Journal of Quantitative Criminology (2019) 35:631–662
https://doi.org/10.1007/s10940-018-9385-x

ORIGINAL PAPER

Quantifying the Likelihood of False Positives: Using Sensitivity Analysis to Bound Statistical Inference

Kyle J. Thomas[1] · **Jean Marie McGloin**[2] · **Christopher J. Sullivan**[3]

Published online: 22 June 2018

Abstract

Objective Criminologists have long questioned how fragile our statistical inferences are to unobserved bias when testing criminological theories. This study demonstrates that sensitivity analyses offer a statistical approach to help assess such concerns with two empirical examples—delinquent peer influence and school commitment.

Methods Data from the Gang Resistance Education and Training are used with models that: (1) account for theoretically-relevant controls; (2) incorporate lagged dependent variables and; (3) account for fixed-effects. We use generalized sensitivity analysis (Harada in ISA: Stata module to perform Imbens' (2003) sensitivity analysis, 2012; Imbens in Am Econ Rev 93(2):126–132, 2003) to estimate the size of unobserved heterogeneity necessary to render delinquent peer influence and school commitment statistically non-significant and substantively weak and compare these estimates to covariates in order to gauge the likely existence of such bias.

Results Unobserved bias would need to be unreasonably large to render the peer effect statistically non-significant for violence and substance use, though less so to reduce it to a weak effect. The observed effect of school commitment on delinquency is much more fragile to unobserved heterogeneity.

Conclusion Questions over the sensitivity of inferences plague criminology. This paper demonstrates the utility of sensitivity analysis for criminological theory testing in determining the robustness of estimated effects.

Keywords Sensitivity analysis · False positives · Unobserved bias · Theory testing

> *"A fragile inference is not worth taking seriously."*
>
> *Leamer (1985)*

✉ Kyle J. Thomas
thomaskj@umsl.edu

[1] University of Missouri-St. Louis, 1 University Blvd, 331 Lucas Hall, St. Louis, MO 63121, USA

[2] University of Maryland, College Park, MD, USA

[3] University of Cincinnati, Cincinnati, OH, USA

Introduction

When testing the validity of theories, criminologists are generally concerned with two related issues. The first concern is *identification*, which relates to whether—in an infinitely large sample—the effect of a theoretical construct can be realistically isolated from other endogenous factors. If this is possible (e.g., through randomized controlled experiments) then researchers can "identify" a clean point estimate of the population parameter that reflects the relationship between two theoretical constructs. Due to both practical and ethical limitations, however, true identification is extraordinarily difficult in the study of crime and deviance (McGloin and Thomas 2013), which ultimately means that there is some degree of uncertainty concerning the accuracy of model parameters' point estimates.

The second concern relates to the statistical *inferences* that can be made given the observational data that we do have. Statistical inference using hypothesis testing is of particular interest to criminologists because, in the case of parametric regression, this translates into how likely it is that the estimate $\hat{\beta}_j$ would be (at least) that size given that the true relationship between two constructs is zero (i.e., $\beta_j = 0$). A rejection of the null hypothesis could reflect at least two possibilities: (1) the null hypothesis is actually false and a particular theory is valid or; (2) the regression is misspecified and the parameter estimate $\hat{\beta}_j$ is biased due to unobserved heterogeneity that correlates residuals to covariates in the model. This latter possibility reflects a "false positive" inference regarding the empirical validity of a theory. In light of the concerns about a replication crisis, which point to a pattern of potential false positives, the possibility that statistical inferences are fragile poses significant problems for the empirical standing of our theories. As a consequence, the sensitivity of inferences to unobserved heterogeneity is a vital concern.

Indeed, concerns about Type I errors are rampant in criminology. Few scholars dispute that criminal behavior is correlated with factors such as delinquent peer associations, school commitment, perceived arrest risk, and marriage, but there is considerable disagreement regarding the meaning of these relationships (e.g., Hirschi and Gottfredson 1993). Typically these disputes stem from different theoretical traditions that make opposing claims. Learning theories, for example, view the statistically significant relationship between deviant peer associations and delinquency as evidence of a causal effect (Akers 1998; Sutherland 1947), whereas control theories view it as spurious (i.e., that the null hypothesis is actually true). Even among control theories, some view school commitment and employment as being causally related to offending (Hirschi 1969; Sampson and Laub 1993) while others suggest that there are omitted factors that can explain both criminal behavior and school/employment effects (Gottfredson and Hirschi 1990). Because the feasibility of leveraging experiments to identify point estimates of many of our core theoretical concepts is limited (Sampson 2010), studies must make concerted efforts to address selection issues when testing theories using observational data and, in turn, be cautious when making inferences about the relationship between two constructs. Even with due diligence, however, both explicit and implicit skepticism about statistical inferences—and thus the validity of theories—remains.

Across the discipline, three types of modeling strategies are typically employed to reduce concerns of Type I error when relying on regression-based approaches. Perhaps most common, researchers include observable constructs from rival theories in models along with factors such as age and gender, under the assumption that accounting for these variables leads to an unbiased estimate of the theoretical construct of interest (e.g., delinquent peers, school commitment, perceived arrest risk). Second, researchers may include a lagged dependent variable into regression models (i.e., prior delinquency), in addition to the relevant controls. The motivation for including a lagged term is that any selection effect that promotes Type I error would inherently be shared with delinquency at the same observation wave of the main independent variable. Thus, accounting for prior measures of the outcome removes the variance from the estimate that is due to selection factors related to delinquency (Haynie and Osgood 2005). A third method involves an estimation of within-individual changes in delinquency through fixed-effects modeling. Fixed-effects models are recognized as "one of the most powerful tools for studying causal processes using nonexperimental data" (Osgood 2010, p. 380) because they allow the researcher to control for observed relevant time-variant risk factors and remove any time-stable sources of selection.

The first two strategies are analogous to a "selection on observables" assumption in econometrics, whereas the latter accounts for concerns of "selection on unobservables." The three approaches often reach similar conclusions regarding the statistical and substantive relationships between theoretical constructs and crime (e.g., Fergusson et al. 2002; Horney et al. 1995; Loughran et al. 2016; McGloin et al. 2014; Sampson and Laub 1993; Thomas 2015). Yet, it remains the case that these are *statistical* attempts to resolve an issue that is inherently tied to *research design*. Because no modeling approach can change the fact that observational studies are vulnerable to questions of whether unobserved selection bias is responsible for observed treatment effects, concerns about false positives and fragile inferences endure (Rubin 2008). This is perhaps best exemplified by Gottfredson and Hirschi's (1990: 156) skepticism about the observed relationship between peers and delinquency: "How much easier would it be to assume that the 'delinquent peer group' is a creation of … the tendency of people to seek the company of others like themselves?" Gottfredson and Hirschi (1990) likely intended this to be a rhetorical and theoretical question, but it nonetheless raises an issue that may be useful when judging the validity of various criminological theories: Given an observed treatment effect, how reasonable is it to assume that the relationship is actually spurious?

Sensitivity analysis was designed precisely to examine how robust observed treatment effects and statistical inferences are to threats of unobserved bias (Rosenbaum and Rubin 1983). These approaches are most commonly used to assess threats to internal validity in experimental designs, but we believe that sensitivity analysis offers a useful strategy to examine the "fragility" of predictions made in different criminological theories and, in turn, assess the relative validity of different perspectives (Hirschi 1979). This approach cannot rectify the limitations of observational research in establishing causality, but it does help estimate the robustness of statistical inferences by identifying the point(s) at which key conclusions would shift.

The current study considers the utility of sensitivity analysis for tests of criminological theory by assessing the fragility of statistical estimates in models used commonly in the existing literature. This paper first considers the different assumptions of these approaches and then performs sensitivity analyses across them using two different thresholds, providing interpretable estimates of "how easy" it would be to assume that unobserved bias explains statistically and substantively significant estimates using two empirical examples: delinquent peers and commitment to school. In the end, this paper highlights the general utility of sensitivity analysis for testing criminological theories.

Inferring a Treatment Effect from Regression Models

The vast majority of research in criminology relies on observational survey data. Using such data, a bivariate regression testing a theory can be modeled using the equation:

$$\hat{y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 W_{it-1} + \eta_i + \delta_{it} + \varepsilon_{it} \tag{1}$$

where $\hat{y}_i$ is respondents' delinquency (written here as continuous), β_0 is a constant, $\hat{\beta}_1$ is the regression weight for the theoretical predictor of interest (W, which here is captured at the previous time point to ensure temporal ordering[1]), η_i is time-stable unobserved heterogeneity endogenous to the treatment, δ_{it} is time varying unobserved heterogeneity endogenous to the treatment, and ε_{it} is truly random error. Assume here that W is a construct that is thought to be positively related to delinquency. If $\hat{\beta}_1 > 0$ and statistically significant, this suggests that W correlates with offending such that higher values of W predict more delinquency. Determining whether this observed effect is causal is difficult, however, due to the endogenous heterogeneity captured by η_i and δ_{it}. In order to interpret $\hat{\beta}_1$ as robust to alternative explanations, one must assume that unobserved heterogeneity is independent and identically distributed (i.i.d.) across individuals. Importantly, if there is an unobserved factor that influences *either* respondent delinquency (y_{it}) or the construct of interest (W), but not both, then the $\hat{\beta}_1$ is unbiased.

That unobserved factors must *simultaneously affect both* the treatment and respondent delinquency is a necessary—but often misunderstood—condition for assuming away an effect. For instance, consider the example of associating with delinquent peers. There is little question that individuals select friends based on a host of factors (e.g., shared interests in music and sports), but this selection process is only problematic for interpreting the peer-delinquency relationship if the selection factors also affect one's delinquent tendencies (e.g., a shared interest in "getting high"; a lack of commitment to prosocial institutions). If such selection exists, then the estimate of $\hat{\beta}_1$ is upwardly biased with respect to β, such that:

$$\text{Cov}[\eta_i + \delta_{it}, W_{it-1}] > 0,$$

and thus,

$$E[\hat{\beta}] > \beta$$

Because it is often reasonable to assume that there are factors that affect both selection into a treatment and one's own delinquency, researchers rely on various regression techniques in estimating these relationships. In general, scholars tend to employ three modeling strategies: controlling for observed factors that are theoretically related to selection; including a lagged dependent variable; or using fixed-effects estimation.

Including Theoretically Relevant Covariates

One common approach adjusts for the likely untenable assumption that unobserved heterogeneity is i.i.d. by accounting for potential confounders observed in data—i.e., the theoretical sources of spuriousness, such that:

$$y_{it} = \hat{\beta}_{0it} + \hat{\beta}_1 W_{it-1} + \hat{\beta}_k \boldsymbol{X}_{it-1} + \eta_i + \delta_{it} + \varepsilon_{it} \tag{2}$$

[1] Of course, scholars may also specify a cross-sectional model in which the key independent variable is measured at the same time point as the dependent variable.

Here, X_{it-1} represents a vector of observed control variables that, based on theory, might render a treatment effect spurious.[2]

This model has several features that researchers may find appealing. First, the dependent variable y_{it} remains the absolute rate (or probability) of delinquency; the equation therefore estimates how the theoretical construct of interest affects individual differences in the overall tendency to be delinquent. Moreover, it is a theoretically-driven and parsimonious attempt at model building. Under the assumption that the vector of specified controls captures much of the individual heterogeneity that promotes selection, researchers can assume that $\hat{\beta}_1$ is an (relatively) unbiased estimate, such that:

$$E[\eta_i + \delta_{it} | X_{it-1}] = 0, \ \mathrm{Cov}[\eta_i + \delta_{it}, W_{it-1}] = 0,$$

and thus,

$$E[\hat{\beta}_1 | \mathbf{X}] \approx \beta$$

Given that existing observational datasets can be limited in the available measures, however, this assumption may not hold in practice. Other unobserved factors, which again may be time stable or time variant, that affect both selection into a treatment and delinquency may still exist and have an impact large enough to render the relationship statistically or substantively non-significant.

Lagged Dependent Variable (LDV) Models

Because enduring concerns about whether $\eta_i + \delta_{it}$ is i.i.d. revolve around unobservable factor(s), controlling for observables captured in existing datasets cannot fully rectify matters. In order to circumvent this problem, scholars often include lagged dependent variables, such that:

$$y_{it} = \hat{\beta}_{0it} + \hat{\beta}_1 W_{it-1} + \hat{\beta}_2 y_{it-1} + \hat{\beta}_k \mathbf{X}_{it-1} + \eta_i + \delta_{it} + \varepsilon_{it} \tag{3}$$

The argument behind adding this term rests in the fact that "any variance in [the treatment] that is attributable to selection factors relevant to delinquency is necessarily shared with the contemporaneous measure of delinquency", and thus this approach eliminates unobserved variance that simultaneously affects both the treatment and the outcome (Haynie and Osgood 2005: 1119; see also Bollen and Curran 2006). Using this model, one can be confident that the treatment estimate is relatively unbiased under the assumptions that:

$$E[\eta_i + \delta_{it} | X_{it-1}, y_{it-1}] = 0, \ \mathrm{Cov}[\eta_i + \delta_{it}, W_{it-1}] = 0,$$

which suggests that:

$$E[\hat{\beta} | \mathbf{X}, y_{it-1}] \approx \beta$$

For $\hat{\beta}_1$ to be biased in favor of a statistically significant effect in this model there must be unobserved factors that influence both treatment selection and delinquency *above and beyond the subject's prior delinquency* and the other covariates.

[2] We specify a condition whereby the independent variable and covariates are captured at the wave prior to measuring the dependent variable because that clearly has become the favored method over time when using this approach. But, the assumptions of the model and the use of sensitivity analysis would not change under a cross-sectional specification.

There are some important limitations in taking this approach, however. First, lagged dependent variables are often an atheoretical way to account for selection bias (Achen 2001). Criminological theories do not posit that prior delinquency *directly* causes future delinquency. Rather, it acts as an indicator of some enduring propensity (Hirschi and Gottfredson 1993), is exogenous to other factors (e.g., identity change; Lemert 1951), or plays an indirect role in some state dependent process (Nagin and Paternoster 1991). In this way, delinquency at time $t-1$ may be among the strongest predictors of delinquency at time *t,* but it is not theoretically clear what this relationship means. This is particularly problematic when prior behavior is caused by the factors for which the researcher is estimating an effect. For example, if prior delinquency is affected by factors such as peer influence or a weak commitment to school, then controlling for delinquency at time $t-1$ can erroneously lead to the conclusion that these constructs do not impact delinquent tendencies. The predictors are effectively cannibalized by the prior delinquency term just as, for example, the effect of education on one's income this year would likely be fully attenuated by one's lagged earnings from last year.

Second, lagged dependent variables are necessarily correlated with the structural error term $\varepsilon_{\mathrm{it}}$, which violates an important assumption of all parametric regression models. This violation leads to systematic downward bias of other coefficients in the model; in fact, Achen (2001) has referred to lagged dependent variables as "kleptomaniacs" that steal the effects of other model estimates. Haynie and Osgood (2005: 1119) even noted that adjusting for prior delinquency is "likely too strong a control for selection factors" when testing criminological theories, meaning that including the lagged term may overcompensate for potential Type I errors by raising the likelihood of false negatives (Type II error). This may be of greater concern because, as noted above, LDVs affect other partial estimates without necessarily providing an explanatory benefit. Finally, accounting for prior delinquency algebraically changes the interpretation of the outcome. Whereas the outcome y_{it} reflected individual *i's* rate or probability of offending at time *t* in Eq. (2), when a lagged term is included, the outcome instead reflects *changes* in the rate or probability of their offending likelihood between time *t* and $t-1.0$.

Under this model unobserved heterogeneity must explain both selection into the treatment and delinquency but also: (1) not be subsumed by observed covariates captured in $\boldsymbol{X}_{it-1}$ and; (2) also not be captured by individuals' prior propensity to be delinquent, measured as $\mathrm{y}_{\mathrm{it-1}}$. Given this, scholars generally have confidence in the robustness of identified treatment effects on delinquency, and in turn the validity of a theory under study, when estimating LDV models. Caution is still necessary, however, as there is a non-zero chance that such bias does exist.

Fixed-Effects Estimates

Although not used as frequently as the prior approaches, fixed-effects models have become an increasingly popular approach to account for selection when testing the predictions of criminological theories (Horney et al. 1995; Loughran et al. 2016; Sampson and Laub 1993). These models estimate effects through the equation:

$$y_{it} - \bar{y}_i = \hat{\beta}\left(W_{it} - \bar{W}_i\right) + \hat{\beta}\left(\mathbf{X}_{it} - \bar{\mathbf{X}}_i\right) + \left(\eta_i - \bar{\eta}_i\right) + (\delta_{it} - \bar{\delta}_i) + \varepsilon_{it} \tag{4}$$

where $y_{i.}$ reflects the rate or probability of delinquency, W is the main theoretical construct of interest, $\boldsymbol{X}$ is a vector of controls, and η_i and δ_{it} are time stable and time varying individual heterogeneity, respectively, that affect both treatment selection and delinquency.[3] Each parameter is estimated as the individual's departure from his or her own mean over the panel, and therefore the model reflects within-individual changes in delinquency. In this approach, all unobserved time stable factors captured in η_i are differenced out of the model given that $\eta_i - \bar{\eta}_i = 0$. Observed time-variant factors potentially captured in δ_{it} are accounted for in the vector of controls $\boldsymbol{X}_i$. Given this, fixed-effects models are considered an effective approach to estimate causal effects (Osgood 2010), though as with the LDV approach, the additional control over isolating an effect might come at the expense of understanding the factors that are at play in a complicated developmental process that is often of interest for criminologists.

The fixed-effects method has been used to test the predictions made by various criminological theories. Ferguson and colleagues (2002) used a fixed-effects model to examine if within-individual changes in deviant peer associations affect changes in substance use. Horney et al. (1995) tested the notion that changes in marital and employment predict within-person changes in offending using fixed-effects (Sampson and Laub 1993). More recently, Loughran et al. (2016) examined if within-person changes in the rewards, costs and risks associated with crime predicted within-individual changes in offending. All of these studies found statistically significant effects and, thus, support for the respective theories. As with the prior regression methods, however, concerns about unobserved bias persist. Although the model differences out potential unobserved heterogeneity that is time invariant (η_i), the assumption that δ_{it} is i.i.d is violated if there is some *time-variant factor* not included in the model that explains both treatment selection and delinquency.

Testing the Robustness of Statistical Inferences

In the absence of experimental designs, claims of spuriousness are easy to make but difficult to refute. In light of the difficulty in identifying a clean point estimate reflecting the effect that some variable has on criminal behavior, it may instead be useful to examine how large the unobserved heterogeneity $\eta_i + \delta_{it}$ would need to be to render the observed treatment effect spurious (according to a particular metric). Sensitivity analysis addresses this specific issue by providing a quantitative indicator(s) of the degree to which unobserved covariates would need to affect both the treatment assignment (e.g., peer delinquency, school commitment) and the outcome (i.e., delinquency) in order to render the treatment effect null. As Young (2014) observes, this is somewhat analogous to the fail-safe N estimate in meta-analysis, which indicates how many null unpublished findings must exist in order to annul a statistically significant effect (Orwin 1983; Rosenberg 2005). As with the fail-safe N, the calculated value means little on its own – it requires some reference point to assess the "reasonableness" of such an unobserved factor. Fortunately, some sensitivity analysis approaches provide intuitive comparisons regarding this reasonableness question while maintaining interpretability.

[3] Note that the temporal subscripts t in Eq. (5) are the same for both the outcome and the treatment (and covariates). Whether to use contemporaneous or lagged predictors in a fixed-effects model is a theoretical issue (Collins 2006), and nothing prohibits the sensitivity analysis described below from utilizing lagged predictors. We use contemporaneous fixed effects models for demonstrative purposes in the current study because they are employed most commonly in criminology (Fergusson et al. 2002; Loughran et al. 2016; Nguyen et al. 2016).

Imbens (2003) builds on Rosenbaum and Rubin's (1983) seminal approach by extending sensitivity analysis to models including control variables. This allows for a consideration of the size of the necessary unobserved bias once accounting for other factors and also provides points of reference for assessing how likely it is that this unobserved bias exists. Consider a study interested in the relationship between school commitment and delinquency. Prior research consistently demonstrates that individual characteristics such as self-control are strongly related both to the treatment (commitment to school) and the outcome (delinquency) and is included as an observed covariate in regressions. Imbens' approach provides an estimate of what portion of the treatment and the outcome are attributable to the observed covariates, and then allows for a comparison of these values, providing a sense of whether the unobserved effect necessary to diminish the focal estimate is plausible in a relative sense. So, if self-control explains 15% of the variation in the treatment and 8% of the variation in the outcome, then an estimate that the hypothetical unobserved covariate must explain at least 45% of the variation in the treatment and 55% of the variation of the outcome in order to nullify the treatment effect suggests that the existence of such a factor may be unlikely. In other words, the treatment effect is robust. If, however, the estimate is that the unobserved bias must explain at least 5% of the variation in the treatment and 7% of the variation of the outcome in order to nullify the treatment effect, then it is reasonable to assume that such a factor may exist and the observed estimate is sensitive to Type I error induced by omitted variable bias. These examples are reflective of another benefit offered by Imbens' method: The estimates of the unobserved bias can be given as partial R-squared values, which provides for easy interpretation and comparison.

Imbens initially designed his sensitivity analysis for cases with continuous outcomes and binary treatments, but Harada (2012) has developed a Stata module that generalizes Imbens' approach to allow for binary outcomes and continuous treatments. This generalized sensitivity analysis specifies two equations, one of which refers to how the unobserved factor affects the treatment and the other to how it affects the outcome. The effect of the unobserved heterogeneity on a binary treatment is estimated using the equation:

$$Log\left(W_i = 1|1 - W_i = 1\right) = \alpha U_i + \gamma \mathbf{X}_i + \varepsilon_{it} \tag{5}$$

where W_i in this case is a binary treatment, U_i is the unobserved covariate that is assumed in this study to be continuously distributed,[4] and $\boldsymbol{X}_i$ is the vector of observed covariates that researchers wish to control for in models. This equation estimates alpha, or the effect of the unobserved bias on the treatment, while accounting for specified observed covariates. The second equation calculates the effect of unobserved bias on the outcome (known as delta), while accounting for specified observed covariates. The formula, which is specified using a binary outcome, is:

$$Log\ \left(Y_i = 1|1 - Y_i = 1\right)\ = \tau W_i + \delta U_i + \beta X_i + \varepsilon_{it} \tag{6}$$

where Log(Y_i|1 − Y_i) is the log odds of engaging in delinquency, τW_i is the estimated effect of the treatment, U_i is the unobserved continuous covariate and βX_i is the impact of the vector of observed covariates. Generalized sensitivity analysis provides a band of joint alpha and delta estimates, where the unobserved covariate must meet a certain, concurrent alpha *and* delta threshold in order to nullify the observed treatment effect on the outcome.

[4] The generalized sensitivity analysis approach allows researchers to specify that the model treats the unobserved heterogeneity as binary as well. A binary characterization of unobserved heterogeneity seems unlikely in the case of criminal propensity.

This can be displayed as a contour line alongside points that reflect the joint estimates of observed covariates (which are calculated as γ in the first formula and β in the second). This contour line essentially acts as a threshold – any unobserved covariate whose impact falls on or above this line would have the specified effect on the treatment estimate (e.g., it would nullify the observed peer influence effect). Thus, the unobserved heterogeneity captured in Eqs. 1–4 represents the joint effects of alpha and delta that influence selection into the treatment and delinquency, respectively.

As noted above, alpha and delta are not single estimates, but rather a series of values that are conjointly estimated and related to one another, so that as the alpha (the unobserved effect on treatment selection) value increases, the complementary delta value necessary to nullify an effect decreases, and vice versa. For instance, in the example of peer influence, if some unobserved factor explains 90% of delinquent peer affiliations, this factor would only need to explain a small percentage of the variation in delinquency before we can reasonably conclude that the observed peer effect is likely null. Similarly, if there is an unobserved factor that can explain 95% of individual delinquent tendencies, this factor would need only to explain a small portion of delinquent peer selection before it nullifies the delinquent peer effect.

The fact that alpha and delta are estimated and reported as partial R-squared values is useful because it allows one to estimate the joint effects that each observable covariate has on both the theoretical treatment and delinquency, and then compare the estimated alphas and deltas of the unobservable to these values. Returning to the example of delinquent peer influence, for models accounting for theoretically-relevant controls, we can estimate the necessary impact of unobserved heterogeneity relative to estimates of factors such as self-control and social bonds. These factors have been consistently found to be strong predictors of both peer associations and delinquency (Thornberry et al. 1994) and have been put forth as primary contributors of the selection argument (Gottfredson and Hirschi 1990; Hirschi 1969). They thus serve as excellent benchmarks to evaluate the robustness of the delinquent peer effect. For LDV models, the partial R-squared of the unobserved heterogeneity can be compared to the prior delinquency covariate. Given that LDVs are considered by many to be "too strong" of a control, Imbens (2003) suggested that if the unobserved heterogeneity required to nullify an effect is greater than the partial R-squared estimated for the LDV, it provides strong evidence that a treatment effect is robust, which sets up a crucial test of their relative effects (see George and Bennett 2005). For the fixed effects models, we can estimate how large the unobserved heterogeneity must be to nullify an effect on changes in delinquent behavior relative to changes in the time-variant controls, while controlling out time-invariant heterogeneity.

Selecting a Threshold for the Sensitivity Analysis

The discussion up to this point has often used traditional statistical significance (i.e., $\alpha = .05$) as the threshold from which to evaluate the robustness of a treatment effect. Some scholars are critical of the overreliance on statistical significance (Cohen 1994; Maltz 1994) and the emphasis on *p*-values is at the root of many of the problems that the current special issue is attempting to address. For this reason, even though the analyses that follow primarily uses traditional statistical significance when assessing the robustness of theoretical predictors, we also offer supplemental analyses focused on effect size. Still, the fact remains that the discipline still often uses $\alpha \leq .05$ as the standard for evaluating the theoretical relevance of criminological constructs,[5] so we believe it is instructive to understand how sensitivity analysis proceeds with this as one of the thresholds.

Nevertheless, it is important to note that the sensitivity analysis demonstrated here does not require the threshold to be set at $\alpha = .05$. Researchers can set the threshold at more conservative levels of statistical significance (e.g., $\alpha = .01$, $\alpha = .001$). In fact, researchers need not conduct a sensitivity analysis with a focus on statistical non-significance. Scholars may be interested in the size of the unobserved heterogeneity necessary before a standardized treatment effect is considered be "weak" by some benchmark (i.e., Cohen's $d < .20$, or partial Pearson $r < .30$). When deciding on the appropriate threshold, scholars should consider the distinction between statistical and substantive importance, the size of the sample employed in the study, and what constitutes an "important" effect based on prior research assessing the specific outcome of interest. This might be generated by an unmeasured variable that has been shown to have an effect of a given size in the previous literature or an effect that would be clinically-relevant given the desire to move someone from an "indicated" diagnostic category to one that would fall in a normal range. Given the flexibility of a generalized sensitivity analysis, the appropriate threshold should be carefully considered—and justified—by the research situation.

Current Study

This study demonstrates the utility of sensitivity analysis for providing a quantitative assessment of how fragile an inference is to unobserved bias. We use delinquent peer influence and commitment to school as examples and estimate the unobserved effects that would be necessary to substantively alter the treatment effects generated by regressions with theoretically-relevant controls, lagged dependent variables, and individual fixed-effects. Then we contextualize the size of that unobserved bias by comparing it to model covariates.

Data and Methods

Our data come from the first Gang Resistance Education and Training Evaluation (G.R.E.A.T.; see Esbensen et al. 1996). The sample is diverse, comprising adolescents from six cities across different regions in the U.S. We selected this sample for several reasons. First, the longitudinal panel design (six waves) allows for proper specification of temporal order and is necessary for LDV and fixed-effects models. Second, the G.R.E.A.T. evaluation contains scales that capture controls important to assessing the robustness of a theoretical construct—self-control, parental attachment, and unstructured socializing. Finally, these data have been used by several scholars in the past to test various criminological theories (Anderson 2002; Carson 2013; Thomas 2015; Thomas and McGloin 2013). The G.R.E.A.T. evaluation began with a sample of around 3500 adolescents. To establish temporal ordering we restrict our analyses to individuals who have valid information on all of the constructs for at least two consecutive waves. The first analytic wave (i.e.,

[5] To be clear, we agree with the notion that the reliance on $p < .05$ to evaluate a theory is inappropriate, and encourage criminologists to read Berk et al. (2017) for more detailed discussion on this and related topics.

Table 1 Descriptive statistics of G.R.E.A.T. sample at wave 1

	Mean (%)	(SD)
Respondent hitting someone (binary)	32.7	(–)
Respondent marijuana use (binary)	8.1	(–)
Respondent violence (scale)	1.5	(2.8)
Respondent substance use (scale)	1.3	(3.0)
Peer hit someone (binary)	56.6	(–)
Peer marijuana use (binary)	20.2	(–)
Peer violence (scale)	4.4	(2.1)
Peer substance use (scale)	5.7	(2.9)
School commitment	4.0	(.7)
Self-control	2.9	(.7)
Maternal attachment	5.3	(1.2)
Unstructured socializing	4.0	(5.8)
Age	12.2	(.7)
Male	48.7	(–)
White	46.1	(–)

wave 2 delinquency regressed on wave 1 predictors) began with 1528 adolescents. This was reduced to 1052 adolescents by the final analytic wave (wave 6 delinquency regressed on wave 5 predictors). It is important to note that attrition over the panel was not random, as non-whites, marijuana users and those without a history of hitting are all more likely to drop out of the sample. Thus, one should take caution to not overstate the representativeness of these data in generalizing empirical inferences to the population. The descriptive statistics of the sample at wave 1 are reported in Table 1.[6]

Measures

Dependent Variable: Delinquent Behavior

At each wave of the G.R.E.A.T., respondents were asked how many times they engaged in various delinquent behaviors in the prior 6 months. For the current study, we look at crime-specific behaviors to determine the robustness of theoretical constructs across crime-types. Specifically, for each model we estimate individual delinquency using items that capture violence and substance use. For the models accounting for theoretically-relevant controls and LDVs we use individual items (one per crime-type): Hitting someone with the idea of hurting them and using marijuana. We chose these acts because they capture behaviors that are deviant but are relatively common among adolescents. The behaviors were initially recorded as open-ended frequencies, but are recoded into binary indicators in the current analyses, distinguishing those who engaged in the behaviors (= 1) from those that did not engage in the behaviors (= 0). The prevalence of each delinquent act follows a trend that is consistent with prior work. 33% of respondents reported hitting someone at wave 1, while

[6] Descriptive information for all other waves is available upon request.

25% report hitting someone at wave 6, and 8% of respondents report using marijuana at wave 1 while 28% report using it at wave 6.

In the models employing fixed-effects we utilize composite count scales, rather than single item dichotomies, that capture violence and general substance use. We utilize composite scales in the fixed effects models to increase variability in the outcome, and to demonstrate the utility of the sensitivity analysis approach across multiple dimensions of use. By prioritizing the illustration of sensitivity analysis under different coding strategies, we consequently limit the ability to directly compare coefficients and their robustness across different regression approaches. The violence scale is comprised of how many times individuals reported that they: hit someone, were involved in a gang fight, attacked someone with a weapon, shot someone, and robbed someone. The substance use scale reflects how many times individuals reported that they: smoked cigarettes, drank alcohol, used marijuana and used other drugs. All of these items were initially recorded as open frequencies, but each was top coded at “5” delinquent acts, which represented about the 95th percentile, in order to reduce skewness. Thus, at each wave the violence scale ranges from 0 to 25, and the substance use scale ranges from 0 to 20. In order to estimate a fixed-effects model, each scale is manually time-demeaned in Stata[7] by first estimating each individual’s mean over the entire panel and then subtracting their count crime rate at each time-point from that panel mean. This results in a normally-distributed outcome (see Allison 1990) that leads to estimates nearly identical to those from a formal fixed-effects command.

First Treatment Variable: Perceived Peer Delinquency

At each wave respondents are asked to report their perceptions of their friends’ delinquent behaviors with the question: “How many of your friends have engaged in________”. For the theoretically-relevant control and LDV models we use the behaviors of hitting someone and using marijuana. The perceived peer deviance measure was initially recorded as an ordinal scale ranging from none of them to all of them, but we recode it into a binary indicator, where a value of 0 indicates that the respondent has no friends who engage in the behavior, and a value of 1 indicates that the respondent has at least one friend who engages. The data indicate that 57% of respondents have at least one friend who hit another person at wave 1, while 53% had at least friend hit another person at wave 5, and 20% had at least one friend using marijuana at wave 1 while 52% did so at wave 5.

Coding peer delinquency as binary—although not the typical approach in peer influence studies—has been done in the past (Britt 1992; Nagin and Smith 1990) and prior research has shown that it leads to substantively the same conclusions as when retaining the ordinal scale (Thomas 2015). Indeed, the largest discrepancy across individuals is the no delinquent/any delinquent friends cut-point. We also code it in this way for practical and illustrative purposes. The Stata module employed here is currently designed for binary or continuous treatments, and does not yet accommodate ordinal treatments or outcomes. Further, many theoretical risk factors of interest are binary (e.g., employed/unemployed, married/unmarried, arrest/no arrest), so we believe it is informative to demonstrate the generalized sensitivity analysis here with a binary treatment. Thus, we use a binary perceived peer

[7] Manually time-demeaning variables when estimating a fixed-effects model leads to unbiased estimates of the regression coefficients, but incorrect standard errors due to the models relying on the wrong degrees of freedom. To account for this concern we used the appropriate Stata commands to adjust the standard errors in all of the models.

deviance treatment effect in two of our regression models and a continuous treatment effect in the fixed effects illustration.

In the fixed effects models we retain the ordinal coding of the individual items and create an average composite scale capturing perceived peer violence and substance use. As with the delinquency outcome items, we use composite scales for perceived peer deviance to increased variability on the measure and to demonstrate the utility of the sensitivity analysis approach across multiple dimensions of use. The perceived peer violence scale is made up of three items asking respondents how many of their friends: hit someone, attacked someone with a weapon, and robbed someone. The perceived peer substance use measures is made up of individuals' perceptions of how many of their friends used the same substances captured in the self-reported delinquency items (cigarettes, alcohol, marijuana, and other drugs). To estimate the fixed effects model, the violence and substance use average scales are all time-demeaned so that within-individual changes in perceived peer delinquency are used to predict changes in delinquency.

Second Treatment Variable: School Commitment

School commitment is captured at each wave using four items that ask respondents how much they agree with the following statements: "I try hard in school," "Education is so important that it's worth it to put up with things about school that I don't like," "In general, I like school," and "I usually finish my homework." Response options were on a Likert scale ranging from 1 = strongly disagree to 5 = strongly agree, and all items were coded so that higher values reflect greater commitment to school. The four items were averaged at each wave to create a *school commitment* scale (Wave 1: $\bar{x} = 4.0$, $SD = .688$; Wave 5: $\bar{x} = 3.9$, $SD = .676$). In the fixed-effect model, the school commitment measure is time-demeaned so that changes in school commitment predict changes in self-reported delinquency.

Covariates

Self-Control The G.R.E.A.T. evaluation contains eight items that are a subset of the Grasmick et al. (1993) self-control scale. Respondents were asked how much they agree with the following statements: "I often act on the spur of the moment," "I do what brings me pleasure here and now," "I don't devote a lot of thought and effort preparing for the future," "I am more concerned with what happens to me in the short run than in the long run," "I like to test myself every now and then by doing something a little risky," "Sometimes I will take a risk just for the fun of it," "I sometimes find it exciting to do things for which I might get in trouble," and "excitement and adventure are more important to me than security." Response options ranged from 1 = strongly disagree to 5 = strongly agree, with higher values indicating lower levels of self-control. We calculated a mean *self-control* score for each individual at each wave (Wave 1: $\bar{x} = 2.9$, $SD = .7$; Wave 5: $\bar{x} = 2.7$. $SD = .7$). In the fixed-effect model, self-control is manually time-demeaned so that changes in self-control predict changes in delinquency.

Maternal Attachment Individuals were told to think about their mother or mother-figure and asked to rate the following statements on a seven-point scale: "can talk about anything," "always trusts me," "knows all my friends," "always understands me," "always ask her for advice," and "always praises me when I do well." Responses ranged from 1 to 7, with higher values indicating higher levels of attachment. Using these six items we calculated a mean *maternal attachment* score for each individual at each wave (Wave 1: $\bar{x}=5.3$, $SD=1.2$; Wave 5: $\bar{x}=4.9$, $SD=1.3$). In the fixed-effect model, maternal attachment is manually time-demeaned so that changes in maternal attachment predict changes in delinquency.

Unstructured Socializing Osgood et al. (1996) argued that delinquency can occur simply when hanging out with friends in unstructured settings. Respondents are asked: "Do you ever spend time hanging around your current friends not doing anything in particular where no adults are present?" and "if yes, how many hours do you spend doing this in an average week?". Responses to the second part of this question are recorded as open-ended frequencies, but we top-code the responses at 20 h (Wave 1: $\bar{x}=4.0$, $SD=5.8$; Wave 5: $\bar{x}=6.1$, $SD=6.5$). In the fixed-effect model, unstructured socializing is manually time-demeaned so that changes in time spent with peers in unstructured settings predict changes in self-reported delinquency.

Demographics *Age* is a continuous measure reflecting respondents' age in years. Respondents were, on average, 12 years of age at wave 1 and 17 years of age at wave 6. *Male* is a dummy indicator reflecting whether the respondent is male (=1) or female (=0). *White* is a dummy indicator reflecting if the respondent is white (=1) or non-white (=0). At wave 1, the sample is, on average, 12 years of age (SD=.65), 49% male, and 46% white.[8]

Analytic Plan

We first estimate regressions using delinquent peers and school commitment to predict delinquency in the models accounting for theoretically-relevant controls (logit), LDVs (logit) and fixed effects (OLS). This offers an initial assessment of the effect size, statistical significance and partial R-squared of the delinquent peer and school commitment effects and other confounders for violence and substance use (e.g., marijuana).[9] In the second stage we predict delinquent peer associations and school commitment using the same covariates for each of the regression strategies and store the effect sizes and partial R-squared estimates. For delinquent friends, we use logit models to predict whether individuals have at least one friend who engages in the delinquent behaviors in the theoretically-relevant controls and LDV models and an OLS model to predict within-individual changes in one's perceived delinquent peer group in the fixed-effects models. Given that school commitment is approximately normally distributed at each wave, we use OLS to predict school commitment in all of the models. Using the results stored in the two stages of the regression analyses we conduct sensitivity analyses that calculate the joint delta and alpha estimates at which the delinquent peer and school commitment estimates would no longer be statistically significant at a .05, two-tailed level (i.e., when the ratio of the estimate to its standard

[8] No missing data were imputed in this study. At the time of this study, the sensitivity analysis employed is not equipped to estimate multiple imputed data sets and, therefore, cases with missing information are listwise deleted.

[9] When estimating the TRC and LDV models, we pool the analytic waves together so there are multiple observations for each respondent.

Table 2 Regression models predicting delinquency in the G.R.E.A.T. data

	Theoretically relevant controls (logit)		LDV (logit)		Fixed-effects (OLS)	
	Hit someone	Marijuana use	Hit someone	Marijuana use	Violence	Substance use
Peer behavior	1.265***	1.681***	.712***	1.187***	.711***	.636***
	(.072)	(.089)	(.075)	(.096)	(.042)	(.022)
School commitment	−.241***	−.289***	−.198***	−.229***	−.187***	−.563***
	(.057)	(.067)	(.055)	(.067)	(.061)	(.075)
Low self-control	.265***	.593***	.189***	.519***	.412***	.275***
	(.058)	(.071)	(.055)	(.053)	(.081)	(.077)
Maternal attachment	−.088**	−.086**	−.076**	−.085***	−.124**	−.104*
	(.029)	(.034)	(.028)	(.032)	(.046)	(.046)
Unstructured socializing	.021***	.044***	.010†	.035***	.024***	.073***
	(.005)	(.006)	(.005)	(.006)	(.007)	(.008)
Age	−.226***	.155***	−.227***	.114***	–	–
	(.025)	(.032)	(.024)	(.033)	(–)	(–)
Male	.342***	−.089	.306***	−.067	–	–
	(.073)	(.090)	(.065)	(.081)	(–)	(–)
White	.018	−.080	.008	−.007	–	–
	(.073)	(.089)	(.064)	(.080)	(–)	(–)
Prior delinquency	–	–	1.377***	1.476***	–	–
	(–)	(–)	(.074)	(.111)	(–)	(–)
Total R^2	.127	.236	.180	.274		
Log likelihood	−3329.093	−.2195.055	−3127.753	−.2087.755		

error (t) $<$ 1.96). We also examine how large the unobserved heterogeneity would need to be before an effect becomes "weak" (i.e., when $d < .20$). These results are then displayed as contour lines which allow for a comparison of how large the unobserved heterogeneity would need to be relative to known confounders in the two stages of the regression models. Our discussion will primarily use respondent self-control and respondent prior delinquency as a benchmark from which to evaluate the robustness of an observed effect. We use these predictors as the benchmarks because: (1) theory and prior research has indicated that self-control confounds the effects on delinquency for both peer influence and school commitment (Gottfredson and Hirschi 1990); (2) as demonstrated below, self-control and prior delinquency are consistently the predictors that are closest to the estimated contour lines and; (3) Imbens (2003) has explicitly stated that contour line surpassing the prior delinquency term provides evidence of a robust effect. We estimate these models using Stata's generalized sensitivity analysis ("gsa") command (Harada 2012).[10] We estimate robust

[10] The full Stata command to estimate how large unobserved bias would need to be before the treatment is $p < .05$ is:
gsa y w x, tstat(1.96) vce(cluster id) nplots 7
where y is the outcome, w is the treatment, x is the vector of controls, tstat(1.96) specifies that the sensitivity test is for a two-tailed test with t < 1.96, vce accounting for the interdependence of observations and nplots specifying how many variables are to be presented on the contour plots. See the Stata do file in the "Appendix" for further information.

Table 3 Regression models predicting delinquent peer associations in the G.R.E.A.T. data

	Theoretically relevant controls (logit)		LDV (logit)		Fixed-effects (OLS)	
	Hit someone	Marijuana use	Hit someone	Marijuana use	Violence	Substance use
School commitment	−.402***	−.516***	−.333***	−.417***	−.179***	−.998***
	(.056)	(.062)	(.059)	(.070)	(.043)	(.067)
Low self-control	.417***	.583***	.270***	.419***	.482***	.372***
	(.053)	(.063)	(.057)	(.066)	(.048)	(.077)
Maternal attachment	−.181***	−.192***	−.144***	−.185***	−.164***	−.370***
	(.029)	(.032)	(.031)	(.034)	(.028)	(.045)
Unstructured socializing	.071***	.069***	.050***	.054***	.026***	.122***
	(.007)	(.006)	(.007)	(.006)	(.005)	(.008)
Age	−.085**	.476***	−.056*	.408***	–	–
	(.024)	(.028)	(.026)	(.029)	(–)	(–)
Male	.712***	−.335***	.604***	−.308***	–	–
	(.074)	(.085)	(.075)	(.092)	(–)	(–)
White	−.268***	−.342***	−.266***	−.225*	–	–
	(.073)	(.083)	(.075)	(.088)	(–)	(–)
Prior delinquency	–	–	2.383***	3.510***	–	–
	(–)	(–)	(.086)	(.171)	(–)	(–)
Total R^2	.125	.192	.259	.310		
Log-likelihood	−3553.167	−3037.992	−3008.339	−2593.040		

standard errors to account for the interdependence in the observations using the "vce" command in Stata, which is compatible with the "gsa" command.

Results

Table 2 provides the results for all models predicting delinquency. When accounting for theoretically-relevant control variables, along with demographic factors, both delinquent peer associations (violence: b = 1.265, *p* < .001; marijuana use: b = 1.681, *p* < .001) and school commitment (violence: b = −.241, *p* < .001; marijuana use: b = −.289, *p* < .001) are significantly related to delinquency. Individuals are 3.54 times more likely to report hitting someone[11] and 5.37 times more likely to use marijuana when they believe that at least one of their friends has also engaged in that same behavior (Models 1 and 2, respectively). Further, a one unit increase in school commitment is associated with a 21% decrease in the likelihood of hitting someone and a 25% decrease in the likelihood of using marijuana.

Accounting for a LDV reduces both the peer deviance (violence: b = .712, *p* < .001; marijuana use: b = 1.187, *p* < .001) and school commitment effects (violence: b = −.198, p < .001; marijuana use: −.229, *p* = .001). In the LDV models, individuals are twice as

[11] All odds ratios calculated based on by exponentiating the logit value (e^B).

Table 4 Regression models predicting school commitment in the G.R.E.A.T. data

	Theoretically relevant controls (OLS)		LDV (OLS)		Fixed-effects (OLS)	
	Hit someone	Marijuana use	Hit someone	Marijuana use	Violence	Substance use
Peer behavior	−.126***	−.178***	−.094***	−.132***	−.024***	−.045***
	(.017)	(.022)	(.018)	(.024)	(.006)	(.004)
Low self-control	−.313***	−.301***	−.308***	−.294***	−.123***	−.110***
	(.014)	(.014)	(.014)	(.014)	(.016)	(.015)
Maternal attachment	.128***	.123***	.126***	.122***	.126***	.105***
	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)
Unstructured socializing	−.003	−.002	−.002	−.002	−.016***	−.010***
	(.002)	(.002)	(.002)	(.002)	(.001)	(.001)
Age	−.040***	−.020**	−.002	−.017*	–	–
	(.007)	(.020)	(.009)	(.007)	(–)	(–)
Male	−.101***	−.140***	−.097***	−.141***	–	–
	(.021)	(.021)	(.021)	(.021)	(–)	(–)
White	−.025	−.023	−.024	−.027	–	–
	(.021)	(.027)	(.021)	(.020)	(–)	(–)
Prior delinquency	–	–	−.079**	−.146	–	–
	(–)	(–)	(.020)	(.021)	(–)	(–)
Total R^2	.289	.289	.291	.293		

likely to report using violence (OR 2.04) and 3.28 times more likely to use marijuana when they report having at least one friend who engages in that behavior, and a one unit increase in school commitment is associated with a 18% decrease in the likelihood of hitting someone and a 20% decrease in the likelihood of using marijuana. Finally, the results of the fixed-effects regressions show that a unit change in perceived peer violence (on an ordinal scale) is associated with a .71 change in one's own count of violent acts ($p < .001$) and a unit change in perceived peer substance use is associated with a .64 change in one's own substance use ($p < .001$). A one unit change in school commitment is associated with a .19 decrease in one's count violent acts and a .56 decrease in one's own substance use.

The control variables are related to all of the outcomes in expected ways. In all of the models in Table 2, self-control, maternal attachment, and unstructured socializing are statistically significant predictors of delinquency. Finally, as expected, prior delinquency is one of the strongest predictors of future delinquency in the LDV models.

We now shift attention to the models that predict treatment assignment (i.e., delinquent peer associations and school commitment). The results in Table 3 indicate that selection into delinquent peer groups is not random. In both the TRC and the LDV models, a one unit increase in school commitment reduces the likelihood of associating with delinquent friends by about 35%, and in the fixed-effects model a within-individual one unit change in school commitment is associated with a .18 to .99 unit change in perceived peer delinquency on the average ordinal scale. Further, in the logistic regressions a one unit increase in low self-control increases one's likelihood of reporting that one's

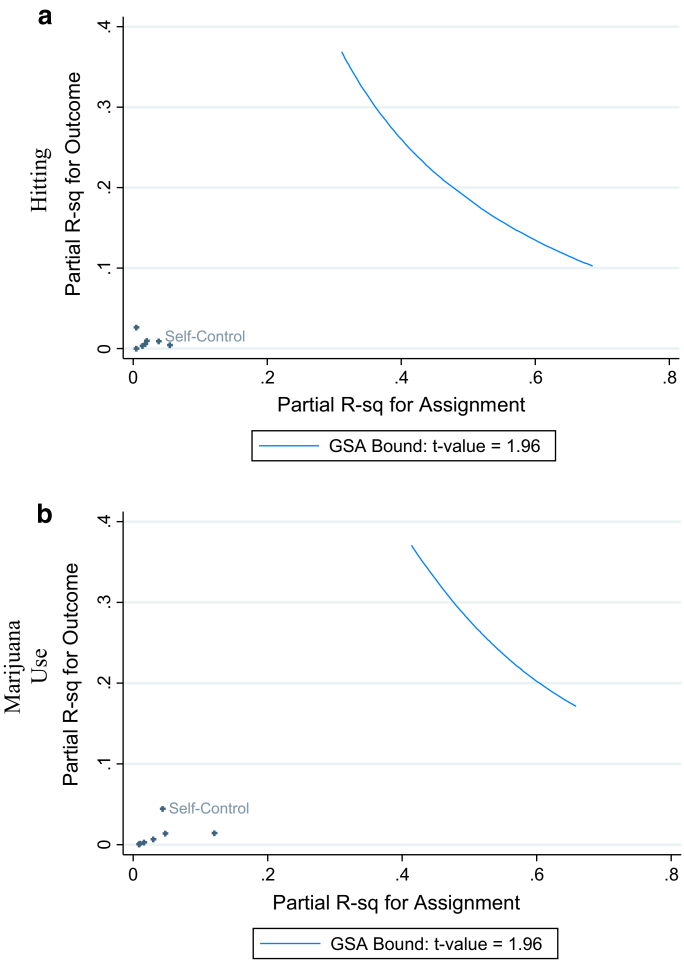

Fig. 1 Contour plots examining robustness of peer effects while controlling for theoretically-relevant controls

friends are delinquent by between 31 and 80%; in the fixed-effects models a within-individual one unit change in self-control is associated with around a .27 to .48 unit change in perceived peer deviance. Finally, the LDV models indicate individuals who are delinquent in the contemporaneous time period are about 10 times more likely to report that at least one friend is involved in violence and 30 times more likely to associate with friends who used marijuana.

We similarly see that commitment to school is not randomly distributed (Table 4). A one unit increase in low self-control is associated with about a .31 unit decrease in school commitment, and a one unit increase in maternal attachment is associated with a .12 unit increase in school commitment. The LDV models indicate that prior hitting is associated with a .08 unit decrease in school commitment, and prior marijuana use is associated with a .15 unit decrease in school commitment. In the fixed-effects models, within-person changes in delinquent peer associations, self-control, and maternal

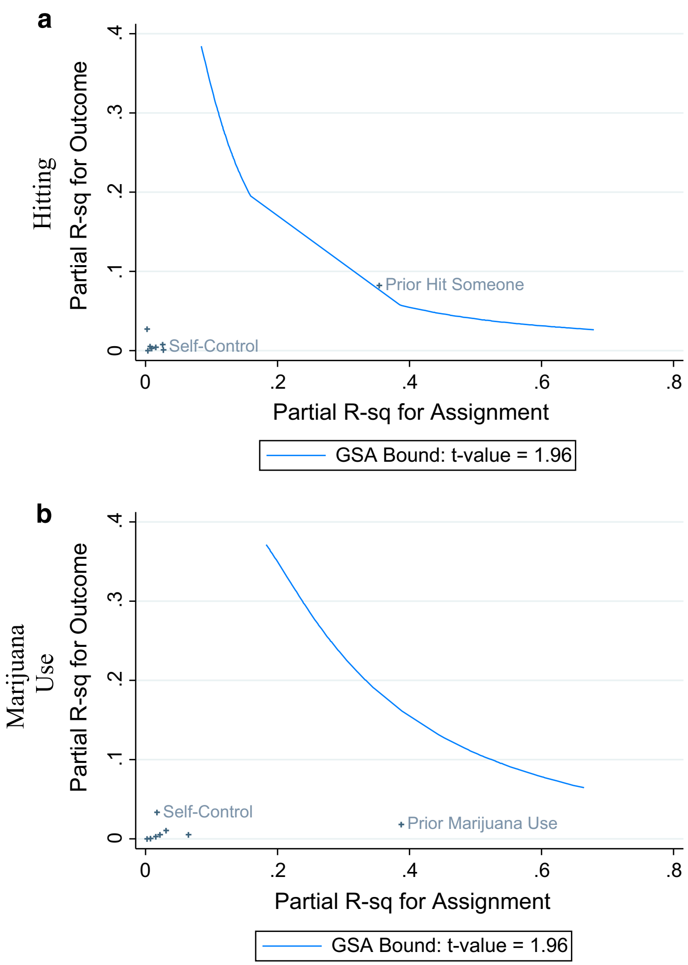

Fig. 2 Contour plots examining robustness of peer effects while controlling for LDVs

attachment are all related to within-person changes in school commitment in the directions that one would anticipate.

Taken together, several factors simultaneously predict both delinquency and delinquent peer associations/school commitment and both associating with delinquent peers and school commitment are strong and statistically significant predictor of delinquency when accounting for these observed (and unobserved time-invariant) confounders. However, the question of whether some further unobserved bias would render the effects null remains open. The next step is to determine how much this hypothetical unobserved factor(s) would have to impact both the treatment and outcome in order to strip the effects of statistical significance.

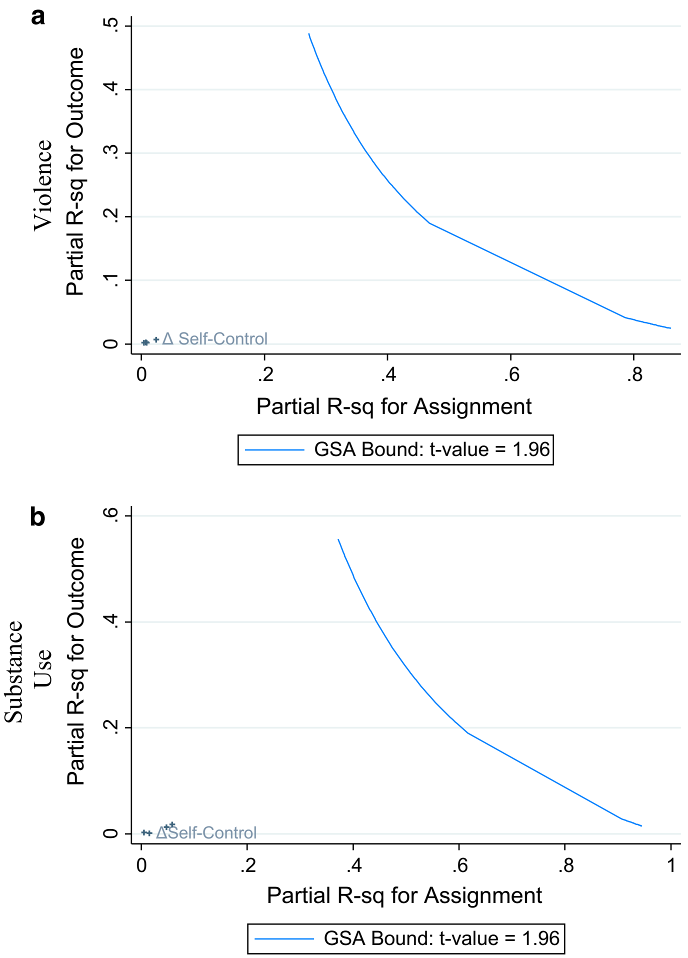

Fig. 3 Contour plots examining robustness of peer effects using fixed-effects

Robustness of Delinquent Peer Effect

Figures 1, 2 and 3 present the contour plots resulting from the sensitivity analyses examining the robustness of the delinquent peer effect. The first set shows the two contour plots for the regressions with TRCs, the second set shows the plots for the LDV models, and the third set shows the plots for the fixed-effects models. In both Fig. 1 the contour line is far past the joint estimates for all of the TRCs. For example, the second contour plot focuses on the effect of perceived peer marijuana use on respondent marijuana use. The confounder with the largest partial R-squared is low self-control, which explains 5% of the variation in perceived peer deviance and approximately 8% of the variation in respondent delinquency. In this model, some unobserved confounder(s) explaining 8% of the variation in delinquency must jointly explain over 50% of the variation in delinquent peer associations—ten times larger than the effect that self-control has on delinquent peer associations—in order to render the peer effect non-significant. Conversely, unobserved heterogeneity that explains 5% of the variation in delinquent peer associations would need to explain around 60% of the variation in respondent delinquency for the peer effect to be rendered null. The results of these contour plots suggest that the size of the unobserved bias necessary to strip the peer effect of its statistical significance for both violence and marijuana use must be dramatically larger than the combined joint effects of all of the observed predictors on peer associations and delinquency.

Turning to the second set of contour plots, Imbens (2003) suggested that if the unobserved heterogeneity required to nullify an effect is greater than the partial R-squared estimated for the LDV, it provides strong evidence that the treatment effect is robust. The results suggest that, for both behaviors, any unobserved heterogeneity affecting both perceived peer delinquency and one's own delinquency must be at least as large as, if not larger than, the respondents' prior delinquency in order to shift the peer effect to statistical non-significance. For instance, prior marijuana use explains around 40% of the variation in perceived peer marijuana use (measured at the same time point) and around 4% of the variation in later marijuana use. Unobserved heterogeneity explaining 10% of the variation in respondent marijuana use—more than twice the variation explained by prior use—would need to jointly explain 50% of the variation in peer marijuana use—over 10% more than explained by prior use—in order to render the peer effect null. For violence, the prior delinquency effect falls almost directly on the contour line. The findings here suggest that the estimate and resultant inference is robust in these data. Given that LDVs are lagged measures of the behavioral outcome, it seems unlikely that there are unobserved factor(s) rivaling their strength that would affect both delinquency and delinquent peer associations.

The final contour plots for the peer effects show the sensitivity analyses for the fixed-effects models. The contour lines in these models can be compared to the time-demeaned theoretically-relevant controls—e.g., changes in social bonds, self-control, and unstructured socializing. The figures show that the delinquent peer effect is robust for both behaviors. None of the confounders come close to crossing the contour line. Most of them explain just a small portion of the joint effects of delinquency and delinquent peer associations. The total R-squared for the time-variant factors for delinquency (including the delinquent peer effect) is .19 for violence and .34 for substance use, while the total R-squared of the time-variant predictors for changes in delinquent peer associations are .06 for violence and .11 for substance use. For the delinquent peer effect to be rendered null for both behaviors, the time-variant heterogeneity would need to be substantially larger than the combined effects of self-control, maternal attachment, school commitment and unstructured socializing—some of the most well-known predictors of delinquency. For example, although the total R-squared in the model for violence suggests the time-variant predictors explain 22% of the variance in changes in delinquency and 6% of the variance in delinquent peer associations, unobserved factor(s) would need to jointly account for over 20% of the variance of changes in delinquency and 40% of changes in peer associations before the delinquent peer effect would be non-significant. This unobserved bias would need to be even larger for substance use. Again, because this sensitivity analysis is employed on fixed-effects models, this unobserved factor(s) must be a *time-varying* influence and not some stable trait or characteristics of individuals.

Robustness of School Commitment Effect

We next examine the robustness of the school commitment effect. The generalized sensitivity analysis requires that the treatment effect be positively related to the outcome;

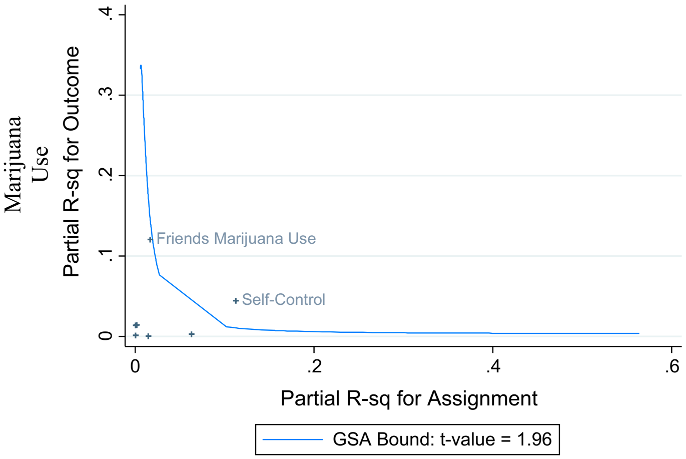

Fig. 4 Contour plots examining robustness of school commitment controlling for theoretically-relevant controls

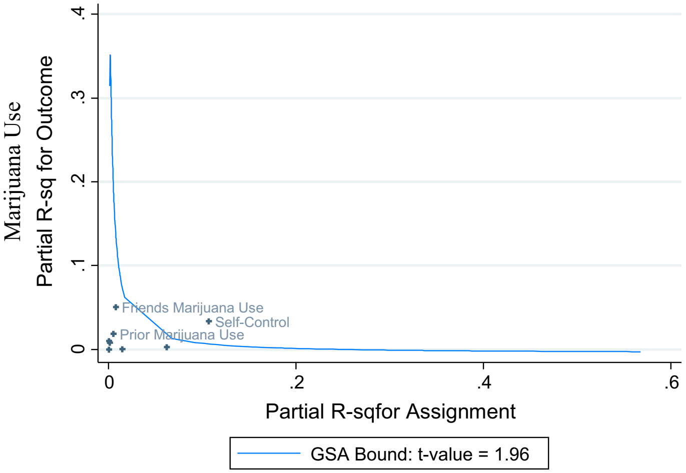

Fig. 5 Contour plots examining robustness of school commitment on controlling for LDVs

Accordingly, we reverse coded the school commitment scale so that higher values reflect *lower* levels of school commitment. We estimated the robustness of the school commitment effect for both violence and substance use, but for space purposes present only the substance use contour plots in text (the contour plots for hitting/violence are presented in the Appendix). The first set of plots represents the models that include the TRCs. For the models predicting hitting someone with the idea of hurting them (presented in Fig. 8 of the Appendix), two of the confounders fall very close to the contour line: Associating with violent friends and low self-control. Low self-control jointly explains about 1.5% of the variation in hitting and 15% of the variation in school commitment; an unobserved factor(s) that also explained 15% of the variation in school commitment would need to explain only about 2% of the variation in delinquency to render school commitment nonsignificant at $p < .05$. For marijuana use, presented in Fig. 4, self-control and delinquent

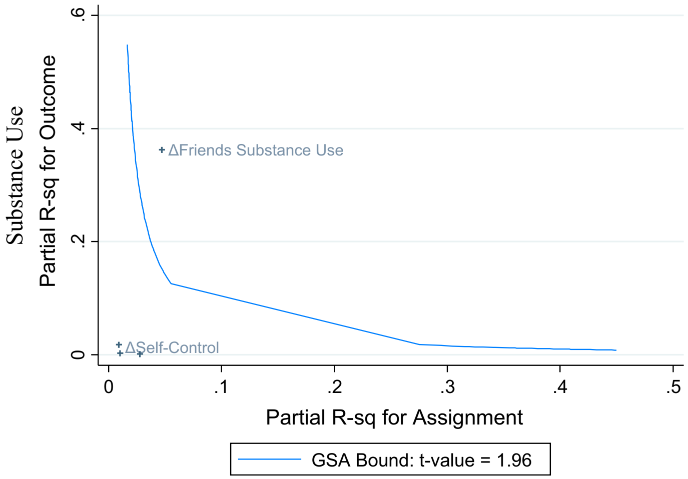

Fig. 6 Contour plots examining the robustness of school commitment using fixed-effects

peer associations fall on or cross the contour line. Self-control explains about 12% of the variation in school commitment and 5% of the variation in marijuana use. To render school commitment statistically non-significant, an unobserved factor explaining 5% of the variation in delinquency would need to explain around 8% of the variation in school commitment. Thus, although school commitment is a statistically significant predictor of hitting and marijuana use in the TRC models, this statistical inference may be susceptible to unobserved bias.

The contour plots assessing the robustness tests of school commitment on hitting and marijuana use in the LDV models are presented in the Appendix (Fig. 9) and Fig. 5, respectively. For hitting someone, four of the confounders—self-control, prior hitting, violent peers, and and maternal attachment—fall on or approach the contour line. In fact, much of the contour line for school commitment falls around a partial R-squared estimate of 0 for both the treatment and the outcome, which suggests that unobserved bias that jointly affects school commitment and hitting even only a little bit (i.e., 1–2% would be enough to render the school commitment effect on hitting statistically non-significant at $p < .05$). For marijuana use, only self-control surpasses the contour line. Self-control explains around 12% of the variation in school commitment and 3% of the variation in marijuana use. An unobserved factor(s) accounting for 12% of the variation in school commitment would need to only explain less than 1% of the variation in marijuana use to render school commitment statistically non-significant at $p < .05$. As with the TRC models, despite school commitment being statistically significant in the LDV models, the sensitivity analysis suggests that these inferences may be quite fragile depending on the nature of the unobserved bias.

The sensitivity analyses examining the robustness of school commitment on violence and substance use for the fixed-effects models are presented in the Appendix (Fig. 10) and Fig. 6, respectively. For both violence and substance use, only within-person changes in peer delinquency surpasses the contour line, while within-individual changes in self-control fall directly on the contour line for the violence model. Nevertheless, and similar to the TRC and LDV models, it is noteworthy that the contour line is close to 0 for both the treatment and outcome measures. A time-varying factor(s) that explains only 5% of the within-person changes in school commitment would need to only explain around 1% of the within-person changes in violence and around 13% of the within-person changes in substance use. Overall, the sensitivity analyses for the school commitment estimates are indicative of an inference that may be susceptible to omitted variable bias. Despite the fact that estimates are statistically significant in all models, we may be more apt to make a type I error in reaching substantive conclusions in the school commitment-delinquency relationship on the basis of results of hypothesis tests. In comparison, the conclusions from the null hypothesis significance tests on the delinquent peer estimates seem to be more robust to hidden biases.

Beyond *P* Values: An Alternative Threshold

The generalized sensitivity analysis approach does not require an exclusive focus on statistical significance. This is important as too much emphasis is sometimes placed on results of inferential tests at the expense of understanding the nature of the relationship(s) of interest (Abelson 1995; Berk et al. 2017). The approach also allows researchers to assess how large unobserved bias must be before the estimated effect size is reduced to some specified value, but it requires a slight modification of the procedure so that the t statistic, which captures the test value, is replaced by τ, which reflect the size of the effect. To demonstrate the application of sensitivity analysis on effect sizes we use just the delinquent peer influence treatment in the TRC and LDV models as an example. Further, we use a well-known standardized effect size measure—Cohen's *d*—as our threshold cut-point. Log odds ratios can be converted into standardized *d* values using the equation:

$$LogOddsRatio = d\frac{\pi}{\sqrt{3}}$$

Using the criteria offered by Cohen (1992) that a "weak" effect is a $d < .20$; we assess how large the unobserved heterogeneity would need to be before the binary delinquent peer treatment falls below the .20 threshold. Using the conversion equation presented above, we calculate that a Cohen's *d* of .20 corresponds to a log odds ratio of .363, and conduct a sensitivity analysis examining how large unobserved bias would need to be before the delinquent peer coefficient falls below that value.[12]

We present the contour plots examining the robustness of the delinquent peer effect size to potential unobserved bias for the TRC and LDV models in the Appendix and Fig. 7, respectively. The results from the TRC models (presented in Fig. 11 in Appendix) shows that the moderate to large effect size of delinquent peer influences is robust to unobserved bias. None of the confounders included in the models come close to the contour line, which suggests that the unobserved heterogeneity necessary to render the delinquent peer effect "weak" would need to be quite large. Figure 7 examines the robustness of the delinquent peer effect size in the LDV models. For the model predicting hitting someone with the idea of hurting them, prior hitting behavior crosses the contour line, indicating that potential unobserved bias that jointly effects violent peer

[12] The Stata command for estimating the tau specification of gsa is:
gsa y w x, tau(.363) vce(cluster id) nplots 7
The command is identical to that presented in footnote 5 but now specifies "tau" instead of "tstat".

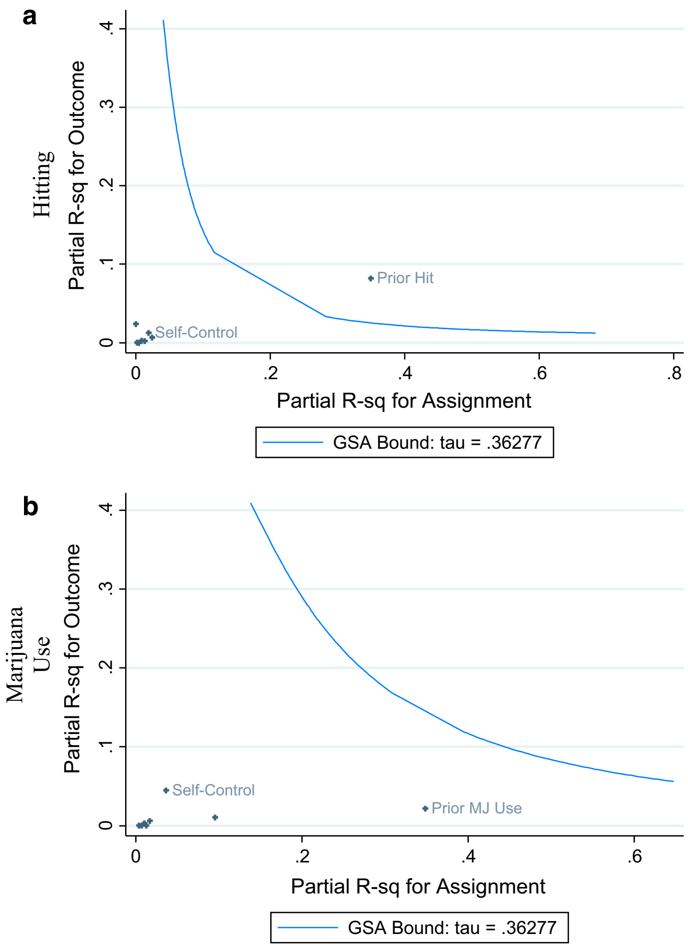

Fig. 7 Contour plots examining robustness of peer effects while controlling for lagged dependent variables at $d < .20$

associations and respondent violence to the same degree as prior violence would render the peer effect "weak" based on $d < .20$. Recall that the sensitivity analysis above found that delinquent peer influence on hitting was quite robust in terms of statistical significance. In contrast, it is reasonable to suspect that possible unobserved biases exist that could diminish the observed effect to the point that it would be considered substantively weak. For marijuana use, however, none of the confounders (including prior marijuana use) approach the contour line. Based on this finding, one could reasonably conclude that the moderate to large effect between perceived marijuana use among friends and respondents' marijuana use is robust to unobserved heterogeneity. It is worth noting that the discrepancy between hitting and marijuana use is consistent with some prior work, which often finds a large and statistically significant peer effect for substance use and a statistically significant, albeit substantively weaker, peer effect on violence (Pratt et al. 2010).

Discussion

Questions of statistical inference are at the heart of the purported "replication crisis" in the social and behavioral sciences (Wasserstein and Lazar 2016), which blend concerns about internal validity and statistical conclusion validity with those of external validity while also drawing in extant debates concerning competing theoretical perspectives. Whereas some scholars view statistically significant relationships as the empirical foundation for theory, others question whether this foundation is built on fragile inferences and omitted variable bias. The empirical questions that correspond with these arguments typically pertain to whether a focal relationship holds after adjusting for potential confounders. In the social sciences this manifests practically in questions about the size and significance of the coefficient(s) in some type of regression model and the sensitivity of the relevant statistical inferences to unobserved heterogeneity.

This process also invokes some competing methods for appropriately controlling for endogeneity in assessing that question. This study considered three of those: The use of theoretically-relevant controls, using a lagged dependent variable (LDV), and fixed effects modeling. Using these regression-based approaches as a base, we demonstrate how researchers might assess the robustness of key estimates from theoretically-informed models. Frequently, reservations about empirical generalizations are less about the nature of accumulated results, and more about the fact that the observational nature of data sets in criminology makes it challenging to statistically identify *the* point estimate for causal effects. This presents obstacles to credible inference that must contextualize our empirical generalizations and their implications for theory (e.g., Manski 2008; Rubin 2008; Weisburd and Piquero 2008).

The growing consensus in disciplines like statistics and econometrics that have wrestled with these issues in greater depth is that the effort expended toward identifying "true" social influence effects is a futile endeavor: With observational data, the potential for unobserved confounders is limitless and scholars will never be satisfied that an observed effect is truly indicative of causality (Manski 1993). From this view, important criminological questions, such as whether peers "cause" delinquency, will never be answered with absolute certainty and further attempts to identify clean point estimates are unlikely to appreciably advance the field.

This does not mean that attempts to empirically validate criminological theories are fruitless. After all, "identification is not an all-or-nothing concept and … models that do not point identify parameters of interest can, and typically do, contain valuable information about these parameters" (Tamer 2010: 168). Instead, given imperfect measures and uncertain estimates, we suggest that considerable insight can be gained if scholars assess the robustness of parameter estimates in systematic ways, are clear about the assumptions that go along with each point in the range of these bounds, and check the sensitivity of estimates to possible alternative influences with increasing levels of control and fewer assumptions about exogeneity (Manski 2008).

This is the process carried out in the current study using a generalized sensitivity analysis approach. Importantly, the method we employed provides useful comparators because we could benchmark the effect necessary to do away with a focal relationship to commonly-used explanatory factors that have fairly intuitive relationships to the core variables involved. Although criminologists have encountered difficulties in explaining crime and other deviant behaviors, as is evidenced by the relatively low explanatory power of predictive models (Weisburd and Piquero 2008), there are nevertheless a host of extensively studied and robust risk factors that can act as informative benchmarks when assessing relationships among criminological constructs (e.g., age, gender, self-control, prior delinquency). Substantively, the results point to a consistent peer effect of at least modest size that would be rendered spurious only with the introduction of unreasonably large unobserved factors lurking above and beyond well-known confounders such as low self-control, weakened social bonds and prior delinquency. In statistical and replication terms, the results do not provide proof that peers *cause* delinquency (i.e., clean point identification), but do provide evidence that the delinquent peer effect is robust in an observational data set with different statistical models that rely on increasing degrees of control. This, in turn, improves our confidence in the statistical inferences that can be made given such data.

Of course, it is important to remember that this study is focused on statistical inference, not theoretical inference. As we stated at the start of this paper, scholars are often faced with the skeptical question of "do peers (really) matter?", which sensitivity analysis can help to address; but we must also wrestle with the important work of assessing how, when and why peers matter (Hedström 2005; Thomas and McGloin 2013; Wikström and Sampson 2003). Even though we document a robust peer effect here, researchers may still wonder whether our use of perceived peer deviance opens up the possibility of projection (Jussim and Osgood 1989; Young 2011; c.f., McGloin and Thomas 2016a, b), and whether a measure of peer deviance captures mechanisms such as opportunity structures, reinforcement contingencies, attitudinal transference or some combination thereof. The point here is that sensitivity analysis provides particular gains on one dimension of inference, but it should be coupled with a sincere effort to better understand process in order to translate into more noticeable gains for the discipline as a whole.

In contrast to the peer effect, the effect of school commitment on delinquency was more fragile. The initial estimate was statistically significant and not trivial in size. Even so, it was only minimally robust to unobserved covariates in each of the models tested. This analysis demonstrated the usefulness of the generalized sensitivity analysis with a variable that has a different track record in the previous literature and which can therefore illustrate its results in different circumstances, as well as confirm that it provides results that correspond to substantive knowledge of different influences on delinquent behavior.

With these findings in mind, as well as the inherent limitations of observational data that make identifying point estimates challenging, we encourage scholars to move beyond simply assessing whether some theoretical construct is related to crime and delinquency at $p < .05$, or specified effect size. Doing this acknowledges that the coefficients estimated in regression models are imprecise approximations of the true parameters, and that even minor sources of bias may dramatically impact the inferences one can make when testing hypotheses. This also, in part, acknowledges the assumption laden process of estimating regression models (Berk et al. 2017).

In light of these limitations, criminologists should routinely estimate both the relationship between a construct and delinquency, as well as the fragility of the inferences to potential unobserved bias. The generalized sensitivity analysis is useful because it can be employed with regression approaches commonly used in criminology. Nevertheless, there are other types of sensitivity analyses and bounding approaches that criminologists may find useful in general and specialized circumstances pertinent to theoretical and practical questions. Rosenbaum (2002) describes the logic and practice of sensitivity analysis in the counterfactual framework, which can be used in conjunction with propensity score matching approaches employed in estimating treatment effects (Gangl 2004). VanderWeele (2011) has developed a sensitivity analysis to examine influence effects using social network data (see also Young 2014) and Ding and VanderWeele (2016) present a general approach to conducting sensitivity analyses in a variety of contexts, including mediation relationships.

Manski (1990) has encouraged scholars to use a nonparametric approach to estimate bounds on treatment effects, which Manski and Nagin (1998) employed to compare the predictions made by labeling and deterrence theories. Although these approaches differ in important ways, they all share the underlying premise that treating parameter estimates as "true values" can be misleading, and that there is utility in examining the sensitivity of inferences to unobserved bias. This approach is buttressed by its integration of the assumptions that are tied to credibility in statistical inference (Manski 2008).

The growing concerns in the social sciences of false positive inferences ultimately speak to anxiety about the robustness of parameter estimates to unobserved heterogeneity—e.g., why is it that we fail to reject the null in some studies but not others? Although we believe that sensitivity analysis offers one approach that can improve our confidence in and understanding of statistical inferences moving forward, we recognize that this approach is not a "cure all" for the replication problem in the social sciences. Sensitivity analysis cannot address questionable research practices such as p-hacking, unethical research practices (Brockman et al. 2015), nor publication biases against studies with null findings. These issues reflect concerns that extend well-beyond model misspecification and omitted variable bias, but they should be of interest to scholars concerned about the presence of false positives in criminological research. The extent to which such factors affect estimates in our discipline is unknown, especially in light of the fact that greater scholarly attention has been focused on the extent to which alternative, unobserved explanations may threaten causal inferences. A more optimistic and parsimonious working explanation suggests that differing conclusions of the relationships between theoretical constructs and delinquency stem primarily from fragile inferences and honest, yet misspecified, regression models. To the extent that this is the case, sensitivity analysis and bounding offer a useful technique for criminologists moving forward.

Although we applied a generalized sensitivity analysis to two sets of relationships relevant to potential confounding explanations, the benefits of sensitivity analyses can be extended to a host of criminological research questions. Given the properties of Imbens's approach—and the fact that much criminological research is based on observational data—this method would certainly be useful in answering questions outside of the scenarios studied here. For example, other theoretical constructs thought to predict delinquency have also been the center of debates in our discipline (e.g., risk perceptions in deterrence), and an analysis of the robustness of the effects using sensitivity analysis could prove informative. Further, sensitivity analysis can be useful for empirical tests in policing, courts and corrections where there is often a need to adjust for legally-relevant factors in order to obtain an unbiased estimate, but also a desire to appropriately contextualize the size and statistical significance of any effect that remains. Indeed, sensitivity analyses offers a useful tool to examine the robustness of statistical effects, and scholars should conduct such analyses before reaching firm conclusions about the effect of a construct on an outcome in their empirical works—whether that is supportive of a given theory or not.

Acknowledgements The authors would like to thank Tom Loughran, the anonymous reviewers, and the Editor, Gary Sweeten, for their helpful comments on earlier versions of this manuscript.

Appendix

See Figs. 8, 9, 10 and 11.

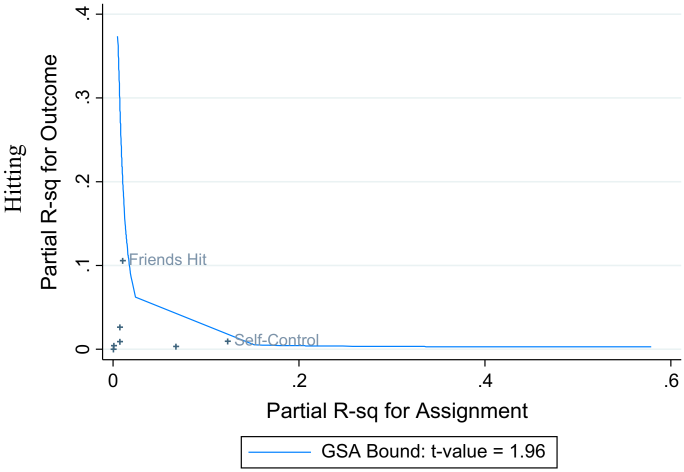

Fig. 8 Contour plot examining robustness of school commitment on hitting when controlling for theoretically-relevant controls

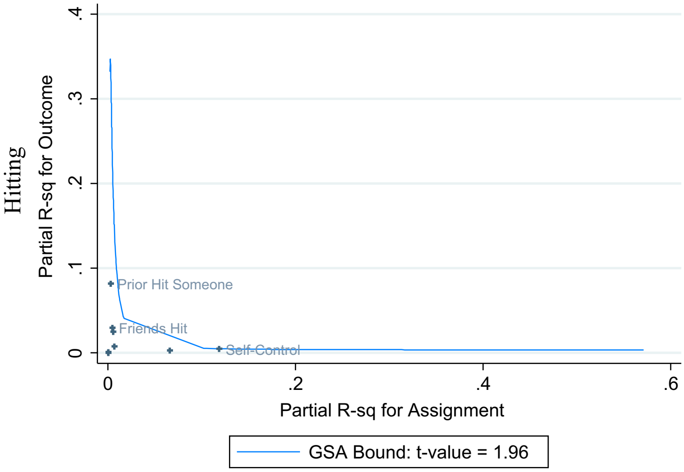

Fig. 9 Contour plot examining robustness of school commitment on hitting when controlling for LDV

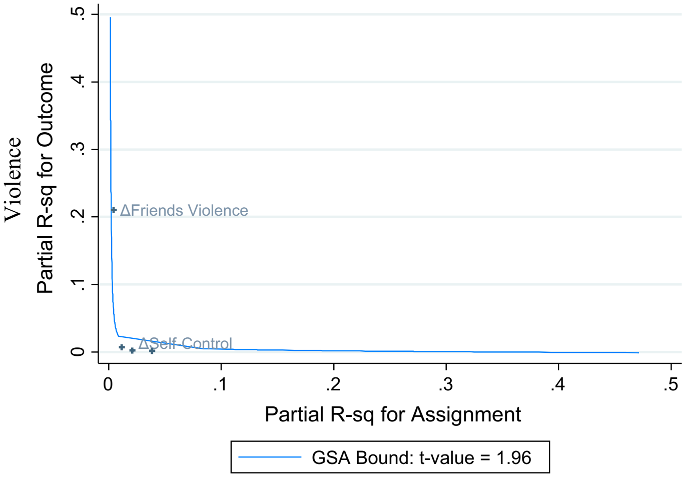

Fig. 10 Contour plot examining robustness of school commitment on violence when using fixed-effects

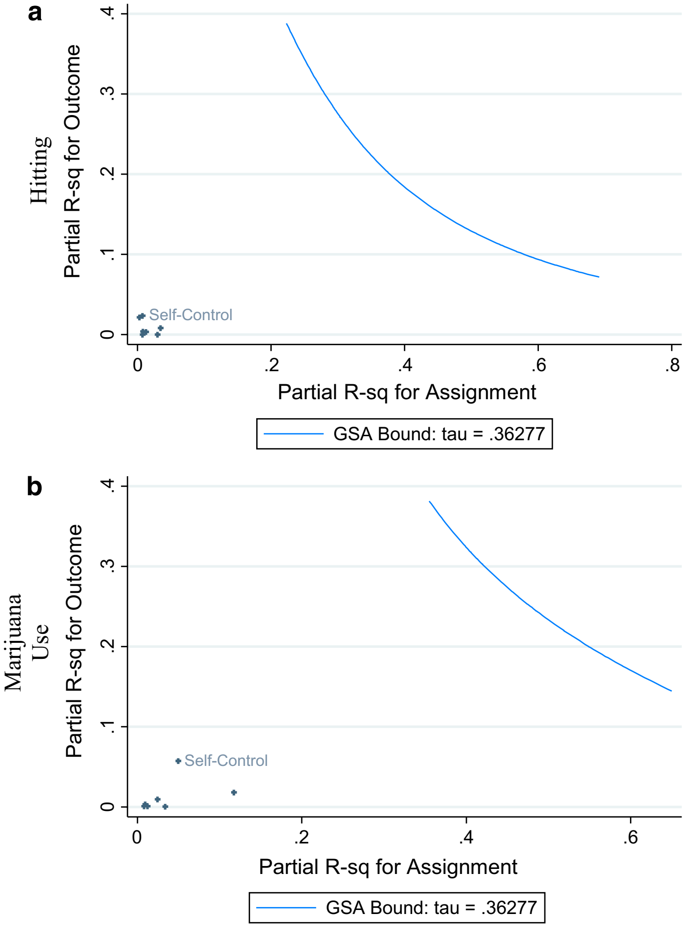

Fig. 11 Contour plots examining robustness of peer effects while controlling for theoretically-relevant controls at $d<.20$

References

Abelson R (1995) Statistics as principled argument. Lawrence Erlbaum Associates, Hillsdale

Achen CH (2001) Why lagged dependent variables can suppress the explanatory power of other variables. University of Michigan, Ann Arbor

Akers RL (1998) Social learning and social structure: a general theory of crime and deviance. Northeastern University Press, Boston

Allison PD (1990) Change scores as dependent variables in regression analysis. Sociol Methodol 20:93–114

Anderson AL (2002) Individual and contextual influences on delinquency: the role of the single-parent family. J Crim Just 30(6):575–587

Berk R, Brown L, Buja A, George E, Zhao L (2017) Working with misspecified regression models. J Quant Criminol. https://doi.org/10.1007/s10940-017-9348-7

Bollen KA, Curran PJ (eds) (2006) Latent curve models: a structural equation perspective. Wiley, Hoboken

Britt CL (1992) Constancy and change in the US age distribution of crime: a test of the "invariance hypothesis". J Quant Criminol 8(2):175–187

Brockman D, Kalla J, Aronow P (2015) Irregularities in LaCour (2014). UC Berkeley, Berkeley

Carson DC (2013) Perceptions of prosocial and delinquent peer behavior and the effect on delinquent attitudes: a longitudinal study. J Crim Just 41(3):151–161

Cohen J (1992) A power primer. Psychol Bull 112(1):155–159

Cohen J (1994) The earth is round (p < .05). Am Psychol 49(12):997–1003

Collins LM (2006) Analysis of longitudinal data: the integration of theoretical model, temporal design and statistical model. Annu Rev Psychol 57:505–528

Ding P, VanderWeele TJ (2016) Sensitivity analysis without assumptions. Epidemiology 27(3):368–377

Esbensen FA, Deschenes EP, Vogel RE, West J, Arboit K, Harris E (1996) Active parental consent in school-based research: an examination of ethical and methodological issues. Eval Rev 20:737–753

Fergusson DM, Swain-Campbell NR, Horwood LJ (2002) Deviant peer affiliations, crime and substance use: a fixed effects regression analysis. J Abnorm Child Psychol 30(4):419–430

Gangl M (2004) RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated. Statistical Software Components

George AL, Bennett A (2005) Case studies and theory development in the social sciences, 4th edn. MIT Press, Cambridge

Gottfredson MR, Hirschi T (1990) A general theory of crime. Stanford University Press, Stanford

Grasmick HG, Tittle CR, Bursik RJ Jr, Arneklev BJ (1993) Testing the core empirical implications of Gottfredson and Hirschi's general theory of crime. J Res Crime Delinq 30(1):5–29

Harada M (2012) ISA: Stata module to perform Imbens' (2003) sensitivity analysis. Statistical Software Components

Haynie DL, Osgood WD (2005) Reconsidering peers and delinquency: how do peers matter? Soc Forces 84(2):1109–1130

Hedström P (2005) Dissecting the social: on the principles of analytic sociology. Cambridge University Press, Cambridge

Hirschi T (1969) Causes of delinquency. University of California Press, Berkeley

Hirschi T (1979) Separate and unequal is better. J Res Crime Delinq 16(1):34–38

Hirschi T, Gottfredson M (1993) Commentary: testing the general theory of crime. J Res Crime Delinq 30:47–54

Horney J, Osgood WD, Marshall IH (1995) Criminal careers in the short-term: intra-individual variability in crime and its relation to local life circumstances. Am Sociol Rev 60(5):655–673

Imbens GW (2003) Sensitivity to exogeneity assumptions in programs evaluation. Am Econ Rev 93(2):126–132

Jussim L, Osgood DW (1989) Influence and similarity among friends: an integrative model applied to incarcerated adolescents. Soc Psychol Q 52(2):98–112

Leamer EE (1985) Sensitivity analysis would help. Am Econ Rev 75(3):308–313

Lemert EM (1951) Social pathology: a systematic approach to the theory of sociopathic behavior. McGraw-Hill, New York

Loughran TA, Paternoster R, Chalfin A, Wilson T (2016) Can rational choice theory be considered a general theory of crime? Evidence from individual-level panel data. Criminology 54(1):86–112

Maltz MD (1994) Deviating from the mean: the declining significance of significance. J Res Crime Delinq 31(4):434–463

Manski CF (1990) Nonparametric bounds on treatment effects. Am Econ Rev 80(2):319–323

Manski CF (1993) Identification of endogenous social effects: the reflection problem. Rev Econ Stud 60:531–542

Manski CF (2008) Identification for prediction and decision. Harvard University Press, Cambridge

Manski CF, Nagin DS (1998) Bounding disagreements about treatment effects: a case study of sentencing and recidivism. Sociol Methodol 28(1):99–137

McGloin JM, Thomas KJ (2013) Experimental tests of criminological theory. In: Welsh BC, Braga AA, Bruinsma GJN (eds) Experimental criminology: prospects for advancing science and public policy. Cambridge University Press, New York, pp 15–42

McGloin JM, Thomas KJ (2016a) Incentives for collective deviance: group size and changes in perceived risk, cost and reward. Criminology 54(3):459–486

McGloin JM, Thomas KJ (2016b) Considering the elements that inform perceived peer deviance. J Res Crime Delinq 53(5):597–627

McGloin JM, Sullivan CJ, Thomas KJ (2014) Peer influence and context: the interdependence of friendship groups, schoolmates and network density in predicting substance use. J Youth Adolesc 43(9):1436–1452

Nagin DS, Paternoster R (1991) On the relationship of past to future participation in delinquency. Criminology 29(2):163–189

Nagin DS, Smith DA (1990) Participation in and frequency of delinquent behavior: a test of structural differences. J Quant Criminol 6(4):335–356

Nguyen H, Loughran TA, Paternoster R, Fagan J, Piquero A (2016) Institutional placement and illegal earnings: examining the crime school hypothesis. J Quant Criminol. https://doi.org/10.1007/s10940-016-9291-z

Orwin RG (1983) A fail-safe N for effect size in meta-analysis. J Educ Stat 8(2):157–159

Osgood DW (2010) Statistical model of life events and criminal behavior. In: Piquero AR, Weisburd D (eds) Handbook of quantitative criminology. Springer, New York, pp 375–396

Osgood DW, Wilson JK, O'Malley PM, Bachman JG, Johnston LD (1996) Routine activities and individual deviant behavior. Am Sociol Rev 61:635–655

Pratt TC, Cullen FT, Sellers CS, Thomas Winfree L Jr., Madensen TD, Daigle LE et al (2010) The empirical status of social learning theory: a meta-analysis. Just Q 27(6):765–802

Rosenbaum PR (2002) Observational studies. Springer, New York

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rosenberg MS (2005) The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. Evolution 59:464–468

Rubin DB (2008) For objective causal inference, design trumps analysis. Ann Appl Stat 2(3):808–840

Sampson RJ (2010) Gold standard myths: observations on the experimental turn in quantitative criminology. J Quant Criminol 26(4):489–500

Sampson RJ, Laub JH (1993) Crime in the making: pathways and turning points through life. Harvard University Press, Cambridge, MA

Sutherland EH (1947) Principles of criminology, 4th edn. J. B. Lippincott, Chicago

Tamer E (2010) Partial identification in econometrics. Annu Rev Econ 2:167–195

Thomas KJ (2015) Delinquent peer influence on offending versatility: can peers promote specialized delinquency? Criminology 53(2):280–308

Thomas KJ, McGloin JM (2013) A dual-systems approach for understanding differential susceptibility to processes of peer influence. Criminology 51(2):435–474

Thornberry TP, Lizotte AJ, Krohn MD, Farnworth M, Jang SJ (1994) Delinquent peers, beliefs, and delinquent behavior: a longitudinal test of interactional theory. Criminology 32(1):47–83

VanderWeele TJ (2011) Sensitivity analysis for contagion effects in social networks. Sociol Methods Res 40(2):240–255

Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. Am Stat 70(2):129–133

Weisburd D, Piquero AR (2008) How well do criminologists explain crime? Statistical modeling in published studies. In: Tonry M (ed) Crime and justice: a review of research. University of Chicago Press, Chicago, pp 453–502

Wikström PH, Sampson RJ (2003) Social mechanisms of community influence on crime and pathways in criminality. In: Lahey B, Moffitt T, Caspi A (eds) Causes of conduct disorder and serious juvenile delinquency. Guilford Press, New York, pp 118–148

Young JT (2011) How do they 'end up together'? A social network analysis of self-control, homophily, and adolescent relationships. J Quant Criminol 27(3):251–273

Young JT (2014) A sensitivity analysis of egocentric measures of peer delinquency to latent homophily: a research note. J Quant Criminol 30(3):373–387