Police, Crime and the Problem of Weak Instruments: Revisiting the "More Police, Less Crime" Thesis

Tomislav Kovandzic
University of Texas at Dallas

Mark E. Schaffer
Heriot-Watt University

Lynne M. Vieraitis
University of Texas at Dallas

Erin A. Orrick
Sam Houston State University

Alex R. Piquero[†]
University of Texas at Dallas

Suggested Running Head: Police, Crime, and Weak Instruments

[†] University of Texas at Dallas, Program in Criminology, University of Texas at Dallas, GR 31, Richardson, TX 75080-3021. Email: apiquero@utdallas.edu.

Abstract

Objectives: A key question in the general deterrence literature has been the extent to which the police reduce crime. Definitive answers to this statement, however, are difficult to come by because while more police may reduce crime, higher crime rates may also increase police levels, by triggering the hiring of more police. One way to help overcome this problem is through the use of instrumental variables (IV). Levitt, for example, has employed instrumental variables regression procedures, using mayoral and gubernatorial election cycles and firefighter hiring as instruments for police strength, to address the potential endogeneity of police levels in structural equations of crime due to simultaneity bias.

Methods: We assess the validity and reliability of the instruments used by Levitt for police hiring using recently-developed specification tests for instruments. We apply these tests to both Levitt's original panel dataset of 59 U.S. cities covering the period 1970 to 1992 and an extended version of the panel with data through 2008.

Results: Results indicate that election cycles and firefighter hiring are "weak instruments" – weak predictors of police growth that, if used as instruments in an IV estimation, are prone to result in an unreliable estimate of the impact of police levels on crime.

Conclusions: Levitt's preferred instruments for police levels—mayoral and gubernatorial election cycles and firefighter hiring—are weak instruments by current econometric standards and thus cannot be used to address the potential endogeneity of police in crime equations.

Keywords: police, crime, endogeneity, instrumental variables

Police chiefs, concerned citizens, and policy makers routinely call for adding more police numbers to the ranks based on the presumption that more police on the street will act as a general deterrent to crime by increasing prospective criminals' perceptions of arrest risk. Prior field experimental studies and correlation studies using the OLS estimator, however, generally find little evidence that enlarging police numbers serves as an effective deterrent to crime (especially when compared to more focused strategies; see Nagin and Weisburd, 2013). Many of the correlation studies have been criticized for failing to properly account for the potential endogeneity of police levels in the police-crime relationship, i.e., higher crime rates may stimulate policymakers to hire more police officers (Fisher and Nagin, 1978; Levitt, 1997, 2002; Marvell and Moody, 1996; but see Kleck and Barnes, 2014).[1] Moreover, increases in police strength can have the unintended consequence of increasing crime if additional officers lead to more offenders being detected, arrested, and processed into the justice system (McPheters and Stronge, 1974, Willis, 1983; Greenberg et al., 1983, but see Levitt, 1995, who found no support that increases in police strength influence the reporting and recording of crime). In either case, the OLS estimate of the impact of police strength on violent crime will not be "consistent" – it will suffer from "endogeneity" bias or "simultaneity" bias. Indeed, some scholars have used the nonsignificant or positive coefficient obtained for police strength in OLS crime equations as indicative of the presence of simultaneity bias.

The standard solution to this problem, at least when working with nonexperimental data, is to estimate the crime equation using the instrumental variables (IV) estimator or its generalization, the Generalized Method of Moments (GMM) estimator. IV estimation requires

[1] Kleck and Barnes (2014) are skeptical as to whether changes in *actual* crime rates lead to increases in police numbers as citizens' and policymakers' perceptions of dangerousness are likely to be more heavily influenced by the volume of crime news, which research suggests is largely unrelated to changes in actual crime rates. Studies linking actual crime and punishment risks to individual perceptions of those risks, more generally, remains poorly understood (see Piquero et al., 2012).

one or more additional variables that are used not as regressors but as "instruments" for *police*. These instruments must satisfy two important criteria. First, they must be correlated with the endogenous regressor. This requirement is often referred to as instrument relevance. Second, the excluded instruments must be uncorrelated with the disturbance term in the equation of interest. This requirement is often referred to as instrument exogeneity (see e.g., Stock and Watson, 2010:480-4). The two requirements together imply that an instrument can affect the outcome only via the endogenous regressor. In the present context, an instrument would be valid if it was correlated with police strength (instrument relevance) and if it affected crime rates indirectly and solely through its association with police strength (instrument exogeneity).

As is well known, however, finding variables that can serve as credible instruments can be quite difficult (see Bushway and Apel, 2010). Many candidate variables are likely to fail the exogeneity requirement because they have a direct impact on crime rates, i.e. they are correlated with the error term in the crime equation. Nevertheless, researchers interested in the study of the causal impact of police strength on crime must confront this potential endogeneity problem (Skogan and Fridell, 2004), and instrumental variable techniques can, in principle, be used to avoid the biases that would contaminate the OLS point estimate of the impact of police levels on crime rates.

Several recent studies have attempted to deal with the potential endogeneity of police levels in structural equations of crime due to simultaneity bias. Two empirical papers by Levitt propose three variables that can be used as valid instruments for police strength—the timing of mayoral and gubernatorial elections (Levitt, 1997) and firefighters per capita (Levitt, 2002). Using traditional IV estimation procedures with two different annual, city-level panel datasets, he instruments police strength with the timing of mayoral and gubernatorial elections (Levitt,

1997) and firefighters per capita (Levitt, 2002) and reports, in both papers, that police strength significantly reduces violent crime, especially homicide, but has little or no impact on most property crimes.[2]

Our study assesses the validity and reliability of the instruments employed by Levitt for police strength. Specifically, we examine more closely the quality of the instrumental variables used by Levitt to identify the violent and property crime equations. Using specification testing procedures available in the recent IV literature to assess instrument validity, we find that while all three instruments appear to be exogenous, they are too weakly correlated with police levels to produce a consistent point estimate of the treatment effect of hiring additional police officers on crime rates. This problem is known as the "weak instruments" problem and it arises when the instruments are weakly correlated with the endogenous explanatory variable – here, police levels.

Over the past twenty years, the weak instruments problem has been studied intensively in the econometrics literature. One important line of research has been to investigate the consequences of weak instruments for IV/GMM estimation, and to develop tests for whether a weak instruments problem is present. Unfortunately for researchers, the consequence of weak instruments for IV/GMM estimation is not simply larger standard errors and less precisely determined coefficient estimates. Simply put, weak instruments render invalid the estimates of the coefficients and its variance, and hence also any subsequent hypothesis testing that use these estimates. Specific tests for the weakness or strength of instruments need to be employed instead.

[2] Several other recent studies have used IV techniques to assess the police/crime hypothesis, including studies focusing on the federal COPS program (e.g., Koper and Roth, 2000; Worrall and Kovandzic, 2007; Zhao et al., 2002). For example, Evans and Owens (2007) used federal COPS spending as an instrument to generate additional estimates of the impact of policing on crime, finding that the police added to the force by COPS generated significant reductions in auto thefts, burglaries, robberies, and aggravated assaults. In another study, Lin (2008) used variations in state tax rates as an instrumental variable for local police numbers using state panel data from 1970-2000 and reported that the police reduced crime.

Tests for the presence of weak instruments were first developed by Staiger and Stock (1997) and Stock and Yogo (2005), but these were limited to the case of identically and independently distributed (i.i.d.) data, a condition not often satisfied in non-experimental data. Montiel Olea and Pflueger (2013) have recently developed tests that are robust to heteroskedasticity, clustering and other i.i.d. violations. We apply these tests and find that the Levitt study suffers from weak instruments, with the implication that reliable estimates cannot be obtained by traditional IV estimation. The choice facing researchers wishing to extend the study is either to find the stronger instruments that IV requires, or to abandon these methods in favor of alternative approaches. The estimation methods and specification tests that we employ are applicable to IV estimation in general, and therefore our study should be of interest to researchers working on other criminological topics who wish to employ IV methods.

The paper is structured as follows. The next section briefly reviews the analytic strategy and findings reported by Levitt (1997, 2002). We also discuss the implications of a programming error discovered by McCrary (2002) because of its impact on the key findings reported in Levitt's (1997) original paper. Section three provides empirical results obtained from the specification testing procedures highlighted above when applied both to the original Levitt city panel data and to an extended version of the dataset. The paper closes with some concluding remarks and implications of the current findings for future research examining the effect of the police levels on crime rates.

Review of Levitt's Police and Crime Studies

In his first paper, Levitt (1997) uses the timing of mayoral and gubernatorial elections as instruments for police strength, arguing some incumbents might decide to hire more police officers in election years with the goal of reducing crime or to promote the perception among the

electorate of being "tough on crime." Using standard IV methods applied to first-differences (except for the election year indicators) for a sample of 59 large U.S. cities with directly elected mayors for the period 1970-1992, Levitt arrives at three main conclusions. First, there is evidence of an electoral cycle in police hiring, although the election-cycle instruments are admittedly weak. Specifically, the point estimates on the mayoral and gubernatorial election-year indicators in the first-stage regression imply current police growth rates of 1.2 and 2.4 percent, respectively, during each electoral cycle (see Levitt's Table 2, Column 3).

Second, increases in police strength produce a strong deterrent effect on overall violent crime (homicide, rape, robbery and aggravated assault), but a weak and nonsignificant effect on most property crimes. For violent crime, the IV estimate of the elasticity of violent crime with respect to police strength is -1.39 with a *t* ratio of 2.53. Importantly, however, the only crime that is significantly related to police strength when IV estimates for each crime type are presented separately is homicide with an elasticity of -3.05 and a *t* ratio above 3.[3] Third, and most importantly, IV estimates for police strength were generally more negative than OLS estimates. The only exceptions occur for rape and larceny in which the OLS estimates for police strength are in the expected negative direction while their IV counterparts are actually positive.

In an important follow-up study, McCrary (2002) finds that a programming error in Levitt's SAS code leads to each crime model being incorrectly weighted by the residual standard deviation instead of its inverse. The net effect of this error is to give crime types with the greatest variance, especially homicide, more weight in both the OLS and IV estimations than those with lower variability. McCrary also re-examines Levitt's mayoral election variable and constructs an

[3] As described later, Levitt stacks (i.e., estimates them jointly) the violent and property crime models in order to obtain more precise estimates for the police strength variable (McCrary, 2002). This approach resulted in coefficient restrictions being placed on the police and exogenous control variables across crime types. He further allows police to have an extended lagged impact on crime by entering both one- and two-year lags of police in the violent and property crime estimations. The elasticities for police strength are based on the sum of the two coefficients.

alternative measure. McCrary re-estimates Levitt's IV estimations using the correct weighting scheme and finds that the error had the effect of severely biasing the point estimates and standard errors for police strength in the violent crime equation. Specifically, the implementation of the correct weighting scheme reduces the point estimate for police strength from -1.39 to -0.79 and increases the standard error from 0.55 to 0.61, resulting in an insignificant *t* ratio of 1.30. Most importantly, McCrary's reanalysis using the correct weights reveals IV estimates for police strength in both the violent and property crime models as well as for each crime category that are never distinguishable from zero and differ little from OLS estimates.

In a response to McCrary, Levitt (2002) acknowledges the weighting error but downplays its importance on the overall conclusions drawn in the paper, i.e., that police strength significantly reduces crime. In so doing, he directs the reader to the still rather large point estimate generated for the police strength variable in the properly weighted IV estimation for violent crime, even though it is not close to being statistically significant, and points out that the point estimates and standard errors for most individual crimes are not greatly impacted by the weighting error. However, this is not the case for homicide which is, as noted above, the only crime type for which Levitt's (incorrectly weighted) IV estimates produce a point estimate for police strength statistically different from zero. Levitt ultimately decides to abandon the use of election cycles as instruments for police levels and replaces them with another instrument which he claims is more strongly correlated with changes in police strength, i.e. changes in the number of firefighters per capita. Levitt's insight is that changes in both firefighter and police hiring are likely to co-vary over time due to factors such as "the power of public sector unions, citizen tastes for government services, affirmative action initiatives, or a mayor's desire to provide spoils" (p. 1245).

Using IV methods with a revised panel dataset consisting of 122 large cities covering the period 1975 to 1995, Levitt reports a strong correlation between the hiring of police and firefighters.[4] In particular, the first-stage estimates produce police elasticities with respect to firefighters ranging from 0.206 (s.e.=0.050) to 0.251 (s.e.=0.050). With respect to crime elasticities for police, the only two crime types for which the IV point estimates are significantly larger than their OLS counterparts occur for homicide and auto theft. For example, the homicide elasticity is -0.914 (s.e.=0.332), over twice as large as the OLS estimate of -0.408 (s.e.=0.088). In all, Levitt's subsequent study indicates that firefighter hiring produces more precise estimates than election cycles regarding the impact of police strength on crime. However, firefighter hiring only induces enough variation in police hiring to generate statistically significant effects for homicide and auto theft. As noted above, the IV estimates for the remaining crime categories are never statistically distinguishable from zero.

Current Study

In this study, we use the weak-instrument-testing procedures outlined above to revisit Levitt's argument that gubernatorial and mayoral election cycles and firefighter hiring are relevant and valid instruments for police levels. Because the mayoral election instrument can

[4] Unlike the original study, Levitt's baseline IV models were estimated using the fixed effects estimator (FE), rather than the first difference (FD) estimator. In addition, the stacked regression approach utilized by Levitt in the original paper was abandoned. The police strength variable is entered into the IV estimation once lagged and is instrumented using only the logged, once-lagged number of firefighters per capita. The electoral variables are not used. Levitt notes (p. 1246, n. 6) that the mayoral variable is unavailable because many cities in the new sample do not have directly-elected mayors. He does not explain why the gubernatorial variable is also dropped. Unlike the procedure in the original paper, Levitt weights all estimations by city population and heteroskedasticity-robust (but not cluster-robust) standard errors are reported. Control variables for potential confounding factors include the effective abortion rate in the state, unemployment rate for the city's MSA, state income per capita, percent black, a one-year lag of the state incarceration rate, and city and year fixed effects.

only be used in cities with directly elected mayors, we apply these techniques to an updated version of the original panel dataset of 59 large U.S. cities compiled by Levitt (1997).[5]

Data

Annual violent crime data and the number of sworn police officers (as of October 1 of each year) for each city are taken from the FBI's *Uniform Crime Reports*.[6] Dates of gubernatorial and mayoral elections for 1993 to 2008 are collected using data sources cited in McCrary (2002) and appended to the original variables created by Levitt. As noted above, McCrary also collected data on the timing of mayoral elections and discovered election dates for many cities differed significantly from those originally collected by Levitt. While not shown, substituting Levitt's mayoral election measure for McCrary's measure has little impact on the substantive results regarding the impact of police strength on violent crime rates. We obtain the firefighter employment data for 1970 to 2008 from the Census Bureau's *Annual Survey of Government Employment and Payroll* and merge it to the updated panel dataset.[7] The remaining exogenous control variables included in Levitt's original analysis include percent of the city's population that is black, percent of city households headed by females, percent of the SMSA population (in which the city is located) between the ages of 15 to 24, combined state and local per capita

[5] The data and computer code used by Levitt (1997) were obtained from McCrary's website at http://eml.berkeley.edu/replications/mccrary/index.html.

[6] Chalfin and McCrary (2013) recently identified some measurement error problems in the analysis of police employment data obtained from the UCR program. Specifically, they found discrepancies in UCR police employment figures as compared to those in police internal reports and those in the Annual Survey of Government administered by the Bureau of Labor Statistics. We do not address or discuss in any detail the measurement error problem in the current study as the principal focus is on specification tests for assessing the weak IV problem. It is worth noting, however, that the IV estimator is generally recognized as a way of dealing with measurement error problems, whether the errors are truly random (i.e. "classic measurement error") or patterned. If, for example, estimates of police manpower levels are measured with relatively moderate amounts of random error, this will cause attenuation of any relationship between police and crime. More importantly, however, moderate to substantial measurement error of the random variety will also attenuate any correlation between the instruments and police levels, walking straight into the weak instruments problem addressed in the current study.

[7] Levitt did not respond to a data request for the firefighter hiring data used in his follow-up paper.

spending on education, combined state and local per capita spending on welfare, and the state unemployment rate.[8] With the exception of the spending variables, we are able to update the dataset through 2008 using data sources listed in Levitt (1997). Because the coefficients on the spending variables are never distinct from zero in any of Levitt's preferred crime specifications, we do not consider their exclusion here to be a problem (though it does remain a possibility). Lastly, we include the logged, once lagged state incarceration rate because of its importance as a predictor of crime in Levitt's (2002) follow-up study as well as other notable macro-panel studies of crime (e.g., Levitt, 1996; Marvell and Moody, 1994; Spelman, 2008).[9] Means and standard deviations are reported in the Appendix Table separately for the three different estimation samples used below.

Estimation Strategy

For parsimony's sake, we focus our attention on violent crime rates, which both Levitt studies indicate are most responsive to increases in police levels. Following Levitt (1997), the models are first transformed by applying the first-differences transformation to eliminate the city fixed effects, and then augmented with additional exogenous covariates:

$$(1) \qquad \Delta \ln (\textit{violent crime}_{it}) = \beta \Delta \ln(\textit{police}_{it-1}) + \alpha X_{it} + u_{it}$$

where *violent crime* and *police* are the numbers of violent crimes and police per 100,000 city population where X_{it} is a Kx1 row vector of values for the it^{th} observation of the K exogenous controls X and α is the corresponding 1xK vector of coefficients.

[8] Following Levitt, percent black, percent female-headed households, and percent ages 15 to 24 are linearly interpolated between decennial census years.

[9] Levitt's (2002) follow-up study also includes a state-level variable capturing the weighted average of the abortion exposure at birth of the crime-aged population. Despite the apparent importance of the abortion rate as a predictor of crime (see Table 3, p. 1248), we do not have access to the measure and do not include it here. Our results should be considered with this caveat.

The three excluded instruments used in the estimation of equation (1) are denoted M_{it-1}, G_{it-1} and $\Delta\ln(F_{it-1})$. M_{it-1} and G_{it-1} are dummies defined =1 if there was a mayoral or gubernatorial election, respectively, and =0 otherwise. F_{it-1} is defined as firefighters per 100,000 population. Note that lagged growth in police levels is instrumented with contemporaneous electoral dummies (as in Levitt 1997) and by lagged growth in firefighter numbers (similar to Levitt 2002[10]). The first-stage estimation corresponding to equation (1) is therefore:

$$(2) \qquad \Delta\ln(police_{it-1}) = \pi_m M_{it-1} + \pi_g G_{it-1} + \pi_f \Delta\ln(F_{it-1}) + \theta X_{it} + \nu_{it}$$

When we specify augmented versions of (1) in which police levels appear twice, once dated (t-1) and once dated (t-2), we also include all three instruments twice in equation (2), again once dated (t-1) and once dated (t-2). We return to this point below.

Levitt (1997) and McCrary (2002) both use the traditional IV covariance estimator to obtain standard errors and test statistics; this covariance estimator requires the assumption of i.i.d. data. Levitt (2002) switches to the well-known Eicker-Huber-White heteroskedastic-robust (HR) estimator to calculate standard errors. However, both the traditional and HR covariance estimators do not address the all too common problems in panel data of either serial correlation or within-state correlation of the disturbance term (Moulton, 1990; Bertrand, Duflo, and Mullainathan, 2004). In this panel dataset, it is reasonable to suspect that cities within the same state are likely to have correlated disturbance terms because several of the explanatory variables are measured at the state-level and thus do not vary within the state.[11]

[10] Levitt (2002) estimates in levels, using levels of firefighters to instrument for levels of police; applying the first-differences transformation to his 2002 estimates implies that first-differences (growth) of police should be instrumented by first-differences of firefighters.

[11] See Angrist and Pischke (2009, chapter 8) for further discussion of this problem.

Cluster-robust (CR) standard errors are a now-standard method of addressing this problem,[12] where clusters are defined here as groups of cities, i.e., states. CR standard errors are valid in the presence of both arbitrary heteroskedasticity and arbitrary within-state correlation of the disturbance term, and only require the assumption of independence across states. This independence assumption is much weaker and more realistic than that required by the traditional HR standard errors, which need independence across observations; within-state correlation, whether it be serial correlation between observations for a single city or correlation between cities in a single state, will render them invalid. The main drawback to this approach, however, is that we have only 30 clusters (states), which is fewer than desirable.[13] Regardless, given the importance of using a variance estimator that is consistent in the presence of heteroskedasticity and within-state clustering, we report CR standard errors.

Specification testing and weak identification

We report tests of overidentification, endogeneity, underidentification, and weak identification. All our tests are robust to heteroskedasticity and clustering. The overidentification test used is Hansen's J statistic; the test of the endogeneity of police levels is a C statistic (variously referred to in the literature as a "difference-in-J", "difference-in-Sargan" or "GMM Distance" test); the test of underidentification is due to Kleibergen and Paap (2006), and is a generalization of Anderson's (1951) canonical correlations test to non-i.i.d. data. The overidentification, endogeneity, and underidentification tests are all now standard in the

[12] See e.g., Wooldridge (2010), chapter 10, sections 10.5 and 10.6 for a detailed presentation.
[13] According to Angrist and Pischke (2009), it would be 42.

literature, and rely on the usual asymptotic (large-sample) justifications. The tests we employ for weak identification, however, are somewhat different.[14]

The earliest formal test for weak instruments is a "rule of thumb" recommended by Staiger and Stock (1997): in the first-stage estimation of IV estimation in which the single endogenous explanatory variable – here, police levels – is regressed on the exogenous regressors and instruments, an F test of the significance of the excluded instruments should be at least 10. Test statistics less than 10 suggest that IV estimates of β will be badly biased, and inference based on the estimate unreliable. The test is later extended by Stock and Yogo (2005) to the case of multiple endogenous regressors and other IV-type estimators, and Stock and Yogo also refine the test to provide critical values for specific hypotheses that quantify the impact of instrument weakness in terms of bias of the IV estimator or in terms of the distortion to hypothesis tests (test size).

These tests, however, suffer from an important practical limitation: they apply to the case of i.i.d. data only. If heteroskedasticity and clustering are present – and we normally expect this in a nonexperimental panel dataset such as Levitt's – the tests' statistics and critical values are invalid. Faced with this problem, researchers would either report non-robust test statistics that were known to be invalid or heteroskedastic-robust versions of the first-stage F statistic that did not have a demonstrated theoretical foundation as a weak instruments test.

This has changed recently with the development of a weak instruments test by Montiel Olea and Pfluger (2013) that can be made robust to non-i.i.d. data. Their "effective F" statistic is

[14] A test for "weak instruments" or "weak identification" is distinct from a test for underidentification. Rejection of the null of underidentification implies that the excluded instruments have some nonzero correlations with the endogenous regressors. Unfortunately, rejection of underidentification is not enough for researchers to have confidence in their IV estimates. The problem is that if the correlation is nonzero but small – intuitively, that the information provided by the instruments in the dataset available is small compared to the noise in the estimates – then IV-type estimators will be badly biased (in the same direction as OLS) and estimated standard errors will be unreliable (Stock and Watson 2010:480-1).

a test of whether the (asymptotic) bias in the estimator exceeds some percentage τ of a "worst-case" benchmark, namely the bias that would arise if the instruments are completely irrelevant and the model unidentified. The critical value for the test depends on the chosen test level (e.g., the traditional 5% significance level) and the chosen percentage τ of the worst-case bias. A limitation of the test is that, at the time of writing, it has been developed only for the case of a single endogenous regressor. We return to this point below.

All of the results presented in this example are estimated in Stata. The main IV/GMM estimation programs, *ivreg2* and *xtivreg2*, are co-authored by one of the authors (BLINDED FOR REVIEW), and can be freely downloaded via the software database of RePEc.[15] The Montiel Olea and Pflueger test is implemented in the program *weakivtest* by Pflueger and Wang (2014), also available from RePEc. For further discussion of how the estimators and tests are implemented, see Baum, Schaffer and Stillman (2003, 2007), Montiel Olea and Pflueger (2013), Pflueger and Wang (forthcoming) and the references therein.

Results

Testing the original Levitt-McCrary estimations for weak identification

We begin by reexamining the original Levitt (1997) study, as replicated and corrected by McCrary (2002), for evidence of weak identification. We focus on the specifications reported in McCrary (2002), Table 1. Levitt's original model specifications have several features that at first glance would appear to make assessing the strength of identification difficult. Fortunately, several simplifications are available that make this assessment straightforward.

Levitt and McCrary consider estimates for 7 different types of crimes separately. Police strength is allowed to differ by type of crime j, and thus Levitt and McCrary obtain estimates of

[15] http://ideas.repec.org/SoftwareSeries.html. *ivreg2* is a general-purpose IV/GMM estimation routine for linear models; *xtivreg2* supports fixed-effects panel data models.

the impact of police β_j that vary across crime j. In an additional set of "pooled" estimates, the constraint is imposed that the impact of police is the same for the 4 types of violent crime, and the same for 3 types of nonviolent crime. Levitt's treatment of covariates is somewhat involved: year, region, and city-size dummies have coefficients that also differ across crime types; 6 state and MSA covariates are constrained to have coefficients that are the same for the 4 violent crime types, and the same for the 3 nonviolent crime types; and city fixed effects (interpreted as city-specific crime growth rates) are constrained to be the same across all 7 categories of crimes. Finally, Levitt also allows for a lag structure in the impact of police strength on crime, by including it as a regressor with lags at time (t-1) and (t-2).

Indexing crimes by j=1..7 and violent/nonviolent crimes by v=1…2 yields the main estimating equation, reported in McCrary's Table 1, Panel A:

(3) $\Delta\ln(crime_{ijt}) = \beta_{1j}\Delta\ln(police_{it-1}) + \beta_{2j}\Delta\ln(police_{it-2}) + \alpha_j X_{ijt} + \gamma_v Y_{ivt} + \delta Z_i + u_{it}$

where X_{ijt} represents covariates that differ across all 7 crime types, Y_{ivt} represents covariates that differ across violent/nonviolent crimes, and Z_i are the city dummies with coefficients that are common across all 7 crime types. The pooled estimates for violent and nonviolent crime are obtained by estimating (3) and imposing the constraint that $\beta_{1j}=\beta_{1violent}$ and $\beta_{2j}=\beta_{2violent}$ for j=1..4 (violent crimes), and $\beta_{1j}=\beta_{1nonviolent}$ and $\beta_{2j}=\beta_{2nonviolent}$ for j=5..7 (nonviolent crimes).

Levitt and McCrary estimate (3) by stacking the data so that each city-year observation appears 7 times and interacting coefficients with crime-type dummies so that the coefficients can vary across crime types. The police, mayoral, and gubernatorial instruments are also interacted with crime-type dummies, and all are used with lags of 1 and 2 years. The main stacked estimation therefore has 14 endogenous regressors (2 lags of police strength x 7 crime types), 28

instruments (2 lags x 2 electoral variables x 7 crime types), and a total of 279 other covariates including a constant term.

The first simplification is to follow McCrary in his reporting of the first-stage estimations (his Table 2) and to drop the constraints on the coefficients on the Y_{ivt} and Z_i variables. That is, we allow the coefficients on all covariates to vary across crime types.[16] The second simplification also follows McCrary in his reporting of the first-stage estimations. The first-stage estimations for all 7 crime types all have the same form, namely OLS regressions of police lagged 1 and 2 periods on the exogenous covariates and the electoral variables, also lagged 1 and 2 periods:

(4) $\Delta\ln(police_{it-1}) = \pi_{11mj}M_{ijt-1} + \pi_{12mj}M_{ijt-2} + \pi_{11gj}G_{ijt-1} + \pi_{12gj}G_{ijt-2} + \theta_{1j}X_{ijt} + v_{1it}$

(5) $\Delta\ln(police_{it-2}) = \pi_{21mj}M_{ijt-1} + \pi_{22mj}M_{ijt-2} + \pi_{21gj}G_{ijt-1} + \pi_{22gj}G_{ijt-2} + \theta_{2j}X_{ijt} + v_{2it}$

where we have simplified by representing all the exogenous covariates by X_{ijt}. Since the first-stage equations are identical for all 7 crime types, we can simply ignore the j subscripts, leaving us with just 2 first-stage equations for 2 endogenous regressors and 4 excluded instruments:

(4') $\Delta\ln(police_{it-1}) = \pi_{11m}M_{it-1} + \pi_{12m}M_{it-2} + \pi_{11g}G_{it-1} + \pi_{12g}G_{it-2} + \theta_{1}X_{it} + v_{1it}$

(5') $\Delta\ln(police_{it-2}) = \pi_{21m}M_{it-1} + \pi_{22m}M_{it-2} + \pi_{21g}G_{it-1} + \pi_{22g}G_{it-2} + \theta_{2}X_{it} + v_{2it}$

The final simplification follows from an implication of the identification strategy that neither McCrary nor Levitt notes. Levitt's strategy is based on a contemporaneous electoral cycle effect: exogenous increases in annual police hiring are linked by Levitt to mayoral and gubernatorial elections in the same year. Levitt does not assert leading or lagged electoral effects on police strength, and indeed the first-stage regressions he and McCrary report are consistent with an electoral effect that is contemporaneous only; the estimates in McCrary's Table 2 show that the coefficients on the lagged (t-2) electoral variables in equation (4') for $\Delta\ln(police_{it-1})$ are

[16] The results of the identification tests change very little when imposing these constraints.

not significantly different from zero, and similarly for the coefficients on the leading (t-1) electoral variables in equation (5') for $\Delta\ln(police_{it-2})$. The coefficients on the contemporaneous electoral variables are, however, generally statistically significant at conventional levels and with the expected positive sign.

The key point is that Levitt's identification strategy in (4') and (5') is essentially "double-or-nothing". The reason is that the identification achieved in equation (5') is basically the same as that in equation (4') with a lag. If the coefficient π_{11m} on mayoral elections at time (t-1) in the police (t-1) equation is nonzero, then the coefficient π_{22m} on mayoral elections at time (t-2) in the police (t-2) equation will also be nonzero. This is because they are essentially the same coefficient: they capture the contemporaneous effect of mayoral elections on police strength. The same point applies to the gubernatorial elections instrument. The only difference between (4') and (5') is the inclusion of the extraneous lags of *M* and *G* in equation (4') and the extraneous leads of the same in equation (5'). These extraneous variables follow from the fact that IV is a single-equation limited information estimator, and the first-stage estimations are reduced-form equations that necessarily include all instruments.[17] Given the reliance on contemporaneous effects for identification, it is not surprising that these extraneous leads and lags are insignificantly different from zero in McCrary's reported first-stage regressions.

We can therefore assess the strength or weakness of identification across all crime types by considering a single first-stage equation with contemporaneous instruments only. We use the first lag of police strength:

(4'') $$\Delta\ln(police_{it-1}) = \pi_{11m}M_{it-1} + \pi_{11g}G_{it-1} + \theta_1 X_{it} + \nu_{1it}$$

[17] We obtain similar results if we use methods that assess underidentification using both lags but take into account the double-or-nothing nature of the identification strategy.

The upshot of these simplifications is that we can apply both the standard Stock-Yogo and the robust Montiel Olea-Pflueger tests in straightforward fashion, based on the single first-stage regression (4").

Table 1 about here

The first-stage regression estimates for the original Levitt (1997) model as replicated by McCrary (2002) are reported in Table 1. Estimates using both Levitt's original mayoral variable and McCrary's alternative are reported. These results correspond directly to those reported in McCrary's Table 2, in which he reports estimates of equations (4') and (5') using both mayoral measures, and which we have replicated. However, whereas McCrary reported only standard non-robust test statistics and standard errors that require the assumption of i.i.d. data, we report both non-robust and cluster-robust statistics.

We consider first the test of underidentification. For the i.i.d. case, this is the Lagrange Multiplier (LM) version of Anderson's (1951) canonical correlations test. The null hypothesis is that the model is underidentified, and this is rejected very convincingly with p-values that are near zero. The situation changes dramatically when we consider an underidentification test that is robust to clustering within states; the version reported in the table is an LM version of the Kleibergen-Paap (2006) test. The p-value for the specifications using either mayoral measure is now about 7%. Once we allow for non-i.i.d. errors, the model appears to be barely identified at all.

The above suggests that a direct test of weak identification that accounts for clustering should also indicate serious problems, and indeed this is what we find. The Cragg-Donald statistic is the first-stage F statistic developed by Staiger and Stock (1997) and Stock and Yogo (2005) for the i.i.d. case. Using this statistic, the original Levitt-McCrary estimation appears to

be only moderately weakly identified: the value of the statistic is 9.87 using Levitt's original measure and 11.40 using McCrary's measure, i.e., close to the Staiger-Stock rule of thumb of at least 10. Stock and Yogo's more formal test of maximal test size distortion, also for the i.i.d. case, leads to a similar conclusion: the Cragg-Donald test statistics lie between the critical values for 15% and 20% maximal test size distortion.[18] However, because this test is not robust to clustering, we cannot use the evidence of the Cragg-Donald-Stock-Yogo tests to draw any firm conclusions about whether weak identification is a problem and instead use cluster-robust test statistics.

This is addressed by the cluster-robust Montiel-Pflueger "effective F" test reported in Table 1. The test statistic is 4.45 for the model using Levitt's original mayoral measure and 6.39 using McCrary's revised measure. The Montiel-Pflueger critical values in Table 1 confirm the problem. For example, the critical value for τ=30% is 4.34 for the Levitt measure specification, and 6.80 for the McCrary version for an M-P test at the 5% significance level. This is clear evidence of a weak instrument problem: there is a 5% chance that the bias in the IV estimator is 30% of the worst case possible.

Based on these results, the appropriate conclusion is that the instrument set recommend by Levitt appears to be only weakly relevant, and the estimation results reported in Levitt (1997) and McCrary (2002) should be regarded as potentially biased in the sense that the excluded instruments are not strong enough predictors of police strength to lead to reliable IV estimates.

[18] The Stock-Yogo test of maximal bias is not available unless the degree of overidentification is at least 2. We use it in our estimations below when firefighters are available as a 3rd instrument.

New Results with Expanded Data: Estimating the Effect of Current-Year Police Growth on Violent Crime Rates using the OLS Estimator

We now turn to the results with the expanded dataset. Looking at Table 2, we consider first the results in column 1 for FD estimation where one lag of the change in police strength is treated as exogenous (i.e., OLS) and that allows for possible heteroskedasticity or clustering of disturbance terms (model 1). Similar to Levitt (1997), who reports significant negative findings for changes in police strength on violent crime rates when using the OLS/FD estimator (as corrected by McCrary, β_{ols}= -0.12/s.e.=0.06), we find police strength to be significantly and negatively related to violent crime rates (β_{ols}= -0.16/s.e.=0.06). We next turn our attention to our main question: can electoral cycles and firefighter hiring be used to mitigate endogeneity bias in the police-crime relationship by serving as valid and reliable instrumental variables of police strength, i.e., do they provide a better means of identifying the causal link between police strength and violent crime?

Table 2 about here

Contrary to Levitt (1997, 2002) but similar to McCrary (2002), when police strength is treated as an endogenous regressor and instrumented with both election-year indicators and firefighter hiring using the standard IV/FD estimator (model 2), the substantive results from OLS estimation disappear. That is, the point estimate obtained for police strength in model 2 indicates police strength is unrelated to violent crime, with a small estimated elasticity of -0.04 and a large standard error of 0.21. We also report the results using the Limited Information Maximum Likelihood (LIML) estimator, because of its greater robustness to weak identification (more on this point below). As seen in Table 2, almost identical results are obtained when substituting LIML for the standard IV estimator (model 3).

One possible explanation for the differing results between the two studies is our failure to follow Levitt's (1997) preferred approach of including two years of lagged police strength in the violent crime specification simultaneously. To examine this possibility, we re-estimate the IV/FD with 1- and 2-year lags of police strength included as predictors of violent crime (model 4), adopting Levitt's "double-or-nothing" identification strategy and using 1- and 2-year lags of the electoral and firefighting instruments. As seen in Table 2, our failure to include lagged police strength as a predictor is not responsible for the divergent findings. The combined 2-year effect police strength is slightly larger again in absolute terms (-0.13) but still small relative to the standard error (0.14) and hence insignificantly different from zero. Following a similar approach to that taken above for the OLS/FD estimations, we estimated a series of follow-up regressions with each model specification becoming more progressively similar to that originally employed by Levitt (1997). Model 5 is the same as model 4 except that we weight the regression by city population size. The combined policing effect remains negative but statistically indistinguishable from zero. In model 6, we estimate a model similar to model 5 except that we limit the study period through 1992. This model specification results in a combined policing effect which is extremely small (-0.01) and with a very large standard error (0.38). Lastly, we re-estimate a model specification which approximates the panel estimation used by Levitt (1997) except that we continue to include firefighter hiring as an instrument for police strength and we do not employ stacked regression methodology (model 7). Specifically, model 7 is the same as model 6 except that we drop the additional predictors (incarceration, alcohol consumption, shall-issue[19] dummy) not included in Levitt (1997). Again, we are unable to replicate Levitt's (1997) IV/FD finding that increases in police strength reduce violent crime.

[19] Shall-issue laws deal with the presumptive right-to-carry a concealed firearm after authorities provide a license to an applicant who meets certain criteria.

Similar to model 6, the combined police strength effect is essentially nil (-0.02), not significantly different from zero, and dwarfed by the large standard error (0.39).

Lastly, we use IV/GMM methods discussed above to examine whether police strength is really an endogenous regressor in violent crime models due to simultaneity bias (i.e., more violent crime, more police), and as a result whether the IV estimator is to be preferred to the OLS estimator in police-crime studies. Because this conclusion rests entirely on the quality of election cycles and firefighter hiring as instruments for police strength, we also assess the reliability and validity of the instruments using the tests described above.

Model 1 is our initial specification and treats police strength as exogenous; it is the OLS estimator with exogenous regressors and a covariance estimator that allows for arbitrary heteroskedasticity and clustering. The corresponding J statistic is a test of the full set of orthogonality conditions, i.e., the exogeneity of police strength, election cycles, and firefighter hiring (plus the control variables). The J statistic in column 1 is 0.583, which is small for a $\chi^2(3)$ statistic ($p = 0.900$), and we therefore fail to reject the null hypothesis that the moment conditions in the police-exogenous FD estimation are satisfied, and take this as evidence that none of the variables – police strength, election cycles, firefighter hiring, and/or the control variables – appears to be endogenous.

In models 2 through 7 police strength is treated as endogenous, and the J statistic is a test of the reduced set of orthogonality conditions, i.e., the exogeneity of election cycles and firefighter hiring. The J statistics for the police-endogenous FD estimations are rather small and never close to statistical significance at conventional levels. Therefore, we cannot reject the null hypothesis that election cycles, firefighter hiring, and the control variables are exogenous; in other words, the evidence indicates both Levitt's (1997, 2002) electoral instruments and our

additional firefighting instrument pass the validity requirement for instrumental variables, i.e., they can appropriately be omitted from the violent crime rate equation.

We next test explicitly whether police strength is endogenous using robust C tests that are based on the difference between the J statistics for police-exogenous and police-endogenous estimations.[20] For models 2-4, the C statistics are very small, with p-values in excess of 50%, suggesting that police strength can be treated as exogenous. For models 5-7, where we weight by city population, the p-values are larger – around 10% – suggesting that in these specifications treating police strength may not be exogenous. We therefore have some evidence that police strength is exogenous, at least for cities with directly elected mayors, and somewhat surprisingly, that IV estimation with FDs is unnecessary in the unweighted specification. The implication is that the estimates for violent crime rates that treat police strength as exogenous (model 1) are to be preferred on efficiency grounds, whereas the estimates that treat police as endogenous (models 2 to 7) are preferable if we wish to avoid a Type II error in concluding they are exogenous.

Crucially, however, all these results – including the tests of endogeneity and instrument exogeneity – are highly dependent on whether the excluded instruments are relevant, i.e., sufficiently correlated with police strength. We consider the various tests of underidentification and weak identification in turn.

Table 3 about here

As seen in Table 3, which reports the first-stage regression results for the excluded instruments for models 2-7, all three variables, when statistically different from zero, are correlated with police strength in the expected positive direction. Specifically, mayoral and

[20] The C statistic for model 2 differs slightly from the difference between the relevant J statistics for models 1 and 2 because we use a version of the C test that guarantees a positive test statistic (see Hayashi, 2000; Baum et al., 2003).

gubernatorial election years, when significant, are associated with an increase in police strength of the order of 1-2%. The effect of firefighter hiring on police strength is also quite large. This is consistent with the underidentification tests reported in Table 2. For models 2 and 3, in which police strength appears for only one year, we report the Kleibergen-Paap cluster-robust LM test comparable to the cluster-robust underidentification tests reported in Table 1 for the original Levitt-McCrary dataset. The test statistic is now considerably larger (12.66 vs. 5.3-5.4 using the original Levitt-McCrary data), and we can now comfortably reject (p=0.005) the null hypothesis that models 2 and 3 are underidentified at conventional significance levels. For models 4-7, where we include two lags of police strength, our identification strategy is the same "double-or-nothing" strategy employed by Levitt. The appropriate underidentification test is therefore one in which the null hypothesis is that the coefficients on both lags of police strength are identified, and where the alternative hypothesis is non-identification (neither is identified).[21] Table 3 shows again that underidentification is comfortably rejected at conventional significance levels (p=0.01-0.03). This is an unsurprising result in light of the test results for models 2-3 because of the "double-or-nothing" strategy: if the three instruments dated (t-1) all identify police strength dated (t-1), as in models 2-3, then the same three instruments dated (t-2) should do an equally good job identifying police strength dated (t-2). We conclude that extending the dataset through 2008 and adding firefighters as an instrument has improved the degree of identification. But has it improved enough? As noted above, it is not enough that the instruments are correlated with the endogenous regressors; the correlations need to be strong. We must therefore consider tests of weak identification.

Table 4 about here

[21] That is, the null hypothesis is that the matrix of reduced form coefficients has rank=2, and the alternative is that it has rank=0.

The "double-or-nothing" identification strategy means again that we can assess weak identification by considering estimations in which police strength and the electoral and firefighting instruments appear with one lag, i.e., once each. Table 4 reports these for the unweighted models 2-4 and city-population-weighted models 5-7. Because the LIML estimator is more robust than the IV estimator to weak identification (Stock and Yogo 2005), the critical values for the tests of model 3 (LIML) in Table 4 are different from those for the otherwise-identical model 2 (IV).[22]

The Cragg-Donald statistic, suitable for the i.i.d. case, suggests that models 2-4 are strongly identified – the first-stage F statistic of 29.5 is large by any benchmark, as a comparison with the Stock-Yogo critical values in Table 3 shows. Model 5 appears moderately strongly identified with a Cragg-Donald test statistic of 16.8, but models 6 and 7 appear weakly identified, with Cragg-Donald F statistics of about 7.3-7.4. However, these tests are not robust to clustering, which is a problem that must be addressed. We cannot use the evidence of the Cragg-Donald-Stock-Yogo tests to draw any firm conclusions about whether weak identification is a problem and instead use cluster-robust test statistics.

The cluster-robust Montiel-Pflueger "effective F" test reported in Table 4 for models 2-4 is only 4.1, far lower than the un-robust Cragg-Donald statistic and well below the Staiger-Stock rule of thumb of 10. The Montiel-Pflueger critical values confirm the problem. For the IV estimator and using the 5% significance level, the critical value for τ=30% is 10.5. In fact, the M-P test statistic is roughly equidistant between the critical value for τ=20% and the critical value for τ=30% at the 50% significance level. This is clear evidence of a weak instrument problem: there is roughly 50:50 chance that the bias in the IV estimator is 20%-30% of the worst case

[22] Stock and Yogo (2005) do not provide critical values for the LIML version of the maximal relative bias test because LIML bias is not well-defined.

possible. The critical values for the LIML estimator (model 3) – which is less prone to weak instruments problems – are only slightly lower than those for the IV estimator (models 2 and 4), and thus our conclusions are the same for model 3.

The results for model 5 with population weights, also reported in Table 4, are similar. The M-P effective F statistic is 3.91, again roughly equidistant between the 50% significance level critical values for τ=20% and τ=30%. The results for models 6-7, with weighting, limiting the sample to data for 1992 or earlier, and using Levitt's original covariates, are worse still; the effective F is about 2.8-2.9, about the critical value for a worst-case Nagar bias of τ=30% at the 50% significance level.

Based on these results, the appropriate conclusion is that the instrument set recommended by Levitt appears to be only weakly relevant, and the estimation results reported here as well as in Levitt (1997) should be regarded as biased in the sense that the excluded instruments appear to be weak predictors of police strength. Extending the data coverage through 2008, and including firefighters as an additional instrument, improves the strength of identification but not enough to avoid serious weak identification problems. Weighting by population, if anything, worsens the strength of identification. This finding applies to the OLS results as well since the test of the endogeneity of police strength is based on a C test calculated using J statistics from the police-exogenous and police-endogenous estimations. Because both J tests are calculated using what appear to be weak instruments, the results of the C test indicating police strength is more appropriately treated as exogenous should also be regarded as uninterpretable. Simply put, mayoral and gubernatorial election cycles and firefighter hiring are all weakly correlated with

police levels, rendering them ineffective as instruments in a standard IV/GMM estimate of the impact of police levels on violent crime.[23]

Discussion

To summarize, since the seminal work of Levitt, there have been various methodological advances relating to IV estimation that have come into general practice. We summarize the key advances and procedures that we have employed above as follows:

1. Levitt's original (1997) work and McCrary's (2002) correction and extension use hypothesis and specification tests that rest on non-robust covariance estimators that require the assumption of i.i.d. data. The i.i.d. assumption is very strong for the panel data setting, and its failure alone would raise significant doubts about these original results. Since then, the "cluster-robust" covariance estimator – a covariance estimator that is consistent in the presence of arbitrary heteroskedasticity and arbitrary within-panel correlation – has come into general use. By employing this covariance estimator, we can improve on the prior work of Levitt and McCrary in that our hypothesis tests about the effect of police levels on crime rates, and our overidentification tests of the exogeneity of instruments, no longer rest on the strong and unrealistic assumption of i.i.d. data.

2. A separate methodological advance since Levitt's work is investigation of the "weak instruments" problem. It is now well understood that IV results are heavily dependent on the relevance of instruments. That is, they must be strongly correlated with the endogenous regressor for IV to be a valid estimator. Specification tests for the weakness of instruments have gradually become available. However, the weak IV tests were not themselves generally valid in the presence of both arbitrary heteroskedasticity and clustering of the disturbance term.

3. This created an unfortunate situation for the applied researcher: there were options for more generally robust inference (i.e., traditional IV with robust SEs) and more robust tests of instrument exogeneity, but no specification tests for assessing the weak IV problem in the presence of arbitrary heteroskedasticity and clustering of the disturbance term.

4. The problem described in #3 has just been solved by Montiel Olea-Pflueger (2013) and is ideally suited for the IV estimations used by Levitt. The results of the Montiel Olea-

[23] An anonymous reviewer noted that one potential contributor to elections being weak instruments is that at least some of the cities in Levitt's sample have an elected mayor, but that mayor does not have the authority to create a budget. So, in some cities, the mayor could advocate for more police, but would have much less authority to increase numbers than cities in which the mayor creates an initial budget. In short, the heterogeneity in city governments, even among cities that elect mayors directly, could contribute to the weakness of using these elections as an instrument (see: http://www.nlc.org/build-skills-and-networks/resources/cities-101/city-structures/forms-of-municipal-government).

> Pflueger test indicate the election cycle and firefighter hiring instruments are too weakly correlated with police levels to reliably produce a consistent point estimate of the treatment effect of hiring additional police officers on crime rates.

So how should criminologists interested in whether increasing the number of police, irrespective of what a police department might do with additional officers, reduces crime rates via deterrence mechanisms proceed? Given the importance of the topical matter, a high priority for future research is to seek out possible instruments for police levels and subject them to the rigorous specification testing described above. Of course, finding suitable instruments for endogenous regressors such as police levels is challenging and like Levitt, most scholars using IV methods will continue to face the weak instruments problem. We therefore also suggest that researchers consider alternatives to traditional IV methods.[24]

An important, relevant and recent advance in econometrics is the extension by Moreira (2003), Kleibergen (2002, 2005), and others of the estimation framework originally introduced by Anderson-Rubin (1949). The advantage of this approach is its robustness to weak instruments. Whereas instrument weakness invalidates IV parameter estimates and hypothesis tests, the "weak-instrument-robust inference" methodology of these authors allows for hypothesis tests that continue to be valid if instruments are very strong, very weak, or anywhere in between. The intuition is that as instruments become weak, confidence intervals for the estimated parameters become wider, reflecting the decreasing precision that weak instruments

[24] Of course, it bears repeating that a "good instrument is correlated with the endogenous regressor for reasons the researcher can verify and explain, but uncorrelated with the outcome variable for reasons beyond its effect on the endogenous regressor" (Angrist & Krueger, 2001, p.73). More generally and in some cases perhaps more importantly, there is a case to be made for theory. In particular, there is always the question about the extent to which the instrument(s) is(are) believable. Good theory and good knowledge about the instrumental variable(s) is always important (Angrist & Krueger, 2001, pp. 73, 76; Rosenzweig & Wolpin, 2000). In the Levitt instance, this revolves around the believability that election cycles have no direct effect on crime except through police hiring. As Kleck and Barnes (2010, p.17) note, mayors seeking reelection may employ several crime-reduction strategies besides the hiring of additional police in an attempt to reduce the crime rate. We would like to thank David Weisburd for this observation.

induce. These methods are still being developed and come with their own limitations,[25] and are not yet in wide use either in economics or in other disciplines, but they hold some promise and researchers in quantitative criminology who face the weak instruments problem may wish to consider employing them to study the police-crime issue or other criminological puzzles.

Our study was not designed to specifically examine the more police, less crime hypothesis, but rather to highlight how future investigations should be carried out with advances made in an IV framework. In this regard, we do not wish to convey that researchers should abandon the use of IV methods to solve potential endogeneity problems likely to be present in any nonexperimental macro-level study of deterrence (cf. Weisburd and Eck, 2004) or in imperfect criminological experiments more generally (Angrist, 2006, p.23). Rather, we encourage researchers who employ IV methods to use the techniques for assessing the validity and reliability of instruments we have outlined and illustrated.

[25] For example: these methods do not yield point estimates of the parameters of interest, only interval estimates; interpretation of the results can be difficult (e.g., disjoint confidence intervals are possible); and estimation with more than one endogenous regressor becomes unwieldy because for K endogenous regressors, the estimation method generate K-dimensional confidence sets (e.g., if K=2, a 2-dimensional confidence region corresponding to the two parameters of interest).

REFERENCES

Anderson, T.W. 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* 22:327-51.

Anderson, T. W. and Rubin, H. 1949. Estimation of the parameters of single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20:46–63.

Angrist, Joshua R. 2006. Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology* 2:23-44.

Angrist, Joshua D., and Alan B. Krueger. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives* 15:69-85.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton: Princeton University Press.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3:1-31.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2007. Enhanced routines for instrumental variables/GMM estimation and testing. *Stata Journal* 7:465-506.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust differences in differences estimates? *Quarterly Journal of Economics* 119:249-275.

Bushway, Shawn D. and Robert J. Apel. 2010. Instrumental variables in criminology and criminal justice. In *Handbook of Quantitative Criminology*, eds. Alex R. Piquero and David Weisburd. New York: Springer.

Chalfin, Aaron and Justin McCrary. 2013. The effect of police on crime: New evidence from U.S. cities, 1960-2010. NBER Working Paper No. 18815. (February 2013).

Evans, William N. and Emily G. Owens. 2007. COPS and crime. *Journal of Public Economics* 91:181-201.

Fisher, Franklin and Daniel Nagin. 1978. On the feasibility of identifying the crime function in a simultaneous model of crime rates and sanction levels. In *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, eds. Alfred Blumstein, Jacqueline Cohen, and Daniel Nagin, Washington, DC: National Academy Press

Greenberg, David F., Ronald C. Kessler, and Colin Loftin. 1983. The effect of police employment on crime. *Criminology* 21:375-94.

Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.

Kleck, Gary and J. C. Barnes. 2014. Do more police lead to more crime deterrence? *Crime & Delinquency* 60:716-738.

Kleibergen, F. 2002. Pivotal statistics for testing structural parameters in instrumental variables Regression. *Econometrica* 70:1781-1803.

Kleibergen, F. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73:1103-1123.

Kleibergen, F. and Paap, R. 2006. Generalized Reduced Rank Tests Using the Singular Value Decomposition. *Journal of Econometrics* 133:97-126.

Koper, Christopher S. and Jeffrey A. Roth. 2000. "Putting 100,000 Officers on the Street: Progress as of 1998 and Preliminary Projections through 2003." In Jeffrey Roth, Joseph F. Ryan, et al. (Eds.), *National Evaluation of the COPS Program -- Title I of the 1994 Crime* 8 (pp. 149-178). Washington, D.C.: U.S. Department of Justice. www.ncjrs.gov/pdffiles1/nij/183643.pdf

Levitt, Steven, D. 1995. Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. Working Paper No. 4991. Cambridge, MA: National Bureau of Economic Research.

Levitt, Steven, D. 1996. The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *The Quarterly Journal of Economics* 111:319-51

Levitt, Steven, D. 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. *The American Economic Review* 87:270-90.

Levitt, Steven, D. 2002. Using electoral cycles in police hiring to estimate the effects of police on crime: Reply. *The American Economic Review* 92:1244-50.

Lin, Ming-Jen. 2008. More police, less crime: Evidence from US state data. *International Review of Law and Economics* 29:73-80.

Marvell, Thomas and Carlisle Moody. 1994. Prison population growth and crime reduction. *Journal of Quantitative Criminology* 10:109-40.

Marvell, Thomas and Carlisle Moody. 1996. Police levels, crime rates, and specification problems. *Criminology* 34:609-46.

McCrary, Justin. 2002. Do electoral cycles in police hiring really help us estimate the effect of police on crime? Comment. *The American Economic Review* 92:1236-43.

McPheters, Lee R. and William B. Stronge. 1974. Law enforcement expenditures and urban crime. *National Tax Journal* 37:633-44.

Montiel Olea, Jose Luis and Carolin E. Pflueger. 2013. A robust test for weak instruments. *Journal of Economic and Business Statistics* 31:358-69.

Moreira, M. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71:1027-1048.

Moulton, Brent R. 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro unit. *Review of Economics & Statistics* 72:334-8.

Nagin, Daniel S. and David Weisburd. 2013. Evidence and public policy: The example of evaluation research in policing. *Criminology & Public Policy* 12:651-679.

Pflueger, Carolin E. and Su Wang. 2014. *weakivtest*: *Stata module to perform weak instrument test for a single endogenous regressor in TSLS and LIML*. http://ideas.repec.org/c/boc/bocode/s457732.html.

Pflueger, Carolin E. and Su Wang. Forthcoming. A robust test for weak instruments in Stata. *Stata Journal* (forthcoming). Draft available at http://www.carolinpflueger.com/WangPfluegerWeakivtest_20141202.pdf.

Piquero, Alex R., Nicole Leeper Piquero, Marc Gertz, Jake Bratton, and Thomas A. Loughran. 2012. Sometimes ignorance is bliss: Investigating citizen perceptions of the certainty and severity of punishment. *American Journal of Criminal Justice* 37:630-646.

Rosenzweig, Mark R. and Kenneth I. Wolpin. 2000. Natural 'natural experiments' in economics. *Journal of Economic Literature* 38:827-874.

Skogan, Wes and Lorie Fridell (Eds.). 2004. *Fairness and Effectiveness in Policing*. Washington, DC: Committee to Review Research on Policy and Practices, National Research Council.

Spelman William. 2008. Specifying the relationship between crime and prisons. *Journal of Quantitative Criminology* 24:149-78.

Staiger, Douglas and James H. Stock. 1997. IV regression with weak instruments. *Econometrica* 65:557-86.

Stock, James H. and Mark M. Watson. 2010. *Introduction to Econometrics*. Prentice Hall.

Stock, James H. and Motohiro Yogo. 2005. Testing for weak instruments in linear IV regression. In D.W.K. Andrews and J.H. Stock, eds. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge University Press. Working paper version: NBER Technical Working Paper No. 284, http://www.nber.org/papers/T0284.

Weisburd, David L. and John E. Eck. 2004. What can police do to reduce crime, disorder, and fear? *The ANNALS of the American Academy of Political and Social Science* 593:42-65.

Willis, Ken. 1983. Spatial variations in crime in England and Wales: Testing an economic model. *Regional Studies* 17:261-72.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Worrall, John L., and Tomislav Kovandzic. 2007. COPS grants and crime revisited. *Criminology* 45:159-190.

Zhao, Jihong "Solomon", Matthew C. Scheider, and Quint Thurman. 2002. Funding community policing to reduce crime: Have COPS grants made a difference. *Criminology & Public Policy* 2: 7-32.

Table 1: Tests of Underidentification and Weak Identification for Original Levitt-McCrary Estimation

First-stage Regressions				
	Levitt mayoral measure		**McCrary mayoral measure**	
	(1)	**(2)**	**(3)**	**(4)**
	i.i.d.	**cluster-robust**	**i.i.d.**	**cluster-robust**
Endog. regressor: Police (t-1)				
Variable:				
Mayoral	0.0108**	0.0108	0.0122***	0.0122**
election year (t-1)	(0.0043)	(0.0065)	(0.0040)	(0.0053)
Gubernatorial	0.0264***	0.0264**	0.0243***	0.0243**
election year (t-1)	(0.0067)	(0.0102)	(0.0066)	(0.0093)
Underidentification test	$\chi^2(2)$=21.02	$\chi^2(2)$=5.25	$\chi^2(2)$=24.20	$\chi^2(2)$=5.44
p-value	0.000	0.072	0.000	0.066
Weak identification F statistics	Cragg-Donald	Montiel-Pflueger	Cragg-Donald	Montiel-Pflueger
	9.87	4.45	11.40	6.39
Critical Values for Weak Identification Tests				
Stock-Yogo Critical Values for				
Cragg-Donald Statistic				
Test level 5%				
Percent maximal IV size distortion:				
15%	11.59		11.59	
20%	8.75		8.75	
Montiel-Pflueger Critical Values for				
Robust Effective F Statistic				
Test level 5%				
Percent of worst-case Nagar bias:				
5%		9.108		18.500
10%		6.453		11.964
20%		4.920		8.226
30%		4.343		6.804
Test level 10%				
Percent of worst-case Nagar bias:				
5%		7.641		16.057
10%		5.245		10.033
20%		3.895		6.662
30%		3.339		5.407

Table 2: Estimates of the Impact of Police Force Size on the Annual Growth Rate of Violent Crime in 59 U.S. Cities with Directly Elected Mayors, 1970 to 2008: Results Using the First-Difference Estimator

	(1) OLS Estimation	(2) Standard IV Estimation	(3) LIML Estimation	(4) Same as Model 2 & Include 2-Year Lag of Sworn Officers	(5) Same as Model 4 & Weight by City Population	(6) Same as Model 5 & Limit Study Sample from 1970-1992	(7) Same as Model 6 & Limit Control Variables to Levitt (1997)
Police treated as:	Exogenous	Endogenous	Endogenous	Endogenous	Endogenous	Endogenous	Endogenous
Variable:							
ln Sworn officers per capita (t-1)	-0.1645***	-0.0405	-0.0403	0.0856	0.4318*	0.4640	0.4181
	(0.0619)	(0.2078)	(0.2082)	(0.1886)	(0.2460)	(0.3269)	(0.3135)
ln Sworn officers per capita (t-2)	-	-	-	-0.2122	-0.5847*	-0.4830	-0.4425
	-	-	-	(0.2105)	(0.3316)	(0.5043)	(0.5085)
Combined effect of (t-1) and (t-2)	-	-	-	-0.1266	-0.1529	-0.0190	-0.0244
	-	-	-	(0.1441)	(0.2186)	(0.3792)	(0.3928)
State unemployment rate	-0.0060	-0.0061	-0.0061	-0.0058	-0.0040	-0.0003	-0.0002
	(0.0064)	(0.0063)	(0.0063)	(0.0062)	(0.0051)	(0.0050)	(0.0048)
Percent black	0.0166	0.0175	0.0175	0.0176	0.0195	-0.0098	-0.0104
	(0.0133)	(0.0128)	(0.0128)	(0.0113)	(0.0165)	(0.0282)	(0.0292)
Percent ages 15-24 in SMSA	0.0568*	0.0549*	0.0549*	0.0478*	0.0487***	-0.0395	-0.0459
	(0.0304)	(0.0306)	(0.0306)	(0.0274)	(0.0179)	(0.0495)	(0.0464)
Percent female-headed households	-0.0501**	-0.0538**	-0.0538**	-0.0486**	-0.0934***	-0.0986*	-0.1083*
	(0.0227)	(0.0245)	(0.0245)	(0.0246)	(0.0333)	(0.0570)	(0.0596)
ln State incarceration rate (t-1)	-0.1703***	-0.1737***	-0.1737***	-0.1626***	-0.2000***	-0.2369***	-
	(0.0378)	(0.0397)	(0.0397)	(0.0416)	(0.0544)	(0.0557)	-
State beer consumption x 1000	0.0643	0.0624	0.0624	0.0295	0.0499	0.1221	-
	(0.1384)	(0.1389)	(0.1389)	(0.1569)	(0.1340)	(0.1795)	-
Shall-Issue Law Dummy	0.0452	0.0466*	0.0466*	0.0538**	0.0786***	-0.0451	-
	(0.0279)	(0.0279)	(0.0279)	(0.0262)	(0.0274)	(0.0744)	-
J statistic (overidentification)	$\chi^2(3)$= 0.583	$\chi^2(2)$=0.162	$\chi^2(2)$=0.162	$\chi^2(4)$= 4.298	$\chi^2(4)$= 3.948	$\chi^2(4)$= 2.145	$\chi^2(4)$= 3.036
p-value	0.900	0.922	0.922	0.367	0.413	0.709	0.552
C statistic (endogeneity)	-	$\chi^2(1)$=0.418	$\chi^2(1)$=0.418	$\chi^2(2)$=1.171	$\chi^2(2)$=4.652	$\chi^2(2)$=4.408	$\chi^2(2)$=4.161
p-value	-	0.518	0.518	0.557	0.098	0.110	0.125
Underidentification test	-	$\chi^2(3)$=12.66	$\chi^2(3)$=12.66	$\chi^2(12)$= 18.67	$\chi^2(12)$= 18.60	$\chi^2(12)$= 17.99	$\chi^2(12)$= 18.15
p-value	-	0.005	0.005	0.013	0.007	0.030	0.031
Number of observations	2078	2078	2078	2005	2005	1129	1129

Notes: *p<0.10, ** p<0.05, *** p<0.01. The table presents estimates of the elasticity of violent crime rates with respect to police growth rates. The sample includes 59 U.S. cities with directly elected mayors from 1970 to 2008. With the exception of the mayoral and gubernatorial election-year indicators, all right-hand-side variables are differenced in columns (1)-(6). Year and city-fixed effects are included in all estimations. Column (1) presents OLS estimates in which the current-year growth rate of police per capita is treated as exogenous. Column 2 presents IV estimates in which the 1-year-lagged growth rate of police per capita is treated as endogenous and instrumented using 1-year-lagged mayoral and gubernatorial election-year indicators and the 1-year-lagged growth rate in firefighters per capita. Column (3) is identical to column (2) except that the LIML estimator is substituted for the traditional IV estimator. Columns (4)-(6) present results similar to those reported in columns (1)-(3) except that the 2-year-lagged election-year and firefighter instruments and the growth rate in police per capita are also included. Cluster-robust standard errors (by state) are presented in parentheses. All tests are robust to clustering except for the Cragg-Donald 1st-stage F statistic, which is valid only for i.i.d. data (see discussion in text).

Table 3: First-Stage Regression Results for Specifications in Table 2

Model:	**(2)-(3) Standard IV and LIML Estimation**	**(4) Same as Model 2 & Include 1-Year Lag of Sworn Officers**		**(5) Same as Model 4 & Weight by City Population**		**(6) Same as Model 5 & Limit Study Sample from 1970-1992**		**(7) Same as Model 6 & Limit Control Variables to Levitt (1997)**	
Endog. regressor:	Police (t-1)	Police (t-1)	Police (t-2)	Police (t-1)	Police (t-2)	Police (t-1)	Police (t-2)	Police (t-1)	Police (t-2)
Variable:									
Mayoral	0.0089**	0.0075**	0.0043	0.0051	0.0058*	0.0070	0.0027	0.0071	0.0023
election year (t-1)	(0.0034)	(0.0036)	(0.0042)	(0.0034)	(0.0032)	(0.0069)	(0.0051)	(0.0070)	(0.0052)
Gubernatorial	0.0218***	0.0200**	-0.0045	0.0235**	-0.0004	0.0227***	-0.0050	0.0226***	-0.0048
election year (t-1)	(0.0078)	(0.0079)	(0.0074)	(0.0086)	(0.0059)	(0.0078)	(0.0091)	(0.0079)	(0.0092)
Firefighter	0.0642*	0.0842**	0.0152	0.0647**	0.0158	0.0427*	0.0033	0.0428**	0.0034
levels (t-1)	(0.0358)	(0.0400)	(0.0136)	(0.0310)	(0.0187)	(0.0209)	(0.0141)	(0.0209)	(0.0141)
Mayoral	-	-0.0027	0.0108***	-0.0043	0.0082***	0.0005	0.0094	0.0006	0.0091
election year (t-2)	-	(0.0031)	(0.0029)	(0.0040)	(0.0026)	(0.0057)	(0.0056)	(0.0057)	(0.0057)
Gubernatorial	-	-0.0045	0.0171***	-0.0047	0.0217**	0.0019	0.0211***	0.0020	0.0211***
election year (t-2)	-	(0.0057)	(0.0056)	(0.0066)	(0.0081)	(0.0061)	(0.0072)	(0.0061)	(0.0070)
Firefighter	-	0.0475***	0.0614*	0.0433***	0.0486	0.0315***	0.0305	0.0315***	0.0306
levels (t-2)	-	(0.0162)	(0.0345)	(0.0137)	(0.0324)	(0.0090)	(0.0253)	(0.0089)	(0.0254)

Notes: *p<0.10, ** p<0.05, *** p<0.01. The table presents the first-stage regressors for the models in Table 2. Specifications and numbers of observations are as in Table 2. Cluster-robust standard errors (by state) are presented in parentheses.

Table 4: Stock-Yogo and Montiel-Pflueger Critical Values

	Models (2, 4) (IV, unweighted)				**Model (3) (LIML, unweighted)**				**Model (5) (IV, city population weights)**			
Cragg-Donald F statistic	29.46				29.46				16.77			
Stock-Yogo critical values for Cragg-Donald F statistic	**Percent maximal IV relative bias** 5%	10%	20%	30%	<maximal bias test not available>				**Percent maximal IV relative bias** 5%	10%	20%	30%
	13.91	9.08	6.46	5.39					13.91	9.08	6.46	5.39
Stock-Yogo critical values for Cragg-Donald F statistic	**Percent maximal IV size distortion** 10%	15%	20%	25%	**Percent maximal IV size distortion** 10%	15%	20%	25%	**Percent maximal IV size distortion** 10%	15%	20%	25%
	22.30	12.83	9.54	7.80	6.46	4.36	3.69	3.32	22.30	12.83	9.54	7.80
Montiel-Pflueger robust effective F statistic	4.12				4.12				3.91			
Montiel-Pflueger critical values for robust effective F statistic	**Percent of worst-case Nagar bias** 5%	10%	20%	30%	**Percent of worst-case Nagar bias** 5%	10%	20%	30%	**Percent of worst-case Nagar bias** 5%	10%	20%	30%
Test level 5%	32.91	20.24	13.14	10.48	30.54	18.93	12.39	9.94	32.39	19.82	12.80	10.18
Test level 10%	29.28	17.43	10.91	8.51	27.05	16.21	10.23	8.02	28.95	17.16	10.70	8.33
Test level 25%	23.69	13.20	7.64	5.68	21.69	12.14	7.08	5.29	23.63	13.15	7.61	5.66
Test level 50%	18.17	9.19	4.70	3.21	16.43	8.32	4.27	2.93	18.35	9.32	4.80	3.31
Number of obs.	2078				2078				2005			

Table 4 (continued): Stock-Yogo and Montiel-Pflueger Critical Values

	Model (6) (IV, weighted, 1970-92 sample)				**Model (7) (IV, weighted, 1970-92, Levitt covariates)**			
Cragg-Donald F statistic	7.33				7.35			
Stock-Yogo critical values for Cragg-Donald F statistic	**Percent maximal IV relative bias**				**Percent maximal IV relative bias**			
	5%	10%	20%	30%	5%	10%	20%	30%
	13.91	9.08	6.46	5.39	13.91	9.08	6.46	5.39
Stock-Yogo critical values for Cragg-Donald F statistic	**Percent maximal IV size distortion**				**Percent maximal IV size distortion**			
	10%	15%	20%	25%	10%	15%	20%	25%
	22.30	12.83	9.54	7.80	22.30	12.83	9.54	7.80
Monteiel-Pflueger robust effective F statistic	2.86				2.82			
Montiel-Pflueger critical values for robust effective F statistic	**Percent of worst-case Nagar bias**				**Percent of worst-case Nagar bias**			
	5%	10%	20%	30%	5%	10%	20%	30%
Test level 5%	26.82	16.57	10.83	8.68	26.82	16.57	10.82	8.67
Test level 10%	23.89	14.30	9.03	7.10	23.89	14.30	9.03	7.09
Test level 25%	19.37	10.88	6.40	4.81	19.37	10.88	6.40	4.82
Test level 50%	14.90	7.64	4.01	2.81	14.91	7.64	4.02	2.82
Number of obs.	1129				1129			

Appendix Table: Means and Standard Deviations of Variables

	Models (1)-(3)		Models (4)-(5)		Models (6)-(7)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Variable:						
ln Sworn officers per capita (t-1)	0.010	0.125	0.010	0.124	0.042	0.124
State unemployment rate	0.006	0.059	0.005	0.058	0.006	0.063
Percent black	0.005	1.088	0.015	1.101	0.094	1.307
Percent ages 15-24 in SMSA	0.192	0.393	0.184	0.390	0.266	0.398
Percent female-headed households	-0.112	0.238	-0.120	0.237	-0.194	0.260
ln State incarceration rate (t-1)	0.059	0.175	0.064	0.172	0.057	0.21
State beer consumption x 1000	0.045	0.070	0.046	0.068	0.060	0.077
Shall-Issue Law Dummy	0.004	0.030	0.003	0.029	0.009	0.034
Mayoral election year (t-1)	0.015	0.094	0.015	0.096	0.004	0.045
Gubernatorial election year (t-1)	0.002	0.148	0.002	0.150	0.001	0.190
Firefighter levels (t-1)	0.303	0.460	0.301	0.459	0.307	0.462
Number of observations	2078		2005		1129	