Research Transparency and Replicability in Criminology

A Replication Case Study of Sweeten et al. (2009)

Candidate number: 1030685. Word Count: 7,770 (Excluding Appendix)

Replication files: github.com/oxfordmphil1030685

Introduction

The recent retractions of five studies published in some of the most prestigious criminology journals has revived the debate on research integrity and replicability in the social sciences. Lack of scientific rigour, data irregularities and potential data fabrication were cited as reasons for the retractions of these studies which dealt with controversial issues, such as antiblack sentiment in the U.S. criminal justice system and racial differentials in punishment preferences (Chawla 2019). One of the retracted articles, entitled *Ethnic threat and social control* (Johnson et al. 2011), was published in the flagship journal of the discipline *Criminology*. Eight years after its publication, Justin Pickett, one of the article's co-authors, requested the retraction of the article after receiving an e-mail that showed data irregularities in the article and in four other articles published by his co-authors. The most striking irregularity was that the published article reported findings for a survey with 1,184 respondents while the original survey only included 500 respondents. In a public letter, Pickett provides re-analyses and explains his reasons for requesting a retraction of the paper, concluding that not only was the sample duplicated, the data was also altered "intentionally or unintentionally" (Pickett 2019).

While most disciplines within the social sciences have taken steps towards increasing research transparency, e.g. by promoting replication studies and requiring authors to share their scripts and data files, concerns have been raised that the field of criminology, as well as its father-discipline sociology, are lagging behind the other social sciences. For instance, no journals in criminology require data sharing or preregistration (see, Savolainen and VanEseltine 2018). In a recent review article on replications in criminology, Pridemore et al. (2018) show that only 0.45% of all studies in the field are replication studies. In other words, no systematic attempts at replication have been carried out to assess the validity of the evidence base that the discipline lies on. This is concerning when we recall the replication crises that have plagued other social science disciplines. In psychology, the Open Science Collaboration (OSC) team, led by Brian Nosek, attempted to replicate 100 studies from three psychological journals and found that the mean effect size of the replications was half the magnitude of the mean effect size of the original studies. Out of the 100 original studies, 97 reported statistically significant results ($p < 0.05$) compared to only 36 of the replication studies (OSC 2015). In political science, similar concerns have been raised following incidents of outright data fabrication (e.g., LaCour and Green 2014), and in sociology Gerber and Malhotra (2008) find strong evidence of publication bias in the discipline's two most prestigious journals (*American Sociological Review* and *American Journal of Sociology*) indicating that published finding might be misleading.

There are at least two good reasons for worrying about replicability in criminology. First, the anal-

yses criminologists use to study complex phenomena such as the causes and consequences of criminal behaviour usually rely on incomplete data sources with many unobserved variables. Compared to the toolbox available to researchers in other disciplines, such as randomised control trials in development economics and laboratory experiments in psychology, the observational methods applied by criminologists are usually ranked at the bottom of the "sciences hierarchy". Empirical findings obtained by these methods are therefore likely to be more fragile, compared to findings in other disciplines which makes replication studies essential in order to verify the evidence base that the discipline lies on (Pridemore et al. 2018). Second, many of the topics studied in criminology, such as the implications of gun control laws, are controversial both to public sentiment and policy makers. Ioannidis (2005) argue that false positive findings are more likely to occur in fields were the hypotheses are controversial because researchers might be driven by political convictions and other interests. The fact that criminologists often deal with questions that are of direct relevance to policy makers should therefore encourage a more transparent research culture to ensure evidence-based policies.

The article I have chosen for this replication study is an example of criminological research of direct relevance to policy makers. In their article, published in the flagship journal *Criminology*, Sweeten, Bushway, and Paternoster (2009) examine the question "Does dropping out of school mean dropping into delinquency?", which is also the title of the article. If it can be proved that youth who drop out of school are more likely to engage in criminal activities, then this information is invaluable to policy makers who might want to implement interventions to reduce the number of dropouts. On the contrary, if dropping out of school is unrelated to crime (or perhaps, only weakly related) then it might be beneficial to intervene in other areas in order to reduce youth delinquency. Regardless of the direction of effects studied by criminologists, obtaining firm evidence is crucial, not only for the theoretical development of the field but also for real life decisions. Pridemore et al. (2018) argue that studies with direct relevance to policy decisions should be prioritised in replication attempts to increase public confidence in evidence-based research. Apart from its policy focus, I have chosen to replicate this particular study for three other reasons.

First, the authors use the National Longitudinal Survey of Youth 1997 in their study, which is an open access data source that can readily be downloaded from the survey webpage after creating a free account. This makes the replication exercise easier and gives the authors a clear incentive to be thorough with their data handling as everyone can, in principle, replicate their work. Second, unlike the retracted articles mentioned above, the authors examine a relatively uncontroversial topic. Again, this gives the authors less of an incentive to manipulate their findings in order to push forward a specific political agenda. Third, the article is overall very clear in its aims and hypotheses and there are no good reasons to believe that the authors engaged in intentional misconduct. This is important because I want to examine how sensitive published research is to researchers' subjective choices, or what Simmons et al. (2011) call "researcher degrees of freedom", rather than pointing out academic misconduct. In other words, the aim of my replication exercise is to examine the validity of "good research" when scrutinised – which is different from showing that some academics engage in questionable research practices.

1 Replicating Research

In this section, I address three questions. Why are replication studies important? What kind of errors in existing research do replications, and in particular my replication case study, seek to assess? And how is the term replication understood in this article?

1.1 The Importance of Replications

> The mores of science possess a methodologic rationale but they are binding, not only because they are procedurally efficient, but because they are believed right and good. They are moral as well as technical prescriptions. Four sets of institutional imperatives—universalism, communism, disinterestedness, organized scepticism—are taken to comprise the ethos of modern science. (Merton 1973, originally published in 1942).

To address the first question, why is replication important, I draw on Merton's essay *The Normative Structure of Science*, from which the quote above is borrowed. This essay is perhaps the most influential work on research ethics in the social sciences (Cole 2004), and while it was written in a very different context in the forties of the last century, it remains surprisingly relevant to the debate on research integrity and the replication crises. The norm of universalism, i.e. the idea that the validity of a research finding hinges on its internal quality and coherence rather than the quality and reputation of the researcher, implies that anyone with access to the same data sources and sufficient research instruments and skills should be able to reproduce a valid finding. With the second norm, communism, Merton highlights that scientific truths are (or should be) products of social collaboration between researchers in a scientific community. Apart from the fact that collaborative research efforts are likely to be more efficient, open sharing between researchers has the positive side-effect that it can be used to detect academic fraud and data irregularities. The value of disinterestedness is obvious when we recall the fact that many of the retractions that involved fraud were of studies of politically controversial topics (e.g., Johnson et al. 2011; LaCour and Green 2014). Finally, organised scepticism is perhaps the most important norm from a replication perspective. While the process of publishing in scientific journals already contains elements of this norm, such as anonymous submissions and peer reviews, further instruments can be implemented in order to further strengthen the validity of research findings. A critical attitude towards existing research is crucial for the progress of science because it helps discovering errors in our current stock of knowledge and improve on previous findings.

Merton's essay sought to outline the moral norms that would ensure the integrity of science as a social institution, rather than describing actual scientists' beliefs and practices. In fact, empirical evidence indicates that while most researchers identify with the norms outlined above, few believe that they are good depictions of the scientific community. In a study of U.S.-based researchers, Anderson et al. (2007) surveyed 3,247 mid- and early-career scientists funded by the U.S. National Institute of Health about their own beliefs and practices, as well as their perceptions of their colleagues. While 90% of the scientists identified with the norms outlined above and 60-70% said that they also lived up to them in their own research, only 5-10% of researchers believed that their colleagues mainly followed these norms, and 60-75% believed that their colleagues indeed practiced counter-norms to the Mertonian norms outlined above. While this finding is ironic, it also uncovers a profound scepticism

towards current research procedures and practices. Improving open access to data and promoting a replication culture can serve as good tools to overcome distrust in research findings (Christensen et al. 2019).

1.2 Forking Paths and Researcher Degrees of Freedom

From my introduction above, one might get the impression that researchers are constantly engaging in questionable research practices, such as fabricating data, p-hacking, or fishing for results, and that replication studies are meant to uncover these malpractices. This is obviously not the case. Most researchers probably do not fabricate data or intentionally distort results. The point is rather that current research procedures and features of the publication process produce fragile findings. Examples are the evident bias towards publishing significant findings (Gerber and Malhotra 2008) or that robustness checks usually only address issues related to a subset of possible specifications anticipated by the authors themselves (Steegen et al. 2016).

Gelman and Loken (2013) discuss the fragility of published research, even when researchers are not intentionally fishing for results, p-hacking, and changing model specification until they find a satisfactory—i.e. publishable—result. Even when the hypothesis is specified clearly from the beginning, researchers are faced with a multiplicity of choices at each stage of the research process which forces them to make decisions about how they delimit the sample, choosing among a set of variables, coding these variables, and so on. The point is that a great number of options might be equally reasonable but in practice researchers pick one among multiple reasonable options. This formulation of the problem is nearly related to the concept of researcher degrees of freedom (Simmons et al. 2011). In statistical terms, researchers thus pick one out of a large number of possible data realisations. If other realisations of the data are reasonable, then inference should not be limited to that one realisation, but the entire set of possible reasonable realisations. Using examples from well-published articles, Gelman and Loken (2013) show that these arbitrary (but reasonable) decisions about how variables are coded, or which subsamples are considered might lead to severely biased estimates.

1.3 Types of Replication Studies

Several authors have pointed out the ambiguous use of the term "replication" which refers to different things, even within the same discipline (e.g., Christensen et al. 2019). One broad classification distinguishes between direct and conceptual replications. Direct replication refers to cases where an author aims to reproduce a finding by using both the same data and methods as the original authors. Conceptual replication refers to cases where the aim is to verify (or falsify) the idea of the original data – for instance, by examining alternative data sources, specifications, or estimation methods (Pridemore et al. 2018). Another helpful classification, specific to quantitative social science, is provided by Freese and Peterson (2017) whose classification scheme is reproduced in Table 1 below.

Table 1: Types of Replications (from Freese and Peterson 2017)

	Similar	**Different**
Old Data	Verifiability	Robustness
New Data	Repeatability	Generalisation

This classification scheme distinguishes replication attempts along two axes: whether the data are old or new, and whether the methods applied are similar or different from the original article. A successful verification of an existing finding is thus showing that one can obtain the same results by using the same data sources and methods. Assessing the robustness of a research finding involves the application of alternative specifications or methods on the same data. Whereas these two replication typologies are restricted to the original dataset, the last two typologies involve extending the data sources in a similar setting (repeatability) or a different setting (generalisation) to assess whether the hypothesis of the original article holds. My replication of Sweeten et al. (2009), in the terminology of Freese and Peterson (2017), mainly considers the verifiability and the robustness of the original article. However, I also extend on the original setup by adding additional waves from the same data source used by the authors to examine the sensitivity of the findings. This last exercise is thus concerned with the repeatability of the original findings in a weak sense.

2 Outline of Original Article

I replicate the article "Does dropping out of school mean dropping into delinquency?" published in *Criminology* by Sweeten, Bushway, and Paternoster (2009). As indicated by the title, the authors examine whether dropping out of school (primary or secondary) affects the likelihood of committing a criminal offense, and whether this effect varies by reason for leaving school. In this article I focus on the first hypothesis, i.e. the general effect of dropout on crime, and do not consider the second part, i.e. whether the effect varies by reasons for leaving school.

The authors use the first seven waves (1997-2003) from the National Longitudinal Youth Survey 1997 (hereafter, NLSY97) which is a publicly available data set that can be downloaded by creating a free account on the Study's webpage[1]. The NLSY97 is carried out by the U.S. Bureau of Labor Statistics and is a nationally representative sample of 8,948 youth aged 12 to 17 at the interview date of the first wave. The same individuals were subsequently re-interviewed every year until 2011, and thereafter every second year. In the sixth wave of the study, 7,897 youths (87.9% of the original sample) were re-interviewed which is considered a low attrition rate. Importantly, the NLSY97 Study oversamples under-represented population groups and it is therefore crucial to use the sampling weights to obtain unbiased inferences.

In order to measure youth delinquency, the authors construct two measures of criminal activity. A count variable, `crime variety` which measures the number of offences the youth committed since the last interview, and a binary indicator, `crime prevalence`, which measures whether the youth committed any offence since the date of the last interview. The main independent variable, `dropout`,

[1] https://www.nlsinfo.org/content/getting-started/accessing-data

is a time-varying binary indicator equal to 1 if the respondent is not enrolled in school (either primary or secondary) in that particular wave, and 0 otherwise. Another important variable is `years since last dropout` which measures the time since the last enrollment was recorded in years. The authors also include a set of control variables, such as age, gender, race, alcohol consumption, and so on. Importantly, the authors lag all time-varying independent variables one year behind the dependent variables in order to avoid issues of reverse causality. In other words, the authors want to ensure that they measure the effect of dropping out of school on subsequent criminal involvement. This is important to stress, since the authors poorly describe this in their article. The time-lag also implies that although data from seven NLSY97 waves are used, the actual number of time periods in the models is six. I describe the data construction in greater detail in Section 3.

2.1 Econometric Model

The focus in my paper is on the replicability of Sweeten et al. (2009) rather than the credibility of their analytic strategy and for identifying the causal effect of school dropout on criminal behaviour. However, since the choice of the statistcal model is one among many choices made by the researcher that influences the final result of the paper, and since my critiques later on deals with some technicalities related to the statistical model, I here shortly outline their model.

The authors exploit the longitudinal structure of the dataset to deal with selection issues and estimate an extension to the standard random-effects model. The prevalence measure of criminal activity, `crime prevalence`, is binary and is therefore estimated using a logistic link function. The variety measure, `crime variety`, is a count variable and the authors therefore use a negative binomial link function for this outcome. The authors' justification for using a random-effects model instead of the more common fixed-effects model is "because of its increased efficiency relative to the fixed-effects model" (Sweeten et al. 2009, p. 61). To clarify thoughts, let us consider the form of the standard random-effects model in their case:

$$y_{i,t+1} = \beta_0 + \beta_1 q_i + \delta_1 x_{it} + \theta_2 D_{it} + \tau_i + \varepsilon_{it} \tag{1}$$

where $y_{i,t+1}$ is the outcome of interest, e.g. `crime prevalence`, lagged one year ahead of the time-varying variables on the right-hand-side, q_i represents time-invariant individual characteristics (e.g., gender and race), $_{it}$ represents time-varying covariates (such as age), and D_{it} is the main regressor of interest, i.e. the dummy variable `dropout`. The random error term is decomposed into time-invariant individual differences represented by τ_i, and indvidual time-varying random error represented by ε_{it}.

Because random-effects models are biased if unobserved confounders are excluded from the model, the authors draw on an extension to the standard random-effects model developed by Bryk and Raudenbush (1992). In the literature, this model is often referred to as a hybrid model, or more recently a Random-Effects-Within-Between (REWB) estimator (e.g., Bell et al. 2019; Schunck 2013). This estimator separates within-individual from between-individual effects by transforming the time-varying variables (here, x_{it} and D_{it}) and estimating a model with the transformed variables instead:

$$y_{i,t+1} = \beta_0 + \beta_1 q_i + \delta_B \bar{x}_i + \delta_W \Delta x_{it} + \theta_B \bar{D}_i + \theta_W \Delta D_{it} + \tau_i + \varepsilon_{it} \tag{2}$$

where $\bar{x}_i = T^{-1}(\sum_{t=1}^{T} x_{it})$ is the average value of the time-varying variable, x_{it} over time for individual i, and $\Delta x_{it} = (x_{it} - \bar{x}_i)$ is the de-meaned form of the time-varying variable. Under assumptions outlined in Bryk and Raudenbush (1992), the two sets of parameters (δ_B, θ_B) and (δ_W, θ_W) then identify the between-individual and within-individual effects, respectively, as indicated by the subscripts $(_{B,W})$ on the parameters.

2.2 Main Findings and Interpretation of Coefficients

The main interest of the article is in the coefficients of the parameters attached to the four variables: time-stable dropout status (θ_B), time-varying dropout status (θ_W), time-stable years since dropout (δ_B), and time-varying years since dropout (δ_W). According to the authors (Sweeten et al. 2009, p. 62), these coefficients should be interpreted in the following way: The time-stable (or between-effects) coefficients "reflect time-invariant differences between dropouts and nondropouts. If the cofficients on these variables differ from zero, it suggests that dropouts are different from nondropouts, but it does not speak to a *causal* effect of dropout". The authors therefore expect that these coefficients are different from zero.

The primary focus of the authors is on the within-effects, i.e. the coefficients on the time-varying components (δ_W, θ_W). The coefficient of the `dropout` variable, according to the authors, reflect the causal effect of dropout on subsequent criminal behaviour, and is therefore the single most important parameter to be estimated. The within-coefficient on the `years since dropout` variable reflect the causal effect of "changes in the amount of time since dropout occured" (Sweeten et al. 2009, p. 62). The inclusion of the `years since dropout` variable changes the interpretation of the dropout status variable to the immediate effect of dropout.

The overall conclusion of the paper is that the evidence does not suggest that dropping out of school is causally related to criminal behaviour, i.e. the within-effects estimates are insignificant. The between-effect estimates for the dropout coefficients are statistically significant which suggests that there is a "baseline difference" between dropouts and non-dropouts in their propensity to engage in criminal activities.

3 Reconstructing the Dataset

In this section, I describe how I constructed my dataset to mimic the data in Sweeten et al. (2009) as closely as possible. In general, this part of the replication exercise was very time-consuming and demanding, mainly because of poor data descriptions in the original article and due to the lack of supplemental replication material, such as codes or even a list of the original variable names and reference numbers in the NLSY97 data. I contacted the corresponding author early on in the project to ask for scripts but he refused to share the original scripts. After informing him that I had difficulties reconstructing even the main independent variable, the author send me a SAS script that generates that variable (even though the article mentions Stata as the statistical package used). Otherwise, I

have relied on the data section in the original article and an Appendix table (Sweeten et al. 2009, pp. 89-92) with some descriptions at the end of the article to reconstruct the dataset.

3.1 Working Sample

The only information on how the authors construct the working sample is comprised in the following sentence "[b]ecause we need at least three adjacent interviews between waves 1 and 7 for each person to be observed, those who did not have this were eliminated from the sample. Our working sample was 8,112 with 4,129 males and 3,983 females, who contributed 45,546 person waves" (Sweeten et al. 2009, pp. 63-4). In general, the authors do not provide the reader with much information on how they constructed the variables or handled missing values – except for three variables: (*i*) `GPA:` missings are replaced with the average `GPA` value. (*ii*) `ASVAB:` which are test scores from the Armed Services Vocational Aptitude Battery. Missings are replaced with a 0. (*iii*) `dropout:` no missings in wave 1. Missings in subsequent waves are replaced with the most recent enrollment record. For all three variables, the authors include a dummy variable indicating missings in the raw data. I have followed this information provided by the authors to construct these measures. If individuals had missing values in other variables, such as the dependent variable, they were dropped from the sample in the waves were they had missing values.

3.2 Constructing the Variables

As mentioned above, not much information is provided on how the authors construct their variables. I have therefore followed the descriptions in their data section and Appendix to reconstruct the variables as closely as possible. Since the NLSY97 reference numbers (which are unique identifiers for variables) are not provided, I have used the search functions on the NLSY webpage to find the variables.

The dependent variables in the article are two measures of criminal activity based on six questions in the NLSY97 survey asking respondents' whether they committed any of the following offences since the data of the last interview: (*i*) intentional destruction of property, (*ii*) theft of items under $50, (*iii*) theft of items greater than $50 (including autos), (*iv*) other property crimes, (*v*) attacking someone with intent to hurt them seriously, and (*vi*) selling illegal drugs (Sweeten et al. 2009, p. 64). The first measure is a count variable measuring the number of different offences the youth committed in each wave. Let C_{ijt} be a variable indicating whether individual i committed the offence $j = 1, ..., 6$, in wave t, then the variety measure of criminal activity, `crime variety`, can be constructed as follows:

$$\texttt{crime variety}_{\texttt{it}} := y_{it}^{V} := \sum_{j=1}^{6} C_{jt}$$

which varies from 0 (if i committed no crimes in wave t) to 6 (if i committed all six crimes in wave t). The prevalence measure of criminal activity, `crime prevalence` can then be constructed in the following way:

$$\texttt{crime prevalence}_{\texttt{it}} := y_{it}^{P} := \mathbb{1}\{y_{it}^{V} > 0\}$$

where $\mathbb{1}\{\cdot\}$ represents the indicator function.

The main independent variable, `dropout`, is constructed using the enrollment status variable from the NLSY97 survey. Following the authors descriptions, an individual is considered a dropout if he is not enrolled in a primary or secondary school, and a dummy variable for being enrolled in high-school, *inhs*, is also constructed. The authors do not provide any information on how they construct the `years since last dropout` variable. In the first wave, I use a question asking about the respondent's last enrollment. This variable is either 0 if the individual is still enrolled or some positive value greater than 0. In the subsequent waves this variable increases by one every time the individual is recorded with a dropout.

Other variables are included by the authors to control for observed differences between dropouts and non-dropouts. The authors include both time-constant covariates such as race and sex, and time-varying covariates such as the number of arrests since the last interview. A full list of control variables used in the original article is provided by the authors in the Appendix table (Sweeten et al. 2009, p. 91). I have provided a full list of variables with their original NLSY97 question names on my Github repository (see title page).

3.3 Comparing the two datasets

In the previous section, I described how I re-constructed the dataset used by Sweeten et al. (2009) in order to perform the replication analyses. How well do the two datasets match? Juding from the numbers provided below: almost perfectly. Table 2 below imitates Table 1 in Sweeten et al. (2009, p. 64) and provides the number of observations in each age-group and each wave of the data. I have highlighted the numbers from my replication data in boldface to make the comparison to the original article easier. As can be read in the table, the differences are very small. The table in the original article also includes a column with the proportion enrolled in primary or secondary school for each age group in each wave. I have replicated that part of the original table in a similar way in Table A.1 in the Appendix – again, the similarities are strikingly close considering that I did not have access to replication scripts from the authors.

Table 2: Sample Size by Age and Wave (Table 1, p. 64 in Sweeten et al. 2009)

	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5		Wave 6	
Age	Rep	Org	Rep	Org	Rep	Org	Rep	Org	Rep	Org	Rep	Org
12	**1,029**	1,021										
13	**1,615**	1,592	**54**	55								
14	**1,673**	1,651	**1,527**	1,506	**65**	65						
15	**1,678**	1,631	**1,619**	1,603	**1,580**	1,576	**26**	26				
16	**1,512**	1,458	**1,631**	1,601	**1,579**	1,582	**1,457**	1,455	**27**	27		
17	**554**	534	**1,610**	1,553	**1,625**	1,626	**1,601**	1,599	**1,477**	1,474		
18	**14**	14	**1,338**	1,287	**1,487**	1,484	**1,524**	1,519	**1,554**	1,546	**1,460**	1,459
19			**112**	112	**1,282**	1,285	**1,480**	1,474	**1,501**	1,495	**1,516**	1,515
20					**77**	76	**1,284**	1,281	**1,490**	1,481	**1,511**	1,504
21							**153**	152	**1,266**	1,259	**1,482**	1,476
22									**134**	130	**1,244**	1,236
23											**126**	126
Total	**8,075**	7,901	**7,891**	7,717	**7,695**	7,694	**7,525**	7,506	**7,449**	7,412	**7,339**	7,316

Notes: Rep = Numbers from replication study. Org = Numbers from original study (Sweeten et al. 2009). Note that this table reproduces the sample-size part of Table 1 in the original article. For the proportion enrolled in school (primary or secondary), please refer to Appendix Table A.1.

With regards to the characteristics of the individuals in the sample, the replication data is very close to the original data. Table 3 below compares summary statistics in the two datasets imitating the full sample summary statistics part of Table 2 in the original article (Sweeten et al. 2009, p. 65). The original article also show summary statistics for the subsamples of males, females, individuals who are ever recorded to dropout, and individuals who never dropped out of school. These statistics are replicated and compared in Table A.2 in the Appendix for the sake of brevity. The first two columns in Table 3 report means and standard deviations for the replication sample and the next two for the original sample. The last columns show the differences in means between the the variables in the replication and original samples. Most importantly, the two dependent variables `crime variety` and `crime prevalence` are virtually identical in the two samples. Similarly the main independent variable, `dropout`, is replicated with the exact mean as in the original paper.

Most of the other variables in the replication sample are also very similar to those in the original article. However, some variables are replicated with some discrepancy. For instance, the individuals in the replication are calculated to be around 4 months younger (`age` variable) and dropout around half a year earlier (`years since dropout` variable) on average than those in the original article. With regards, to the race variables, three of my reconstructed measures are very close to the original variables (`White, Black, Hispanic`). The `Other race` variable seems to be off because of an error in the original article, because the proportions do not sum to 1. The remaining differences are relatively trivial, which indicates that the replication exercise was very successful.

Table 3: Comparison of Summary Statistics for The Total Sample

	Replication		Original		Differences in means
	Mean	SD	Mean	SD	
Crime variety	0.34	0.83	0.35	0.85	-0.01
Crime prevalence	0.20		0.20		0.00
Dropout	0.11		0.11		0.00
Years since dropout	1.15	1.48	1.72	1.55	-0.57
Male	0.51		0.51		0.00
Age	17.45	2.38	17.77	2.38	-0.32
White	0.71		0.72		-0.01
Black	0.15		0.16		-0.01
Other race	0.01		0.12		-0.11
Hispanic	0.13		0.13		0.00
Live with biological parents	0.54		0.54		0.00
Arrests	0.10	0.55	0.11	0.56	-0.01
Smoking prevalence	0.42		0.42		0.00
Years sexually active	2.36	2.50	2.42	2.75	-0.06
Antisocial peer scale	1.71	1.65	1.76	1.66	-0.05
Middle-school GPA	2.87	0.53	2.87	0.86	0.00
Ever suspended	0.26		0.32		-0.06
Ever retained	0.18		0.17		0.01
ASVAB: arithmetic reasoning	0.08	0.86	0.03	0.90	0.05
ASVAB: word knowledge	0.05	0.81	0.02	0.91	0.03
ASVAB: paragraph comprehension	0.09	0.82	0.03	0.91	0.06
ASVAB: math knowledge	0.12	0.87	0.03	0.90	0.09
Mother dropout	0.17		0.16		0.01
Father dropout	0.17		0.15		0.02
Received federal aid	0.40		0.35		0.05
Outside: nice	0.65		0.65		0.00
Outside: fair	0.27		0.27		0.00
Outside: poor	0.06		0.07		-0.01
N (Person waves unweighted)	45,974		45,546		
N (Person waves weighted)	44,248				
N (Individuals)	8,376		8,112		

Notes: This table only compares summary statistics for the total sample to replicate the first column of Table 2 in the original paper because of space concerns. Note that standard deviations (SD) are not reported for dummy variables in this table or in the original paper because the standard deviation for a dummy variable can easily be calculated as $\sqrt{(m - m^2)}$ where m is the mean. The original table also includes columns for the following subsamples: Males, Females, Ever Dropout, and Never Dropout. These columns are replicated in Appendix Table A.2.

4 Verification of Estimation Results

In Section 3 I showed that the constructed replication dataset is almost identical to the original dataset used by the authors, in terms of both number of observations in each age group and wave (Table 2), the proportion enrolled in school by age group and wave (Table A.1), and in terms of summary statistics (Table 3). Therefore, the estimation results for the effect of dropout on criminal activity should be similar to those in the original paper, given that the datasets are highly similar, and the models estimated are the same. Recall from Section 2 that the authors estimate REWB models for the two dependent variables. The link functions are a negative binomial for the count variable `crime variety` and a logit for the binary variable `crime prevalence`. I have performed the same transformations on the time-varying independent variables as the authors (outlined in Section 2) and used the same control variables including a linear time trend as reported by the authors in the original article (Sweeten et al. 2009, Tables 5 and 6 on p. 72, and Appendix Table A on p. 89). The models are estimated using the `me` package in Stata (commands: `melogit` and `menbreg`) specifying the random effects estimator and using the panel data survey probability weights from the NLSY97 Survey.

Table 4: Replication of Main Estimation Results for the Total Sample[a]

	Between-Effects				Within-Effects			
	Replication		Original		Replication		Original	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Negative Binomial RE Model								
Dropout	**0.303**	**0.227**	0.569	0.137**	**0.058**	**0.078**	0.044	0.057
Years since dropout	**-0.135**	**0.088**	-0.231	0.057**	**-0.037**	**0.032**	-0.037	0.026
Logistic RE Model								
Dropout	**0.439**	**0.324**	0.661	0.199**	**0.180**	**0.126**	0.132	0.095
Years since dropout	**-0.192**	**0.123**	-0.296	0.080**	**-0.082**	**0.050**	-0.108	0.042*
N (person-waves)	**45,974**		45,546		**45,974**		45,546	
N (individuals)	**8,376**		8,112		**8,376**		8,112	

Notes: Estimates and adjusted standard errors calculated using the Stata commands `menbreg` and `melogit`. Survey weights from wave 1 used. Original estimates copied directly from original paper (Sweeten et al. 2009, Tables 5 and 6, p. 72). [a]Replication results for subsamples in Appendix Tables A.3 and A.4. *$p < .05$,** $p < .01$,*** $p < .001$.

Table 4 above compares estimation results for the full sample across the replication (in boldface) and original datasets. Recall that the between-effects coefficients reflect baseline differences between dropouts and non-dropouts, and these coefficients are therefore of secondary importance according to the authors. In both the negative binomial and the logit models, the between effects for both the `dropout` and the `years since dropout` variables are substantially smaller in the replication data compared to the original estimates. In addition, the standard erros are substantially smaller in the original data which yields significant results for the between-effects coefficients in the original paper. These coeffcients are all statistically insignificant in the replication data.

Turning to the within-effects, which are the most important according to the authors themselves, as

these reflect the causal effect of dropout on criminal activity, the replication results are closer to that of the original paper. Both the negative binomial and the logit models in the original paper report positive but insignificant effects for the dropout variable. These effects are successfully reproduced as they are close in magnitude and also not significantly different from zero in the replication paper. In the logit model, the within-effects coefficient for `years since dropout` is significantly negative in the original paper but statistically insignificant in the replication data. Apart from this discrepancy, the broad conclusion is that the most important effects were successfully replicated. It is also worth to note that all estimates have the same direction across the two datasets. Despite some discrepancies, the overall conclusion from this verification exercise is that the most important estimates (according to the authors themselves), i.e. the within-effects, did indeed replicate.

5 Extensions and Sensitivity Analyses

While the aim in Section 4 was to replicate the exact results in the original paper to verify its validity by using the same methods and data sources, this section goes beyond the original data and methods to scrutinize how stable the results are. In the first extension, I investigate how sensitive the parameter estimates are when adding additional data. In the second extension exercise, I consider the sensitivity of the estimates to the particular dataset I have constructed by using a bootstrap method. Finally, I have included a technical note at the bottom of this section on a possible mistake in the original article's estimation of standard errors.

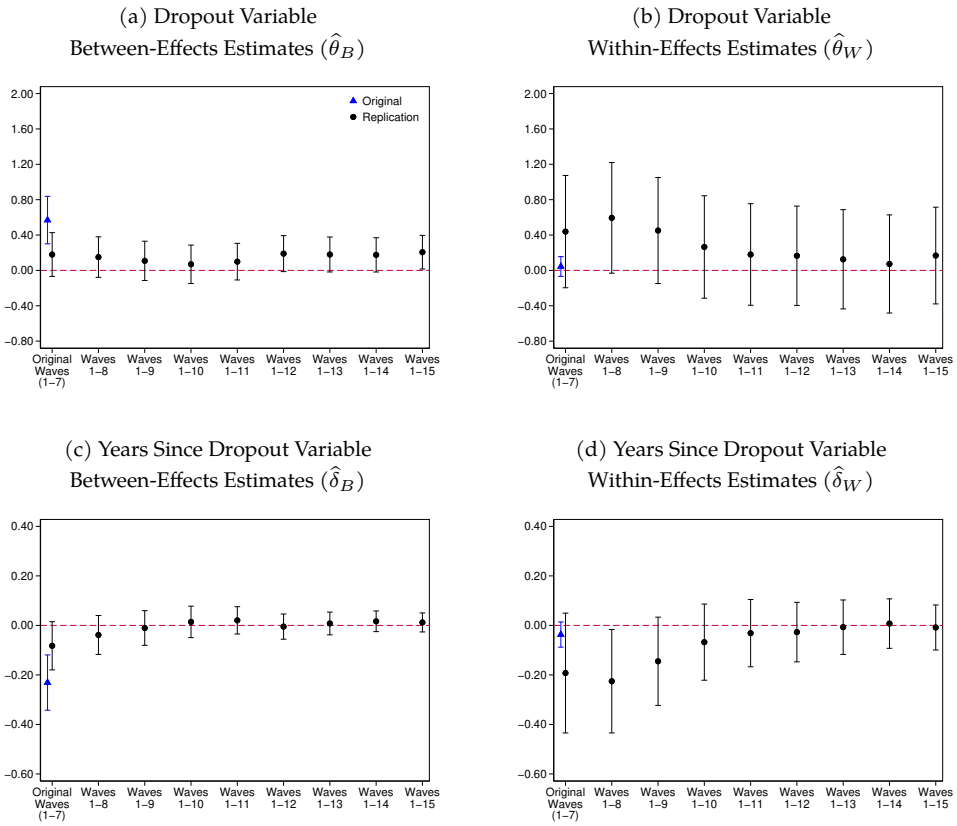
5.1 Adding Additional Data Waves

In their original dataset, Sweeten et al. (2009) use waves 1 through 7 of the NLSY97 study (see Sections 2 and 3 for details). In my first extension on the original paper, I analyse the sensitivity of the estimates when more data are added to the model. Intuitively, the estimates should become more precise (i.e. smaller standard errors) when the number of observations grows. I add one additional wave to my replication dataset at a time and re-estimate the model. Wave 15 (year 2011) is the last wave where the dependent variable is available in the NLSY97 study. This means that I re-estimate the model 8 times. The 0^{th} iteration is the original setup that only includes waves 1-7, iteration 1 includes waves 1-8, and so on until the last iteration which includes all available waves (1-15). Figure 1 below shows the estimation results with (95% confidence intervals) of this exercise for the logit estimates of the four main coefficients. The blue triangles represent the original estimates from Sweeten et al. (2009). The black dots represent estimates from the replication data.

Which conclusions can be drawn from this exercise? First, recall that the most important estimates according to the authors were the within-effects estimates, as these reflect the causal effect of dropout on crime. All within-effects estimates in the original paper are not significantly different from zero. Figure 1 shows that even when more data are added this result holds. Second, the within-effects estimates are still very noisy (large standard errors) after adding more data. By the last iteration (waves 1-15), the decrease in the standard errors are not noticeable. This indicates that the standard

errors on the within-effects estimates in the original paper might have been estimated incorrectly (see Section 5.3). Third, with regards to the between-effects estimates, which reflect baseline differences between dropout and non-dropouts over all time periods, this exercise further confirms the replication results from Section 4. Even after adding more precision to the estimation, the between-effects in the replication are still insignificant – unlike the finding in the original article. Bearing in mind that the most important findings in the original article are related to the within-effects, the overall conclusion is that the Sweeten et al. (2009) replicates pretty well even after this sensitivity analysis.

Figure 1: Sensitivity Analysis Adding One Additional NLSY97 Wave at a Time to the Original Data.



Notes: Original estimates (blue triangles) for the random-effects logistic regressions are copied from Sweeten et al. (2009, Table 6, p. 72). Replication estimates (black dots) calculated using the Stata command `melogit`. At each iteration a new wave was added to the data. Estimation results for the random-effects negative binomial models can be found in the Appendix, Figure A.1

5.2 Bootstrap Samples

While the broad conclusion is that Sweeten et al. (2009) replicates well, there are still some discrepancies between the replication results and the original paper. This is of course expected because I did not have access to the authors' scripts and not all steps and details of the data construction is thoroughly described in the original article. In consequence, the replication dataset has a larger number

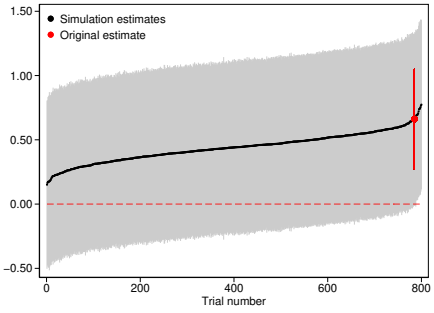
of observations ($8,376$ vs. $8,112$ in the original) and some of the summary statistics do not correspond to those in the original article. Most importantly, the between-effects estimates are significant in the original article but very robustly insignificant in the replication. A natural question here is whether my replication results are just driven by the fact that I am analysing a different dataset?

As described in Section 3, I have tried to mimic the original data set as closely as possible. And the fact that my replication dataset has a larger number of observations might indicate that the dataset used by Sweeten et al. (2009) is one possible sample of my replication data – realised by the choices made by the authors. Many different samples could have been realised by making other choices. We can therefore ask the question: how sensitive are the replication findings to different data realisations? One way of addressing this issue is to randomly draw new datasets based on the constructed replication datasets by using a bootstrap approach. I draw 800 random samples from my replication data, dropping around 10% of the observations at each trial. I choose 800 trials because the maximum matrix size in the Stata student version (IC) is 800×800.

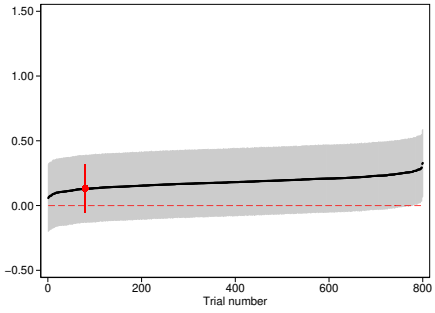
Figure 2 below shows the bootstrap results for the coefficients on the `dropout` variable using the logit model for 800 samples (this exercise is very heavy computationally – and I have therefore not repeated this exercise for the negative binomial model). The red dots (and 95% CIs coloured in red) represent the estimates from the original article (i.e., from Sweeten et al. 2009). The black dots (with grey 95% CIs) are the results from the bootstrap trials. The results are very re-assuring. The first important thing to note is that virtually all estimates (both within- and between-effects) are insignificant. The second important finding is that the authors' estimates are located at the tails of the distributions. The estimates are sorted by effect size – which means that points further away from the center represent unlikely estimates. With regards to the within-effects estimate, this exercise adds to my confidence in the finding that this coefficient should be zero. With regards to the between-effects estimate, it appears that the significant result in Sweeten et al. (2009) is driven by deflated standard errors.

Figure 2: Sensitivity Analysis: Parameter Estimates from 800 Bootstrap Trials

(a) Dropout Variable
Between-Effects Estimates ($\widehat{\theta}_1$)



(b) Dropout Variable
Within-Effects Estimates ($\widehat{\theta}_2$)



Notes: Original estimates for the random-effects logistic regressions are copied from Sweeten et al. (2009, Table 6, p. 72). Replication estimates calculated using the Stata command `melogit`.

5.3 Different Model Specifications and Sampling Weights – Some Technical Notes

One interesting finding when examining the data, that I wish to highlight here, is that the deflated standard errors in the original paper – that I have referred to a number of times in my article – might be caused by a small technical error in the estimation procedure of the original article. When the original article was written in 2009, the available Stata commands for estimating panel data regressions were in the `xt` package. For instance, panel data models for binary outcomes can be estimated with a logit model using the command `xtlogit` or a probit model using `xtprobit`. However, a limitation of the `xt` package is that it does not allow for survey weights when the random effects estimator is used. Instead, so-called importance weights are allowed. According to the official documentation, these weights are *not* probability weights, but weights used by programmers – and do not have any clear definition (StataCorp 2013). However, if one specifies a weight, for instance by writing:

```
xtlogit y x z [w=w], re
```

then the command would automatically assume importance weights instead of probability weights, because the random-effects option is specified. The command only returns an error if one explicitly specifies that the weights should be probability weights. Therefore, it might be the case that the authors made this mistake in their coding. On the contrary, the new `me` package, which I have used in my replication, allows for probability weights even with random effects models (StataCorp 2019).

I am not sure whether this is indeed the reason behind the discrepancies. Another possible reason could be due to clustering of standard errors. In the original article, the authors are very unclear about whether they do cluster the standard errors on the household level—because some siblings exist in the survey— or on the individual level which is the standard option in most statistical software's panel data methods (Sweeten et al. 2009, see footnote 6 on p. 63). To explore the results' sensitivity to these issues, I have re-estimated all possible combinations of standard errors and weights in Table 5 below. Note that there are 6 columns for the `me` command because it allows for all combinations of (no weights, probability weights, importance weights) and (individual clustered standard errors, and household clustered standard errors), whereas the `xt` command allows for either importance weights with individually clustered standard errors (same as none), or any combination of clustered standard errors if weights are not applied.

Column 1 shows the main specification, i.e. an `me` model with probability weights and clustered standard errors on the individual levels. These estimates are identical to those reported in table 4. The interesting thing is that when the `me` command is used, the model only gives statistically significant estimates when it is misspecified, e.g. by specifying no weights or importance weights (columns 3-6). Similarly, in the `xt` package, estimates are statistically significant because probability weights are not allowed (columns 7-9) and when importance weights are specified (column 10). Finally, note that clustering on the household level does not change the results much when the model is correctly specified (columns 1-2), and that the importance weights massively deflate the standard errors which might explain the significant results for the between-effects estimates in the original paper.

Table 5: Estimation Results for the Random-Effects Logistic Model with All Possible Combinations of Choices for Standard Errors and Weights in the `me` and `xt` Packages.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
crim_prvl										
Dropout										
(within-effect)	0.180	0.180	0.180***	0.180	0.237*	0.237*	0.237*	0.237*	0.237*	0.180***
	(0.126)	(0.127)	(0.000)	(0.127)	(0.105)	(0.113)	(0.105)	(0.113)	(0.113)	(0.000)
Dropout										
(between-effect)	0.439	0.439	0.439***	0.439	0.463	0.463	0.463	0.463	0.463	0.439***
	(0.324)	(0.323)	(0.001)	(0.323)	(0.285)	(0.281)	(0.285)	(0.281)	(0.282)	(0.001)
Years since										
last dropout										
(within-effect)	-0.082	-0.082	-0.082***	-0.082	-0.098*	-0.098*	-0.098*	-0.098*	-0.098*	-0.082***
	(0.050)	(0.050)	(0.000)	(0.050)	(0.039)	(0.043)	(0.039)	(0.043)	(0.043)	(0.000)
Year since										
last dropout										
(between-effect)	-0.192	-0.192	-0.192***	-0.192	-0.176	-0.176	-0.176	-0.176	-0.176	-0.192***
	(0.123)	(0.122)	(0.000)	(0.122)	(0.107)	(0.104)	(0.107)	(0.105)	(0.104)	(0.000)
Stata command	me	me	me	me	me	me	xt	xt	xt	xt
Weights	Probabilty	Probabilty	Importance	Importance	None	None	None	None	None	Importance
SE Cluster	Individual	Household	Individual	Household	Individual	Household	None	Individual	Household	None

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Discussion

In 2018, one year before the retractions of five criminology articles from prestigious journals in the discipline, the *Journal of Contemporary Criminal Justice* published a special issue in its 34th volume under the title "Replication and Research Integrity in Criminology". The issue included five replication studies of highly cited and influential articles, such as "Neighborhoods and Violent Crime" by Sampson, Raudenbush, and Earls (1997), which according to Google Scholar has been cited more than 11,000 times as by February 2020. The re-analysis of this article by Maxwell et al. (2018) replicated the results in Sampson et al. (1997) almost accurately which is rare in replication studies. In fact, four of the five articles replicated in the special issue confirmed the main results in the original studies, although with varying degrees of precision (see, Savolainen and VanEseltine 2018). My replication study of Sweeten et al. (2009) further adds to this pile of successful replications in the field of criminology and the social sciences. Despite some discrepancies and technical issues, discussed in Section 5, the overall results in the original article were successfully replicated. Furthermore, Section 4 showed that I was able to construct a very similar dataset with comparable summary statistics for almost all variables included by the authors, which is a clear indicator that the authors did not engage in any intentional scientific misconduct or data fabrication.

First, no systematic replication attempts have been carried out in criminology of the sort that have been carried out in other disciplines, such as the Open Science Collaboration in psychology (OSC 2015). As shown by Pridemore et al. (2018) only 0.45% of all articles published in criminology journals are replication studies. A large proportion of the remaining 99.55% of published articles are therefore likely articles that, in some way or the other, test a new hypothesis or an established one in a different context. Therefore, the current publishing culture might lead to a discipline with a large number of poorly tested hypotheses, rather than one with a foundation based on strong, unrefuted hypotheses—in a Popperian sense—that have been tested over and over again. Encouraging researchers to test existing theories and hypotheses by increasing the proportion of replication studies is therefore an essential step towards strengthening the foundation of the discipline.

Second, the retractions of the five criminology studies mentioned in the introduction of my article show that the discipline of criminology is certainly not immune to a replication crisis and that scientific misconduct does exist in criminology and sociology, as it does in any other discipline. Replication studies will help uncover incidents of fraud in existing research disincentivise scientific misconduct in future research.

Third, promoting a replication culture is beneficial for the progress of the discipline, even in the absence of an actual replication crisis in criminology or sociology. Christensen et al. (2019) demonstrate that there are numerous examples of procedures that can easily be implemented to improve research integrity and transparency in the social sciences which do not necessarily involve unrealistic policies from a social science perspective. These easily implementable steps are improving open access to data sources, requiring sharing of scripts (and datasets when possible) from authors upon publication and higher standards on data descriptions and reporting. Today, this last step is even easier to implement because authors can submit supplemental material online.

Although my replication of Sweeten et al. (2009) was successful overall, writing a replication study like this is probably not feasible for most researchers due to the high time cost. Had the authors of the original article submitted their scripts with their article—or described their data construction thoroughly—then the verification part of this exercise would have been straightforward and probably also been a minimal part of my article. This holds for other replication attempts as well. Open access, data sharing, and an improved replication culture means that researchers can focus less on making the numbers match and guessing which variables the authors of the original article used, and more time on sensitivity analyses, alternative specifications, and so on. It will probably also encourage more researchers to carry out replication studies.

7 Conclusion

The objective of this article was to replicate the article "Does dropping out of school mean dropping into delinquency?" by Sweeten, Bushway, and Paternoster (2009). The original article found that there are significant baseline differences between dropouts and non-dropouts in their propensities to engage in criminal activities. That is, individuals who end up dropping out of school at some point are more likely to commit a crime, than individuals who never drop out of school. These differences were captured by the "between-effects"-part of the random-effects model used by the authors. Importantly, when baseline differences were accounted for, the authors main finding was that dropout has no causal effect on crime. In short, dropout does not in itself cause crime – a fact that is confused by comparing otherwise incomparable dropouts and non-dropouts. The causal (zero) effect was captured by the "within-effects"-part of the model. My replication study verified the latter part of the conclusion and showed that there is indeed no effect of dropout on crime—at least in the NLSY97 survey and for the crime and dropouts measured used here— which is the main conclusion of the original article. Like in most other replication studies, some discrepancies were found. The most important one is that my replication study showed that the between-effects (i.e. the baseline differences) are not significant either. My overall conclusion is thus that the main finding of the original article did replicate – even when scrutinised using several sensitivity analyses.

References

Anderson, Melissa S., Brian C. Martinson, and Raymond De Vries (2007). "Normative Dissonance in Science: Results from a National Survey of U.S. Scientists". *Journal of Empirical Research on Human Research Ethics* 2(4), pp. 3–14.

Bell, Andrew, Malcolm Fairbrother, and Kelvyn Jones (2019). "Fixed and random effects models: making an informed choice". *Quality and Quantity* 53(2), pp. 1051–1074. URL: https://doi.org/10.1007/s11135-018-0802-x.

Bryk, Anthony S. and Stephen W. Raudenbush (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, NJ: Sage.

Chawla, Dalmeet Singh (2019). *Quintet of study retractions rocks criminology community*. URL: https://www.sciencemag.org/news/2019/11/quintet-study-retractions-rocks-criminology-community (visited on 02/09/2020).

Christensen, Garret, Jeremy Freese, and Edward Miguel (2019). *Transparent and Reproducible Social Science Research: How to Do Open Science*. 1st ed. Oakland, California: University of California Press.

Cole, Stephen (2004). "Merton's contribution to the sociology of science". *Social Studies of Science* 34(6), pp. 829–844.

Freese, Jeremy and David Peterson (2017). "Replication in Social Science". *Annual Review of Sociology* 43(1), pp. 147–165.

Gelman, Andrew and Eric Loken (2013). "The garden of forking paths: Why multiple comparisons can be a problem, when when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time".

Gerber, Alan S. and Neil Malhotra (2008). "Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?" *Sociological Methods and Research* 37(1), pp. 3–30.

Ioannidis, John P.A. (2005). "Why most published research findings are false". *PLoS Medicine* 2(8), pp. 0696–0701.

Johnson, Brian D., Eric A. Stewart, Justin Pickett, and Marc Gertz (2011). "RETRACTED: Ethnic Threat And Social Control: Examining Public Support For Judicial Use Of Ethnicity In Punishment". *Criminology* 49(2), pp. 401–441.

LaCour, Michael J and Donald P Green (Dec. 2014). "When contact changes minds: An experiment on transmission of support for gay equality". *Science* 346(6215), 1366 LP –1369. URL: http://science.sciencemag.org/content/346/6215/1366.abstract.

Maxwell, Christopher D., Joel H. Garner, and Wesley G. Skogan (2018). "Collective Efficacy and Violence in Chicago Neighborhoods: A Reproduction". *Journal of Contemporary Criminal Justice* 34(3), pp. 245–265.

Merton, Robert K. (1973). "The Normative Structure of Science: Theoretical and Empirical Investigations". In: *The Sociology of Science*. Ed. by Robert K. Merton. Chicago and London: The University of Chicago Press. Chap. 13, pp. 267–278.

OSC, Open Science Collaboration (Aug. 2015). "Estimating the reproducibility of psychological science". *Science* 349(6251), aac4716. URL: http://science.sciencemag.org/content/349/6251/aac4716.abstract.

Pickett, Justin T. (2019). *Why I asked the editors of Criminology to retract Johnson, Stewart, Pickett, and Gertz (2011)*. URL: https://osf.io/preprints/socarxiv/9b2k3/ (visited on 02/13/2020).

Pridemore, William Alex, Matthew C. Makel, and Jonathan A. Plucker (2018). "Replication in Criminology and the Social Sciences". *Annual Review of Criminology* 1(1), pp. 19–38.

Sampson, Robert J, Stephen W Raudenbush, and Felton Earls (1997). "Neighborhoods and Violent Crime : A Multilevel Study of Collective Efficacy Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/2892902". *Science* 277(5328), pp. 918–924.

Savolainen, Jukka and Matthew VanEseltine (2018). "Replication and Research Integrity in Criminology: Introduction to the Special Issue". *Journal of Contemporary Criminal Justice* 34(3), pp. 236–244.

Schunck, Reinhard (2013). "Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models". *Stata Journal* 13(1), pp. 65–76.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant". *Psychological Science* 22(11), pp. 1359–1366.

StataCorp (2019). *Stata: Release 16.* Statistical Software. College Station, TX: StataCorp Lp.

— (2013). *Stata. Release 13.* Statistical Software. College Station, TX: StataCorp Lp.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel (2016). "Increasing Transparency Through a Multiverse Analysis". *Perspectives on Psychological Science* 11(5), pp. 702–712.

Sweeten, Gary, Shawn D. Bushway, and Raymond Paternoster (2009). "Does dropping out of school mean dropping into delinquency?" *Criminology* 47(1), pp. 47–91.

Appendix A Supplementary Tables and Figures

Table A.1: Proportion Enrolled in School (Primary or Secondary) by Age and Wave (Table 1, p. 64 in Sweeten et al. 2009)

	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5		Wave 6	
Age	Rep	Org	Rep	Org	Rep	Org	Rep	Org	Rep	Org	Rep	Org
12	**0.996**	0.996										
13	**0.993**	0.992	**0.963**	0.964								
14	**0.992**	0.993	**0.967**	0.953	**1.000**	1.000						
15	**0.980**	0.982	**0.967**	0.954	**0.970**	0.965	**1.000**	0.962				
16	**0.954**	0.952	**0.937**	0.923	**0.949**	0.944	**0.934**	0.930	**0.926**	0.926		
17	**0.890**	0.891	**0.884**	0.860	**0.854**	0.856	**0.878**	0.883	**0.890**	0.889		
18	**0.929**	0.929	**0.445**	0.443	**0.366**	0.369	**0.410**	0.414	**0.384**	0.390	**0.373**	0.370
19			**0.196**	0.205	**0.080**	0.086	**0.070**	0.072	**0.069**	0.074	**0.059**	0.057
20					**0.039**	0.039	**0.019**	0.019	**0.020**	0.020	**0.019**	0.020
21							**0.007**	0.007	**0.013**	0.013	**0.008**	0.009
22									**0.015**	0.015	**0.008**	0.006
23											**0.000**	0.000
Total	**0.976**	0.976	**0.844**	0.832	**0.667**	0.667	**0.471**	0.473	**0.280**	0.283	**0.093**	0.093

Notes: Rep = Numbers from replication study. Org = Numbers from original study (Sweeten et al. 2009). Note that this table reproduces the "proportions columns" from Table 1 in the original paper. The "numbers part", i.e. sample sizes by age and wave, is reproduced in Table 2.

Table A.2: Comparison of Summary Statistics. All Subsamples (Columns) from Table 2 in Sweeten et al. (2009, p. 65) Included.

	Total Sample		Males		Females		Ever Dropout		Never Dropout	
Crime variety	0.34	(0.83)	0.43	(0.95)	0.24	(0.67)	0.54	(1.09)	0.29	(0.74)
Crime prevalence	0.20	(0.40)	0.25	(0.43)	0.15	(0.36)	0.29	(0.45)	0.18	(0.38)
Dropout	0.11	(0.31)	0.11	(0.32)	0.10	(0.29)	0.51	(0.50)	0.00	(0.00)
Years since dropout	1.15	(1.48)	1.15	(1.47)	1.16	(1.50)	1.15	(1.48)	.	(.)
Male	0.51	(0.50)	1.00	(0.00)	0.00	(0.00)	0.55	(0.50)	0.50	(0.50)
Age	17.45	(2.38)	17.45	(2.39)	17.44	(2.38)	17.42	(2.35)	17.45	(2.39)
White	0.71	(0.46)	0.70	(0.46)	0.71	(0.45)	0.61	(0.49)	0.73	(0.44)
Black	0.15	(0.36)	0.15	(0.36)	0.16	(0.36)	0.22	(0.41)	0.14	(0.34)
Other race	0.01	(0.11)	0.01	(0.11)	0.01	(0.11)	0.01	(0.11)	0.01	(0.11)
Hispanic	0.13	(0.33)	0.13	(0.34)	0.12	(0.33)	0.16	(0.37)	0.12	(0.32)
Live with biological parents	0.54	(0.50)	0.55	(0.50)	0.53	(0.50)	0.33	(0.47)	0.60	(0.49)
Arrests	0.10	(0.55)	0.15	(0.66)	0.06	(0.39)	0.28	(0.97)	0.06	(0.35)
Smoking prevalence	0.42	(0.49)	0.42	(0.49)	0.42	(0.49)	0.60	(0.49)	0.38	(0.48)
Years sexually active	2.36	(2.50)	2.43	(2.59)	2.28	(2.40)	3.44	(2.70)	2.08	(2.36)
Antisocial peer scale	1.71	(1.65)	1.53	(1.61)	1.89	(1.68)	2.14	(1.74)	1.59	(1.61)
Middle-school GPA	2.87	(0.53)	2.79	(0.52)	2.95	(0.52)	2.59	(0.53)	2.94	(0.50)
Ever suspended	0.26	(0.44)	0.34	(0.47)	0.18	(0.39)	0.53	(0.50)	0.19	(0.39)
Ever retained	0.18	(0.39)	0.21	(0.41)	0.15	(0.36)	0.42	(0.49)	0.12	(0.33)
ASVAB: arithmetic reasoning	0.08	(0.86)	0.10	(0.91)	0.06	(0.81)	-0.37	(0.84)	0.20	(0.83)
ASVAB: word knowledge	0.05	(0.81)	0.05	(0.83)	0.04	(0.78)	-0.35	(0.78)	0.15	(0.78)
ASVAB: paragraph comprehension	0.09	(0.82)	-0.01	(0.85)	0.18	(0.78)	-0.37	(0.78)	0.20	(0.79)
ASVAB: math knowledge	0.12	(0.87)	0.07	(0.89)	0.18	(0.85)	-0.45	(0.78)	0.27	(0.83)
Mother dropout	0.17	(0.38)	0.17	(0.38)	0.17	(0.37)	0.34	(0.47)	0.13	(0.33)
Father dropout	0.17	(0.38)	0.18	(0.38)	0.16	(0.37)	0.31	(0.46)	0.13	(0.34)
Received federal aid	0.40	(0.49)	0.39	(0.49)	0.41	(0.49)	0.60	(0.49)	0.35	(0.48)
Outside: nice	0.65	(0.48)	0.64	(0.48)	0.66	(0.47)	0.44	(0.50)	0.70	(0.46)
Outside: fair	0.27	(0.44)	0.28	(0.45)	0.27	(0.44)	0.39	(0.49)	0.24	(0.43)
Outside: poor	0.06	(0.25)	0.06	(0.24)	0.07	(0.25)	0.15	(0.36)	0.04	(0.20)
N (Person waves unweighted)	45,974		23,204		22,770		10,949		35,025	
N (Person waves weighted)	44,248		22,264		21,984		10,268		33,980	
N (Individuals)	8,376		4,268		4,108		2,021		6,355	

Notes: All numbers in the table are from the replication data. For original numbers, please refer to the original paper.

Table A.3: Replication Results for Random-Effects Negative Binomial Estimates for the Variety of Involvement in Delinquent Behaviour (Sweeten et al. 2009, Table 5, p. 72)

	Total Sample	Males	Females
crim_vrty			
Dropout (within-effect)	0.058	0.057	0.027
	(0.078)	(0.093)	(0.143)
Dropout (between-effect)	0.303	0.161	0.416
	(0.227)	(0.270)	(0.428)
Years since last dropout (within-effect)	-0.037	-0.019	-0.066
	(0.032)	(0.037)	(0.064)
Year since last dropout (between-effect)	-0.135	-0.086	-0.185
	(0.088)	(0.104)	(0.166)
N (person waves unweighted)	45,974	23,204	22,770
N (person waves weighted)	44,248	22,264	21,984
N (individuals)	8,376	4,268	4,108

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.4: Replication Results for Random-Effects Logistic Estimates for the Prevalence of Involvement in Delinquent Behaviour (Sweeten et al. 2009, Table 6, p. 72)

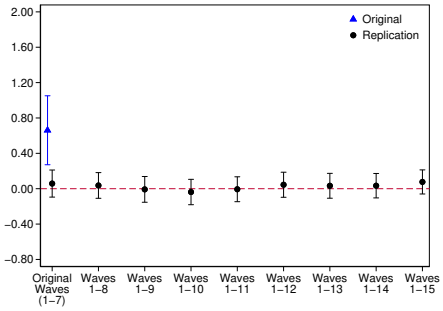
	Total Sample	Males	Females
crim_prvl			
Dropout (within-effect)	0.180	0.236	0.055
	(0.126)	(0.155)	(0.217)
Dropout (between-effect)	0.439	0.127	0.878
	(0.324)	(0.404)	(0.551)
Years since last dropout (within-effect)	-0.082	-0.050	-0.131
	(0.050)	(0.060)	(0.089)
Year since last dropout (between-effect)	-0.192	-0.096	-0.351
	(0.123)	(0.153)	(0.211)
N (person waves unweighted)	45,974	23,204	22,770
N (person waves weighted)	44,248	22,264	21,984
N (individuals)	8,376	4,268	4,108

Standard errors in parentheses

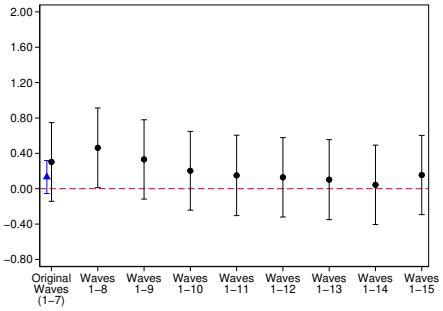
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.1: Sensitivity Analysis Adding One Additional NLSY97 Wave at a Time to the Original Data. Estimation Results for Negative Binomial Model

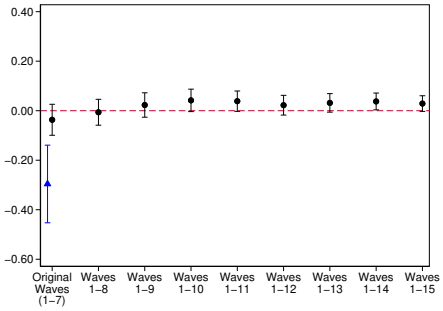
(a) Dropout Variable
Between-Effects Estimates ($\widehat{\theta}_B$)



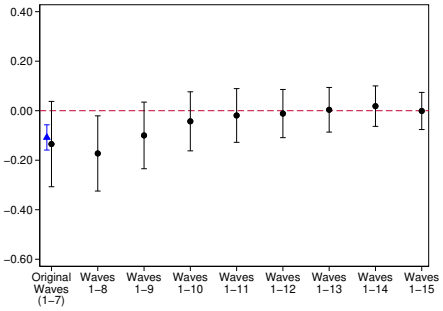
(b) Dropout Variable
Within-Effects Estimates ($\widehat{\theta}_W$)



(c) Years Since Dropout Variable
Between-Effects Estimates ($\widehat{\delta}_B$)



(d) Years Since Dropout Variable
Within-Effects Estimates ($\widehat{\delta}_W$)



Notes: Original estimates (blue triangles) for the random-effects negative binomial regressions are copied from Sweeten et al. (2009, Table 5, p. 72). Replication estimates (black dots) calculated using the Stata command `menbreg`. At each iteration a new wave was added to the data.