P}[D_{1i} = D_{0i} = 1] = \mathrm{E}[D_i|Z_i = 0] \tag{11}$$

$$\mathrm{P}[D_{1i} = D_{0i} = 0] = 1 - \mathrm{E}[D_i|Z_i = 1] \tag{12}$$

We can use the student numbers provided by PHL on p. 203 to estimate these probabilities. Equation 11 states that the probability of being an always-taker is the expected value of treatment participation for those who were not assigned to treatment. In PHL's application, this is the proportion of summer bridge non-participants who participated in a learning community, which equals 19.6% (359/(2193-363). Never-takers are calculated by 1 minus the expected value of treatment participation for those who were assigned to treatment, which equals 45.2% (1-199/363).[7]

Chism et al. (2008, p. 9) describe most entering IUPUI students as disengaged:

> They become "parking lot" student[s], who drive to campus, talk to no one, eat their lunches in their cars, and drive home. College is viewed as a necessary burden that they must endure so that they can acquire a career and make money in the future, but they do not invest themselves in the life of the campus nor engage in its intellectual environment.

Given the makeup of the student body and the nature of the treatment, which requires a significant investment of time and engagement, it is not surprising that with this instrument almost half of incoming students are never-takers.

Conclusion

As the previous discussion has demonstrated, PHL's analysis meets only one of the five assumptions required to interpret instrumental variable estimates as causal effects:

- Given the nature of the treatment and communication between treated and control units on a single campus, violation of SUTVA is likely.
- With student selection into summer bridge programs and major declaration, and the limited set of covariates used in their analysis, their instruments cannot be considered as good as randomly assigned.
- Causal paths from the instruments to the dependent variable outside of the endogenous regressor are very likely.
- Their instruments can be considered strong.
- Presence of defiers is possible, and the small set of compliers is a limitation of their analysis.

[7] These proportions use raw probabilities from the distributions of Z and D, rather than the estimated probability of compliance from PHL's first-stage equation; the compliance probabilities differ slightly.

Of these, violations of Assumptions 2 and 3 are by far the most serious. Without unrealistic assumptions about unobservable student characteristics and alternate causal paths between their instruments and academic performance, their instruments are invalid. Unfortunately, PHL find themselves in one of the two most likely states for researchers trying to conduct an IV analysis: potential instruments tend to be strong and invalid (the case here), or weak and valid.

The key lesson for researchers seeking to use IV is Murray's sage advice: "... subject candidate instruments to intuitive, empirical, and theoretical scrutiny to reduce the risk of using invalid instruments. We can never entirely dispel the clouds of uncertain validity that hang over instrumental variable analyses, but we should chase away what clouds we can" (2006, p. 130). Unfortunately, PHL make little effort to dispel the clouds hanging over their evaluation of learning communities. Given what we know about college students, academic performance, and the postsecondary environment, it is unlikely they can clear away the clouds hanging over their study.

Recommendations for future research on college students

From the previous discussion, it may seem as if the technique of instrumental variables is unlikely to be useful for researchers studying college students, but this is not the case. I offer comments in four areas for researchers seeking to use IV to study college students.

The importance of description

Besides a lack of attention to the assumptions underlying IV, PHL provide very little information about their strongest instrument, summer bridge participation, as well as their treatment. Such detail is needed for other researchers to understand the validity and strength of the instrument. It is not clear, for example, how students enter the summer bridge program at IUPUI. Is is compulsory for some, and voluntary for others? If technically compulsory, is it truly compulsory in practice?

More importantly, we know very little about the treatment, learning community participation. From the brief description provided, it seems as if these are actually many different treatments, as the themes and majors vary across the learning communities. Based on what I have learned from learning communities at other institutions, I suspect that different individuals are in charge of each learning community and have significant leeway in how each one is structured. For example, Chism et al. (2008) reports that at IUPUI, learning communities are discipline-based, and that “Program offerings range from seven to twelve credit hours, tailored to a variety of majors and interests ” (p. 10). It is difficult to argue that there is a single treatment on this campus, as the treatment is in practice defined by the differing actions of learning community personnel. Moreover, such heterogeneity of treatment raises serious questions about fidelity of implementation. If we find null results, how do we know that poor implementation in some learning communities did not overwhelm the positive effects of well-run learning communities? Any research design that is used to estimate causal effects must carefully define what both the treatment and control groups experience. It is difficult to discuss what works, if we do not first know exactly what “what” means.

SUTVA

I would argue that any analysis of postsecondary programs and causal effects should include a discussion of SUTVA and whether it holds for a particular application, regardless of whether IV is specifically used for estimation. As we move away from analyses that measure associations to analyses that measure causal effects, it is vital that we understand exactly how an intervention does or does not work. Thinking about SUTVA can help us think more broadly about interventions, their theories of action, and whether they are implemented with fidelity. Such thinking can help us address an important critique of the causal inference revolution: we focus solely on what works, and do not address the equally important question of *why* something works (Deaton, 2010).

Grappling with SUTVA is especially important for college student researchers, because of the likelihood of spillover effects. Besides selection bias, SUTVA violations are likely to pose a major problem for researchers seeking causal inferences about college students. The best randomization strategy in the world is problematic if students in the control condition are exposed to elements of the treatment. Keeping SUTVA in mind as we think about analyses can help us strengthen our research designs. For example, SUTVA violations are likely to be more common at residential colleges than at commuter institutions, given the greater amount of student interaction at residential colleges. This suggests testing interventions at commuter institutions, although an increase in internal validity comes at the expense of external validity (Imbens, 2010).

Examples of good instruments

Admittedly, it is difficult to find instruments that are both strong and valid. Table 2 lists some instruments that have been used in the literature, although I should emphasize that there is (as always) some debate about their validity. Murray (2010) provides an excellent detailed review of good instruments, so I will only focus on their common characteristic: they are largely outside the control of the specific individuals undergoing treatment.

Random assignment is an obvious candidate for a good instrument, which is why it is often used to salvage a random experiment when units do not comply with their treatment assignment. It is as good as randomly assigned, and given the randomization, it is difficult to think of alternate causal paths from the assignment to the outcome. Most importantly, it generally cannot be influenced by individuals facing treatment. Other variables used as instruments, such as distance and laws, are similar, in that it is often difficult for individuals to affect the instrument in ways that lead to violations of Assumptions 2 and 3. This points to the main shortcoming of PHL's instruments: they are easily altered by student choices. Postsecondary researchers should keep this in mind as they seek instruments for their analyses. Likely instruments for postsecondary interventions include distance, institutional and state policies, and local and state laws. For example, Porter and Stephens (2010) use state union laws to estimate the effect of faculty unionization on institutional decision-making.

An instrument for college student interventions

Random assignment may not be appealing to many postsecondary researchers. Besides ethical issues of withholding treatment from some students (which is somewhat overblown, given that we withhold treatments in randomized medical trials), there is the very real issue of explaining to students why some must undergo the treatment, and others must not. In the context of developmental education, this is easily justified, but such situations are rare in postsecondary education.

Instead, one related instrument that holds promise for postsecondary research is what I term a *randomized invitational instrument*. Suppose that a campus has an intervention, such as first-year seminars, that they would like to evaluate. Half of incoming students would receive an email and/or printed material describing the benefits of the intervention and inviting students to register for the course, while half would not. The invitation could also include incentives for participation, such as compensation. This instrument can take two different forms, one weak and one strong. The weak version would allow any student to participate in the intervention, while the strong version would not permit members of the control group to participate. The choice of approaches depends on the constraints facing the researcher. The weak version overcomes ethical objections to withholding treatment, as all students would still be able to participate in the treatment. The price paid is the strength of the instrument: so many students could either ignore the treatment or partake of the treatment that the researcher is faced with a weak instrument problem. Conversely, the probability of a strong instrument is increased with the strong approach, as all of the control students are by definition restricted from treatment participation.

Would such an instrument meet the five criteria of LATE?

- SUTVA violations are still possible, depending on the nature of the treatment.
- The instrument is obviously as good as randomly assigned.
- Alternate causal paths are difficult to envision. One would have to argue, for example, that not receiving an invitation to treatment would so upset a student that it would affect the outcome being studied, such as academic performance or persistence. Such a causal path could occur if the treatment is perceived to be very valuable. Careful thought about whether Assumption 3 holds would be necessary for any application using this instrument.
- There is a possibility that the instrument could be weak, given the nature of the instrument. This is probably the biggest drawback to randomized invitations.
- Defiers are difficult to imagine with this instrument.

Some researchers may scoff at the idea, but such randomized invitational instruments are already being used in postsecondary research. For example, Castleman and Page (2012) wish to understand the impact of summer college counseling after high school graduation on subsequent college enrollment. The problem they address is that students during this time period may need counseling, but do not have access to their high school counselors. Although they focus on students who have expressed the intent to attend college, there is still likely heterogeneity in this group, so simply estimating college enrollment rates for those counseled and those not counseled will not yield the causal effect of summer counseling.

Their solution is the weak version of the random invitational instrument:

> Students in both the treatment and control groups were told in advance that individualized counseling would be available from ACCESS over the summer. The treatment group additionally received proactive outreach from an ACCESS advisor over the course of the summer, while the control group did not. Advisors made multiple attempts to contact each member of the treatment group to offer counseling, and used a variety of outreach methods: phone, email, and text and Facebook messaging. Advisors had substantive communication with 76 percent of the treatment group, in which advisors asked students about how they were feeling about their college plans, and offered each of them a $25 gift card incentive to attend an assessment meeting with an advisor at the ACCESS Center for College Affordability (CCA) in Boston's city center.

Note that their research design was such that members of the control group could access the treatment, alleviating concerns that some students would be excluded from a potentially useful treatment. Using the invitation as an instrument for undergoing summer counseling, their IV estimates suggest that counseling increased college enrollment by 10 percentage points.

Instrumental variables has the potential to allow postsecondary researchers to estimate program effects, by taking into account unobservable student characteristics that drive both treatment participation and college outcomes. Successful use of the technique, however, requires careful reading of the IV literature, and thoughtful analyses of the validity and strength of the proposed instrument.

References

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., and Pathak, P. A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics*, 126(2):699–748.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative record. *American Economic Review*, 80(3):313–336.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455.

Angrist, J. D. and Kreuger, A. B. (1991). Does compulsory school attendance affect schooling and earnings. *Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D., Lang, D., and Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–163.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Bronars, S. G. and Grogger, J. (1994). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *American Economic Review*, 84(5):1141–1156.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Working paper, National Bureau of Economic Research.

Castleman, B. L. and Page, L. C. (2012). The forgotten summer: Does the offer of college counseling the summer after high school mitigate attrition among college-intending low-income high school graduates? Papre presented at the annual conference of the Association for Educational Finance and Policy, Boston, MA.

Chism, L. P., Baker, S. S., Hansen, M. J., and Williams, G. (2008). Implementation of first-year seminars, the summer academy bridge program, and themed learning communities. *Metropolitan Universities*, 19(2):8–17.

Currie, J. and Yelowitz, A. (2000). Are public housing projects good for kids? *Journal of Public Economics*, 75(1):99124.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48:424–455.

Dee, T. S. and Evans, W. N. (2003). Teen drinking and educational attainment: Evidence from two-sample instrumental variables estimates. *Journal of Labor Economics*, 21(1):178–209.

Evans, W. and Ringel, J. (1999). Can higher cigarette taxes improve birth outcomes? *Journal of Public Economics*, 72(1):135–154.

Gordon, V. N. (2007). *The undecided college student: An academic and career advising challenge*. Charles C. Thomas, Springfield, IL.

Hansen, M. J., Evenbeck, S. E., and Williams, G. A. (2008). The influence of a summer bridge program on college adjustment and success: The importance of early intervention and creating a sense of community. Paper presented at the annual conference of the Association for Institutional Research, Seattle, WA.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–970.

Hoxby, C. M. (2000a). Does competition among public schools benefit students and taxpayers? *American Economic Review*, 90(5):1209–1238.

Hoxby, C. M. (2000b). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, 115(4):1239–1285.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2):399–423.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):586.

Laguilles, J. S., Williams, E. A., and Saunders, D. B. (2011). Can lottery incentives boost web survey response rates? Findings from four experiments. *Research in Higher Education*, 52:537–553.

Levitt, S. D. (1996). The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *Quarterly Journal of Economics*, 111(2):319–351.

Long, B. T. and Kurlaender, M. (2009). Do community colleges provide a viable pathway to a baccalaureate degree? *Educational Evaluation and Policy Analysis*, 31(1):30–53.

McClellan, M., McNeil, B. J., and Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *Journal of the American Medical Association*, 272(11):859–866.

Miguel, E., Satyanath, S., and Sergenti, E. (2004). Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy*, 112(4):725–753.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives*, 20(4):111–132.

Murray, M. P. (2010). The bad, the weak, and the ugly: Avoiding the pitfalls of instrumental variables estimation. Unpublished paper, Bates College.

Pike, G. R., Hansen, M. J., and Lin, C.-H. (2011). Using instrumental variables to account for selection effects in research on first-year programs. *Research in Higher Education*, 52:194–214.

Porter, S. R. and Stephens, C. M. (2010). The causal effect of faculty unions on institutional decision-making. Paper presented at the meeting of the Association for the Study of Higher Education, Indianapolis, IN.

Shea, J. (1997). Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics*, 79:348–552.

Stock, J. H. and Watson, M. W. (2007). *Introduction to Econometrics*. Pearson Education, Inc.

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In Andrews, D. W. K. and Stock, J. H., editors, *Identification and Inference for econometric models: Essays in Honor of Thomas Rothenberg*, chapter 5, pages 80–108. Cambridge: Cambridge University Press.

Table 1: Proportion of Students Who Are Compliers Due to Summer Bridge Participation and Major Declaration

	Assigned to treatment P$[Z = 1]$	Treated P$[D = 1]$	Compliers $P[D_1 > D_0]$	Compliers among treated P$[D_1 > D_0 \vert D = 1]$	Compliers among controls P$[D_1 > D_0 \vert D = 0]$
Summer bridge	.17	.25	.36	.25	.40
Major declaration	.92	.25	.06	.22	.01

Table 2: Examples of Instruments

Construct	Instrument used	Endogenous regressor	Dependent variable	Study
Chance	Charter school lottery	Years in charter school	Achievement test scores	Abdulkadiroğlu et al. (2011)
	Draft lottery	Military service	Earnings	Angrist (1990)
	Experimental assignment	Program participation	College outcomes	Angrist et al. (2009)
Biology	Birth of twins	Number of children	Economic outcomes	Bronars and Grogger (1994)
	Season of birth	Years of education	Earnings	Angrist and Kreuger (1991)
	Sex mix of children	Public housing usage	K-12 outcomes	Currie and Yelowitz (2000)
Geography	Miles from college	Attend community college	BA degree completion	Long and Kurlaender (2009)
	Miles from hospital	Cardiac catheterization	Mortality	McClellan et al. (1994)
	Rainfall	Economic growth	Civil war	Miguel et al. (2004)
	Streams	School choice	Student achievement	Hoxby (2000a)
Laws & rules	Drinking age	Teen drinking	Educational attainment	Dee and Evans (2003)
	Maximum class size rule	Class size	Student achievement	Hoxby (2000b)
	Prison litigation	Prisoner population	Crime	Levitt (1996)
	State cigarette taxes	Maternal smoking	Birth weight	Evans and Ringel (1999)

Figure 1: Correlation Between Instruments and Determinants of Academic Performance

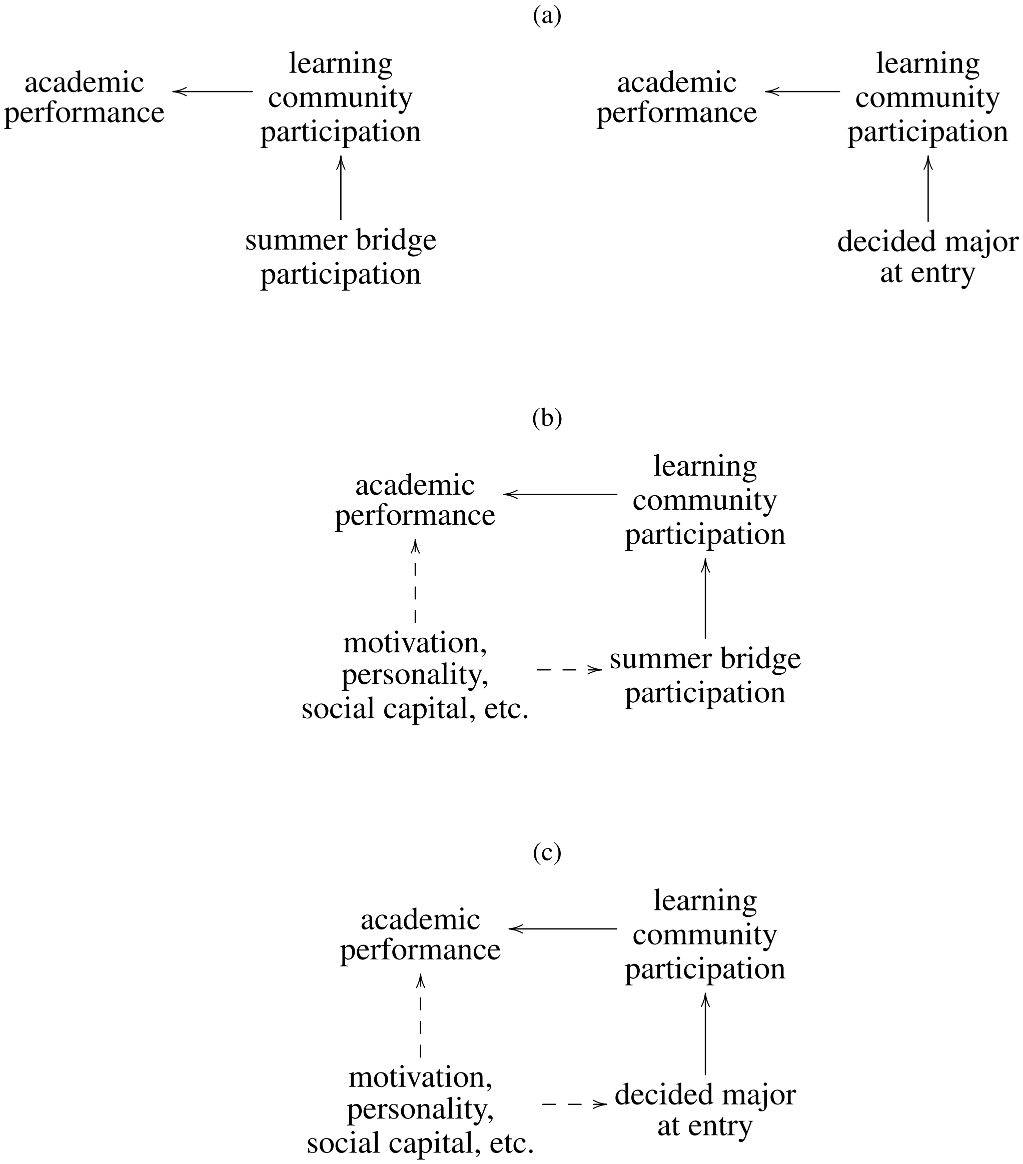

(a)
academic performance
learning community participation
summer bridge participation
academic performance
learning community participation
decided major at entry
(b)
academic performance
learning community participation
motivation, personality, social capital, etc.
summer bridge participation
(c)
academic performance
learning community participation
motivation, personality, social capital, etc.
decided major at entry

Figure 2: Causal Paths Between Instruments and Academic Performance

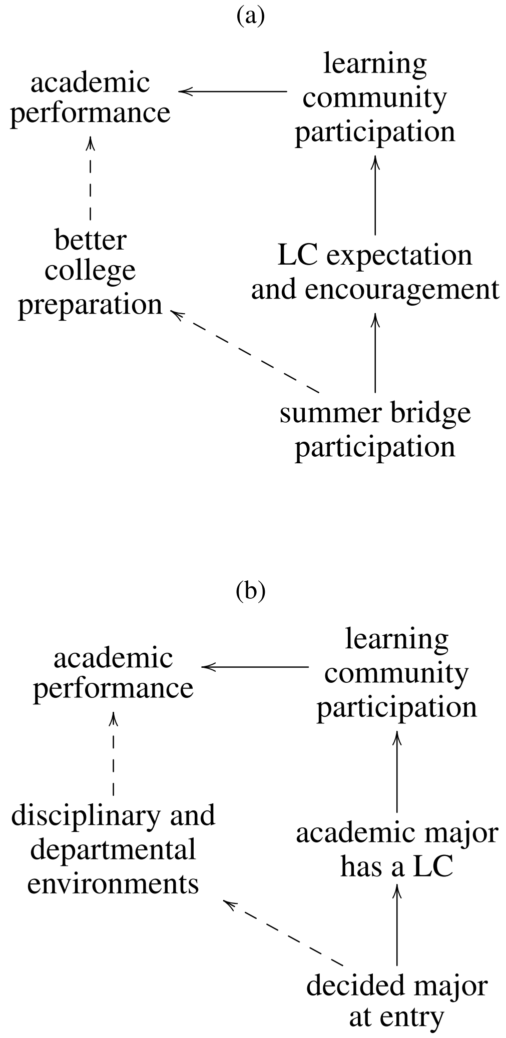
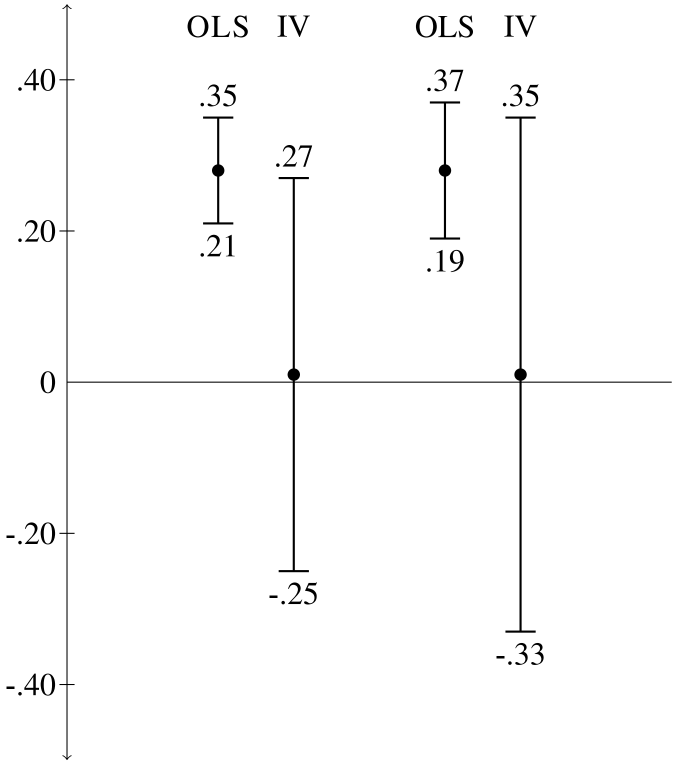

(a)
academic performance
learning community participation
better college preparation
LC expectation and encouragement
summer bridge participation
(b)
academic performance
learning community participation
disciplinary and departmental environments
academic major has a LC
decided major at entry

Figure 3: Confidence Intervals for OLS and IV Estimates of the Effect of Learning Communities



Note: Left pair of confidence intervals use $\alpha = .05$; right pair use $\alpha = .01$.