Alabama Law Review, 2011, V 63 #1, 43-95

DO FAITH-BASED PRISONS WORK?

Alexander Volokh[*]

ABSTRACT

This Article examines everything we know about the effectiveness of faith-based prisons, which is not very much.

Most studies cannot be taken seriously because they are tainted by the "self-selection problem." It is hard to determine the effect of faith-based prison programs because they are voluntary, and volunteers are more likely to be motivated to change and are therefore already less likely to commit infractions or be re-arrested. This problem is the same one that education researchers have struggled with in determining whether private schools are better than public schools.

The only credible studies done so far compare participants with non-participants who volunteered for the program but were rejected. Some studies in this category find no effect, but some do find a modest effect. But even those that find an effect are subject to additional critiques: for instance, participants may have benefited from being exposed to treatment resources that non-participants were denied.

Thus, based on current research, there is no strong reason to believe that faith-based prisons work. However, there is also no strong reason to believe that they do not work. I conclude with thoughts on how faith-based prison programs might be improved, and offer a strategy that would allow such experimentation to proceed consistent with the Constitution.

* Assistant Professor, Emory Law School, avolokh@emory.edu. I am grateful to Steve Aos, Scott D. Camp, Aimee Chin, Hashem Dezhbakhsh, Griffin Sims Edwards, Esfandiar Maasoumi, Marna Miller, Thomas P. O'Connor, Elena Pesavento, Joanna E. Saul, George B. Shepherd, Joanna M. Shepherd, Christopher R. Taber, and Eugene Volokh. I am also grateful to Margaret Ada Sapozhnikov, Matthew E. Uretsky, and the librarians at Emory Law School for their able research assistance.

I. INTRODUCTION

There are five things one should know about faith-based prisons:

(1) There are a lot of them out there. As of 2005, eighteen states and the federal government had some sort of residential faith-based program,[1] aimed at rehabilitating participating inmates by teaching them subjects like "ethical decision-making, anger management, victim restitution,"[2] and substance abuse[3] in conjunction with religious principles.

1. U.S. Dep't of Justice, National Institute of Corrections, *Residential Faith-Based Programs in State Corrections*, SPECIAL ISSUES IN CORRECCTIONS, Sept. 2005, http://static.nicic.gov/library/020820.pdf.
2. *Programs*, U.S. DEP'T OF JUST. ARCHIVE, http://www.justice.gov/archive/fbci/progmenu_programs.html#3 (last visited Oct. 26, 2011).
3. Ams. United for Separation of Church & State v. Prison Fellowship Ministries, 509 F.3d 406, 415–16 (8th Cir. 2007).

(2) One of them—the InnerChange Freedom Initiative program in Iowa—was struck down on Establishment Clause grounds in 2006,[4] but various faith-based prison programs still exist,[5] including InnerChange programs in other states.[6] InnerChange programs, which are explicitly motivated by Christian and Biblical principles,[7] are probably more vulnerable to constitutional challenges; programs that are more interfaith and have less explicitly religious content, like Florida's Faith- and Character-Based Institutions[8] or the federal Life Connections Program,[9] are probably less so.

(3) Faith-based prisons continue to be promoted as promising avenues for reform, chiefly on the grounds that they improve prison discipline and reduce recidivism.[10]

(4) However, most of the empirical studies of the effectiveness of faith-based prisons have serious methodological problems and, to the extent they find any positive effect of faith-based prisons, can't be taken at face value.

(5) Those few empirical studies that approach methodological validity either fail to show that faith-based prisons reduce recidivism or provide weak evidence in their favor.

In what follows, I explain and critically evaluate the empirical studies of the effectiveness of faith-based prisons. The reader who gets through this Article will know everything that we currently know about whether they "work," by which I mean, chiefly, whether they reduce in-prison infractions or some measure of post-release recidivism, such as time to re-arrest, probability of re-arrest, or probability of reconviction.

As the summary above indicates, we don't know much about the effectiveness of faith-based prisons. This is a shame, because the empirics of faith-based prisons are important, both to the legal community and to policymakers generally.

---

4. Ams. United for Separation of Church & State v. Prison Fellowship Ministries, 432 F. Supp. 2d 862, 933 (S.D. Iowa 2006), *aff'd*, 509 F.3d 406, 425 (8th Cir. 2007).

5. *E.g.*, *Florida Expanding Faith-Based Prisons*, UPI.COM (May 11, 2009, 9:41 AM), http://www.upi.com/Top_News/2009/05/11/Florida-expanding-faith-based-prisons/UPI-59611242049261/.

6. *E.g.*, *Programs*, INNERCHANGE FREEDOM INITIATIVE, http://www.ifiprison.org/state-programs (last visited Oct. 26, 2011).

7. *Ams. United for Separation of Church & State*, 509 F.3d at 413–14.

8. *Faith- and Character-Based Institutions*, FLA. DEP'T OF CORRECTIONS, http://www.dc.state.fl.us/oth/faith/ci.html (last visited Oct. 11, 2011).

9. U.S. DEP'T OF JUST. ARCHIVE, *supra* note 2.

10. *See, e.g.*, Marti W. Harkness, Staff Dir., Fla. Legislature Office of Program Policy Analysis & Gov't Accountability, Review of the Dep't of Corrections' Faith-Based Prisons (Jan. 25, 2011), http://www.oppaga.state.fl.us/monitordocs/reports/pdf/DOC_Faith-based_prisons.pdf.

First, we should care about the empirics because faith-based prisons and rehabilitative services are, and will continue to be, on the cutting edge of Establishment Clause litigation,[11] and empirics matter in the law. One might think that whether a program works shouldn't matter to whether it's consistent with the Establishment Clause cases; but in fact, there are several areas in Establishment Clause doctrine that seem to allow targeted uses of statistical evidence.[12]

In any event, regardless of whether empirics *should* matter in Establishment Clause cases, reading judicial opinions suggests that they *do*.

In *Americans United for Separation of Church & State v. Prison Fellowship Ministries*, which invalidated Iowa's contract with the InnerChange Freedom Initiative, District Judge Robert Pratt, immediately before launching into his Establishment Clause analysis, concluded his description of the faith-based program with the following complaint:

> More significant [than the warden's personal testimony about the program's beneficial in-prison effect], however, is the lack of evidence presented by the Defendants about the effect of InnerChange on recidivism. Aside from anecdotes, the Defendants offered no definitive study about the actual effects the InnerChange program has on recidivism rates. [The warden's] predecessor . . . communicated his desire early on in the initial RFP process that accountability for the program be included in the contractual agreement between the parties. Specifically, he requested "at least annual program evaluations to include, but not limited to, re-incarceration rates and other measurable outcomes." But, in fact, there was no information presented at trial about whether InnerChange participants are more or less prone to recidivism than other inmates.[13]

Of course, finding empirical language in opinions isn't definitive evidence that the empirics are relevant. Maybe judges who claim to care about the effectiveness of faith-based programs are just indulging in legally meaningless rhetoric designed to support a conclusion they already arrived at by strictly legal means. On the other hand, if judges discuss empirical data,

11. *See, e.g.*, *Ams. United for Separation of Church & State*, 509 F.3d 406; Teen Ranch, Inc. v. Udow, 479 F.3d 403 (6th Cir. 2007); Freedom from Religion Found., Inc. v. McCallum, 324 F.3d 880 (7th Cir. 2003); Williams v. Lara, 52 S.W.3d 171 (Tex. 2001); *see also* Ira C. Lupu & Robert W. Tuttle, *The Faith-Based Initiative and the Constitution*, 55 DEPAUL L. REV. 1 (2005); Ira C. Lupu & Robert W. Tuttle, Zelman*'s Future: Vouchers, Sectarian Providers, and the Next Round of Constitutional Battles*, 78 NOTRE DAME L. REV. 917 (2003); Ira C. Lupu & Robert Tuttle, *Sites of Redemption: A Wide-Angle Look at Government Vouchers and Sectarian Service Providers*, 18 J.L. & POL. 539 (2002).

12. In a voucher context, they could help establish that a choice was "genuine." In a coercion context, they could help establish whether there was subtle pressure to join a program, or whether instead participants joined a program because of valid quality reasons. Effectiveness studies could help establish a secular purpose in cases where the purpose is doubtful. *Cf.* Perry v. Schwarzenegger, 704 F. Supp. 2d 921, 931–32 (N.D. Cal. 2010).

13. 432 F. Supp. 2d 862, 914 (S.D. Iowa 2006) (citation omitted).

chances are that they believe the extra rhetorical force is useful in persuading *someone* and thus makes their opinion more influential and less subject to reversal.[14]

Moreover, if we put our Legal Realist hats on,[15] maybe it's the judges' views of effectiveness that are driving their legal conclusions (at least for *some* judges, who aren't in favor of or against faith-based prisons on ideological grounds). At the very least, if judges believe that a program is effective, they may think it's a shame if the program were to be found unconstitutional and might therefore be extra careful in their legal analysis to avoid striking it down. (If we disagree with their conclusion, we might replace the word *careful* in the previous sentence with the word *imaginative*.) Conversely, if they believe a program is ineffective, they may feel no particular pressure to uphold it.

Perhaps a better piece of evidence that the empirics matter is that lawyers spend time discussing empirical studies in their briefs. Perhaps in response to Judge Pratt's concern, when the case came up to the Eighth Circuit, the Alliance Defense Fund and others submitted an amicus brief arguing that "InnerChange's faith-based rehabilitative prison programs are proven to reduce recidivism."[16] (Unfortunately, of the two studies cited favorably in their brief, one has serious methodological problems,[17] while the other, properly interpreted, shows no beneficial effect of the program.[18])

Empirical data also seems important in other Establishment Clause contexts. In *Zelman v. Simmons-Harris*, which upheld an Ohio program that allowed publicly funded vouchers to be redeemed at religious schools, Justice Souter's dissent cited statistics on the academic performance of schools to judge whether the parents' choices were "genuine."[19] Justice O'Connor's concurrence similarly used performance statistics to rebut Justice Souter's argument.[20]

Justice Thomas, for his part, noted that "the success of religious and private schools is in the end beside the point, because the State has a constitutional right to experiment with a variety of different programs to promote educational opportunity."[21] But that didn't stop him from marshalling statistics to argue that "[r]eligious schools, like other private schools, achieve far

14. *See, e.g.*, Alexander Volokh, *Choosing Interpretive Methods: A Positive Theory of Judges and Everyone Else*, 83 N.Y.U. L. REV. 769, 792–97 (2008) (discussing how judges use rhetoric to, among other things, maximize persuasiveness and minimize reversal).

15. *E.g.*, Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457 (1897).

16. Brief for Alliance Defense Fund et al. as Amici Curiae Supporting Defendant-Appellants at 15, *Ams. United for Separation of Church & State*, 509 F.3d 406 (No. 06-2741), 2006 WL 2923982 (citing JOHNSON & LARSON, *infra* note 282; Johnson, *infra* note 201). The Alliance Defense Fund was joined on this brief by the National Association of Evangelicals, the Center for Neighborhood Enterprise, Teen Challenge, Time to Fly, the Center for Public Justice, Evangelicals for Social Action, and the Coalition to Preserve Religious Freedom.

17. *See infra* text accompanying notes 201–12 (discussing Johnson, *infra* note 201).

18. *See infra* text accompanying notes 282–290 (discussing JOHNSON & LARSON, *infra* note 282).

19. 536 U.S. 639, 702–03 n.10 (2002) (Souter, J., dissenting).

20. *Id.* at 673–76 (O'Connor, J., concurring).

21. *Id.* at 681 (Thomas, J., concurring).

better educational results than their public counterparts."[22] "That Ohio's program includes successful schools," Justice Thomas wrote, "simply indicates that such reform can in fact provide improved education to underprivileged urban children."[23] Here, too, seasoned appellate litigators spent several pages of their Supreme Court brief discussing the history and performance of voucher programs[24]—and this was in a party's argument, not just an amicus brief[25]—though this discussion was ostensibly *not* "to convince the Court that parental choice is proper public policy."[26]

But perhaps more importantly, we should care about the empirics because, whether or not they should matter in the law,[27] they obviously should matter in policy. If faith-based prisons don't reduce recidivism, the case for their funding is correspondingly weakened (though they may still be supported by other arguments).[28] If they do reduce recidivism, or if they have other policy advantages, then even if judges are wrong to stretch the law to find them constitutional, *we* aren't wrong to try to find ways to allow them to function constitutionally.[29] After all, even the non-religious have an interest in the rehabilitation of prisoners, and if religion can play a positive role in rehabilitation, this may be good news not only to the irreligious but even to those who are hostile to religion.

It is thus unfortunate that the legal literature hasn't done a good job evaluating the empirical evidence on faith-based prisons. The law reviews are devoid of any comprehensive, critical discussion of the existing studies. Most legal articles on the subject simply choose not to bother with empirical

22. *Id.*

23. *Id.*

24. Brief for Petitioners, *Zelman*, 536 U.S. 639 (No. 00-1779), 2001 WL 1663809 at *19–25. The seasoned appellate litigators were former Solicitor General and Harvard Law School professor Charles Fried, Institute for Justice litigators Clint Bolick and William H. Mellor, and Cleveland appellate litigator David Tryon. *Id.*

25. The other side of the empirical voucher debate in *Zelman* was, however, represented by amici. *E.g.*, Brief for Nat'l Sch. Bds. Ass'n et al. as Amici Curiae Supporting Respondents at *9–14, *Zelman*, 536 U.S. 639 (Nos. 00-1751, 00-1777, 00-1779), 2001 WL 34092026 (relevant pages missing in Westlaw), *also available at* http://nsba.org/SecondaryMenu/COSA/Search/AllCOSAdocuments/SimmonsHarrisvZelmanUSSupCtdup.aspx.

26. Brief for Petitioners, *supra* note 24, at *25; *cf.* Brief of Respondents at 40, *Zelman*, 536 U.S. 639 (Nos. 00-1751, 00-1777, 00-1779), 2001 WL 1636772, *available at* http://supreme.lp.findlaw.com/supreme_court/briefs/00-1751/00-1751.mer.simmons.pdf.

27. *Cf., e.g.*, Brief for Respondents, *Zelman, supra* note 26, at 41.

28. *See infra* note 370.

29. I discuss such a way briefly in the Conclusion. *See generally* Alexander Volokh, *The Constitutional Possibilities of Prison Vouchers*, 72 OHIO ST. L.J. (forthcoming December 2011).

30. *See* Lynn S. Branham, *"The Devil Is in the Details": A Continued Dissection of the Constitutionality of Faith-Based Prison Units*, 6 AVE MARIA L. REV. 409, 433–35 (2008); James A. Davids, *Putting Faith in Prison Programs, and Its Constitutionality Under Thomas Jefferson's Faith-Based Initiative*, 6 AVE MARIA L. REV. 341, 350–63 (2008); Marc O. DeGirolami, *The New Religious Prisons and Their Retributivist Commitments*, 59 ARK. L. REV. 1, 19–21 (2006); Alex J. Luchenitser, *"InnerChange": Conversion as the Price of Freedom and Comfort—A Cautionary Tale About the Pitfalls of Faith-Based Prison Units*, 6 AVE MARIA L. REV. 445, 470–73 (2008); Nathaniel Odle, *Privilege Through Prayer: Examining Bible-Based Prison Rehabilitation Programs Under the Establishment Clause*, 12 TEX. J. C.L. & C.R. 277, 302–05 (2007).

studies. Some legal articles do address the empirical evidence,[30] but their discussions are generally quite cursory.[31]

Outside of the legal literature, a few review articles do take a broader approach.[32] Some even single out which articles they believe are methodologically more valid than others.[33] But, unfortunately for other scholars, these generally have no in-depth discussion of the studies being reviewed, nor do they discuss *why* the studies identified as superior really are superior. The rest of us are thus left to either take them at their word (I myself disagree with some of their assessments)[34] or track down the studies (many of which are hard to find) and read them ourselves—a daunting task for those without empirical training.[35]

This Article fills the void. More specifically, this Article makes three distinct contributions.

First, I provide a detailed critical discussion of (to my knowledge) every existing empirical study on the effectiveness of faith-based prison programs.

The word *every* in the last sentence masks a few critical decisions. In making such a survey, I have chosen to include certain kinds of studies and exclude others. Because the relevant legal issues center around "immersion"-style faith-based prison units "that seek to immerse prisoners in an almost monastic or total experience of religiously based living,"[36] I exclude studies that explore more general issues like the effect of "religiosity."[37] The

31. The most comprehensive law review treatment of the empirical studies discusses five different empirical studies of faith-based prison programs. Davids, *supra* note 30, at 350–63 (citing Johnson et al., *infra* note 193; Johnson, *infra* note 201; JOHNSON & LARSON, *infra* note 282; JEANETTE HERCIK ET AL., DEVELOPMENT OF A GUIDE TO RESOURCES ON FAITH-BASED ORGANIZATIONS IN CRIMINAL JUSTICE 161 (2004), www.ncjrs.gov/pdffiles1/nij/grants/209350.pdf (discussing the results of HERCIK ET AL., *infra* note 307); *id.* at 32 (citing O'Connor et al., *infra* note 342)). But it accepts their positive conclusions largely uncritically, whereas only three of these studies approach methodological soundness. (I discuss JOHNSON & LARSON, *infra* note 282, HERCIK ET AL., *infra* note 307, and O'CONNOR ET AL., *infra* note 331, in Part III *infra*.) Moreover, of those three, two, properly interpreted, show no effect, *see infra* text accompanying notes 282 and 307, and the third shows, at most, weak effects. *See infra* text accompanying note 331; *see also infra* Part IV (resources problem).

32. *E.g.*, Daniel P. Mears et al., *Faith-Based Efforts to Improve Prisoner Reentry: Assessing the Logic and Evidence*, 34 J. CRIM. JUST. 351 (2006).

33. *E.g.*, STEVE AOS ET AL., EVIDENCE-BASED ADULT CORRECTIONS PROGRAMS: WHAT WORKS AND WHAT DOES NOT (Wash. State Inst. for Pub. Policy, Jan. 2006), http://www.wsipp.wa.gov/rptfiles/06-01-1201.pdf; Thomas P. O'Connor, *What Works, Religion as a Correctional Intervention: Part II*, 14 J. CMTY. CORRS. 4, Winter 2004–05, *available at* http://www.oregon.gov/DOC/TRANS/docs/pdf/rs_whatworks2.pdf?ga=t.

34. Tom O'Connor, *see* O'Connor, *supra* note 33, is himself a contributor to the literature and evaluates the methodology of some of his own studies. I disagree with O'Connor's assessment of one of the studies, which he ranks "good," *see* O'Connor et al. *infra* note 65, but which I lump in with the studies marred by "naked self-selection." I also disagree with some of Aos's assessments. *See* AOS *infra* notes 111, 135, & 201.

35. Several of these studies are hard to find, and I've had to obtain them directly from the authors.

36. Thomas P. O'Connor & Jeff B. Duncan, *Religion and Prison Programming: The Role, Impact, and Future Direction of Faith in Correctional Systems*, 11 OFFENDER PROGRAMS REP. 81, 86 (2008); *see also* Thomas P. O'Connor et al., *Criminology and Religion: The Shape of an Authentic Dialogue*, 5 CRIMINOLOGY & PUB. POL'Y 559, 563 (2006) (similar).

37. Todd R. Clear et al., *Does Involvement in Religion Help Prisoners Adjust to Prison?*, NCCD FOCUS, Nov. 1992 (measuring religiousness by the Prisoner Values Survey); Todd R. Clear & Marina Myhre, *A Study of Religion in Prison*, 6 INT'L ASS'N RES. & CMTY. ALTS. J. ON CMTY. CORRS., no. 6, 1995 at 20 (attitudinal measures of religion); Todd R. Clear & Melvinia T. Sumter, *Prisoners, Prison,*

measures of religiosity, in these studies, are either general measures of how religious an inmate feels or how many times an inmate attends religious services in prison. These studies may be valuable for some purposes, but they don't help in evaluating immersion-style faith-based units since ordinary prison worship services are both widespread and uncontroversial.

I instead focus on studies of the effectiveness of specific faith-based interventions on variables of interest like the likelihood of recidivism. I also include studies of religious after-care for released inmates, even though these aren't technically in-prison programs.[38]

Second, I provide a detailed discussion of the methodological issues involved in evaluating faith-based prisons generally. In statistics, methodology is everything; it's a shame that the legal community, which often relies on these empirical studies, isn't as sophisticated as it could be at telling valid studies apart from invalid ones.

Roughly speaking, the studies fall into four categories. Three of them—naïve comparisons of participants to non-participants, comparisons with some controls, and matching based on propensity scores—aren't credible because they don't account for what is, in my view, the most serious obstacle to effective assessment: the "self-selection problem." Inmates who are motivated enough to choose to participate in a rehabilitative program are already less likely to reoffend. So any study that compares voluntary participants and voluntary non-participants may just be picking up the effect of being a good person, not the effect of the program itself. (Some of these

and Religion: Religion and Adjustment to Prison, 35 J. OFFENDER REHAB., nos. 3 & 4, 2002, at 125 (measuring religiousness by the Prisoner Values Survey); Byron R. Johnson, *Religiosity and Institutional Deviance: The Impact of Religious Variables upon Inmate Adjustment*, 12 CRIM. JUST. REV. 21 (1987) (using self-reported religiosity, chaplain-reported religiosity, and church attendance); Byron R. Johnson et al., *A Systematic Review of the Religiosity and Delinquency Literature: A Research Note*, 16 J. CONTEMP. CRIM. JUST. 32 (2000) (meta-analysis concentrating on church attendance, prayer, religious salience, and Bible study, though including some other "religious activities"); Thomas P. O'Connor & Michael Perreyclear, *Prison Religion in Action and Its Influence on Offender Rehabilitation*, 35 J. OFFENDER REHAB., nos. 3 & 4, 2002, at 26–27 ("rate of religious participation" doesn't distinguish between ordinary services and retreats); Michael G. Pass, *Religious Orientation and Self-Reported Rule Violations in a Maximum Security Prison*, 28 J. OFFENDER REHAB., nos. 3 & 4, 1999, at 119 (attitudinal measures of religion); Thomas P. O'Connor, A Sociological and Hermeneutical Study of the Influence of Religion on the Rehabilitation of Inmates (2003) (unpublished Ph.D. dissertation, Catholic University of America) (on file with author) (count of religious attendance generally); Melvina T. Sumter, Religiousness and Post-Release Community Adjustment (1999) (Ph.D. dissertation, Florida State University), https://www.ncjrs.gov/pdffiles1/nij/grants/184508.pdf (using Prisoner Values Survey and similar attitudinal variables).

38. I have also excluded programs whose religiosity is unclear. Transcendental Meditation, for instance, is considered religious by some and non-religious by others. *Compare, e.g.*, Amy Karasz & Meaghan Midgett, *Transcendental Meditation*, http://web.archive.org/web/20060831081613/religious movements.lib.virginia.edu/nrms/tm.html (last modified Jan. 12, 2001), *with The Technique*, THE TRANSCENDENTAL MEDITATION PROGRAM, http://www.tm.org/meditation-techniques (last visited Oct. 26, 2011). *Cf.* Debbie Elliott, *At End-of-the-Line Prison, An Unlikely Escape*, NPR, http://www.npr.org/2011/02/08/133505880/at-end-of-the-line-prison-an-unlikely-escape (last visited Oct. 26, 2011) ("The Vipassana technique, though secular, is based on the teachings of Buddha."). In any event, Transcendental Meditation is quite different from the other sorts of religious programs canvassed in this survey. Nonetheless, a number of studies purport to find that Trancendental Meditation reduces recidivism. Below, *see infra* note 161, I cite two studies that belong in the "Studies with Some Controls" Part—that is, studies that do not adequately control for self-selection.

studies are subject to even further sources of bias. For instance, in addition to self-selection in the decision whether and how intensively to participate, there can be selection by the program staff in the decision of whom to admit or whom to kick out,[39] as well as "success bias" in the consideration only of those who completed the program without dropping out.[40])

In my view, the only credible studies so far fall into a fourth category—those that compare (voluntary) participants in faith-based programs with people who volunteered for the program but were rejected.[41] And even these studies are subject to the "resources problem": they compare participation in the program either with the alternative of no program at all or with the "business as usual" alternative of whatever other programs happened to be available, rather than with participation in a comparably funded secular program. Thus, even if a religious program is better than nothing at all, it could be because of the greater access to treatment resources—for instance, mentors and counselors—and not because of the religious content of the program.[42]

Third, I set the empirical debate on faith-based prisons side by side with a parallel empirical debate: whether private schools are better than public schools. One striking aspect of the faith-based prisons research is how much it looks like the private school research. (Some of this research studies not private schools as such, but Catholic schools, since most private schools are religious, and Catholic schools are a homogeneous enough group to be susceptible to generalization.) Both faith-based prisons and private schools are subject to self-selection—any naïve comparison between, on the one hand, secular prisons or public schools and, on the other hand, faith-based prisons or private/Catholic schools is subject to the critique that the private or religious options have to be affirmatively chosen, and people who are motivated enough—or whose parents are motivated enough—to make that choice are already more likely to be high-achieving students or low-recidivism inmates.[43]

While sophisticated researchers in both areas are aware of the self-selection problem, the education literature has addressed it far more con-

39. *E.g.*, the Brazil study, *infra* text accompanying notes 52–64; the theology study, *infra* text accompanying notes 65–68; and the Texas InnerChange studies, *infra* text accompanying notes 282–290.

40. *E.g.*, the Brazil study, *infra* text accompanying notes 52–64; the Florida DOC's Kairos Horizons study, *infra* text accompanying notes 77–84; the Texas InnerChange studies, *infra* text accompanying notes 282–290; Wilson et al.'s Detroit TOP study, *infra* text accompanying notes 319–330; and O'Connor et al.'s TOP study, *infra* text accompanying notes 331–342.

41. There are other promising statistical techniques, but so far they apparently haven't been used in the faith-based prison context. *See infra* Part III.A.

42. *See infra* Part IV.

43. An analogous problem occurs in various studies of the effectiveness of private prisons. A comparison of public to private prisons is biased if the private prisons studied, rather than getting a random sample of inmates, were sent a sample of inmates that was systematically healtheir or less dangerous than average. *See, e.g.*, Richard A. Oppel Jr., *Private Prisons Found to Offer Little in Savings*, N.Y. TIMES, May 18, 2011 ("'It's cherry picking,' said State Representative Chad Campbell, leader of the House Democrats. 'They leave the most expensive prisoners with taxpayers and take the easy prisoners.'").

scientiously and has progressed further than the prisons literature in advancing our empirical knowledge. Perhaps prison researchers could learn something from school researchers' example.

In the end, this Article has bad news and good news.

The bad news, as I've said above, is that most studies are low quality, and the results of the higher quality studies aren't promising. There seems to be little empirical reason to believe that faith-based prisons work.

The good news is that there's also no proof that they *don't* work. The absence of statistically valid or statistically significant findings isn't the same as the presence of negative findings. And while the self-selection problem is real and important, the resources problem may not even be a problem at all: maybe the "zero alternative" or the "business as usual" alternatives really are proper empirical baselines, since they reflect both reality and, perhaps, political feasibility. So the picture isn't uniformly bleak: there are *some* programs that seem to show *some* statistically significant effects, even if they are weak and even if we're not sure how well they compare to the hypothetical effects of a hypothetical, comparably funded secular program.

Perhaps future research will shed light on these questions. In the meantime, clearly some groups want to have such prisons, some inmates want to attend them, and they probably do little if any harm. If some programs don't work, this is an indication to future practitioners that something needs to be changed; if some programs work, maybe they can be replicated elsewhere. Better results won't emerge unless they're allowed to emerge by a process of experimentation. As I've suggested above, it would be a shame if this process is cut off for constitutional reasons, provided there is a constitutionally valid way for them to proceed.

At the end of this Article, I suggest such a way. Faith-based prisons, as currently constituted, are probably unconstitutional under modern Establishment Clause doctrine. But they would become fully constitutional under a system of prison vouchers that would allow inmates to choose their own prisons, whether secular or religious. I develop this idea at length elsewhere,[44] but the bottom line is that, despite the current weak evidence, some version of faith-based prisons may still work, and there is a way for that version to emerge consistent with the Constitution.

II. THE SELF-SELECTION PROBLEM

The most serious problem with studies of the effectiveness of faith-based prisons is the self-selection problem. Prisoners obviously select into faith-based prisons voluntarily.[45] And the factors that would make an inmate

44. Volokh, *supra* note 29.

45. At least in the United States, this would be so clearly required by the Establishment Clause that, even if the programs are ultimately found to be unconstitutional, this is a feature that the designers of such programs would be sure to include.

select a faith-based prison may also make him less likely to commit crimes in the future.[46] One such factor might be religiosity itself.[47] In addition, an inmate who takes the trouble to choose to join a rehabilitative program may be more motivated and more open to change, and this may itself make him more likely to change—regardless of whether the program actually "works."

The following Parts illustrate three types of studies that don't adequately control for self-selection, both for faith-based prisons and for the analogous context of private/Catholic schools.[48]

The first type of study shows the self-selection problem in its most naked form: it simply compares the results of participants in a faith-based program with those of non-participants.

The second type of study accounts for some of the differences between participants and non-participants by comparing the group of participants with a matched group of non-participants, where the matching is based on various observable factors like race, age, criminal history, and the like.[49] But, of course, such a procedure can't control for *unobservable* variables, like motivation to change.

The third type of study uses a more sophisticated statistical technique called "propensity score" matching. Participants are matched to non-participants not based on observable factors directly, but based on their propensity score, that is, their estimated probability of participating in the program. While propensity scores are a useful technique in some applications, they don't alleviate the self-selection problem in the faith-based prison context.

A. Naked Self-Selection

The studies in this section purport to find a positive effect of faith-based prisons based on comparing, say, recidivism rates of participants in faith-based units and prisoners in the general population or in different prisons. But these sorts of studies aren't credible because they make no effort to control for self-selection. "Without knowledge of the selection process, there is no way to determine whether observed differences between program participants and 'comparisons' are due to actual program effects or are an

46. Scott D. Camp et al., *An Exploration into Participation in a Faith-Based Prison Program*, 5 CRIMINOLOGY & PUB. POL'Y 529, 534 (2006).

47. *E.g.*, sources cited *supra* note 37. I don't want to oversell this point, since it takes no imagination at all to imagine religiously inspired violence; moreover, one observes associations between white prison gangs and forms of Christianity or neopaganism (Odinism or Asatru), and between black prison gangs and forms of Islam.

48. *See supra* text accompanying note 42.

49. Matching—a method from epidemiology—is commonly used in faith-based prison studies. In education studies, instead of matching, generally we have a large number of observations of participants and non-participants; no one is "matched" to anyone else, but the researchers "control" for observable factors in a statistical sense, by including those factors as independent variables in a regression.

artifact of preexisting differences between the groups."[50] Rather than giving us the effect of faith-based *prisons*, these studies may be giving us the effect of faith-based *prisoners*.[51]

1. Johnson's Brazil Study

Byron Johnson compared recidivism among inmates in two Brazilian prisons: Humaitá, a faith-based facility, and Bragança, a secular facility with vocational training programs.[52] Data wasn't available for 46% of the inmates, though the data loss didn't differ significantly between the two prisons.[53] High-risk Humaitá inmates had significantly[54] lower recidivism—12% of the high-risk Humaitá inmates were re-arrested after three years, versus 38% of the high-risk Bragança inmates.[55] The average number of re-arrests was also significantly lower for Humaitá prisoners[56]—even though on average the original offenses of the Humaitá prisoners had been more serious, they were more likely to be violent, and they had possibly served more time in prison.[57]

The main problem with this study is that prisoners apply to be in Humaitá, prisoners' families must be "involved in the prisoner's recuperation process," prisoners aren't accepted without sufficient "motivation and commitment to change," and prisoners don't stay unless they and the prison agree after an initial 60-day assessment period.[58] The results are thus tainted by multiple sources of bias: self-selection, selection by the prison itself, and success through the assessment period.

Moreover, among low-risk inmates, recidivism rates weren't significantly different between the two prisons.[59] There was no significant differ-

50. Camp et al., *supra* note 46, at 529; *see also* James J. Heckman & Richard Robb, *Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes*, *in* DRAWING INFERENCES FROM SELF-SELECTED SAMPLES 63, 77–78 (Howard Wainer ed., 1986) (explaining why a mere comparison of sample means does not yield the treatment effect when the decision to get treatment is correlated with unobservable variables).

51. *Cf.* Volokh, *supra* note 14, at 775 (because of failure to account for self-selection, "many statements about *textualism* may really only be statements about *textualists*").

52. Byron R. Johnson, *Assessing the Impact of Religious Programs and Prison Industry on Recidivism: An Exploratory Study*, TEX. J. CORR., Feb. 2002, at 7. O'Connor rates this study as having "fair" methodological quality (on a poor-fair-good-excellent scale). O'Connor, *supra* note 33, at 23 tbl.3. The study doesn't say how inmates arrived in the faith-based facility, but one suspects that they chose to be there.

53. Johnson, *supra* note 52, at 9.

54. In this Article, I use the term *significant* to mean "statistically significant at the 5% level."

55. Johnson, *supra* note 52, at 9 tbl.1. The p-value of the differences between high-risk inmates (that is, the level of statistical significance) is less than 0.01 (that is, it's highly significant). *Id.*

56. *Id.* at 9. Humaitá prisoners averaged 0.23 re-arrests, compared to 0.53 for Bragança prisoners. *Id.* The p-value is less than 0.01. *Id.*

57. *Id.* at 8.

58. ANGUS CREIGHTON, HUMAITA PRISON 5 (1998), http://www.pfi.org/cjr/apac/where1/reports/brazil/humaita2/at_download/file; Jonathan Burnside, *The Prison That Started It All*, *in* JONATHAN BURNSIDE ET AL., MY BROTHER'S KEEPER: FAITH-BASED UNITS IN PRISONS 1, 13 (2005).

59. Johnson, *supra* note 52, at 9 tbl.1. The p-value of the difference between re-arrest rates for low-risk inmates is 0.129. *Id.*

ence between times to re-arrest[60] or the severity of the subsequent offense.[61] The reincarceration rate was lower among Humaitá inmates, but "the validity of this finding is questionable due to extensive data loss."[62] Moreover, many relevant background factors, like age or criminal history, weren't considered, perhaps because the data wasn't available.[63]

Finally, Humaitá differs from other Brazilian prisons (possibly including Bragança) in many ways unrelated to religion. The environment is more pleasant, prisoners and their families are treated better, there are more (non-religious) activities, and so on.[64] Any improvements in recidivism could therefore have been caused not only by selection, but also by better secular prison conditions.

2. *O'Connor et al.'s Theology Study*

Thomas O'Connor and his coauthors compared recidivism between 54 inmates who participated in a master's program in theology at Sing Sing prison and 402 non-participants.[65] Completion of the ministry program was associated with a significantly lower risk of re-arrest in the first 28 months out of prison—only 9% of participants were re-arrested, compared to 37% of non-participants.[66]

However, both self-selection and selection by program administrators taint these results. The students were selected by "a highly competitive application and reference process"; the program was open only to inmates with a college degree, who read and wrote well, and who had "references from chaplains and other inmates attesting to their religious commitment" and showed "a deep willingness to turn their lives around."[67] In fact, according to the president of the seminary that ran the theology program, the program had "built-in success" because they made sure to accept applicants

---

60. *Id.* at 9. Humaitá reoffenders were arrested, on average, 20.2 months after release, and Bragança reoffenders were arrested, on average, 18.3 months. *Id.* The p-value of this difference is 0.68. *Id.*

61. *Id.* The severity ranking of the subsequent offense was 3.8 for Humaitá prisoners, vs. 3.0 for Bragança prisoners; the p-value of this difference was 0.10. *Id.*

62. *Id.* "Six of 10 Humaita inmates returned to prison, while 16 of 17 former prisoners from Braganca returned to prison." *Id.* This difference was significant at the 5% level based on a one-tailed Fisher's Exact Test. *Id.*

63. Gerry Rose, *The Impact of Kainos and Christian-Based Units on Recidivism*, *in* BURNSIDE ET AL., *supra* note 58, at 294, 299.

64. Burnside, *supra* note 58, at 14 tbl.1.1.

65. Thomas O'Connor et al., *Theology and Community Corrections in a Prison Setting*, CMTY. CORR. REP., July-Aug. 1997, at 67, 68. O'Connor (one of the authors of this study), in a separate work, rated this study as having "good" methodological quality (on a poor-fair-good-excellent scale). O'Connor, *supra* note 33, at 23 tbl.3; *see also* TOM O'CONNOR ET AL., RECIDIVISM AND THE MASTER IN PROFESSIONAL STUDIES PROGRAM AT SING SING PRISON: AN EXPLORATORY STUDY (Ctr. for Soc. Research, 1996).

66. O'Connor et al., *supra* note 65, at 75. In a logistic regression analysis of re-arrest within 28 months, controlling for number of months out of prison, participation in the program had a coefficient of −7.2683 with a standard error of 1.3078, which made it very highly significant, with a reported significance level of "0.0000." O'CONNOR ET AL., *supra* note 65, at 7 tbl.2.

67. O'Connor et al., *supra* note 65, at 67.

"who want to learn who they are, what they value and what they believe in."[68]

3. Kerley et al.'s Religiosity Study

Kent Kerley and his coauthors examined the relationship between religiosity and negative prison behaviors at the Mississippi State Penitentiary in Parchman, Mississippi.[69] First, they measured inmates' religiosity using a survey.[70] Most of these measures are irrelevant for our purposes because they don't involve specific programming[71]—for instance, inmates were asked whether they had experienced a conversion and whether they believed in God.[72] But one of the measures was attendance at a one-day Prison Fellowship Ministries event called Operation Starting Line, "which included Christian musicians, comedians, professional athletes, and other speakers," and which was held about six months before the survey.[73]

Participation in Operation Starting Line predicted a significantly reduced rate of *arguing* with other inmates—52.5% of participants argued with other inmates once or more per month, as opposed to 60.0% of non-participants.[74] But participants and non-participants didn't differ statistically significantly in their likelihood of fighting once or more per month—18.9% for participants versus 19.3% for non-participants.[75]

Inmates, of course, self-selected into the Starting Line events. In addition, the data was collected by a survey distributed to inmates, where both religiosity and negative behaviors were self-reported, where participation in the survey was voluntary, and where the response rate was 45%.[76]

4. The Florida DOC's Kairos Horizons Study

The Florida Department of Corrections, which ran a faith-based dorm, Kairos Horizons, at its Tomoka Correctional Institution, performed an unpublished study of the effectiveness of the program.[77] To be eligible for the dorm, an inmate had to have had no disciplinary reports in the previous six

68. O'CONNOR ET AL., *supra* note 65, at 4.
69. Kent R. Kerley et al., *Religiosity, Religious Participation, and Negative Prison Behaviors*, 44 J. FOR SCI. STUDY RELIGION 443, 446 (2005).
70. *Id.*
71. *See supra* text accompanying note 37.
72. Kerley et al., *supra* note 69, at 447.
73. *Id.*
74. *Id.* at 450 tbl.2. The p-value of this difference was less than 0.05. *Id.*
75. *Id.* The p-value of this difference was above 0.05. *Id.*
76. Id. at 446.
77. BUREAU OF RESEARCH & DATA ANALYSIS, FLA. DEP'T OF CORR., COMPARING TOMOKA CI'S FAITH-BASED DORM (KAIROS HORIZONS) WITH NON-PARTICIPANTS (2000). Note that "Kairos" (this program and that described in text accompanying notes 307–318 *infra*) and "Kainos" (the program described in text accompanying notes 111–116 *infra*) are different programs. Jonathan Burnside, *From Cursillo to Prison: The Story of Kairos*, *in* BURNSIDE ET AL., *supra* note 58, at 34, 36. The correct name of the dorm is apparently "Kairos Horizon," not "Kairos Horizons." Jonathan Burnside, *Navigating by the Heavens: Horizon Communities*, *in* BURNSIDE ET AL., *supra* note 58, at 196, 196.

months.[78] The 59 inmates who spent the entire six-month program at the faith-based dorm were compared to 8 inmates who didn't complete the six months, 741 inmates at Tomoka who didn't participate at all, and 54,997 inmates at other Florida prisons.[79] (The comparison groups were also limited to inmates without disciplinary reports in the previous six months.)[80]

Inmates who completed the six-month program had lower rates of disciplinary reports than non-participants or inmates at other Florida prisons; about 5% of completers received disciplinary reports, compared to 37.5% of non-completers, 17% of non-participants, and 12% of inmates at other prisons.[81] If—to see the effect of participation rather than the effect of program completion—we lump non-completers and completers together, the rate becomes about 9%,[82] which isn't significantly different from the rate among non-participants at Tomoka or at other prisons.[83]

A similar faith-based program in England also reports greater disciplinary improvement among program participants.[84]

5. *Denny's Kairos Horizon Study*

Dan Denny analyzed in-prison misconduct and post-release recidivism rates for participants in a Kairos Horizon program at the Davis Correctional Facility, a private, medium-security prison in Oklahoma.[85]

---

78. BUREAU OF RESEARCH & DATA ANALYSIS, *supra* note 77, at 2. However, four inmates were allowed into the program even though they had disciplinary reports in the previous six months. *Id.*

79. *Id.* at 2, 9.

80. *Id.* at 9.

81. *See id.*

82. *Id.* Of the 8 non-completers, 3 got disciplinary reports, while of the 59 completers, there were also 3 that got disciplinary reports. *Id.* Together, that makes 6 inmates receiving disciplinary reports out of 67 participants. *Id.*

83. The paper doesn't report statistical significance. But we can calculate it ourselves based on the table, *id.* at 9, using Pearson's chi-square test. Among the 67 participants, 6 inmates received disciplinary reports and 61 didn't. Among the 741 non-participants at Tomoka, 124 received disciplinary reports and 617 didn't (for a rate of 17%). Among the 54,997 non-participants at other prisons, 6,614 received disciplinary reports and 48,383 didn't (for a rate of 12%). The difference between the participants' disciplinary report rate of 9% and the Tomoka non-participants' rate of 17% has a p-value of 0.08, so it isn't significant at the 5% significance level. The difference between the participants' rate of 9% and the other-prison non-participants' rate of 12% has a p-value of 0.45, so it isn't significant at any reasonable significance level.

84. *Id.* at 13. There are also reports of an earlier study, conducted in 1995, evaluating the Kairos program at Union Correctional Institution in Florida. The study examined recidivism among 505 inmates who had attended Kairos over 10 years or who had attended 11 Kairos Weekends. The non-Kairos control group had a 23.4% recidivism rate; the Kairos group had a 15.7% recidivism rate; and those who had participated in a Kairos follow-up program in addition to attending a Weekend had a recidivism rate of 10%. Burnside, *From Cursillo to Prison*, *supra* note 77, at 62 (citing *Profile of Kairos*, KAIROS NEWSL., (Kairos Prison Ministry, Winter Park, Fla.) Dec. 1, 1998); *see also Kairos Fact Sheet*, KAIROS PRISON MINISTRY INT'L, http://www.mykairos.org/templates/System/details.asp?id=23761&PID=148702 (last visited Oct. 6, 2011). However, I haven't been able to obtain this report (Kairos Prison Ministry International doesn't use it anymore, *see* E-mail from Ann M. Kreiler, Exec. Admin. Assistant, Kairos Prison Ministry Int'l, to author (Jan. 26, 2011) (on file with author)), so it's unclear whether the results are statistically significant and whether they're tainted by self-selection.

85. Dan Denny, Individual and Organizational Impact of Kairos Horizon, a Faith-Based Adult Learning Program, in a Correctional Setting 41 (May 2006) (unpublished D.Ed. thesis, Oklahoma State University) (on file with the Oklahoma State University Library), *available at* http://digital.library.ok

Denny examined three cohorts of participants, from "Year One" (2002), "Year Two" (2003), and "Year Three" (2004).[86] The 36 Year One participants had 89% fewer misconduct reports after the program than before; the drop for the 51 Year Two participants was 80%; and the drop for the 51 Year Three participants was 84%.[87] The average drop was 86%.[88] Misconduct reports in the entire facility fell from 901 to 308 (a 66% drop) from Year One to Year Three,[89] which is presumably comparable to the 80% before-to-after drop for the Year Two participants. It's unclear from the paper how many inmates there were at the facility during this time, so it's unclear whether the drop in misconduct among program participants is significantly different from the total decrease facility-wide.

When the paper was written, only seven participants had been released, the longest-released graduate had only been out for one year, and no graduate had been re-arrested.[90] So the author couldn't report "true recidivism rates" by Oklahoma standards, which require a three-year post-release history.[91]

6. Education Studies

Some education studies also use this approach, neither addressing self-selection nor controlling for observable variables.

One example is Janet Beales and Maureen Wahl's assessment of the Partners Advancing Values in Evaluation (PAVE) program in Milwaukee,[92] a privately funded voucher system that functioned parallel to the publicly funded voucher system, the Milwaukee Parental Choice Program (MPCP).[93] Beales and Wahl found that 63.2% of PAVE students scored above the 50th percentile in reading (60.4% in math), which was much higher than the corresponding percentages for MPCP students, Milwaukee public school low-income students, or all Milwaukee public school students.[94] (These percentages were all between 16% and 35%.) PAVE students were similarly above the three comparison groups in reading and math test score medians and means.[95]

state.edu/etd/umi-okstate-1723.pdf.

86. *Id.* at 70.

87. *Id.* at 92 tbl.15.

88. *Id.*

89. *Id.* at 94 tbl.16.

90. *Id.* at 96.

91. *Id.*

92. Janet R. Beales & Maureen Wahl, *Private Vouchers in Milwaukee: The PAVE Program*, *in* PRIVATE VOUCHERS 41 (Terry M. Moe ed., 1995). Similar results are reported in JANET R. BEALES & MAUREEN WAHL, GIVEN THE CHOICE: A STUDY OF THE PAVE PROGRAM AND SCHOOL CHOICE IN MILWAUKEE (Reason Found. Pol'y Study No. 183, Jan. 1995).

93. On the publicly funded voucher program in Milwaukee, see sources cited *infra* notes 151, 154–156, 349–354.

94. Beales & Wahl, *supra* note 92, at 61 tbl.10.

95. *Id.*

However, the PAVE group differed from the other groups in various ways. Most obviously, the PAVE group, like the MPCP group, was self-selected, since one had to apply for a voucher; the public school students weren't self-selected.[96] But the PAVE group and the MPCP group weren't comparable either: the PAVE scores were the test results of seventh-grade students, while the MPCP scores were test results from multiple grade levels, so the authors weren't even comparing the same test.[97] Finally, the authors couldn't control for income, parental education, or other variables.[98]

B. Studies with Some Controls

The studies in the previous section aren't credible because participants in religious programs are just so different from non-participants. One possible fix would be to control for observable differences between participants and non-participants.[99] This is what the studies reported in this section do: participants are matched with non-participants with observable characteristics that are as similar as possible.

But these studies are still vulnerable. An unobserved variable—motivation to change—affects *both* whether the inmate participates and whether he reoffends.[100] Because motivation and success (avoiding rearrest) are positively correlated, any effect we find is probably biased upward (ignoring any other sources of bias in one direction or another). A true zero effect may look like a positive effect because we're measuring the effect of motivation.[101]

In other words, if two prisoners are perfectly matched on the observables, but one of them chose to participate and the other didn't, these two prisoners aren't really well matched. Any study that finds better results among participants is thus still subject to self-selection bias.[102]

1. La Vigne et al.'s Florida Study

Nancy La Vigne and her coauthors reported on six- and twelve-month recidivism rates of participants in two Florida "faith- and character-based institutions" (FCBI)—one male (Lawtey) and one female (Hillsborough).[103]

---

96. *Id.* at 60. There may have been other self-selection biases: the PAVE test scores were voluntarily revealed by some of the participating parents, and perhaps parents would be more willing to reveal their children's test scores if they were high.

97. *Id.* at 61 tbl.10 nn.a-b.

98. *Id.* at 60.

99. Heckman & Robb, *supra* note 50.

100. *E.g.*, Paul R. Rosenbaum & Donald B. Rubin, *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score*, 79 J. AM. STAT. ASS'N 516, 516 (1984); Heckman & Robb, *supra* note 50, at 78.

101. *Cf.* Caroline Minter Hoxby, *Are Efficiency and Equity in School Finance Substitutes or Complements?*, J. ECON. PERSP., Fall 1996, at 51, 64.

102. Moreover, the variables that are controlled for have no obvious connection to motivation, so their inclusion may not alleviate this particular selection problem. Mears et al., *supra* note 32, at 360.

103. NANCY G. LAVIGNE ET AL., EVALUATION OF FLORIDA'S FAITH- AND CHARACTER-BASED

Participants were matched with a control group based on "sex, age, race, primary offense type, violent/non-violent offense, number of prior incarcerations, time incarcerated for current offense, time to expected release, and pre-study disciplinary report rate."[104]

At first, male FCBI participants had lower recidivism rates than their control group—none of the 189 male inmates from Lawtey were reincarcerated after six months, compared to four of the 189 male comparison inmates (2.1%).[105] There was no significant difference for females,[106] and twelve months out, there was no significant effect at all for either males or females.[107] There was also no significant difference between average time to reincarceration for the faith-based inmates and the comparison inmates, for either males or females.[108] The results here are thus extremely weak.[109]

A later report by Diana Brazzell and Nancy La Vigne, using new data, continued to find "no statistically significant difference . . . in the proportion of FCBI and non-FCBI inmates returned to prison within 12, 18, 24, and 26 months of release," for either males or females.[110]

2. Rose's Kainos Community Study

Gerry Rose evaluated the effect on reconviction of participation in the Kainos Community, a faith-based prison chiefly operating out of The Verne prison in England.[111] The 84 participants were compared against a sample of 13,832 prisoners; the comparison sample was composed of all adult sen-

INSTITUTIONS: FINAL REPORT 1, 42–43 (Urban Inst. Justice Pol'y Ctr., Oct. 2007), http://www.urban.org/uploadedPDF/411561_fcbi_evaluation.pdf.

104. *Id.*

105. *Id.* at 45 tbl.G. This was statistically significant at $p \leq 0.05$. *Id.*

106. *Id.* None of the 100 Hillsborough females were reincarcerated within 6 months, as opposed to one female in the comparison group. *Id.*

107. *Id.* at 45. Of the 56 male Lawtey inmates who were released at least a year before the study end date, one was reincarcerated, as opposed to two of the 82 in the comparison group. *Id.* Of the 54 female Hillsborough inmates who were released at least a year before the study end date, one was reincarcerated, as opposed to four of the 62 in the comparison group. *Id.*

108. *Id.* at 44–45. Among reincarcerated males, mean time to reincarceration was 371 days for the Lawtey inmates and 262 days for the comparison group. *Id.* For females, the difference is 385 days versus 318 days. *Id.* Neither of these differences is significant at $p<0.05$. *Id.*

109. *Id.* at 46–47.

110. Diana Brazzell & Nancy La Vigne, Evaluating the Potential of Faith-Based Correctional Models: A Case Study of Florida's Faith-Based and Character-Based Institutions, presented at the Faith-Based and Community Initiatives Conference on Research, Outcomes, and Evaluation 245–46 (June 2008), http://aspe.hhs.gov/fbci/comp08/Brazzell.pdf.

111. Gerry Rose, *Kainos Community and Reconviction Rates*, *in* JONATHAN BURNSIDE ET AL., KAINOS COMMUNITY IN PRISONS: REPORT OF AN EVALUATION 42 (presented to Res. Dev. & Stats. Directorate, Home Office, HM Prison Service England & Wales & Kainos Community, Dec. 2001), http://webarchive.nationalarchives.gov.uk/20110218135832/rds.homeoffice.gov.uk/rds/pdfs/kainos_finalrep.pdf. Steve Aos and his coauthors considered this study to be one of the few that were of good enough quality to include in their review of evidence-based adult corrections programs. AOS ET AL., *supra* note 33, at 19. The program also had some participants at Highpoint North, Highpoint South, and Swaleside prisons, though there were apparently no Swaleside participants in Rose's empirical study. Rose, *supra*, at 34; Jonathan Burnside, *Introduction*, *in* BURNSIDE ET AL., *supra*, at 16. Substantially the same study was printed, with two years of follow-up data instead of one, as in Rose, *supra* note 63. There was no significant difference between the results of these two studies.

tenced prisoners released from prisons in England and Wales in 1996 and 1997 who were British nationals, had served sentences of six months to 15 years, had been released from particular categories of prisons, and satisfied a few additional restrictions.[112] In the Kainos sample, 22.6% of the participants were reconvicted within a year of release; among non-participants, the percentage was 25.9%.[113] This difference wasn't significant.[114]

So far, this didn't control for any variables. But Rose then went further, comparing the actual reconviction rates of Kainos participants with their own *predicted* reconviction rates. The predicted rates were based on a statistical model that controlled for observable factors such as their sex, offense category, age at first conviction, age at sentence, months spent in prison after sentence, and number of custodial sentences before age 21.[115] Thus, rather than comparing participants and non-participants, he compared *actual* participants with *hypothetical* participants whose recidivism was predicted based on factors that didn't include their participation in a faith-based program.

There, too, Rose found no significant effect: 25.0% of the Kainos sample was reconvicted, while the expected percentage would have been 26.0% or 24.2% (depending on which prediction model one used).[116]

3. *Young et al.'s Prison Ministry Study*

Mark Young and his coauthors investigated "long-term recidivism among . . . federal inmates trained as volunteer prison ministers" as part of Prison Fellowship Ministries' Washington D.C. Discipleship Seminars.[117] Participants were sent to Washington for a two-week faith and leadership seminar, and their recidivism was compared to that of a control group.[118] The control group was selected to match the experimental group with respect to race, gender, age at release, and the "salient factor score"[119] (an estimate of a prisoner's likelihood of recidivism[120]).

Participants' recidivism rate was 40%, while the control group's recidivism rate was 51%.[121] Participating women had a recidivism rate of 19%, compared to 47% for the control women, and participating men had a reci-

112. Rose, *supra* note 111, at 44.
113. *Id.* at 47 tbl.10.
114. *Id.* at 48 n.17. The p-value was close to 0.50. *Id.*
115. *Id.* at 49, 138–39 app. 2.
116. *Id.* at 50. The same results held when the Kainos sample was divided by prison, and, within one of the prisons, by amount of exposure to the Kainos program. *Id.* The prediction model came in two forms—a second version also took account of which prison the offender was released from. *Id.* at 49.
117. Mark C. Young et al., *Long-Term Recidivism Among Federal Inmates Trained as Volunteer Prison Ministers*, 22 J. OFFENDER REHAB. nos. 1 & 2, 1995, at 97. O'Connor rates this study as having "fair" methodological quality (on a poor-fair-good-excellent scale). O'Connor, *supra* note 33, at 23 tbl.3.
118. Young et al., *supra* note 117, at 101.
119. *Id.* at 101–02.
120. Peter B. Hoffman et al., *Salient Factor Score and Releasee Behavior: Three Validation Samples*, 2 LAW & HUM. BEHAV. 47, 49 tbl.1 (1978).
121. Young et al., *supra* note 117, at 107. This difference was significant at p=0.04. *Id.* at 106.

divism rate of 45%, compared to 52% for the control men.[122] When the groups were further broken down by gender and race, participants had lower recidivism rates for all subgroups except black men.[123]

As in the theology study above,[124] these results are subject to both self-selection and selection by program administrators, in this case prison chaplains, who chose which inmates could participate.[125]

4. O'Connor et al.'s Lieber Prison Study

Tom O'Connor and his coauthors reported on rates of in-prison infractions among participants in Prison Fellowship (PF) programming at Lieber Prison in South Carolina.[126] Their data set of 1,597 included both participants and non-participants; 302 inmates attended at least one out of 47 Prison Fellowship meetings.[127]

Participants had lower infraction rates than non-participants: "9.9% of PF inmates had an infraction since attending at least one PF program compared to the 23.2% of Non PF inmates who had an infraction."[128] The more an inmate participated in PF programs, the lower his chance of having an infraction.[129]

Controlling for prior violent convictions, age, marriage status, and days spent in the prison, whether an inmate participated in PF programs strongly predicted lower infraction rates.[130] "Non PF inmates were still 2.5 times more likely than PF inmates to have an infraction."[131]

The rate of participation in PF programs, controlling for the same variables, likewise strongly predicted lower infraction rates.[132] But controlling for the rate of participation isn't useful. Given a valid control group, the only valid comparison is between the control group and the *entire* treatment group.[133] If we compare the control group to isolated, self-selected subsets of the treatment group, like those who participated the most in PF programs, we are merely reintroducing another layer of self-selection bias. Even if high participation reduces infraction rates (which is doubtful, given that the high participants may already be better people), the relevant question from a *policy* perspective, that is, from the perspective of someone wondering

122. *Id.* at 107–08 tbl.2.
123. *Id.*
124. *See supra* text accompanying notes 65–68.
125. Young et al., *supra* note 117, at 113.
126. THOMAS P. O'CONNOR ET AL., THE IMPACT OF PRISON FELLOWSHIP ON INMATE INFRACTIONS AT LIEBER PRISON IN SOUTH CAROLINA (1997).
127. *Id.* at 2, 5.
128. *Id.* at 8–9.
129. *Id.* at 9.
130. *Id.*
131. *Id.* The p-value was $p<0.0001$. *See id.* at 9–10, tbl.1 (providing a Chi-square of 202.342 and 5 degrees of freedom).
132. *Id.* at 10. The p-value was $p<0.0001$. *See id.* at 10, tbl.2. (providing a Chi-square of 204.085 and 5 degrees of freedom).
133. Camp et al., *supra* note 46, at 532.

whether to introduce the program, is how well it works overall, including for those who choose not to participate much.[134]

5. Wilson et al.'s COSA Study

Robin Wilson and coauthors examined the effect on recidivism of the Circles of Support and Accountability (COSA) program in south-central Ontario.[135] Unlike the programs discussed so far, COSA isn't an in-prison program; rather, it's a support network, largely staffed by religious volunteers, to support the reintegration of released sex offenders into society.[136] A group of 60 sex offenders assigned to COSA were compared against a group of non-participants who were similarly detained, had similar recidivism risk categories, were released around the same time, and had similar "prior involvement in sexual offender treatment programming."[137] The COSA group had significantly lower recidivism rates: the COSA group had a 5% rate of sexual recidivism and a 15% rate of violent recidivism, as compared to 17% and 35% among the comparison group.[138]

Robin Wilson and coauthors found similar results in a follow-up study of COSA participants across Canada. There, too, the comparison group of 44 COSA participants from assorted Canadian cities was matched, according to similar control variables, to a group of sexual offenders who didn't participate.[139] The COSA group had lower rates of sexual recidivism (2.27%), violent recidivism (9.09%), and overall recidivism (11.36%) than the control group (13.67%, 34.09%, and 38.64%, respectively).[140]

6. Self-Selection in Prisons and Schools

As I've pointed out above,[141] self-selection also plagues studies of the effectiveness of private schools.[142]

134. There is apparently a similar study (which I haven't been able to obtain), also by Thomas O'Connor and coauthors, and with a similar title, T.P. O'CONNOR ET AL., THE IMPACT OF RELIGIOUS PROGRAMS ON INMATE INFRACTIONS AT LIEBER PRISON IN SOUTH CAROLINA (1997), *cited in* Stephen T. Hall, *Faith-Based Cognitive Programs in Corrections*, CORRECTIONS TODAY, Dec. 2003, at 108, 111–12. According to the Hall article, this study *doesn't* find any difference in infractions between the PF and non-PF groups, but a difference does emerge when the PF group is divided by levels of religious attendance. *Id.* at 111. This study is thus even less supportive of an effect of the religious program than is the study discussed in the text.

135. ROBIN J. WILSON ET AL., CIRCLES OF SUPPORT & ACCOUNTABILITY: AN EVALUATION OF THE PILOT PROJECT IN SOUTH-CENTRAL ONTARIO (2005). Steve Aos and his coauthors considered this study to be one of the few that were of good enough quality to include in their review of evidence-based adult corrections programs. *See* AOS ET AL., *supra* note 33, at 16 (citing WILSON ET AL., *supra*).

136. WILSON ET AL., *supra* note 135, at 1–3.

137. *Id.* at 20–21.

138. *See id.* at 23-24 tbl.3. The sexual recidivism difference was significant at $p<0.05$, and the violent recidivism difference was significant at $p<0.01$. *Id.* at 24 tbl.3.

139. Robin J. Wilson et al., *Circles of Support & Accountability: A Canadian National Replication of Outcome Findings*, 21 SEXUAL ABUSE: J. RES. & TREATMENT 412, 417–18 (2009).

140. *Id.* at 421 tbl.2. These differences were significant at $p<0.05$, $p<0.01$, and $p<0.01$, respectively. *Id.*

141. *See supra* text accompanying notes 42–43.

Early work by James Coleman and his coauthors estimated the effect of private schooling on sophomore scores, controlling for various background characteristics.[143] Coleman et al. recognized that selection was a potentially serious problem,[144] but noted that it was impossible to properly solve the problem "in the absence of random assignment to treatments, or something approximating it," and that one had to proceed regardless.[145]

Other studies found a weaker effect. Jay Noell,[146] Doug Willms,[147] Karl Alexander and Aaron Pallas,[148] and William Morgan[149] analyzed the same data with different specifications and different control variables and found a much weaker effect of private schools.[150] John F. Witte and his coauthors found that students in the Milwaukee voucher program didn't "differ in any predictable way on achievement tests" [151] from Milwaukee public-school students over the first four years of the program.[152] And, in a recent study,

142. Caroline M. Hoxby, *The Effects of School Choice on Curriculum and Atmosphere*, *in* EARNING AND LEARNING: HOW SCHOOLS MATTER 281, 290–92 (Susan E. Mayer & Paul E. Peterson eds., 1999).

143. JAMES S. COLEMAN ET AL., HIGH SCHOOL ACHIEVEMENT: PUBLIC, CATHOLIC, AND PRIVATE SCHOOLS COMPARED 137–39, 138 tbl.6–7 (1981). Most of this book is based on a previous report by the same authors, PUBLIC AND PRIVATE SCHOOLS (Nat'l Op. Res. Ctr. 1981), *id.* at xxii, and various critiques of Coleman et al.'s work cite the original report directly, *see, e.g.*, Arthur S. Goldberger & Glen G. Cain, *The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer and Kilgore Report*, 55 SOC. OF EDUC. 103, 103 (1982).

144. COLEMAN ET AL., *supra* note 143, at 201; James Coleman et al., *Cognitive Outcomes in Public and Private Schools*, 55 SOC. OF EDUC. 65, 68–69 (1982). Coleman et al.'s methodology has been criticized on this and other grounds. *Compare* Goldberger & Cain, *supra* note 143, at 109–13, *with* James Coleman et al., *Achievement and Segregation in Secondary Schools: A Further Look at Public and Private School Differences*, 55 SOC. OF EDUC. 162, 168–76 (1982).

145. COLEMAN ET AL., *supra* note 143, at 202; *see also infra* note 176.

146. Jay Noell, *Public and Catholic Schools: A Reanalysis of "Public and Private Schools"*, 55 SOC. OF EDUC. 123, 124–27 (1982).

147. DOUG WILLMS, ACHIEVEMENT OUTCOMES IN PUBLIC AND PRIVATE SCHOOLS: A CLOSER LOOK AT THE HIGH SCHOOL AND BEYOND DATA 1–8 (1982).

148. Karl L. Alexander & Aaron M. Pallas, *Private Schools and Public Policy: New Evidence on Cognitive Achievement in Public and Private Schools*, 56 SOC. OF EDUC. 170, 172-73 (1983).

149. William R. Morgan, *Learning and Student Life Quality of Public and Private School Youth*, 56 SOC. OF EDUC. 187, 192–94 (1983).

150. Noell, *supra* note 146, at 127; *see* Alexander & Pallas, *supra* note 148, at 178; WILMS, *supra* note 147, at 9–12; Morgan, *supra* note 149, at 211. Another study apparently didn't do a formal regression, but analyzed means of achievement scores between public and private schools. After adjusting for the different socioeconomic status of students attending private schools, it found that the private-school advantage either decreased substantially or became statistically insignificant. NAT'L ASSESSMENT OF EDUC. PROGRESS, EDUC. COMM'N OF THE STATES, READING AND MATHEMATICS ACHIEVEMENT IN PUBLIC AND PRIVATE SCHOOLS: IS THERE A DIFFERENCE? 4–5, 8 (1981). Andrew Greeley came to conclusions broadly similar to Coleman's. *See* ANDREW M. GREELEY, CATHOLIC HIGH SCHOOLS AND MINORITY STUDENTS 109–14 (1982); Thomas Hoffer et al., *Achievement Growth in Public and Catholic Schools*, 58 SOC. OF EDUC. 74, 75–82 (1985) (extending Greeley analysis to accommodate follow-up data). Another study used Hierarchical Linear Modeling, which also doesn't account for selection bias. *See* HENRY BRAUN ET AL., COMPARING PRIVATE SCHOOLS AND PUBLIC SCHOOLS USING HIERARCHICAL LINEAR MODELING (2006). *But see* PAUL E. PETERSON & ELENA LLAUDET, ON THE PUBLIC-PRIVATE SCHOOL ACHIEVEMENT DEBATE, 10–13 (2006) (giving a strong critique of this study).

151. JOHN F. WITTE ET AL., FOURTH-YEAR REPORT: MILWAUKEE PARENTAL CHOICE PROGRAM (1994) *available at* http://www.lafollette.wisc.edu/publications/workingpapers/MilwaukeeChoice4YR/fourthyear.html.

152. *Id*; *see also* JOHN F. WITTE ET AL., FIFTH-YEAR REPORT: MILWAUKEE PARENTAL CHOICE PROGRAM (1995),. *available at* http://www.lafollette.wisc.edu/publications/workingpapers/MilwaukeeChoice5YR/fifthYear.html. *But see* JAY P. GREENE ET AL., THE EFFECTIVENESS OF SCHOOL CHOICE IN MILWAUKEE: A SECONDARY ANALYSIS OF DATA FROM THE PROGRAM'S EVALUATION, 16–26 (1996)

Harold Wenglinsky similarly controlled for various observable variables and followed students over time, and found no positive effect for private schools.[153]

Various studies found effects that differed according to the precise outcome variable or the precise population being studied. Cecilia Rouse, comparing Milwaukee voucher students with Milwaukee public school students, found a substantial effect on math scores, but no effect on reading scores, of being selected to attend a voucher school in Milwaukee.[154] Jeffrey Grogger and Derek Neal found significant effects on high-school graduation rates, college attendance rates, and math test scores.[155] Gains for urban minorities were especially large, but there was "little evidence of math-achievement gains for suburban minorities in Catholic schools."[156]

Private-school researchers have also investigated whether the public versus private choice affects the *growth* of test scores from the sophomore to the senior year. Coleman and his coauthors did this by comparing two different cohorts—a sophomore class and a senior class in the same year.[157] Later, John Chubb and Terry Moe,[158] as well as Douglas Willms[159] and Karl Alexander and Aaron Pallas,[160] who had the benefit of follow-up data, compared the sophomore and senior scores of the same students. But these methods also don't control for selection bias if one believes (as is plausible, and as Coleman et al. agree) that selectivity affects growth rates in addition to levels.[161]

It should be clear that prison and education studies share common methodological problems. We can discount any positive results of these studies

---

(giving a sharp critique on the methodology of WITTE ET AL., *supra* note 151).

153. HAROLD WENGLINSKY, ARE PRIVATE HIGH SCHOOLS BETTER ACADEMICALLY THAN PUBLIC HIGH SCHOOLS? 2 (Ctr. on Educ. Policy 2007).

154. Cecilia Elena Rouse, *Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program*, 113 Q.J. ECON. 553, 580–84 (1998).

155. Jeffrey Grogger & Derek Neal, *Further Evidence on the Effects of Catholic Secondary Schooling*, BROOKINGS-WHARTON PAPERS ON URBAN AFFAIRS, 2000, at 151, 158–67.

156. *Id.* at 166.

157. COLEMAN ET AL., *supra* note 143, at 141–43, 204; Coleman et al., *supra* note 144, at 71–72.

158. JOHN E. CHUBB & TERRY M. MOE, POLITICS, MARKETS, AND AMERICA'S SCHOOLS 21–22 (1990). Chubb and Moe also model selection *out* of school, that is, dropping out of high school, and their "Selection bias correction" variable, *id.* at 126 tbl. 4-8, measures this. *Id.* at 248–51.

159. J. Douglas Willms, *Catholic-School Effects on Academic Achievement: New Evidence from the High School and Beyond Follow-Up Study*, 58 SOC. OF EDUC. 98, 101 (1985).

160. Karl L. Alexander & Aaron M. Pallas, *School Sector and Cognitive Performance: When Is a Little a Little?*, 58 SOC. OF EDUC. 115, 115 (1985).

161. Willms, *supra* note 159, at 143; Coleman et al., *supra* note 144, at 70–71; *see also* Goldberger & Cain, *supra* note 143, at 114–17 (critiquing the Coleman et al. strategy). Coleman et al. also tested the effect of different school policies (homework, absenteeism, discipline, etc.) *within* sectors, on the theory that if one finds positive effects of attending a public school that "looks like" a private school in terms of school policies, that would support the findings of private-school superiority, since selection effects probably aren't great within the public sector. COLEMAN ET AL., *supra* note 143, at 171 tbl.6-21, 204–05; Coleman et al., *supra* note 144, at 73–76. *But see* Goldberger & Cain, *supra* note 143, at 117–20 (critiquing this strategy as well, partly on technical grounds, and partly by pointing out that some of the "school policies" variables are in fact indicators of student-specific traits). Chubb and Moe do a similar regression within public schools, to determine the importance of particular school policies. CHUBB & MOE, *supra* note 158, at 259–77.

as being potentially artifacts of self-selection.[162] But what about the studies that found *no* effect—for instance, in the faith-based prison case, the La Vigne studies,[163] and the Rose study?[164] Surely, if positive results are overstated by some unknown amount, zero results must prove that faith-based prisons don't work at all, and that the true effect is, if anything, negative?

This is tempting, but we should resist this conclusion for the following reasons:

- The self-selection bias overstates results, but there may be other empirical problems that tend to understate results. For instance, there may be other unobserved variables that are *negatively* correlated with success. (Perhaps people also tend to participate in programs if they feel they need it more? Perhaps programs that provide additional resources to inmates and that are selective also attract inmates who are good at lying to the program administrators about their suitability for the program?[165] Perhaps, if participation in a program contributes to parole decisions, the program attracts problem inmates who are more likely to need the good points on their record? Generally, there is always a problem with insincere inmates who take advantage of religious programs to "gain protection," "meet other inmates," "interact with volunteers," and "gain access to prison resources,"[166] quite apart from any desire to reform.) Or there may be measurement error in the dependent variable (i.e., some of the inmates who are re-arrested are wrongly coded as not having been re-arrested and vice versa), which tends to reduce the measured effect. So just as a positive measured effect could hide a true zero effect, a zero measured effect could hide a true positive effect.

- Every program is different, and some programs may only have a zero measured effect because they were badly designed or badly run. Their failure needn't reflect badly on other programs that are done well—in fact, even if only a handful of programs "work," but if those programs, once they have been shown to work, can be replicated, the whole process of experimentation can be thought to have been a success.

162. Also, a few studies of Transcendental Meditation, *see* Karasz & Midgett, *supra* note 38; Elliott, *supra* note 38, fall into this category. *See also* Charles N. Alexander et al., *Walpole Study of the Transcendental Meditation Program in Maximum Security Prisoners III: Reduced Recidivism*, 36 J. OFFENDER REHAB. nos. 1–4, 2003, at 161, 174 (2003); Maxwell V. Rainforth et al., *Effects of the Transcendental Meditation Program on Recidivism Among Former Inmates of Folsom Prison: Survival Analysis of 15-Year Follow-Up Data*, 36 J. OFFENDER REHAB., nos. 1–4, 2003, at 198.

163. *See supra* notes 103–110 and accompanying text.

164. *See supra* notes 111–116 and accompanying text.

165. The programs in the La Vigne and Rose studies didn't seem to be selective, *see* LAVIGNE ET AL. *supra* note 103; Brazzell & La Vigne *supra* note 110; Rose, *supra* note 111, but this is a general consideration that is valid for other studies.

166. O'Connor & Duncan, *supra* note 36, at 88.

- Alexander and Pallas noted that the effect of private schools appeared much smaller when the follow-up data was analyzed and students' previous test scores were used as controls for their current performance.[167] This dramatic change from a background-controls-only specification to a background-controls-and-test-scores specification, they argued, showed "that background proxies are simply inadequate when attempting to assess the impact of school organization on cognitive outcomes."[168] This is a modest moral of the "background proxy" studies: when one's empirical method is subject to an important source of bias, the precise specification can have a large effect on the results.

Perhaps most importantly, we now have other studies that are methodologically more valid. We thus don't need to spend too much time interpreting the results of the less valid studies.

C. Matching on the Propensity Score

In this Part, I discuss a technically more sophisticated way of dealing with selection problems: propensity score matching.[169]

In propensity score matching, the researchers first identify the observable variables that best predict whether someone will participate in the program.[170] This first-stage estimation generates a "propensity score" for each inmate; this is essentially an estimated probability of participating in the program.[171] One inmate may have participated and the other may have not, but they may both have propensity scores of, say, 70%, so that they are estimated to be equally likely, ex ante, to have chosen to participate.[172]

The matching process then matches each participant to another participant with a similar propensity score; a 70% propensity participating inmate is matched with a 70% propensity non-participating inmate, even if these inmates may differ on various individual characteristics.[173]

167. Alexander & Pallas, *supra* note 160, at 123.
168. *Id.*
169. PAUL R. ROSENBAUM, OBSERVATIONAL STUDIES 296–302 (2d ed. 2002); PAUL R. ROSENBAUM, DESIGN OF OBSERVATIONAL STUDIES 72–75, 165–68 (2010); Paul R. Rosenbaum & Donald B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, 70 BIOMETRIKA 41 (1983); Rosenbaum & Rubin, *supra* note 100; Paul R. Rosenbaum & Donald B. Rubin, *Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score*, 39 AM. STATISTICIAN 33 (1985) [hereinafter Rosebaum & Rubin, *Constructing a Control Group*]; LORI S. PARSONS, REDUCING BIAS IN A PROPENSITY SCORE MATCHED-PAIR SAMPLE USING GREEDY MATCHING TECHNIQUES (2001).
170. *E.g.*, Rosenbaum & Rubin, *supra* note 100, at 516.
171. *See generally* Rosenbaum & Rubin, *Constructing a Control Group*, *supra* note 169, at 34–35.
172. *See* ROSENBAUM, DESIGN OF OBSERVATIONAL STUDIES, *supra* note 169, at 166. (showing that if one used a coin to determine exposure, then both groups would have a propensity score of 0.5).
173. ROSENBAUM, DESIGN OF OBSERVATIONAL STUDIES, *supra* note 169, at 166.

Practitioners of propensity score matching point to certain advantages of the method over trying to match on observable variables directly.[174] Given a participant with particular observable characteristics, it is often hard or impossible to find a non-participant with identical, or nearly identical, values of those same variables; by contrast, it is easier to match according to a single number.[175]

But propensity score matching can't overcome the problems of selection bias in the case of faith-based prisons. To see this, suppose that there were so many non-participating prisoners that *exact* matching on observables was always possible; every participating inmate would be matched with a non-participant who looked exactly identical. Because these two inmates would have identical observable characteristics, they would also have identical propensity scores. Matching on propensity scores would then produce exactly the same control group as the previous set of studies, which matched on observables directly.

Thus, if the direct matching studies weren't credible, the propensity score matching studies aren't credible either. Using the propensity score may improve the efficacy of matching, but it doesn't alleviate the self-selection problem.[176]

More technically, the problem is that propensity score methods give the correct result *if nonobservables play no role in the selection mechanism*,[177] or more precisely, if the unobserved determinants of participation play no role in ultimate success (that is, low recidivism). This assumption is quite false in the case of faith-based prison programs, where motivation to change, and possibly religiosity itself, *both* determine participation in the program *and* play a large role in whether an inmate reoffends. James Heckman and Richard Robb argue that "[t]he propensity score methodology solves a very special problem . . . that is of limited interest to social science data analysts."[178] Whether Heckman and Robb are right about the interest of propensity score studies in general, faith-based prison evaluation certainly seems like one area where the method doesn't seem credible.

1. O'Connor et al.'s New York Study

Tom O'Connor and his coauthors analyzed the effect on prison infractions and recidivism of participation in Prison Fellowship programs in New York prisons.[179] The participating group of 225 inmates was matched with a

174. Rosembaum & Rubin, *supra* note 100, at 516.

175. *E.g.*, *id.*; Rajeev H. Dehejia & Sadek Wahba, *Propensity Score-Matching Methods for Nonexperimental Causal Studies*, 84 REV. ECON. & STAT. 151, 153 (2002).

176. Coleman et al. also point out that propensity score models have the disadvantage of being "sensitive to alternative specifications." Coleman et al., *supra* note 144, at 172.

177. *E.g.*, Dehejia & Wahba, *supra* note 175, at 151 & n.1, 152–53 (2002); Rosenbaum & Rubin, *supra* note 100, at 517; ROSENBAUM, DESIGN OF OBSERVATIONAL STUDIES, *supra* note 169, at 73.

178. Heckman & Robb, *supra* note 50, at 101–02.

179. Tom O'Connor, *The Impact of Religious Programming on Recidivism, the Community and Prisons*, INT'L ASS'N RES. & CMTY. ALTS. J. ON CMTY. CORR., no. 6, 1996 at 13, 14–16. O'Connor, the

control group based on race and a propensity score calculated using six variables—"age, religion, county of residence, military discharge, minimum sentence and initial security classification."[180]

The study found no significant difference between participants and the control group in prison infractions, number of re-arrests, or time to re-arrest.[181] Among participants, 37% had infractions: 28% had security infractions, 16% had nonviolent infractions, and 15% had violent infractions.[182] In the control group, the percentages were 32%, 23%, 18%, and 11%, respectively.[183] None of these differences were significant.[184] Nor was there any significant difference in the frequency of re-arrest (36% for participants versus 34% for non-participants),[185] though a difference emerged when arrests were broken down by type of charge.[186] Participants were "more likely to be re-arrested for a violent offense" (28% versus 16%), but "less likely to be re-arrested for a drug offense" (21% versus 44%).[187] There were also significant differences when re-arrests were broken down by region—for whatever reason, a re-arrest of a participant was more likely to occur in upstate New York (and less likely to occur in New York City or suburban New York) than the re-arrest of a non-participant.[188]

The authors then divided the group into high-participating and low-participating groups. There was still no significant difference between high and low participants in infraction or re-arrest rates.[189] The authors then computed a score from 0 to 3 for each inmate, based on the "Level of Supervision Inventory" that measured their estimated risk of being re-arrested, and then classified inmates by PF participation level (none, low, or high) *and* risk score (0, 1, 2, or 3).[190] When they did this, they found that among high-risk PF inmates—that is, inmates with a risk level of 3—high-participating inmates were significantly less likely to be re-arrested than low-participating inmates.[191]

---

author, rates his own study as having "fair" methodological quality (on a poor-fair-good-excellent scale). O'Connor, *supra* note 33, at 23 tbl.3. The same analysis, with more details, can be found in TOM O'CONNOR ET AL., THE NEW YORK STUDY OF PRISON FELLOWSHIP PROGRAMMING: EXECUTIVE SUMMARY AND FINAL REPORT 5–20 (1994), and in Thomas O'Connor et al., Religion and Prisons: Do Volunteer Religious Programs Reduce Recidivism? 4–12 (Aug., 1996) (unpublished paper) (on file with author).

180. O'Connor, *supra* note 179, at 15, 47 n.8; O'CONNOR ET AL., *supra* note 179 at 6.

181. O'Connor, *supra* note 179, at 15; O'CONNOR ET AL., *supra* note 179, at 9–10.

182. O'CONNOR ET AL., *supra* note 179, at 8–9, tbls.2, 3, 4, 5 & 6.

183. O'Connor, *supra* note 179, at 15; O'CONNOR ET AL., *supra* note 179, at 8–9 & tbls.2, 3, 4, & 5.

184. O'CONNOR ET AL., *supra* note 179, at 9. The significance level chosen was 10%, so all these differences had $p>0.10$. *Id.*

185. *Id.* at 10 & tbl.6.

186. O'CONNOR ET AL., *supra* note 179, at 10–11, tbls.7, 8, & 9.

187. *Id.* at 10–11 & tbls.7 & 8. The significance of these differences is $p<0.06$ and $p<0.01$. *Id.* at 10.

188. *Id.* at 11 & tbl.9. The breakdown of re-arrests for PF inmates was 71% in New York City, 6% in suburban New York, and 23% in upstate New York; the same breakdown for non-PF inmates was 75%, 14%, and 11%. *Id.* at 11, tbl.9. The significance of these differences is $p<0.04$. *Id.* at 11.

189. O'CONNOR ET AL., *supra* note 179, at 11–12.

190. *Id.* at 12; O'Connor, *supra* note 179, at 15

191. O'Connor, *supra* note 179, at 15; O'CONNOR ET AL., *supra* note 179, at 12. The significance of this difference is $p<0.10$. *Id.*

However, as I have explained above,[192] we shouldn't read anything into this last set of results. Any analysis that divides inmates by levels of participation merely reintroduces self-selection bias. One can't compare the control group against a self-selected sample of the treatment group, nor can one compare self-selected parts of the treatment group (high Bible study participants) against other self-selected parts (low Bible study participants). Even if this told us the effect of high participation (which it probably doesn't), the proper question for a policymaker deciding whether to introduce such a program is how well it works for everyone, including those who choose not to participate much.

2. *Johnson et al.'s New York Study*

Byron Johnson and his coauthors reanalyzed this data,[193] using only 201 inmates instead of the original 225.[194] They found substantially the same results.[195] There was no significant difference between participating and non-participating inmates in rates of infractions (36% versus 31%), serious infractions (8% versus 9%), or re-arrest (37% versus 36%).[196]

When inmates were broken down by level of participation (low, medium, or high), there continued to be no significant difference between Prison Fellowship (PF) and non-PF inmates, except that high-participating PF inmates were re-arrested at lower rates than their non-PF counterparts (14% versus 41%).[197] High-participating PF inmates were also significantly less likely to be arrested than low- or medium-participating PF inmates.[198] The authors also further broke down inmates by risk level and found that high participation continued to be associated with a lower re-arrest rate.[199]

---

192. *See supra* text accompanying notes 132–134.

193. Byron R. Johnson et al., *Religious Programs, Institutional Adjustment, and Recidivism Among Former Inmates in Prison Fellowship Programs*, 14 JUST. Q. 145, 149 (1997). O'Connor rates this study as having "fair" methodological quality (on a poor-fair-good-excellent scale). O'Connor, *supra* note 33, at 23 tbl.3.

194. Johnson et. al., *supra* note 193, at 143. Johnson's control group was a subset of O'Connor's control group after some of the participants were dropped for data-quality reasons. Email from Tom P. O'Connor, Administrator, Religious Serv., Oregon Dept. Corrections, to Alexander Volokh, Assistant Professor of Law, Emory Univ. School of Law, (Mar. 8, 2010).

195. Johnson et al., *supra* note 193, at 161.

196. *Id.* at 154 tbl.2. The p-values of these differences are 0.342, 0.727, and 0.756, respectively. *Id.*

197. *Id.* at 156 tbl.3. The p-value of this difference is 0.042; all other p-values are above 0.10, usually substantially above it. *Id.* at 156 tbl.3. Similarly, Rose states that the recidivism differences are significant at p=0.058 if the PF group is subdivided into low, medium, and high participation. *See* Rose, *supra* note 63, at 296–97, 316 n.5.

198. Johnson et al., *supra* note 193, at 157. Similarly, in a regression, the variable for "Bible study," measuring whether an inmate had attended ten or more Bible studies, had a significant statistical effect on re-arrest. *Id.* at 157.

199. *Id.* at 157. Rose calls the discussion around Johnson et al.'s Tables 3 and 4 "somewhat difficult to follow" and "not at all clear." Rose, *supra* note 63, at 297. Rose reanalyzed the data, and found that, after controlling for risk, there was no statistically significant effect of PF participation. *Id.* at 358–59 app. A.

But, as discussed above,[200] we shouldn't divide the sample based on participation level, since this introduces a new source of self-selection bias.

When Johnson did a follow-up evaluation on these same inmates seven years later,[201] he again found no significant difference in median time to re-arrest or in reincarceration rates between participating and non-participating inmates.[202] When the sample was divided into high- and low-participating groups, high-participating inmates had a lower two-year probability of re-arrest than low-participating ones, but this effect disappeared after three years.[203]

3. *Camp et al.'s Life Connections Program Study*

Scott Camp and his coauthors analyzed the effect on prison misconduct of participation in the Life Connections Program.[204] They estimated the probability of participation (i.e., propensity score) using a number of models; the fit of these models was reasonably good.[205] Variables used included a "scale of motivation for change," frequency of spiritual experiences and religious observance, religious affiliation, "feelings of self-worth," custody risk, previous incarceration, age, ethnicity, "race, sex, education, marital status, and months of current incarceration" so far.[206]

There was generally no significant association between participation and misconduct in general, and no association between participation and less serious misconduct.[207] However, there was a significant association between participation and serious misconduct: in some of the models, "slightly over 5 percent of the inmates in the LCP had an instance of serious misconduct, where for the comparison group, the number was closer to 11 percent."[208] Other models on serious misconduct produced differences that were smaller, but still significant.[209]

This article has a significant advantage that the others in this Subpart don't have. I've argued that the problem with comparative studies, even ones based on propensity scores, is that they don't get at the unobserved motivation to change. As I've noted above, though, Camp et al. explicitly

---

200. *See supra* text accompanying notes 132–134, 192–193.

201. Byron R. Johnson, *Religious Programs and Recidivism Among Former Inmates in Prison Fellowship Programs: A Long-Term Follow-Up Study*, 21 JUST. Q. 329, 335–36 (2004). O'Connor rates this study as having "fair" methodological quality (on a poor-fair-good-excellent scale), O'Connor, *supra* note 33, at 23 tbl.3, and Steve Aos and his coauthors considered this study to be one of the few that were of good enough quality to include in their review of evidence-based adult corrections programs, *see* AOS ET AL., *supra* note 33, at 19.

202. Johnson, *supra* note 201, at 342–43.

203. *Id.* When "high participation" was redefined as five or more Bible studies rather than ten, the effect did persist through the third year. *Id.* at 344 tbl.3.

204. Scott D. Camp et al., *The Effect of Faith Program Participation on Prison Misconduct: The Life Connections Program*, 36 J. CRIM. JUST. 389, 389 (2008).

205. *Id.* at 391–92.

206. *Id.* at 392.

207. *Id.* at 393–94.

208. *Id.* at 394.

209. *Id.* at 393–94.

include "a scale of motivation for change" in their first-stage propensity model.[210] If this scale accurately measures motivation for change, then it can potentially solve the selection problem. Unfortunately, this scale, developed by Prochaska and DiClemente,[211] is derived from inmates' own self-reported views,[212] so it should be taken with a grain of salt.

4. Education Studies

As with the previous set of studies, this ground has already been trodden by education researchers, with similar methodological vulnerabilities. Unobserved motivation is as problematic with private or Catholic schools as with faith-based prisons—a student's (or his parents') motivation is correlated both with a decision to choose a different school and with success on outcomes like test scores.[213]

Thomas Hoffer and his coauthors (including James Coleman) predicted the probability that a student would choose a Catholic school using the background measures used in his base-year analysis and a measure of sophomore achievement.[214] Then they "stratified the sample into quintiles of the propensity score[s] and estimated Catholic-school effects within each of these homogeneous groups."[215] They found that controlling for selection using this method didn't change the results much relative to the results earlier in their paper, which they had estimated without propensity scores.[216]

Stephen L. Morgan similarly estimated propensity scores and stratified the sample into quintiles.[217] He found that "there is considerable variation in estimates of the average causal effect for Catholic school students with different propensities for attending Catholic schools";[218] "the Catholic students who are least likely to be enrolled in Catholic schools . . . are the most likely to benefit from having attended a Catholic school."[219] Overall, he found that students in Catholic schools benefited from attending those schools, and—

210. *Id.* at 392.

211. *See* SCOTT D. CAMP ET AL., FED. BUR. OF PRISONS, AN EXPLORATION INTO PARTICIPATION IN A FAITH-BASED PRISON PROGRAM 13 (April 25, 2006), http://www.bop.gov/news/research_projects/published_reports/gen_program_eval/volunteer_cpp.pdf.

212. *See* Carlo C. DiClemente & James O. Prochaska, *Toward a Comprehensive, Transtheoretical Model of Change: Stages of Change and Addictive Behaviors*, *in* TREATING ADDICTIVE BEHAVIORS 3, 8–10 (William R. Miller & Nick Heather eds., 2d ed. 1998) (describing how readiness to change is assessed through various questionnaires).

213. *See* Dan D. Goldhaber, *School Choice: An Examination of the Empirical Evidence on Achievement, Parental Decision Making, and Equity*, EDUC. RES., Dec. 1999, at 16, 16; Paul Teske & Mark Schneider, *What Research Can Tell Policymakers about School Choice* 20 J. POL'Y ANALYSIS & MGMT. 609, 623–24 (2001); Goldberger & Cain, *supra* note 143, at 109–10; Noell, *supra* note 146, at 128–32.

214. Hoffer et al., *supra* note 150, at 88 n.8.

215. *Id.* at 88.

216. *Id.*

217. Stephen L. Morgan, *Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning*, 74 SOC. OF EDUC. 341, 352–54 (2001).

218. *Id.* at 359.

219. *Id.*

unlike Hoffer et al.—the effect he estimated was *larger* than the standard regressions that didn't control for selection into Catholic schools.[220]

In any event, because these studies don't account for selection on unobservables, it isn't worth dwelling on them at length. Since there are more valid studies that *are* able to control for selection on unobservables, let's move on to those.

III. POTENTIALLY VALID STUDIES

The only credible studies of faith-based prisons done so far have been those where the comparison group of inmates was made up of those who volunteered for the faith-based program but were rejected. However, before describing those studies, I discuss a few empirical strategies that have been used for private schools but, for whatever reason, haven't been attempted for faith-based prisons: the instrumental variables method and identification by exogenous policy shocks.

A. The Roads Not Taken

The empirical literature on education is extremely large, and there has been a lot of debate on appropriate empirical methods. Here, I focus on two widely used approaches that can deal with selection: the instrumental variables approach and the exogenous policy shock approach.[221]

1. Instrumental Variables

Standard regression models (the "ordinary least squares" method)[222] take as given that we won't be able to explain *all* of the variable of interest, whether that variable is ex-prisoners' recidivism or students' test scores.[223] There will always be some error, as is recognized by the ε term in the stan-

220. *Id.* at 359.

221. Altonji et al. suggest another way of dealing with selection. They estimate the effect of Catholic school on high school graduation and on college entrance *assuming that "selection on the* []*observables is the same as selection on the* [*un*]*observables*." Joseph G. Altonji et al., *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools*, 113 J. POL. ECON. 151, 154 (2005) (emphasis added). Their method

> requires some strong assumptions . . . (1) that the set of observed variables is chosen at random from the full set of variables that determine Catholic school attendance and high school graduation and (2) that the number of observed and unobserved variables is large enough that none of the elements dominates the distribution of school choice or graduation.

Id. However, Altonji et al. "argue that these assumptions are no more objectionable than the assumptions needed to justify the standard ordinary least squares (OLS) or univariate probit requirement that the index of unobservables that determine graduation has no relationship with Catholic school attendance." *Id.* Altonji et al. find little evidence that Catholic school affects test scores, but conclude that "Catholic high schools substantially increase the probability of graduating from high school and, more tentatively, attending college." *Id.* at 152.

222. JEFFREY M. WOOLDRIDGE, INTRODUCTORY ECONOMETRICS 30 (4th ed. 2009).

223. *Id.* at 22.

dard notation, $y = X\beta + \varepsilon$.[224] The models do, however, demand that the average value of the error term, ε, not depend on the explanatory variables in X.[225] It's useful to think of the error as embodying not just whatever inherent randomness may exist in the world, but also every omitted variable. The requirement that ε, on average, not depend on X, can thus be interpreted as a rule that one can harmlessly omit variables (either by choice or because they're unobservable) as long as the omitted variables are uncorrelated with the included ones.[226]

This is precisely the problem with selection bias: the inmates' or the student's parents' motivation (which is an omitted variable and therefore part of the error term) is correlated with the main explanatory variable—whether or not the inmate signs up for a faith-based prison program or the student attends a private or Catholic school.

If we ran an ordinary least squares regression on the equation above, we would get biased estimates of β. But there are ways around this. Suppose we could find some other variable, Z, that predicted X but was uncorrelated with unobservable motivation. For instance, suppose Catholic religion (Z) predicted whether someone attended Catholic school (X) (this seems true, since Catholics are more likely to attend Catholic school) but was uncorrelated with the unobservable determinants of scholastic success (this seems possible, since why would Catholics do better in school?).[227] We would call Z an *instrument* for X.

We would then use a two-stage process, called the instrumental variables (or IV) method.[228] Initially, we would use Z to obtain a predicted value of X—call it X'.[229] Instead of having a 0 or 1 value of whether someone attended a faith-based prison program or Catholic school, we'd have their *predicted* value based on Z; this would typically be a number between 0 and 1,[230] and we could think of it as their probability of attending the program.

Once this first stage was done and we had our predicted X', we would replace X with X' in the regression, and estimate the regression $y = X'\beta + \varepsilon$. We would then use the resulting estimate of β. (This method thus has the *flavor* of matching based on propensity scores, as discussed above,[231] but it has the advantage of being able to handle selection on unobservables.[232])

224. *Id.* at 23.
225. *Id.* at 25.
226. *See id.* at 24–26, 69.
227. *See infra* text accompanying note 234. *But see infra* text accompanying notes 238–243 (noting how Catholic religion may not be a good instrument after all, because it *can* affect scholastic success).
228. *See generally* WOOLDRIDGE, *supra* note 222, at 506–25.
229. *See id.* at 507–08.
230. The value can be forced to be between 0 and 1, for instance if the initial regression of X on Z is in the probit form, a form often used for predicting probabilities.
231. *See supra* Part II.C.
232. As applied to the selection problem, it's primarily based on the work of James Heckman. James J. Heckman, *Sample Selection Bias as a Specification Error*, 47 ECONOMETRICA 153 (1979). There's a more complicated version of Heckman's selection correction, which involves estimating two regressions, one for participants and the other for non-participants; this method has the advantage of not as-

Mathematically, it turns out that, unlike the naïve estimation, this two-stage IV process gives us an unbiased estimate of β. The advantage of using X' instead of X is that, because X' is just predicted off of Z (which is uncorrelated with the error term), it isn't "contaminated" by whatever is in the error term, like unobservable motivation.[233] In essence, using the two-stage process has "purged" X of the pernicious effects of unobserved motivation.

Of course, whether the IV method works depends on whether we can find a true instrument—something that really predicts X and is really uncorrelated with ε. Good instruments are hard to find. We can test whether Z predicts X—just try doing it and see how well it works—but we can't directly test whether Z is correlated with ε, since the true error term is unknown; this is unfortunate, since even moderate correlations can introduce substantial bias into the IV estimates.[234]

These potential problems haven't stopped education researchers from using IV methods.

James Coleman and his coauthors used two strategies. First, they used religion together with region (Northeast or other) as instruments for Catholic school attendance; then they used religion together with income and educational expectations in the eighth grade.[235] They rejected both of these models because the resulting Catholic-school effect was implausibly large.[236] But note that even if religion is a valid instrument,[237] it seems that income and previous educational expectations *should be* correlated with the unobservable determinants of scholastic success, which makes them invalid instruments.

Other authors, using different specifications, have found conflicting results. Jay Noell, in the reanalysis of Coleman's work discussed above,[238] also used Catholic religion as an instrument for Catholic school attendance;[239] this made the Catholic-school effect insignificant.[240] Richard Murnane and his coauthors, on the other hand, used Catholic religion as an instrument and determined that Catholic school attendance had a significant effect on Hispanic students', and possibly also on black students', achievement.[241]

Using Catholic religion as an instrument seems to have fallen out of fashion, after various researchers suggested that being Catholic is unfortunate-

suming a constant effect of treatment. But I don't discuss it here because (1) it's basically a generalization of the Heckman method that I *do* discuss, and (2) neither faith-based prison nor private/Catholic education studies, to my knowledge, have used it.

233. *See* WOOLDRIDGE, *supra* note 222, at 506–08.

234. *Id.* at 508–10, 514–15.

235. COLEMAN ET AL., *supra* note 143, at 214.

236. *Id.* at 214 & n.8.

237. *See infra* text accompanying note 241.

238. *See supra* text accompanying note 146.

239. Noell, *supra* note 146, at 130 n.3.

240. *See id.* at 131–32 (finding only a "small advantage" on sophomore reading tests for Catholic schools).

241. Richard J. Murnane et al., *Comparing Public and Private Schools: The Puzzling Role of Selectivity Bias*, 3 J. BUS. & ECON. STAT. 23, 27–28, 29 & tbl.3 (1985).

ly correlated with the unobserved determinants of scholastic success.[242] The same goes for a related variable, frequency of church attendance.[243]

A better instrument might be a variable unrelated to one's own characteristics—perhaps the Catholic share of the population of one's county, which could affect Catholic school attendance just because Catholic-heavy counties have more Catholic schools and possibly lower tuitions because they're more heavily subsidized by their local congregations.[244]

Thus, William Evans and Robert Schwab used, among other variables, Catholic county population as an instrument.[245] This strategy didn't change the high-school graduation results much compared to a naïve specification without instruments, though the college entrance results were more sensitive to the choice of specification.[246]

Jeffrey Grogger and Derek Neal used the county's Catholic school density and the county's percentage of Catholic population.[247] They found Catholic-school effects on high-school graduation for urban minorities that were even larger than in the models without selection.[248] They also found significant effects for urban whites, though no effects for suburban students (whether white or minority).[249] There were no significant effects of Catholic school on college entrance.[250]

Derek Neal used these same variables—county Catholic school density and county Catholic concentration—but not at the same time.[251] He estimated two different models since the validity of the instruments seemed to differ as between urban minorities and urban whites.[252] The analysis of minorities used only Catholic school density as an instrument, while the analysis of whites used only local Catholic population density.[253] A positive effect of Catholic school attendance on high-school graduation rates remained after this correction for selection bias and, in fact, even increased.[254]

---

242. William N. Evans & Robert M. Schwab, *Finishing High School and Starting College: Do Catholic Schools Make a Difference?*, 110 Q.J. ECON. 941, 965 (1995); Murnane et al., *supra* note 241, at 30–31; Derek Neal, *The Effects of Catholic Secondary Schooling on Educational Achievement*, 15 J. LABOR ECON. 98, 104 (1997); *see also* Altonji et al., *supra* note 221, at 153 n.2.

243. William Sander & Anthony C. Krautmann, *Catholic Schools, Dropout Rates and Educational Attainment*, 33 ECON. INQUIRY 217, 221 (1995). Interactions between these religious variables and other variables, as in William Sander, *Catholic Grade Schools and Academic Achievement*, 31 J. HUM. RES. 540, 544, 545 tbl.2 (1996), should likewise be defective instruments.

244. Neal, *supra* note 242, at 105.

245. Evans & Schwab, *supra* note 242, at 968 tbl.VII, 969. They initially used Catholic religion and then also used interaction terms involving Catholic religion and religious attendance. *Id.* at 962, 963 tbl.VI. 966.

246. *Id.* at 968–71, 968 tbl.VII.

247. Grogger & Neal, *supra* note 155, at 178–79.

248. *Id.* at 179. However, they couldn't exclude these instruments from the attainment models for minorities. The estimates for the minority samples were "identified only by the assumptions concerning the functional form of the model and the joint normality of the errors." *Id.*

249. *Id.* at 179–80, 182.

250. *Id.* at 180 tbl.6A, 181 tbl.6B.

251. Neal, *supra* note 232, at 105.

252. *Id.* at 110.

253. *Id.* at 112.

254. *Id.* at 105, 110, 112, 113 tbl.6. Neal's estimates of the effect of Catholic school on college

Other studies use instruments unrelated to Catholicism. William Sander and Anthony Krautmann used, among other variables, "urban" interacted with region and concluded that Catholic schooling has a highly significant negative effect on the probability that a sophomore drops out before his senior year, but no effect on educational attainment beyond high school.[255]

Dan Goldhaber also used a number of variables, including controls for the cost and availability of private schools, dummy variables for region and urbanicity, and percent of white students at the students' school.[256] He found no positive sectoral effect favoring private schools.[257]

David Figlio and Joe Stone predicted sector choice using, among other factors, whether the state had "duty to bargain" or "right-to-work" laws, as well as median county income.[258] They found that private schools, whether religious or nonreligious, had no relation to math test scores, but were significantly related to two years of college enrollment, as well as enrollment in a selective college.[259]

All these models use different specifications, have different choices of instruments, and yield different results. Some find an effect of private or Catholic schools; some don't. The moral, though, is that finding a good instrument is hard. Many instrumental-variables studies have been sloppy about why the instrument Z is correlated with X and why it's uncorrelated with ε.[260] Pretty much any *individual* attribute, whether Catholic religion, or income, or race, probably has some correlation with the unobserved determinants of success.[261] *Aggregate* variables, like perhaps the Catholic population density in the child's county, *may* work better, but of course aggregate variables may also affect achievement. Moreover, the aggregate approach only works as an estimation strategy if we observe children from a large number of different aggregates: If all the children in the study come from the same county, we won't be able to use the local Catholic population density as an instrument since it will be the same for each child.

This is a problem for faith-based prison studies as well. So far, almost all faith-based studies have analyzed the results of a *single* faith-based pro-

graduation rates and on future wages didn't control for self-selection. *Id.* at 117–18.

255. Sander & Krautmann, *supra* note 243, at 221–24, 225 tbl.III, 226 tbl.IV, 227 tbl.V (using personal religious variables and then also using "urban" interacted with Catholic).

256. Dan D. Goldhaber, *Public and Private High Schools: Is School Choice an Answer to the Productivity Problem?*, 15 ECON. OF EDUC. REV. 93, 96 (1996).

257. *Id.* at 98. He also predicted attendance at private school using various student and family background variables including religion, parents' schooling, family size, day care attendance, amount of money set aside for future education, gender, race, learning disability, the number of high-school courses taken by the student, family income, and eighth-grade test score. *Id.* at 101.

258. David N. Figlio & Joe A. Stone, *Are Private Schools Really Better?*, 18 RES. LABOR ECON. 115, 121 (1999).

259. *Id.* at 121, 131. They also found that private schools, whether religious or nonreligious, had a significant *negative* effect on high-school completion, but rejected this conclusion after running an alternative specification. *Id.*

260. Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics*, 24 J. ECON. PERSP. 3, 11 (2010).

261. Altonji et al., *supra* note 221, at 152.

gram at *one* prison. Only a small number deal with more than one prison.[262] Perhaps an IV approach wouldn't have been very useful in most of these cases, but it would be worth exploring the IV method when there is a data set with inmates from several prisons.

2. *Exogenous Policy Shocks*

Other studies have identified the effect of educational policies using exogenous shocks. Some of these are natural shocks; some are policy shocks when a policy is first introduced; some are policy shocks when an already existing policy is applied in a particular context for a random reason.

Here are some examples, unrelated to the public versus private school debate:

- Caroline Hoxby identified the effect of class size on student achievement using two strategies.[263] First, she used natural randomness in the population, which makes certain classes larger or smaller from year to year.[264] Second, she used "the fact that class size jumps abruptly when a class has to be added to or subtracted from a grade because enrollment has triggered a maximum or minimum class size rule."[265] (This is the "regression discontinuity" approach.)[266] Both strategies showed little or no effect of class size on achievement.[267]

- Joshua Angrist and Victor Lavy used a similar regression discontinuity approach to study the effect of class size on achievement in Israel.[268] Unlike Hoxby, they found a negative effect of class size on achievement.[269] (Of course, in all studies of this type, we want to guard against parents' ability to game the system by choosing schools with enrollments just above the cutoff, which would bias the results.)[270]

- Martin West and Paul Peterson examined the effect on a Florida public school of receiving an F grade on the state's A+ Accounta-

262. La Vigne et al.'s Florida study, *see supra* notes 103–110 and accompanying text, deals with two different prisons, but even here, there is only one per gender. The O'Connor et al. and Johnson et al. studies, *see supra* notes 179–203 and accompanying text, do discuss a few prisons. So does OPPAGA's FCBI study, *see infra* notes 291–301 and accompanying text.
263. Caroline M. Hoxby, *The Effects of Class Size on Student Achievement: New Evidence from Population Variation*, 115 Q.J. ECON. 1239, 1241 (2000).
264. *Id.* at 1241–42.
265. *Id.* at 1242.
266. *Id.* at 1254.
267. *Id.* at 1280. *But see infra* text accompanying notes 268–270.
268. Joshua D. Angrist & Victor Lavy, *Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement*, 114 Q.J. ECON. 533, 533 (1999).
269. *Id.* at 569.
270. Angrist & Pischke, *supra* note 260, at 13.

> bility Plan.[271] Students at schools that received an F twice would get a voucher for private school; F schools were also assigned a team to write an intervention plan for the school.[272] In 2002, Florida changed its evaluation system so that most schools received a different grade than the previous year.[273] To isolate the effect of an F—separate from the effect of being subject to a voucher threat—the authors focused only on the 24 schools that hadn't previously gotten an F, that *wouldn't* have gotten an F under the old system, but that *did* get an F under the new system.[274] They compared these schools to all D schools whose scores were close to those of the 24 F schools.[275] They found that getting an F had a significant positive effect on student achievement.[276] The same was true for D schools, as compared to C schools.[277]

There are many more examples.[278] Exogenous policy shocks are another way of dealing with self-selection: If we compare an *entire* prison before the introduction of a faith-based program with the *same* prison after the introduction of the program, we don't have to deal with self-selection issues as long as people don't choose which prison they go to, and as long as the assignment mechanism didn't change once the program was introduced.

Or we could compare a prison with a faith-based program to a prison without one, though one would want to be sure that the two prisons are really comparable. Again, the comparison would have to be between *entire* prisons since limiting the set at one prison to participants would introduce self-selection issues.

Or one could merge the two approaches and observe how the *difference* between two prisons changed when a faith-based program was introduced at *one* of them. This would essentially be a differences-in-differences approach.

So the exogenous policy shock approach seems promising for faith-based prisons. This is another area where prison researchers could learn from education researchers.

B. Using Rejected Volunteers

I will now discuss the studies that use rejected volunteers as the control group. Unsuccessful applicants seem like the best control group, but in fact

271. Martin R. West & Paul E. Peterson, *The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments*, 116 ECON. J. C46, C48–49 (2006).
272. *Id.* at C48.
273. *Id.* at C47.
274. *Id.* at C51.
275. *Id.*
276. *Id.* at C53 tbl.2.
277. *Id.* at C54 tbl.3.
278. *E.g.*, Brian A. Jacob, *Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools*, 89 J. PUB. ECON. 761 (2005).

they're not completely ideal. For instance, there may be nonrandom attrition from the program. "[I]f the more motivated parents among the unsuccessful applicants were more likely to enroll their child in a private school outside of the choice program"—where statistics aren't being kept—the unsuccessful applicants group would look worse, and the estimate of the effect of getting a voucher would be inflated.[279]

There may be other issues, like exceptions to random assignment—a "sibling" rule for schools,[280] or allowing some rejected students to enter from waiting lists after the beginning of the year—or just lack of oversight of the random selection process. Some analysts, like John F. Witte, have therefore concluded that the rejected applicants approach is worse than instrumental variables or even than standard approaches that don't control for selection.[281]

But even if one concludes that rejected applicant groups aren't ideal for schools, the problems seem much less in prisons. The attrition in favor of schools outside the system doesn't seem so problematic in the prison context since both the successful and the rejected applicants are, so to speak, a captive audience. The same goes for sibling rules. Oversight of the random selection process is still important, but overall, it seems like rejected applicant studies of faith-based prisons are substantially better than the other studies to date. (And other methods that take selection on unobservables into account, like instrumental variables or exogenous policy shocks, simply haven't been attempted for faith-based prisons.)

The first few studies below find no positive effect of faith-based programs; the next few do find some effect.

1. The Texas InnerChange Studies

Byron Johnson and David Larson conducted a preliminary evaluation of a Texas-based InnerChange Freedom Initiative program (IFI).[282] (This report was based on data in an earlier report by Brittani Trusty and Michael Eisenberg.)[283] They compared the 177 IFI participants against three differ-

279. Rouse, *supra* note 154, at 563.

280. *E.g.*, Beales & Wahl, *supra* note 92, at 47.

281. John F. Witte, Dep't of Political Sci., Robert La Follette Inst., Univ. of Wis.-Madison, Achievement Effects of the Milwaukee Voucher Program (Jan. 4–6, 1997) (transcript available at http://www.disc.wisc.edu/choice/aea97.html).

282. BYRON R. JOHNSON & DAVID B. LARSON, THE INNERCHANGE FREEDOM INITIATIVE: A PRELIMINARY EVALUATION OF A FAITH-BASED PRISON PROGRAM (Ctr. for Res. on Rel. & Urban Civil Soc'y, 2003). O'Connor rates this study as having "good" methodological quality (on a poor-fair-good-excellent scale). *See* O'Connor, *supra* note 33, at 23 tbl.3.

283. BRITTANI TRUSTY & MICHAEL EISENBERG, INITIAL PROCESS AND OUTCOME EVALUATION OF THE INNERCHANGE FREEDOM INITIATIVE: THE FAITH-BASED PRISON PROGRAM IN TDCJ (Crim. Just. Policy Council, Feb. 2003). Steve Aos and his coauthors considered the Trusty and Eisenberg study to be one of the few that were of good enough quality to include in their review of evidence-based adult corrections programs. AOS ET AL., *supra* note 33, at 19. (A preliminary paper describing the program is MICHAEL EISENBERG & BRITTANI TRUSTY, OVERVIEW OF THE INNERCHANGE FREEDOM INITIATIVE: THE FAITH-BASED PRISON PROGRAM WITHIN THE TEXAS DEPARTMENT OF CRIMINAL JUSTICE (Crim. Just. Pol'y Council, Feb. 2002).) Johnson and Larson further break participants down by time spent in

ent groups: (1) a "match group" of 1,754 inmates who "met IFI selection criteria but did not participate in the program," (2) a "screened group" of 1,083 inmates who "were screened as eligible for the program but did not volunteer or were not selected for program participation," and (3) a "volunteer group" of 560 inmates who "actually volunteered for the IFI program, but did not participate, either because they were not classified as minimum-out custody, their remaining sentence length was either too short or too long to be considered, or they were not returning to the Houston area following release."[284] Of these three groups, only the third avoids selection bias.

IFI participants did no better than the other groups in either two-year re-arrest or reincarceration rates. Two-year re-arrest rates were 36.2% for the IFI group, compared to 35% for the match group, 34.9% for the screened group, and 29.3% for the volunteer group.[285] Two-year reincarceration rates were 24.3% for the IFI group, compared to 20.3% for the match group, 22.3% for the screened group, and 19.1% for the volunteer group.[286]

It's true that IFI *graduates* had lower re-arrest (17.3%) and reincarceration (8.0%) rates.[287] But IFI's definition of "graduation" is "quite restrictive" and includes completing 16 months in the IFI program, completing 6 months in aftercare, and holding a job and having been an active member in church for the 3 months before graduation.[288] Inmates could be removed from the program "for disciplinary purposes," "at the request of IFI staff," "for medical problems," and "at the voluntary request of the applicant."[289] The set of inmates who "graduated" from the program is thus tainted by self-selection (the decision to participate), selection by the program staff (the decision not to expel), and "success bias" (the decision to finish the program, which in this case even includes a post-release component).[290]

2. OPPAGA's FCBI Study

Florida's Office of Program Policy Analysis and Government Accountability (OPPAGA) published a report on several "faith- and character-based programs" in Florida prisons.[291]

the program and whether early parole was granted.

284. JOHNSON & LARSON, *supra* note 282, at 12–13.

285. *Id.* at 17 tbl.3.

286. *Id.*

287. *Id.* at 20 tbl.6.

288. *Id.* at 15.

289. *Id.* at 18.

290. *See also* Camp et al., *supra* note 46, at 535; Rose, *supra* note 63, at 302; Mark A.R. Kleiman, *Faith-Based Fudging*, SLATE (Aug. 5, 2003, 12:35 PM), http://www.slate.com/id/2086617/; *cf.* DOUGLAS MCDONALD ET AL., PRIVATE PRISONS IN THE UNITED STATES: AN ASSESSMENT OF CURRENT PRACTICE, appx. 2, at 29 (Abt Assocs., 1998). Rose also points out that the extremely high recidivism rate among IFI participants who didn't complete the program because they were paroled early (62.7%, JOHNSON & LARSON, *supra* note 282, at 20 tbl.6)—and who were naturally excluded from the group of IFI "graduates"—is "odd" and "puzzl[ing]," and should be investigated further. Rose, *supra* note 63, at 302–04.

291. OFFICE OF PROGRAM POL'Y ANALYSIS & GOV'T ACCOUNTABILITY, FAITH- AND CHARACTER-BASED PRISON INITIATIVE YIELDS INSTITUTIONAL BENEFITS; EFFECT ON RECIDIVISM MODEST 2 (Oct.

Some of these programs were institution-wide, "offered to all inmates," and "incorporated into the facility's mission."[292] These programs included Bible study groups, Native American prayer, parenting skills, and yoga classes,[293] so they really don't count as "faith-based prisons" as we are using the term here.[294]

Other programs were dorm-based; the dorms were "established as . . . enclave communit[ies] within the prison compound."[295] The dorm-based programs "provide a more intensive experience than the prison-wide programs"[296] and look more like the faith-based prisons that we have been discussing.

The authors compared 1,293 inmates released from a faith- and character-based *institution* with 2,283 inmates who had requested transfer to such an institution but weren't placed there before their release.[297] They also compared 1,311 inmates released from a faith- and character-based *dorm* with 9,988 inmates who had requested transfer to such a dorm but weren't placed there before their release.[298] (The study doesn't say why the comparison inmates weren't accepted.)[299]

For the institution-wide programs, the study found that inmates' relative risk of reoffending ranged from 0.85 to 0.95 relative to the comparison group, depending on the institution.[300] The authors found no positive effect of the dorm-based programs—on the contrary, the relative risk of reoffending for inmates released from such dorms was 1.03 relative to the comparison group.[301]

3. *Hall's Putnamville Study*

Stephen Hall examined the effect of the Biblical Correctives to Thinking Errors program on in-prison infraction rates of inmates at the Putnamville Correctional Facility in Indiana.[302] The study was open to volunteers who weren't participating in other treatment programs, who regularly participated in chapel programs, and who had graduated from the chapel's Chris-

2009), www.oppaga.state.fl.us/MonitorDocs/Reports/pdf/0938rpt.pdf.

292. *Id.*

293. *Id.*

294. *See supra* text accompanying notes 36–38 discussing types of "immersion" programs relevant to this article.

295. OFFICE OF PROGRAM POL'Y ANALYSIS & GOV'T ACCOUNTABILITY, *supra* note 291, at 2.

296. *Id.* at 2, 6.

297. *Id.* at 8.

298. *Id.*

299. *Id.*

300. *Id.* at 9.

301. OFFICE OF PROGRAM POL'Y ANALYSIS & GOV'T ACCOUNTABILITY, *supra* note 281, at 9. The authors didn't report statistical significance, writing that because the results were "based on the entire population of cases . . . , inferential analyses using p-values and confidence intervals were not appropriate and [their] analysis addressed the magnitude of the differences between treatment and control groups for both the faith- and character-based institutions and the dorms." *Id.* at 9 exhibit 1 note.

302. Hall, *supra* note 134, at 112–13, 120, 137.

tian twelve-step program.[303] After 46 inmates responded and 8 of these were transferred or discharged, the remaining 38 were divided into a treatment group of 10 and a control group of 28.[304]

There were no infractions in the treatment group, and 17 infractions in the control group (all from 6 of the 28 members).[305] The difference was significant, but the authors wrote that "the sample size in this study is too small to make a case for validity."[306]

4. Hercik et al.'s Kairos Horizon Study

Jeanette Hercik and her coauthors evaluated the effect of participation in the Kairos Horizon Communities in Prison program at Florida's Tomoka Prison.[307]

The authors considered 413 inmates who participated in any of the first five classes of the program.[308] (Class One ran from November 1999 to October 2000; Class Two ran from May 2000 to April 2001; and so on.)[309]

First, participants were compared against their previous selves. After the treatment started, the proportion of participants with at least one discipline report dropped from 24.4% to 12.3%, and this proportion remained in the 12–17% range through three years after the start of treatment (two years after the end of treatment).[310] Similarly, the proportion of participants with at least one segregation stay dropped from 20.6% to 10.6%, and this proportion hovered around 15–16% through three years after the start of treatment, with a blip up to 18.2% in the 25–30-month range.[311]

Next, the 157 participants in Classes Four and Five were compared against two different groups: a "Matched Comparison" group of 157 inmates who were eligible but didn't apply, and a "Waiting List Comparison" group of 248 inmates who were eligible and *did* apply.[312] From the start of treatment, the proportion of the treatment sample with at least one discipline report was lower than for either of the comparison samples (14% versus 25% and 31%, respectively), and the proportion stayed lower through two years after the start of treatment,[313] though this difference wasn't significant

303. *Id.* at 112.
304. *Id.* The article doesn't say why these eight were transferred or discharged or whether assignment to the two groups was random. *Id.* at 112–13, 120, 137.
305. *Id.* at 120.
306. *Id.*
307. JEANETTE HERCIK ET AL., COMPASSION CAPITAL FUND EVALUATION OF THE KAIROS HORIZON COMMUNITIES IN PRISON PROGRAM: FINAL REPORT (2004). A shorter version of this report is OFFICE OF CMTY. SERVS., U.S. DEP'T OF HEALTH AND HUMAN SERVS., REDISCOVERING COMPASSION: AN EVALUATION OF KAIROS HORIZON COMMUNITIES IN PRISON (Horizon Communities in Prison, n.d.), *available at* http://www.horizoncommunities.org/images/Horizon3%20-compassion%20.pdf.
308. HERCIK ET AL., *supra* note 307, at IV-35 tbl.4-1.
309. *Id.* at IV-33.
310. *Id.* at IV-43 fig.4-1.
311. *Id.* at IV-43 fig.4-1, IV-47 fig.4-4.
312. *Id.* at IV-45 fig.4-2.
313. *Id.*

past the 12-month mark.[314] Similarly, the proportion of the treatment sample with at least one segregation stay was lower than for either of the comparison samples (13% versus 26% and 25%, respectively), and the proportion stayed lower through two years after the start of treatment; these differences were all significant.[315]

The probability of re-arrest of participants during the follow-up period (19.0% among those released during the study period) was greater than that of the matched comparison group (15.2%) and basically the same as that of the waiting list group (19.6%).[316] Program participation may be associated with a somewhat longer time for re-arrest (3.5 months for the treatment sample, 1.4 months for the matched comparison group, and 3.2 months for the waiting list comparison group), but the standard deviations are so large that I doubt that these differences are significant.[317]

The matched comparison sample is subject to self-selection bias,[318] and the comparison of participants to their previous selves is probably also biased because those who choose to participate probably have a greater responsiveness to the material. So the waiting list comparison group is the most valid control group. For this group, while the difference in discipline reports and segregation stays may be significant, participation seems to confer no significant advantage in the probability of re-arrest.

5. Wilson et al.'s Detroit TOP Study

Leon Wilson and his coauthors prepared an unpublished report on an ex-prisoner aftercare program, the Detroit Transition of Prisoners (TOP) program.[319]

A group of 135 former inmates who participated[320] was compared to a 139-member designated control group,[321] mainly composed of former in-

314. This isn't reported in the study but can be calculated from the authors' raw data. The difference between the treatment group and the waiting list comparison group is only significant at the 5% level through the 12-month mark. The difference between the treatment group and the matched comparison group is only significant at the 5% level through the 12-month mark and in the 19–24-month period, but the matched comparison group isn't a good comparison group in any event. *See infra* text accompanying note 322.

315. HERCIK ET AL., *supra* note 307, at IV-49 fig.4-5. Significance can be calculated from the authors' raw data.

316. *Id.* at IV-56 tbl.4-9. The significance level (p=0.92) can be calculated from the raw data.

317. *Id.* The standard deviations for the months to first re-arrest are 3.907, 0.973, and 3.205, respectively.

318. *Id.* at IV-33 to -34.

319. LEON C. WILSON ET AL., PROMISING EFFECTS ON THE REDUCTION OF CRIMINAL RECIDIVISM: AN EVALUATION OF THE DETROIT TRANSITION OF PRISONER'S FAITH BASED INITIATIVE (n.d.). Steve Aos and his coauthors considered this study to be one of the few that were of good enough quality to include in their review of evidence-based adult corrections programs. AOS ET AL., *supra* note 33, at 19. The data here is the same as that in Leon C. Wilson, Detroit TOP Evaluation Report (2000) (unpublished report) (on file with author).

320. The definitions of the treatment and control groups are a little confusing. "[T]he control group is comprised of ex-prisoners who had some contact with the Detroit TOP program but received no services. Contact is defined as having completed the application process and the screening materials inclusive of the LSI-R inventory." WILSON ET AL., *supra* note 319, at 11. However, there is a subcategory of the

mates who applied but were turned down because they didn't meet the inclusion criteria.[322] The TOP program was trying to take people it believed to be high risks, so the treatment group was actually estimated to be at higher risk for recidivism than the non-treatment group.[323]

The recidivism rate was 18% for graduates, as compared to 57% for the control group.[324] However, the set of "graduates" is the result of a significant weeding-out process.[325] Of the 124 initial participants, only 66 remained in the program for six months and only 47 remained after a year.[326] Only 40 graduated from the program; others didn't complete the one-month probationary period, were terminated for rule violations, didn't participate, or just lost contact with the program after applying.[327] These groups all had recidivism rates much higher than 18%, and even mostly higher than the 57% of the control group.[328]

The study doesn't give the recidivism rate for the entire population of participants. Using the authors' data from their adjusted regressions, we can estimate the recidivism rate at roughly 52% for participants and 57% for the control group, which isn't a significant difference.[329] But once we use the adjusted recidivism rate, which the authors obtain after controlling for risk rating, age, and education—so the treatment and control groups are more comparable—the recidivism rate comes out at roughly 54% for participants and 68% for the control group, which *is* a significant difference.[330]

---

treatment group called "Lost Contact after application," *id.* at 28 tbl.5, even though the previous quotation seems to define these as part of the control group. I assume that, in reality, the control group was composed of inmates who applied and received no services because they were *rejected*; those who applied and were accepted but received no services because they lost contact with the program are considered part of the treatment group.

321. *Id.* at 28 tbl.5. While most of the paper uses these numbers, another table in the paper has other numbers: 38 graduates, 113 inmates who had had some contact with the group, and 120 control inmates. *Id.* at 29 tbl.6. The authors write that "because of missing data on key variables, not all [274] cases are included in every analysis." *Id.* at 17.

322. *Id.* at 11.

323. *Id.* at 11, 18.

324. *Id.* at 30 fig.1.

325. *Id.* at 26 tbl.3.

326. *Id.*

327. *Id.* at 11–12, 26 tbl.3, 28 tbl.5, 30 fig.1.

328. According to Figure 1, *id.* at 30, recidivism rates were 18% for graduates (of which there were 40 according to Table 5, *id.* at 42), 81% for those terminated for a rule violation (16), 56% for those who didn't complete the probationary period (20), 65% for those who were terminated for lack of participation (39), and 68% for those who lost contact after application (20). This makes a recidivism rate of approximately 52% overall, or 49% if one excludes those who lost contact after application. (I'm not sure whether those who lost contact after application are part of the control group. *See id.* at 40.) Using the numbers from Table 6, *id.* at 43, instead of Table 5, yields 50% or 48%, which isn't much different.

329. Using Pearson's chi-square test, the significance level is about p=0.32 or p=0.4, depending on which numbers one uses.

330. Using Pearson's chi-square test, the significance level is approximately p=0.02.

6. O'Connor et al.'s Detroit TOP Study

Tom O'Connor and his coauthors examined the same Detroit TOP program.[331] They compared the 60 men who applied for and were accepted into TOP with two control groups—a set of 109 rejected applicants and a random sample of 174 non-applicants who were at the pre-release centers involved in the program.[332] The rejected applicants were rejected for various reasons: some were rejected because they wouldn't be living in Detroit, some because they had insufficient prior church involvement, and some because they had too much time left to serve at the time they applied.[333] Demographic data suggested that the participating group had the highest risk of recidivism, the rejected volunteer group had the next highest risk, and the random sample of non-applicants group had the lowest risk of the three groups.[334]

First, the authors looked at the likelihood of being returned to prison for escaping from the pre-release center.[335] At least when looking at those with three or more felonies, participants did better than rejected volunteers, who did better than the random non-applicants.[336] (On the other hand, the participants had, on average, more church involvement than the rejected applicants.)[337] However, participants with less than three prior felonies did *worse* than the rejected volunteers and no better than non-applicants.[338]

Next, the authors looked at the likelihood of being returned to prison for a parole violation or a new crime.[339] Unfortunately, at this point the authors divided participants into those who stayed with the program and those who were discharged, whether for lack of participation, inappropriate conduct, or escape.[340] This reintroduces selection. We don't know what the results would have been if the group hadn't been subdivided. But even with the subdivision of the participating group into those who continued and those who didn't, the continuing group and the rejected volunteers group were both "two times less likely to have a parole violation or new crime than the general population of ex-offenders."[341] Thus, if the group hadn't been sub-

331. Tom O'Connor et al., Detroit Transition of Prisoners: Final Evaluation Report (Ctr. for Soc. Res., 1997). O'Connor, one of the authors, rates this study as having "good" methodological quality (on a poor-fair-good-excellent scale), O'Connor, *supra* note 33, at 23 tbl.3, and Steve Aos and his coauthors considered this study to be one of the few that were of good enough quality to include in their review of evidence-based adult corrections programs, Aos et al., *supra* note 33, at 19.

332. O'Connor et al., *supra* note 331, at 6.

333. *Id.* at 5–6. A small number were rejected because they weren't being released through a pre-release center. *Id.*

334. *Id.* at 9.

335. *Id.* at 10–14.

336. *Id.* at 13.

337. *See supra* text accompanying note 322.

338. O'Connor et al., *supra* note 331, at 13.

339. *Id.* at 14–18.

340. *Id.* at 15.

341. *Id.* at 17.

divided, we probably would have found that participation conferred no benefit over the rejected volunteers.[342]

7. Education Studies

Private school studies have also been able to use control groups of rejected applicants, thanks to the advent of small-scale voucher programs with a limited number of spots.

Some voucher programs distribute vouchers on a first-come, first-served basis, so the rejected applicants—the ones who applied too late—likely differ systematically from those who were accepted.[343] (Some of the faith-based prison programs above, which don't say how people made it off the wait list,[344] are potentially vulnerable to this problem.)

The most recent studies use data from school voucher programs with limited slots *and* random selection of students off the waiting list.[345] In principle, voucher programs could be problematic ways of testing private versus public school effectiveness. If voucher programs, through the threat of competition, encouraged public schools to improve,[346] a comparative analysis would understate any positive effect of private schools. Fortunately (for private school researchers), voucher programs have, for political purposes, tended to be extremely limited.[347] Some studies have argued that vouchers improve public schools,[348] but clearly the extent of any improvement is

342. Another paper by O'Connor and coauthors on the TOP examined a group of 19 TOP participants who had completed 18 months in the program. Thomas O'Connor et al., *A Model Program for Churches and Ex-Offender Reintegration*, 28 J. OFFENDER REHAB., nos. 1 & 2, 1998, at 107. Over the course of their participation in the program, these participants' mean LSI-R score (which estimates the probability of recidivism) decreased. *Id.* at 120. But this study didn't compare participants against any other group. *Id.* Moreover, to the extent it focuses on participants who completed 18 months, it incorporates two layers of self-selection bias—once for the choice to participate and once more for the choice not to drop out.

343. *E.g.*, Michael Heise et al., *Private Vouchers in Indianapolis: The Golden Rule Program*, *in* PRIVATE VOUCHERS, *supra* note 92, at 100, 102.

344. In the Johnson and Larson study, applicants were rejected "because they were not classified as minimum-out custody, their remaining sentence length was either too short or too long to be considered, or they were not returning to the Houston area following release." JOHNSON & LARSON, *supra* note 282, at 13; *see also* TRUSTY & EISENBERG, *supra* note 283, at 20. O'CONNOR ET AL., *supra* note 331, is explicit about why applicants were rejected, *see supra* text accompanying note 337. Note that, because one reason for rejection was insufficient prior church involvement, there is still an element of self-selection. WILSON ET AL., *supra* note 319, says that applicants were rejected because they didn't meet the program criteria—and one reason for being rejected was that the program was seeking out worse risks. *See supra* text accompanying notes 320–323. The OPPAGA and the Hercik et al. studies say nothing about why rejected applicants were rejected. OFFICE OF PROGRAM POL'Y ANALYSIS & GOV'T ACCOUNTABILITY, *supra* note 291; HERCIK ET AL., *supra* note 307.

345. *See, e.g.*, Rouse, *supra* note 154.

346. *E.g.*, West & Peterson, *supra* note 271; Caroline M. Hoxby, *School Choice and School Productivity: Could School Choice Be a Tide That Lifts All Boats?*, *in* THE ECONOMICS OF SCHOOL CHOICE 287 (Caroline M. Hoxby ed., 2003).

347. Caroline M. Hoxby, *School Choice and School Competition: Evidence from the United States*, 10 SWED. ECON. POL'Y REV. 9, 43 (2003).

348. *See* sources cited *supra* note 346. *But see* Helen F. Ladd, *Comment on Caroline M. Hoxby: School Choice and School Competition: Evidence from the United States*, 10 SWED. ECON. POL'Y REV. 67, 74–75 (2003).

much less than it would be if vouchers were more widely adopted. So this methodological concern shouldn't worry us much.

Several papers have analyzed the Milwaukee Choice program, using unsuccessful applicants as their comparison group. Jay Greene and his coauthors found that private schools produced significant gains in math scores in students' third and fourth years in the program, though no significant effects for reading.[349] (It's plausible that school reforms would improve math more than reading, since math is learned primarily in school while reading is also practiced outside of school.)[350] John F. Witte found no significant effects for reading and weak effects for math.[351] Cecilia Rouse found no consistent effects for reading; for math, there seemed to be some effects, but not until two years after application, and some other specifications yielded no significant differences until the fourth year.[352] However, all three of these papers were apparently based on inaccurate test score data.[353] Greene and his coauthors, using a corrected data set, found significant effects on math scores starting three years in and significant effects on reading scores three or four years in.[354]

Paul Peterson and his coauthors analyzed the New York City School Choice Scholarships program.[355] They found that being offered a voucher had a positive and significant effect on both math and reading scores, at least in grades four and five.[356]

Outside of the public-private school debate, Alan Krueger used a rejected-applicants approach in concluding that smaller class sizes increased average performance on standardized tests.[357]

A few studies have merged the unsuccessful-applicants approach and the instrumental-variables approach. Not all successful applicants enroll in choice schools,[358] so if one uses a rejected-applicants approach, one shouldn't compare the rejected applicants with people who *actually* use the program—that would reintroduce self-selection. Rather, one should compare the rejected applicants with the successful applicants, regardless of whether they used the program. The measured effect isn't the true effect of

---

349. GREENE ET AL., *supra* note 152, at 32 tbl.4.

350. Jay P. Greene & Marcus A. Winters, *Getting Ahead by Staying Behind: An Evalution of Florida's Program to End Social Promotion*, EDUC. NEXT, Spring 2006, at 65, 68.

351. Witte, *supra* note 281; *see also* WITTE ET AL., *Fourth-Year Report*, *supra* note 151. *But see* GREENE ET AL., *supra* note 152, at 14–16.

352. Rouse, *supra* note 154, at 557, 575–78 & 576–77 tbls.Va & Vb.

353. Jay P. Greene et al., *School Choice in Milwaukee: A Randomized Experiment*, *in* EVALUATION IN PRACTICE: A METHODOLOGICAL APPROACH 329, 331 (Richard D. Bingham & Claire L. Felbinger eds., 2002).

354. *Id.* at 337 tbl.3 (reading effects were only significant at $p<0.05$ when the third and fourth year effects were tested jointly).

355. PAUL E. PETERSON ET AL., AN EVALUATION OF THE NEW YORK CITY SCHOOL CHOICE SCHOLARSHIPS PROGRAM: THE FIRST YEAR (Harvard Univ. Educ. Policy & Governance, Oct. 1998).

356. *Id.* at 53 tbl.18. The effects were significant for all grades (jointly) at the 10% level in math and at the 5% level in reading. *Id.*

357. Alan B. Krueger, *Experimental Estimates of Education Production Functions*, 114 Q.J. ECON. 497 (1999).

358. Rouse, *supra* note 154, at 561.

actually attending a choice school. Instead, one should interpret the estimate as an "intention-to-treat" effect rather than as a "treatment" effect.[359]

This is a good approach, since *offering* the voucher is "the only policy instrument available to policy makers," who after all can't force parents to remove their children from public schools.[360] (This point also applies to faith-based prisons.) Still, one may be interested in the actual effect of attending a choice school, particularly if one is a parent. To solve the self-selection problem in the choice whether to attend a choice school, one can use an instrumental-variables approach, using whether one gets a voucher to predict whether one attends a choice school.

Cecilia Rouse took this approach with the Milwaukee voucher program and found that attending a voucher school raised math scores by about 3 percentile points (an estimate she thought overstated the true effect of the program) and had no effect on reading scores.[361] Paul Peterson and William Howell took the same approach with the voucher programs in New York City, Dayton, and Washington, D.C., and found significant achievement gains among African-Americans, immediately in the case of New York City and in the second year in the case of Dayton and Washington, D.C.[362] In other work, Peterson and his coauthors found that switching to a private school had a significant effect, at least after the first year, for African-Americans, but no significant effect for other ethnic groups.[363]

IV. CONCLUSION

So, after discarding the faith-based prison studies tainted by self-selection bias, we're left with two studies that find no effect of faith-based programs,[364] one study that's too small to be meaningful,[365] and three studies that find *some* effect, even if the effect that a few of these find is quite weak.[366] And of those three, two aren't about prisoners at all, but about after-care of released prisoners,[367] and the remaining one shows no significant effect once the prisoners have been released.[368] So we have no study that

359. *Id.* at 559–60.
360. *Id.* at 561.
361. *Id.* at 569–70, 586–88 & 587 tbl.VIII.
362. Paul E. Peterson & William G. Howell, *Exploring Explanations for Ethnic Differences in Voucher Impacts on Student Test Scores*, at 10–12 & 30 tbl.1 (2001) (unpublished manuscript) (on file with Harvard University, Taubman Center for State and Local Government); *see also* PETERSON ET AL., *supra* note 355, at 14, 30 n.12 app. at 53–54.
363. WILLIAM G. HOWELL ET AL., THE EDUCATION GAP: VOUCHERS AND URBAN SCHOOLS 49, 146 tbl.6-1 (2002). Results for African-Americans were significant when the results from New York, Dayton, and Washington were averaged. *Id.* New York results for African-Americans were significant by themselves, Dayton results generally weren't, and Washington results were only significant in year two. *Id.*
364. *See* discussion *supra* Part III.B.1–2.
365. *See* discussion *supra* Part III.B.3.
366. *See* discussion *supra* Part III.B.4–6.
367. *See* discussion *supra* Part III.B.5–6.
368. *See* discussion *supra* Part III.B.4.

actually finds a significant effect of an in-prison faith-based program on recidivism.[369]

The picture looks fairly bleak for faith-based prisons. (Of course, this analysis has no bearing on arguments in favor of religious prisons that don't hinge on their value in reducing recidivism.)[370]

Several literature reviews agree. One says the faith-based prison research

> has shown only moderate effects of faith-based programs on outcome measures such as institutional adjustment, incidence of prison infractions, and likelihood of postrelease arrest. This literature also suffers, in some cases, from important limitations such as: small and nonrepresentative samples, a reliance on anecdotal evidence, an absence of theoretical context, limited statistical analysis, and self-evaluation by prison ministry providers.[371]

Another review says that "research on intentional religion is remarkably underdeveloped" and that "[c]ase studies and descriptive studies," dominated by "subjective interpretations," are "clearly over-represented," and calls for the use of "more rigorous methodologies."[372]

One major contributor to the faith-based prison literature is even more pessimistic about the "more in-depth faith-based programs," calling "the current trend . . . to put large sums of money and a great deal of public, political, and criminological emphasis" on such programs "mistaken" and arguing that there is no evidence that such programs are better than other correctional treatment programs.[373]

I agree, though I emphasize the self-selection issue more than the other problems. (Indeed, I haven't even included any anecdotal evidence here because I find it to be of extremely limited [i.e., no] value.)

Is there any hope for the future? I think there may be. Not all faith-based prison programs are the same. Tom O'Connor and Jeff Duncan write that various programs that showed no effect[374] "probably did a good job of faith development but failed to reduce recidivism because they did not fol-

369. In addition, there is Camp et al.'s propensity score study described above, *see* text accompanying *supra* notes 204–212, which finds a small but significant effect of participation on serious in-prison misconduct. Whether this study is credible depends on how much one trusts the self-reported motivation score that the authors use.

370. Marc O. DeGirolami, *The New Religious Prisons and Their Retributivist Commitments*, 59 ARK. L. REV. 1, 3 (2006) ("[R]eligious programming can be justified in theory by reference to its potential for a special manifestation of retribution that might not otherwise exist."); *see also id.* at 21.

371. Kerley et al., *supra* note 69, at 445.

372. BYRON R. JOHNSON ET AL., OBJECTIVE HOPE: ASSESSING THE EFFECTIVENESS OF FAITH-BASED ORGANIZATIONS: A REVIEW OF THE LITERATURE 20–21 (Ctr. for Res. on Religion & Urban Civil Soc'y, 2002).

373. O'Connor & Duncan, *supra* note 36, at 86.

374. O'Connor is referring specifically to BURNSIDE ET AL., *supra* note 58, by the same authors of the larger study containing Rose, *supra* note 111; Johnson et al., *supra* note 193; O'CONNOR ET AL., *supra* note 331; TRUSTY & EISENBERG, *supra* note 283. O'Connor & Duncan, *supra* note 36, at 89.

low what are known as the principles of effective correctional treatment, such as criminogenic risk/need, responsivity, family/community context, program integrity, and program delivery type."[375] On the other hand, they write, the COSA[376] and TOP[377] programs worked because they did follow many of these principles.[378] Similarly, Daniel Mears and his coauthors write that many faith-based programs have failed because they haven't "articulate[d] a clear statement of program goals and how exactly specific activities will contribute to these goals"; activities have been inconsistently implemented; the different organizations and agencies involved have been uncoordinated; and funding has been insufficient or inconsistent.[379]

If this is the case, the failures of certain faith-based programs may not indicate that the faith-based agenda itself is flawed. Later programs may do better; experimentation may result in the discovery of a more effective model; the programs that have worked well may be replicated in more places. The weak evidence supporting faith-based prisons so far may mean that more experimentation is in order, provided such experimentation can be done consistent with constitutional constraints. "The process of accumulating empirical evidence is rarely sexy in the unfolding, but accumulation is the necessary road along which results become more general."[380]

Moreover, I'm hopeful that—now that studies are available in each of these categories, and now that a number of critical review articles (including this one) have appeared—faith-based prison researchers will get the hint and pursue valid empirical techniques. They'll be following in a venerable tradition. According to economists Joshua Angrist and Jörn-Steffen Pischke, empirical economics is going through a "credibility revolution."[381] "[T]he primary engine driving improvement," Angrist and Pischke write, "has been a focus on the quality of empirical research designs."[382] Randomized assignment has been part of the story[383]—we've now seen a few faith-based prison studies that used this very approach in the form of using rejected volunteers as a control group. Another part of the story is "natural experiments" or "quasi-experimental" designs where we observe people's responses to random institutional flukes; this approach hasn't been used so far for faith-based prisons, but we've seen it for education in the "exogenous policy shocks" literature.[384] Even the instrumental-variables literature is

375. O'Connor & Duncan, *supra* note 36, at 89.
376. WILSON ET AL., *supra* note 135.
377. WILSON ET AL., *supra* note 319.
378. "Furthermore, one of the . . . studies that did not show a significant impact on recidivism was a study of the Transition of Prisoners program in Detroit in its early stages. . . . [, when] the program was in a learning mode." O'Connor & Duncan, *supra* note 36, at 89 (citing O'CONNOR ET AL., *supra* note 331).
379. Mears et al., *supra* note 32, at 360–62.
380. Angrist & Pischke, *supra* note 260, at 24.
381. *Id.*
382. *Id.* at 4.
383. *Id.* at 1–4.
384. *Id.* at 12–13.

better than it was: researchers proposing an instrument now typically take seriously the need to justify why a particular instrument is valid.[385]

* * *

Can such experimentation be done consistent with constitutional constraints? As I've mentioned above,[386] one faith-based prison program, Prison Fellowship Ministries' InnerChange Freedom Initiative, was struck down on Establishment Clause grounds in 2006,[387] so at least some faith-based programs are vulnerable.

As I explain elsewhere,[388] a constitutional faith-based prison program will have to comply with the following requirements:

- Its religious content must be significantly watered down, so that one can't find "religious indoctrination."
- It must be chosen by a process that is neutral as between religious and non-religious programs. Thus, the process that chose it must have been capable of selecting a secular program.
- There must be at least one, and possibly several, comparable secular programs.
- The program must not only be formally voluntary but also not offer significantly greater benefits—for instance, a greater possibility of parole or a safer environment—than secular alternatives.
- Program officials must, at a minimum, not play any role in maintaining order or meting out discipline. But even divesting oneself of these governmental roles may not be good enough to avoid unconstitutionality.

The more ecumenical programs—like the Federal Bureau of Prisons' Life Connections program, which looks much more like an outgrowth of traditional chaplaincy programs together with visits by "spiritual guides of different faiths"[389]—may yet turn out to be constitutional, but it's easy to imagine that some of the more "hard-core" advocates of faith-based prisons will be disappointed at this prospect. To many, salvation (and, in this world, rehabilitation) comes through Christ (for instance) alone, so that a rehabilitative program that isn't allowed to use specifically Christian material may not even be worthwhile. Moreover, program administrators may want to hold on to the ability to expel inmates who aren't engaging constructively

385. *Id.* at 12.
386. *See supra* text accompanying notes 4–9.
387. Ams. United for Separation of Church & State v. Prison Fellowship Ministries, 432 F. Supp. 2d 862 (S.D. Iowa 2006), *aff'd*, 509 F.3d 406 (8th Cir. 2007).
388. *See* Volokh, *supra* note 29.
389. Patrick B. Cates, *Faith-Based Prisons and the Establishment Clause: The Constitutionality of Employing Religion as an Engine of Correctional Policy*, 41 WILLAMETTE L. REV. 777, 824–25 (2005); *see also* O'Connor & Duncan, *supra* note 36, at 87.

sources, not the effect of the program's religious content. It may be that a faith-based program is better than *nothing*—this is important because "the lack of reentry programming constitutes a common criticism of reentry practices to date."[394] But, at the same time, the program may be no better than a comparably funded secular program—which is problematic for policymakers deciding which of several programs to fund. To answer this question, we would need a comparison group of volunteers who were rejected from the religious program and instead assigned to a comparable secular program. To my knowledge, such a study hasn't been done.

The reality, of course, is that most prisoners *don't* get comparable programs. Nor do they all get nothing. Many prisoners "in fact participate in one or more community-based services, even if the intensity of these services may be nominal."[395] So perhaps the best comparison is to this "business as usual" approach.[396] But that, too, is difficult since the "business as usual" baseline differs from place to place and over time.[397]

This problem, while real, is less serious than the self-selection problem. The trouble with self-selection is that a program that seems to work may in fact be worthless. In fact, a program may appear to work even when it's positively harmful, as long as the self-selection effect in the other direction is strong enough.

On the other hand, the resources problem tells us that a faith-based program may be no better, and possibly worse, than some secular program. But where will this program come from? The reality for many prison administrators is that the alternative to a religious program sometimes *is* nothing at all. Religious providers may just be more available than secular providers, and may also cost less to the prison system to the extent that they're more likely to be subsidized by donations from the outside.

Moreover, if one did compare a religious program against a specific secular program, it would thus answer a *specific* question of comparative effectiveness, and would be useful to people considering the hypothetical question of where social resources could, in principle, be best spent. But this result would be hard to generalize to comparisons with other secular programs, and moreover, it would be false to the actual choices faced by prison administrators today.[398]

* * *

Let's take the broad view and come back to the education studies that I've been using as a point of comparison throughout this Article.

394. Mears et al., *supra* note 32, at 354.
395. *Id.*
396. *Id.*
397. *Id.* at 354–55.
398. *Id.*

Finally, after decades of research, we have some credible studies estimating the effect of private schools. The best evidence, taken from studies comparing accepted and rejected applicants, indicates that private schools *do* have a positive effect on the students who attend them, at least for black students and at least for math scores.[399]

On the one hand, one can observe that, next to these results (modest as they are), it's all the more disappointing that faith-based prisons haven't shown much in the way of significant positive effects.

But on the other hand, it took decades of research and debate by different groups, each using a slightly different empirical approach—and many finding little to no effect—before we got even the mild results we have on private education.[400] This suggests that we should encourage more research on the matter, in different contexts, using a variety of different empirical techniques.

The result is that, if there's no strong reason to believe that faith-based prisons work at all, and even less reason to believe that they work better than comparably funded secular programs, there's also little reason to believe that they *don't* work, and in many cases they may be the only available alternative. It's probably sensible to allow such programs to operate and to allow the process of experimentation to work its course, provided that all this can be done constitutionally.

399. *See supra* Part II.A.6.

400. One could also be somewhat depressed that so many empirical studies have failed to yield consistent results across the board. Oh well, empirical research is messy.