American Journal of Criminal Justice (2019) 44:430–456
https://doi.org/10.1007/s12103-018-9466-2

Calibrating Student Perceptions of Punishment: a Specific Test of General Deterrence

Timothy S. Nixon[1] · **J. C. Barnes**[1]

Received: 25 June 2018 / Accepted: 19 November 2018 /
Published online: 5 December 2018

Abstract
General deterrence theory assumes objective risks of punishment and citizens' perceptions of punishment risks are closely calibrated. Yet little empirical attention has been devoted to testing this assumption. Of the few studies that exist, most have tested the calibration with county-level indicators of objective punishment risk. This strategy has been criticized for being too far removed from the individual citizen: why should we expect citizens to know the punishment risks in such a large geographic unit? We estimated the calibration between objective punishment levels and individuals' perceptions of those punishment levels by analyzing data drawn from a large sample of students ($n = 11{,}085$) from 44 schools in Ohio. Multi-level models found the calibration between objective punishment and students' perceptions is weak and not statistically significant. More than half of our calibration estimates were in the wrong direction (i.e., they were negative) and results from interaction tests did not indicate that the calibration is any stronger among those with the highest levels of self-reported offending. We discuss the implications of these findings for policies rooted in general deterrence theory.

Keywords Deterrence · Policy · Perceptions · Punishment

Justifications for punishment often fit into four broad categories: retribution, rehabilitation, incapacitation, and deterrence. Of these, deterrence has probably received the most scholarly attention because findings from deterrence-based research can be directly translated into policy action. *General deterrence* is the familiar idea that punishing deviant individuals will deter other members of the public (i.e., would-be

✉ Timothy S. Nixon
nixonth@mail.uc.edu

J. C. Barnes
barnejr@ucmail.uc.edu

[1] School of Criminal Justice, University of Cincinnati, 2160 McMicken Circle, Cincinnati, OH 45221, USA

offenders) from committing crime. But the question is not whether punishment works for reducing any one person's criminal behavior; instead, the question is whether the criminal justice system, through the threat and use of formal sanction, can measurably alter the prevalence of deviant behavior in a population (Cook, 1980). The prevailing assumption is that changes in behavior among those in the population can be realized by affecting the cost/benefit calculations carried out by would-be offenders (Nagin, 1998; Nagin, Solow, & Lum, 2015).

The deterrent effect of punishment is the primary justification for formal sanctioning that can be found in the writings of Beccaria (1995[1764]), the proverbial father of deterrence theory. Beccaria, concerned with long-term societal change through punishment reform, assumed that the systematization of punishment in society would result in general deterrence. He wrote that a punishment could only be just if it was willed by the people and was executed with the goal of maintaining social order. In his view punishment promotes social order by being "as effective as possible on others and as lenient as possible [on the criminal]" (Beccaria, 1995, p. 48). In other words, punishment "must be pitched at just that level of intensity which suffices to deter men from crime" (Beccaria, 1995, p. 68).

This focus leads naturally to the proposition that increasing objective levels of punishment risk (e.g., increasing the number of police or increasing the number of arrests they make) will decrease criminal activity due to the general deterrent effect of that increased risk (see Pogarsky & Loughran, 2016). Indeed, this is the primary hypothesis espoused by general deterrence theorists (Paternoster, 2010). But Ball (1955) notes that would-be offenders' knowledge of punishment risks is among the most important elements of the general deterrence process. Taking these points together suggests that researchers should place an emphasis on tests of the calibration between objective punishment levels and citizens' perceptions of those punishment risks (for early examples see Grasmick & Bryjak, 1980; Tittle, 1977).[1] Yet, this connection has received only scant empirical attention (Pickett & Roche, 2016a). In other words, general deterrence theory implies that individuals' *perceptions* of punishment are closely calibrated with *objective* levels of punishment, but relatively few empirical studies speak to the strength (or lack thereof) of that calibration (Apel & Nagin, 2017). Nagin (2013a, p. 97) noted that "A major theoretical and empirical gap involves how active criminals and people on the margin of criminality perceive the sanction regime. Deterrence is the behavioral response to perceptions of sanction threats. Establishing the linkage between risk perceptions and actual sanction regimes is imperative." And Apel and Nagin (2017, pp. 125–126) remarked that, "An important research tradition therefore entails estimation of the correspondence between area-level measures (city, county, or state) of criminal punishment and individual perceptions of the certainty and severity of sanction risk." In other words, there remains a need to better understand

[1] While studies of the certainty of apprehension have dominated tests of deterrence theory, it is interesting to note that Beccaria had something very different in mind when he wrote his oft-cited passage that certainty was a more effective deterrent than severity (see, for example, Paternoster, 2010, p. 769). Beccaria believed that when an individual was brought before a magistrate and was certain the magistrate would enforce punishment, certainty was a better deterrent than severity. Since that time, the operationalization of certainty has morphed into what researchers use today, which is a much more general conceptualization that usually includes the probability of being apprehended.

whether and to what extent individuals' perceptions of punishment risk are calibrated with objective measures of punishment risk.

The Calibration Between Objective Punishment Risk & Perceptions of Punishment Risk

The calibration between objective punishment risk and citizens' perceptions of those risks is necessarily a function of the criminal justice system's ability to communicate those risks to the population (Cook, 1980). In other words, if changes in punishment risk are to affect citizens' perceptions of those risks (and, downstream, their behavior), then it is imperative that those citizens be made aware of said changes. This point was illuminated by Cook (1980) more than 30 years ago, but surprisingly there has been little empirical attention paid to it since (except for the four studies discussed below). Cook suggested three mechanisms for threat communication: the media, the visible presence of enforcers, and personal experience and observation. While several studies examine the deterrent effect of arrests (Anwar & Loughran, 2011; Horney & Marshall, 1992; Paternoster, Saltzman, Waldo, & Chiricos, 1983; Sherman & Berk, 1984; Sherman, Schmidt, Rogan, & Smith, 1992), focused deterrence (Braga & Weisburd, 2012), police crackdowns (Rosenfeld & Fornango, 2017; Sherman, 1990), security measures in schools (Tanner-Smith, Fisher, Addington, & Gardella, 2018), or school-based law enforcement officers (Zhang, 2018), periodic reviews indicate few studies have tested the role of threat communication in the formation of citizens' perceptions of punishment risk (Apel, 2013; Nagin, 1998, 2013b; Paternoster, 2010).

Indeed, to date there are four studies that have bearing on this point and can be used to paint an image—albeit perhaps a preliminary version—of the degree to which the criminal justice system impacts citizens' perceptions of punishment risk (Apel, Pogarsky, & Bates, 2009; Kleck, Sever, Li, & Gertz, 2005; Kleck & Barnes, 2014; Lochner, 2007). In one of the first studies to directly examine the issue, Kleck et al. (2005) estimated the correlation (i.e., calibration) between objective punishment risk and perceptions of punishment risk by drawing on a random sample of 1500 respondents from 54 counties across the US. These scholars sought to determine how strongly correlated objective indicators of punishment risk (i.e., the county-level arrest rate, sentencing rate, punishment severity, and punishment swiftness) were with citizens' perceptions of those same outcomes. To gauge perceptions of punishment severity, for example, Kleck et al. (2005) asked the respondents to report the number of crimes—out of 100—in their county that eventually ended in a jail or prison sentence. Their analysis revealed no consistent association between objective sentencing rates and perceptions of those sentencing rates.[2]

[2] Kleck et al. (2005) also studied the calibration between measures tapping into certainty and celerity. The substantive pattern of findings was identical to the results reported in the text when those other constructs were the analytic focus. Moreover, wondering if offenders may be more in tune with objective punishment risks than non-offenders, Kleck et al. (2005) analyzed whether the relationship between objective punishment and perceptions of punishment was moderated by arrest history. The results did not support this hypothesis either: for arrestees, only the assault clearance rate significantly predicted perceptions of assault arrest rates, but the association was in the opposite direction of what would be hypothesized by deterrence theory.

In a related analysis, Lochner (2007) estimated the calibration between county-level objective and perceived punishment risk among respondents in the National Longitudinal Survey of Youth. The analysis revealed a statistically significant, but substantively weak, positive relationship between the clearance rate for motor vehicle theft and individual perceptions of risk. But that relationship disappeared when covariates were added to the model, suggesting the original bivariate association may have picked up on other selection factors that caused perceptions to spuriously correlate with county-level indicators of punishment risk.

But Lochner (2007) took his analysis one important step further. Research from behavioral economics suggests that independent assessments of something like the risk of punishment for a specific crime may have only a loose association with the objective value being estimated. But when asked to rate things in a relative manner—as in comparing the risk of punishment between two crimes—individuals' perceptions are organized coherently (Ariely, Lowenstein, & Prelec, 2003; Thomas, Hamilton, & Loughran, 2017). Building on these observations, one might suspect that changes in objective risk will align with changes in perceptions of risk. Citizens may not be well informed about the prevailing levels of punishment risk, but they may be in tune enough to pick up on shifts in risk levels (Pogarsky & Loughran, 2016). Lochner (2007) sought to address this research question, but his statistical models failed to detect a statistically significant association between changes in objective levels of punishment and changes in perceptions of punishment risk.

The evidence gleaned from these two studies (Kleck et al., 2005; Lochner, 2007) suggests there is a weak-to-nonexistent calibration between objective levels of punishment risk and citizens perceptions of punishment risk, at least when objective risk is measured at the county level. But this does not rule out other ways that objective punishment risks might be connected to perceptions. For instance, rather than assuming individuals are in tune with an esoteric statistic like the arrest or sentencing rate—something many police officers or even criminologists are unlikely to know with any accuracy—one might ask whether there are other indicators of punishment risk that the public *is* in touch with. This was the motivating research question underlying the analysis carried out by Kleck and Barnes (2014). Using the same sample of individuals from the Kleck et al. (2005) study, Kleck and Barnes (2014) analyzed whether perceptions of arrest risk were associated with the number of sworn police officers in their county. Based on the idea that police officers act as sentinels—which is the term Nagin (1998) used to define the role of police as capable guardians—it is logical to argue that citizens living in areas with relatively high levels of police strength (i.e., more police) will report higher perceptions of arrest risk compared to citizens living in areas with low levels of police strength. But, like the results reported above, Kleck and Barnes (2014) found no evidence of a statistical association between the number of police in an area and local citizens' perceptions of arrest risk.

Drawing on the evidence from these three studies, one might be inclined to conclude there is no association between objective punishment risks and citizens' perceptions of punishment risk (Kleck, 2016). But some scholars are hesitant to draw this conclusion because of certain limitations with the above-mentioned studies (Apel, 2013; Nagin, 2013b). For instance, the research relied on measures of punishment risk that were collected at the county-level. But a citizen, criminal or not, may have little reason to be aware of county-wide punishment risks. In addition, there is reason to suspect a county does not have a single sanctioning policy, meaning a county-level measure may mask important variation. Pogarsky (2009) points out that police activity is organized at the municipal, not the county level. Thus, the units of analysis may not have been well suited to answering the research questions posed in those prior studies.

There is another reason to anticipate that the calibration might be weaker at the county-level than it would at a smaller level of aggregation. As theorized by Geerken and Gove (1975), an "open system"—a system with uncontrolled communication and, therefore, one that has a difficult time sending the deterrence message to the population—is unlikely to see much of a deterrent effect. The county-level would seem to reflect an open system given the difficulties of communicating accurate information to such a large group of people. A "closed system"—one where official communication can be effectively and efficiently disseminated to the target population—is much more likely to realize general deterrence because the gap between perceptions and reality should be smaller.

This argument concerning the communication of punishment risks provides the key point upon which the present effort will build. If the county represents an "open system", then we should not anticipate much of a calibration between objective punishment risks and perceptions of punishment risk. We might, however, anticipate a strong calibration when the unit of analysis is smaller; for example, a school.

Apel and colleagues (Apel et al., 2009, p. 220) recognized the appeal of studying the calibration between objective punishment risks and students' perceptions at the school-level:

> [Schools] have clearly defined members, distinct spatial boundaries, formal rules for behavior, well known sanctions for wrongdoing, and efficient avenues for transmitting information from authorities to subjects. They also occupy the time of their members for a nontrivial number of hours each day. On the other hand, schools are not necessarily a microcosm of the community at large, as they are obviously of smaller size and less organizational complexity. Additionally, the subjects of deterrence messages (students) are, presumably, more impressionable by virtue of their immaturity. Although schools as settings for the study of deterrence are obviously limited in a number of important ways, what is gained is an ability to "control" for a variety of confounding influences that would otherwise contaminate any estimate of the correlation between sanctions and perceptions.

Drawing on data from the National Education Longitudinal Study (NELS), Apel et al. (2009) analyzed the relationship between school-level indicators of punishment risk and the students' perceptions of those risks. Their cross-sectional models indicated schools with more disciplinary practices led to a small ($b = 0.04$, where the outcome variable ranged from 0 to 4) increase in perceptions of sanction risk among students. But respondents who were enrolled in schools that were more likely to suspend students for first time violations did not report higher perceptions of risk when compared to students in schools with more lenient policies. Nonetheless, after controlling for a host of possible confounders, Apel and colleagues (Apel et al., 2009) reported a small but statistically significant association between a composite measure of sanction enforcement and student perceptions of rule strictness.

It thus appears there may be some truth to the argument that the calibration between objective punishment risks and perceptions of risk is stronger in a "closed system" (e.g., school, the peer group, or even the home environment) compared to an "open system" (e.g., county, state). But with only one study analyzing a closed system (Apel et al., 2009), more evidence is needed before scholars can say with any certainty that closed systems induce stronger calibrations. The present study will address this need by analyzing the calibration between objective punishment risks and individual perceptions of punishment risks among a large body of students attending schools in a large county in Ohio.

The Current Study

We seek to provide an estimate of the calibration between objective levels of punishment risk and individuals' perceptions of those punishment risks. The degree of calibration that might exist was recently the source of a scholarly debate between Pickett and Roche (2016a, 2016b) and Nagin (2016). Pickett and Roche (2016a) noted that much previous work has assumed a perfect calibration, but the findings from the four studies discussed above (Apel et al., 2009; Kleck et al., 2005; Kleck & Barnes, 2014; Lochner, 2007) suggest that assumption may not be tenable or, at a minimum, requires qualification. One such qualifier may be that the strength of calibration is contingent on the level of aggregation in the objective risk measure. As the level of aggregation decreases (e.g., moving from a county to the school-level), we might expect to see the calibration increase.

We submit that with only four studies available, the evidence for or against the conclusion that perceptions are only weakly (or not at all) related to objective risk levels is somewhat premature. We seek to offer an additional piece of evidence that can be used to sort out how closely calibrated individuals' perceptions are with objective risk levels. But our study provides more than just an additional data point in the debate over the salience of deterrence-based policy. Based on arguments against using county-level data, we should expect a study carried out in a small unit like a school system to set something of a high-water mark for what the calibration might be.

With these points in mind, we offer Fig. 1 as a simplified theoretical model that captures the goal of our study and can help to situate some of the complexities that will be revisited later. The figure is divided into two levels—the institutional level and the individual level. The institutional level captures changes that occur in an institution like a school. Here, it is expected that actions taken by the institution will change the objective levels of punishment risks that prevail in that institution. Some actions may increase these risks (e.g., more cops on the street or more school resource officers in the halls) and others may decrease them.

The level of punishment risk must then impact individual-level perceptions of those same punishment risks through communication of the risks to the citizens. Such communication can take several forms such as direct advertisement (e.g., an announcement made over an intercom) or indirectly through vicarious communication (e.g., students hear about increased risks by others who have recently been punished). The diagram in Fig. 1 captures this relationship with the arrow from punishment risk to perception of punishment risk. Notice that we label this relationship with ϕ. The ϕ

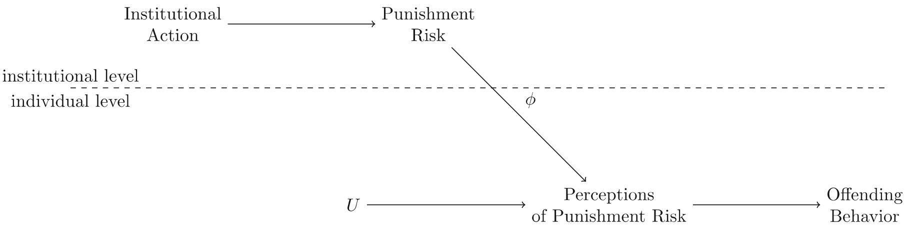

Fig. 1 Theoretical diagram

coefficient reflects the average level of calibration between actual punishment risks and perceptions of those punishment risks. The primary goal of the current study is to garner an estimate of ϕ at the school level. Recall that a few prior studies have estimated ϕ and shown it to be weak when analyzing county-level punishment risks (Kleck & Barnes, 2014; Kleck et al., 2005; Lochner, 2007). And one study suggested ϕ is statistically significant but substantively weak at the school-level (Apel et al., 2009). With only one prior study estimating ϕ at the school-level, the present study contributes much needed additional information.

Two other points from Fig. 1 are worth mentioning. First, we are not arguing that aggregate punishment risks are the only source of variation in perceptions of punishment risk. On the contrary, Fig. 1 includes U as a variable to catch all other sources of variance for individual-level perceptions. In this case, U would capture all other factors that impact one's beliefs such as his/her baseline level of risk seeking and risk perceptiveness (e.g., Jacobs, 2010), along with the impact of any heuristics and cognitive biases that might affect one's judgments (Pogarsky, Roche, & Pickett, 2017). By including U, we are explicitly stating that many factors impact perceptions of risk. Our study simply asks whether actual punishment risks are one such source of variance and, if so, how large of an impact they have. Note that our model assumes these individual-level sources of variance U are uncorrelated with actual punishment risks (i.e., there is no correlation between U and punishment risk in Fig. 1). This assumption is reasonable because actual punishment risks, at least as they are defined in this study, will come from school-level policies about the typical response to a student's misbehavior. Thus, the actual punishment risks in our study are assumed exogenous to all individual-level factors U that might determine student perceptions. Nonetheless, we will include controls for a host of factors that might be captured by U.

The second point to note is that the diagram in Fig. 1 terminates in offending behavior. A primary premise of deterrence theory—and, indeed other theories like Gottfredson and Hirschi's (1990) discussion of self-control—is that perceptions of punishment risk will inversely predict offending behaviors such that individuals who perceive lower risks will be more likely to offend, all else being equal (see also Jacobs, 2010). The present study will not address that last link but there are numerous other reviews that provide insight into the strength of this association (e.g., see Apel & Nagin, 2017, pp. 125–128; Nagin, 2013b, pp. 246–252).

Based on the literature discussed above and the basic assumptions of general deterrence theory, we test the following hypothesis:

> School-level sanctions are significantly correlated with student-level perceptions of those sanctions.

Methods

Data

Data for the current study were drawn from a survey of 44 junior and senior high schools representative of a large urban county in Ohio. In February 1994, half of the students attending these schools were randomly selected and administered a questionnaire.[3] Principals were given a similar questionnaire that included closed and open-ended queries regarding demographics, victimization rates, gang violence, and disciplinary policies. Of the public schools approached, 15 of 17 junior high schools and 16 of 18 senior high schools agreed to participate. Private schools participated at a lower rate: 4 of 6 for high schools but 8 of 33 junior high schools. The 44th school was a private, nondenominational school. Approximately 35% of all students in the schools participated, for a total of $n = 11{,}085$ completed student surveys. Only one principal failed to return a questionnaire, but data for that survey was supplemented—where possible—by district files for the school. These data were originally collected by Lab and Clark (2006) and more information about the data and measures can be found in their publication or on the ICPSR page that houses the data (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/2587?searchSource=revise&q=2587).

Measures

Student-Level Perceptions of Punishments All students were asked, "Most of the time, what happens to a student who is caught…" followed by prompts for six behaviors: 1) disrupting class; 2) being disrespectful to a teacher; 3) cutting classes; 4) fighting with other students; 5) drinking, being drunk, or using illegal drugs at school; and 6) with a weapon at school. Response options were provided in an ordinal list where the first option was the least severe disciplinary (non)action and the last was the most severe: 1 = *nothing*, 2 = *discipline by teacher*, 3 = *sent to the principal's office*, 4 = *parents are notified*, 5 = *detention*, 6 = *suspension or expulsion*, and 7 = *police are called*.[4]

Our use of an ordinal measure to tap into perceptions of risk might be criticized for lacking variation or even because it may be considered a measurement strategy that will capture high levels of measurement error (see, generally, Braga & Apel, 2016). With this in mind, it is worth pointing out that similar ordinal response variables have been used to capture perceptions of risk in previous analyses (e.g. Apel et al., 2009). Nonetheless, we report on several sensitivity analyses and consider various

[3] It is important to acknowledge the age of these data. But it is equally important to point out that there is little reason to suspect the age of the data will systematically strengthen or weaken the calibration (ϕ) between objective punishment risk and students' perceptions. True, modern advances in technology will accelerate the speed with which information is communicated (e.g., peers may find out that a friend has been punished more quickly now due to social media), but we are reluctant to say that the accuracy of the information communicated has changed in the last two decades. These data, despite their age, afforded several unique benefits that were not available in other data sources. Thus, we believe the benefits outweigh these potential concerns.

[4] Students were also allowed to select "something else", which for the present purposes was coded as a missing value. Five or fewer students selected this option for each of the disruptive behaviors referenced in the text.

perspectives when interpreting the findings presented below. These sensitivity analyses are described in the "Results" section.

In addition, one might consider the ordinal measure to be one of the closest representations of what Beccaria (1995[1764]) actually had in mind when he spoke of the certainty of punishment. To summarize Beccaria's position, he was concerned about the certainty of punishment *after* one is apprehended for a crime. Certainty of being punished after apprehension—not certainty of apprehension itself—was most important. Aligning this with our present analysis, it might be argued that our measure closely represents this conceptualization because students were asked to report on the most likely outcome for a hypothetical student *who had been caught*.

School-Level Predictors (Objective Punishments) Information was drawn from the principal survey to measure objective levels of punishment risk. The principal at each of the $j = 44$ schools included in the analysis was asked questions identical to those asked to the students, providing a unique opportunity to directly compare the two. Principals were asked, "indicate the most common [penalty] assessed by the school" for the same behaviors in the student questionnaire: 1) disrupting class; 2) being disrespectful to a teacher; 3) cutting classes; 4) fighting with other students; 5) drinking, being drunk, or using illegal drugs at school; and 6) with a weapon at school. Response options were provided in an ordinal list where the first option was the least severe disciplinary (non)action and the last was the most severe: 1 = *nothing*, 2 = *discipline by teacher*, 3 = *sent to the principal's office*, 4 = *parents are notified*, 5 = *detention*, 6 = *suspension*, and 7 = *police are called*.[5]

This measurement strategy assumes principals are aware of the punishment practices that take place in their school and that they accurately reported them on the surveys. We are not the first to rely on that assumption. Other studies have drawn on similar strategies to tap into the level and severity of sanctions in schools (Apel et al., 2009). For example, Maimon, Antonaccio, and French (2012) drew on school administrator reports of the most likely response to a student caught fighting with another student and the response options ranged from no policy in place to expulsion.

Self-Reported Delinquency Scholars have long noted the possibility of an experiential effect on the formation of risk perceptions (Kahneman, 2011; Lochner, 2007; Stafford & Warr, 1993). Individuals who commit crime are more likely to be aware of the reality that most offenses go unpunished, thus allowing offenders to have a closer calibration between objective punishment risk and perceptions of punishment risk. To account for this possibility, we constructed a measure of self-reported delinquency to include as a student-level control variable. We also use this measure to test for a statistical interaction between self-reported delinquency and objective punishment risks. There is reason to suspect calibration (ϕ) will be higher among those students who are more involved in delinquent activity.

Students were asked to indicate how often they engaged in 14 different delinquent acts. The behaviors ranged from minor theft and school vandalism to assault with a weapon and assault of a teacher. Responses to each item were coded such that 0 =

[5] Principals also had an "other" option. For the purposes of this analysis "other" was coded as missing. Only one variable (bring weapons to school) had cases where "other" was chosen.

never, 1 = *once or twice*, 2 = *once a month*, 3 = *once every 2–3 weeks*, 4 = *once a week*, and 5 = *more than once a week*. Responses were summed to create an index of self-reported delinquency.

Self-Reported Drug Use Students were asked to self-report the frequency with which they used eight different illicit drugs. Questions covered relatively minor substances (such as alcohol and marijuana) as well as serious drugs (such as cocaine). The response options were coded such that 0 = *never*, 1 = *once or twice*, 2 = *once a month*, 3 = *once every 2–3 weeks*, 4 = *once a week*, and 5 = *more than once a week*. Responses to each item were summed to create an index of self-reported drug use.

Self-Reported Victimization Students who have previously been victimized may have different risk perceptions than students who have not been victimized (Pickett, Mancini, Mears, & Gertz, 2015). Thus, we accounted for each student's victimization experiences. Students reported the frequency with which they experienced four types of victimization including robbery, theft, assault, and being teased or bothered. Responses were coded such that 1 = *never*, 2 = *once*, 3 = *2–3 times*, 4 = *4–5 times*, and 5 = *6+ times*. These items were then summed to generate an index of victimization experiences.

Female Respondent sex was coded 0 = *male* and 1 = *female*.

Age Student age was coded as a count variable reflecting the student's age at the time of the interview.

Race Dummy variables were created to capture the racial/ethnic background of each student. The following racial/ethnic categories were captured: American Indian, Asian-American, Black, Hispanic, White, and other.

Grades Students were asked to report their typical grade performance. Responses were coded 1 = mostly *F's*, 2 = mostly *D's*, 3 = mostly *C's*, 4 = mostly *B's*, 5 = mostly *A's*.

School Size As Apel and colleagues (Apel et al., 2009) astutely noted, school size may confound students' perceptions of punishment risk. To account for this, we included a measure of school size that was taken from the principal questionnaire. The variable is a count of the total number of students in the school.

High School Schools were dummy coded *0 = middle school* and *1 = high school*. A school was identified as a high school if it served students in the 9th through 12th grades.

Number of Disrespecting Incidents Principals were asked to use official records to report the frequency of students threatening teachers since the start of the school year. Responses were coded so that 0 = *never*, 1 = *once or twice*, 2 = *once a month*, 3 = *once every 2–3 weeks*, 4 = *once a week*, and 5 = *more than once a week*.

Number of Fights Principals were asked to use official records to report the frequency of gang fights, assaults on students, assaults on teachers, and assaults with a weapon in their school since the start of the school year. Responses were coded so that 0 = *never*, 1 = *once or twice*, 2 = *once a month*, 3 = *once every 2–3 weeks*, 4 = *once a week*, and 5 = *more than once a week*. The four items were summed to create a scale of officially recorded fights and assaults.

Weapons Brought to School Principals were asked to report the frequency with which students brought weapons to school. Responses were coded so that 0 = *never*, 1 = *once or twice*, 2 = *once a month*, 3 = *once every 2–3 weeks*, 4 = *once a week*, and 5 = *more than once a week*.

Analysis Plan

The structure of our data were such that students ($i = 11{,}085$) were nested within schools ($j = 44$). As such, a multi-level modeling strategy was the most appropriate analytic approach. We performed several preliminary analyses to determine which of the generalized linear models (GLM) was most appropriate for our ordinal-coded dependent variables. These analyses indicated that a multi-level model fit with the identity link function did not substantively alter the conclusions compared to a binomial model with a logit link. Additionally, recognizing the identity link allowed for a direct interpretation of the intercept and retained the substantive meaning of the coefficients led us to determine that the multi-level models were best fit with the identity link function.

The analyses presented below provide estimates from multi-level models where the level-1 outcome is the student-level perceptions of punishment and the primary predictor of focus is the school-level objective punishment. Generally, then, the multi-level models can be expressed as two linked equations:

$$Y_{ijk} = \pi_{0jk} + \sum_{q=1}^{Q} \pi_{qjk}\left(q_{ijk}\right) + \varepsilon_{ijk} \tag{1}$$

$$\pi_{0jk} = \phi_{00k} + \phi_{01k}\left(ObjPunish_j\right) + \sum_{x=1}^{X} \phi_{xk}\left(x_{jk}\right) + u_{0jk} \tag{2}$$

where the response Y is captured for each respondent i from school j for the behavioral item k of focus. We have six different behaviors that will be used as dependent variables, so $k = \{1, 2, \ldots, 6\}$. The first equation reveals the level-1 model, which is a linear model that allows Y_{ijk} to be predicted as function of π_{0jk}, which is the model intercept but in this context captures the average level of risk reported by students across all schools when the covariates are set to zero. The term $\sum_{q=1}^{Q} \pi_{qjk}\left(q_{ijk}\right)$ captures the collective influence of all the level-1 covariates on Y.

The second equation forms the school-level (i.e., level-2) portion of the model. Here, we see the intercept from the level-1 model, π_{0jk}, is the outcome that will be predicted by objective punishment risks through estimation of ϕ_{01k}. In words, ϕ_{01k}reveals the degree to which the average student perception of punishment risk captured by the intercept of the level-1 model, π_{0jk}, covaries with objective punishment risk. The coefficient estimate for ϕ_{01k}will be the primary estimate of focus for this analysis because it provides a direct estimate of the calibration between perceptions of punishment risk and objective punishment risk. Because the outcome variables were coded on the same ordinal scale, we can interpret the coefficient estimates for ϕ_{01k}in terms of how close to (or far from) 1.00 they are. The closer to 1.00 they are, the more closely calibrated are objective punishment risk and perceptions of punishment risk. The closer to 0.00, the weaker the calibration.

Before moving to the results it is important to note that we experimented with different centering strategies for the right-hand side variables. All school level variables were grand-mean centered and student-level variables were group-mean centered (with the exception of race). Substantive results for the focal relationships were unchanged when alternative centering strategies were employed.

Results

Sample descriptive statistics are provided in Table 1. The sample closely matched expectations for a midwestern school system. Respondents were 15 years old on average, while the youngest students were 11 and the oldest were 19. The vast majority of students were White (70%), with Black students comprising the next largest racial/ethnic group (15%). Each perception of punishment risk variable had a standard deviation around 1.5, revealing that there was sufficient variation in risk perceptions to justify the analysis.

Principals had slightly more homogenous responses for punishment risks. None reported calling the police (a score of 7) would occur for disrupting class, disrepecting a teacher, or cutting class. None reported nothing would happen (a score of 1) for disrespecting a teacher, fighting, alcohol/drug use, or bringing a weapon to school. But the pattern of means reveals the principals responded as expected. The least serious offense—disrupting class—had the lowest average objective punishment risk and the most serious offense—bringing a weapon to school—had the highest. Perhaps most critical, though, is the point that the principal data had substantively meaningful levels of variation. As the histograms in Fig. 2 reveal, principals did not always report the same values of objective punishment. This finding is consistent with prior work that shows meaningful variation in punisment risk at the between-school level (Welch & Payne, 2010). In other words, the measures of objective punishment are not acting as constants. Note, however, that the variation does become restricted for the most severe forms of misbehavior (i.e., alcohol/drug use and carrying a weapon to school). This is to be expected, but it also suggests the results from those models should be interpreted with due caution because of the relatively limited variation in objective punishment risks.

The first stage of the analysis examined the bivariate relationship between the average student perception (within each school) of punishment levels and objective punishment levels. Only one of six correlations—for fighting—was statistically significant ($r = 0.415$, $P < 0.05$). A heatmap (Fig. 3) facilitates the visualization of these bivariate results. In the figure, dark shades represent locations where there was a

Table 1 Descriptive statistics

Variable	n	Mean	Standard Deviation	Minimum	Maximum
Student-level perceptions					
Disrupt class	10,659	3.00	1.56	1	7
Disrespect teacher	10,744	3.60	1.66	1	7
Cut class	10,655	4.63	1.50	1	7
Fighting	10,678	5.46	1.28	1	7
Alcohol/Drug use	10,353	5.80	1.29	1	7
Bring weapon to school	10,303	6.08	1.23	1	7
School-level predictors					
Disrupt class	42	3.81	1.53	1	6
Disrespect teacher	40	4.40	1.45	2	6
Cut class	39	4.54	1.41	1	6
Fighting	40	5.38	1.21	2	7
Alcohol/Drug use	38	6.03	1.24	2	7
Bring weapon to school	33	6.58	1.12	2	7
Student-level controls					
Self-reported delinquency	9373	2.88	9.37	0	68
Self-reported drug use	9668	4.22	7.26	0	40
Self-reported victimization	9140	6.64	2.83	4	20
Female	10,860	0.45	–	0	1
Age	10,848	15.08	1.72	11	19
American Indian	10,793	0.02	–	0	1
Asian-American	10,793	0.02	–	0	1
Black	10,793	0.15	–	0	1
Hispanic	10,793	0.05	–	0	1
White	10,793	0.70	–	0	1
Other	10,793	0.06	–	0	1
School-level controls					
School size	44	669.36	471.31	33	1593
High school	44	0.48	–	0	1
Number of disrespect incidents	44	0.52	0.73	0	3
Number of fights	42	2.21	1.31	0	5
Weapons brought to school	44	0.36	–	0	1

relatively high proportion of responses. A significant correlation—that is, general agreement on punishment levels—would appear as a diagonal line of dark squares starting in the bottom-left corner and extending to the top-right corner of each plot. None of the panels in Fig. 3 demonstrate this relationship. Instead, each of the heatmaps show just how little agreement there was between student perceptions and principal reports. This can be seen by the scattering of dark panels throughout each of the plots.

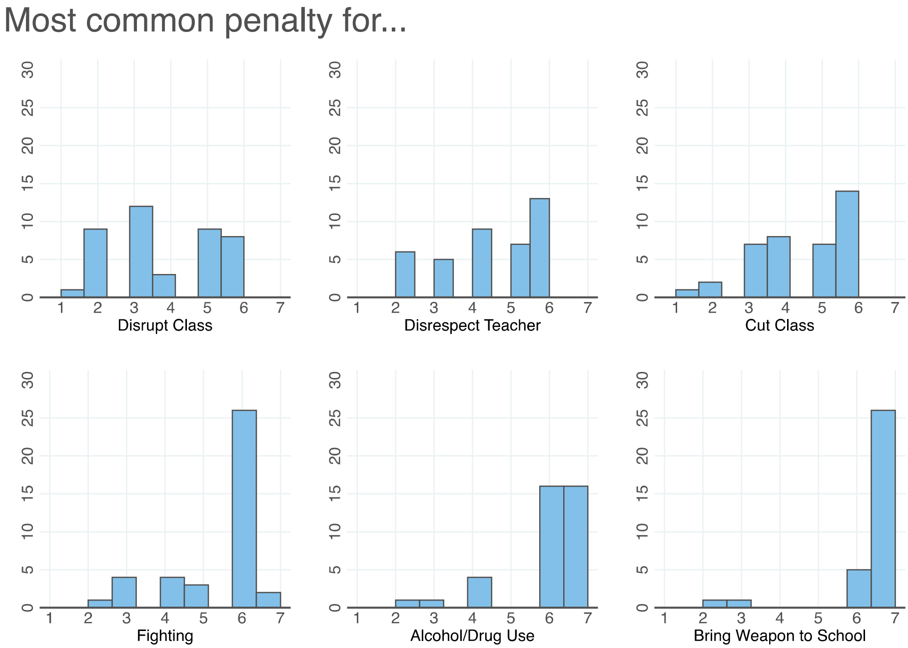

Fig. 2 Histograms of principal reports of punishment risks

Note also that the significant bivariate relationship for the fighting variable is driven by the student/principal agreement that suspension was a common punishment.

Before turning to the results of the multi-level models, it is important to note that we performed a series of additional preliminary analyses to assess whether the outcome variables were appropriate for analysis. Most important is that we analyzed each of the dependent variables to ensure there was sufficient variation at level-2 (i.e., the school-level) to warrant a multi-level analysis. We relied on significance levels and the intra-class correlation coefficient (ρ) calculated from an unconditional (meaning no covariates were included) random intercept model. In this context, ρ provides an estimate of the proportion of variance in the outcome variable that is attributable to the between-school (i.e., level-2) portion of the model. The estimates of ρ for each outcome were as follows: disrupt class ρ= 0.034 (95% confidence interval = 0.02, 0.06); disrespect teacher ρ= 0.02 (95% confidence interval = 0.01, 0.03); cut class ρ= 0.03 (95% confidence interval = 0.02, 0.05); fighting ρ= 0.07 (95% confidence interval = 0.04, 0.11); alcohol/drug use ρ= 0.02 (95% confidence interval = 0.01, 0.03); and bring weapon ρ= 0.03 (95% confidence interval = 0.01, 0.05).

An important point emerged: while the amount of variation between schools is statistically significant, the bulk of the variation appears *within*-schools. Put a different way, variation in perceptions of punishment risks is largely a between-student concern, not a between-school concern. This preliminarily indicates there is little variation in perceptions of punishment risk that can be attributed to actual variation in objective punishment risk because objective punishment risk is a *between*-school factor. Nonetheless, rather than relying on this extrapolation, it was important to estimate the full multi-level model(s) to garner an actual estimate of the calibration.

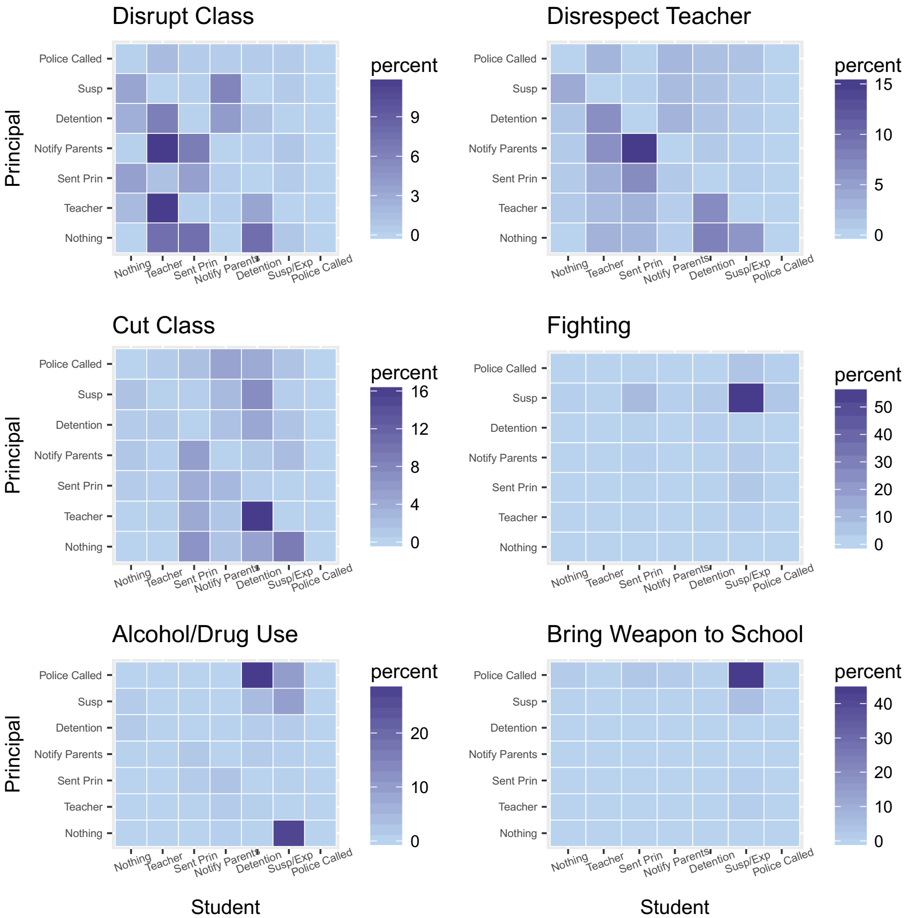

Fig. 3 Heat maps showing the relationship between principal and student reports

The results of these models can be found in Table 2 and the estimates for the calibration can be found under the school-level predictor heading. Each measure of objective punishment risk was included as an independent variable in the model that corresponded with the behavior referenced by the outcome measure.

Table 2 reveals that none of the six multi-level models revealed a statistically (or substantively) significant association between objective punishment risk and average student perceptions of punishment risk.

But, despite this pattern of null results, it remains important to substantively interpret certain aspects of these models. Doing so may serve to discount various arguments for why we report null results for the calibration estimates. First, we direct attention to the estimated intercept values. For example, the first model, which uses disrupting class as the outcome, produced an intercept value of 3.335. Given the centering scheme described in the analysis plan, this value reflects the expected value of perceptions of punishment risk for a Black student in a middle school of average size and average objective punishment risk who scored his or her school's average for delinquency, drug use, victimization, sex, and age. Notice that a clear pattern emerges, such that the

Table 2 Multi-level model results of the regression of student perceptions on principal reports and covariates

	Disrupt class		Disrespect teacher		Cut class		Fighting		Alcohol/Drug use		Bring weapon to school	
	b	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error
Intercept	3.335***	[0.072]	3.958***	[0.068]	4.717***	[0.069]	5.494***	[0.076]	5.732***	[0.052]	6.125***	[0.064]
School-level predictors												
Disrupt class	−0.026	[0.036]										
Disrespect teacher			0.008	[0.031]								
Cut class					−0.016	[0.034]						
Fighting							0.014	[0.060]				
Alcohol/Drug use									−0.011	[0.027]		
Bring weapon to school											−0.070	[0.041]
Student-level controls												
Self-reported delinquency	−0.042	[0.043]	0.073	[0.046]	−0.027	[0.042]	0.033	[0.036]	−0.020	[0.036]	0.031	[0.038]
Self-reported drug use	0.002	[0.003]	0.002	[0.003]	-0.005*	[0.003]	−0.007**	[0.002]	−0.018***	[0.002]	−0.027***	[0.002]
Self-reported victimization	0.008	[0.007]	0.022**	[0.008]	0.004	[0.007]	−0.009	[0.006]	−0.039***	[0.006]	−0.034***	[0.006]
Female	−0.018	[0.037]	0.096*	[0.039]	0.095**	[0.036]	−0.007	[0.030]	0.001	[0.030]	−0.046	[0.032]
Age	−0.039*	[0.015]	−0.040*	[0.016]	−0.071***	[0.015]	0.002	[0.012]	−0.037**	[0.013]	−0.019	[0.013]
Grades	−0.069***	[0.019]	−0.141***	[0.020]	−0.068***	[0.019]	−0.006	[0.016]	−0.047**	[0.016]	−0.006	[0.016]
American Indian	−0.142	[0.142]	−0.360*	[0.151]	0.020	[0.138]	0.158	[0.118]	0.323**	[0.117]	−0.028	[0.128]
Asian-American	−0.332*	[0.168]	−0.550**	[0.178]	−0.096	[0.165]	0.075	[0.138]	−0.145	[0.141]	−0.112	[0.152]
Hispanic	−0.302**	[0.096]	−0.375***	[0.102]	−0.192*	[0.094]	0.061	[0.080]	0.169*	[0.079]	0.022	[0.088]
White	−0.453***	[0.061]	−0.463***	[0.065]	−0.079	[0.060]	0.043	[0.052]	0.165***	[0.050]	−0.018	[0.057]

Table 2 (continued)

	Disrupt class		Disrespect teacher		Cut class		Fighting		Alcohol/Drug use		Bring weapon to school	
	b	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error
Other	−0.195*	[0.096]	−0.288**	[0.103]	−0.097	[0.095]	−0.011	[0.080]	0.076	[0.080]	−0.116	[0.086]
School-level controls												
School size	0.000	[0.000]	0.000	[0.000]	0.000*	[0.000]	0.001**	[0.000]	0.000*	[0.000]	0.000**	[0.000]
High school	−0.097	[0.143]	−0.057	[0.116]	−0.214	[0.130]	−0.167	[0.190]	−0.266**	[0.091]	−0.262*	[0.132]
Number disrespect incidents												
Number fights												
Weapons brought												
Random effects	Standard Deviation	Standard Error	Standard Deviation	Standard Error	Standard Deviation	Standard Error	Standard Deviation	Standard Error	Standard Deviation	Standard Error	Standard Deviation	Standard Error
Variance in intercept	0.281	[0.039]	0.21	[0.035]	0.266	[0.04]	0.324	[0.042]	0.164	[0.026]	0.201	[0.034]
Level-1 error variance	1.495	[0.012]	1.583	[0.013]	1.434	[0.012]	1.184	[0.01]	1.96	[0.01]	1.11	[0.01]
Rho, ρ (unconditional model)	0.034		0.017		0.031		0.069		0.016	0.028		
Rho, ρ (full model)	0.043		0.022		0.038		0.144		0.013	0.034		
n (students) [schools]	(7419) [42]		(7310) [40]		(7235) [39]		(6941) [38]		(7052) [38]	(5704) [33]		

$^{*}p < 0.05$, $^{**}\ p < 0.01$, $^{***}p < 0.001$

a. Standard errors in brackets; b. School delinquency is a model-specific item (see text); c. Black students served as the reference category for the race/ethnic variables intercept values increase steadily from left to right, indicating that students' perceptions of punishment track with the seriousness of the offense in question. This pattern would not be expected if students either failed to understand the question, if they intentionally undermined the survey, or if they answered at random. Instead, this pattern, coupled with the absence of a calibration, suggests there may be some truth to the argument that individuals are not adept at judging individual risks but that they can make relative assessments (Thomas et al., 2017). We return to this point in the "Discussion" section.

Second, we should assess the degree to which certain covariates have effects that would be expected. Age was a negative and statistically significant factor in four of six models. One might argue that age in this analysis serves as a proxy for the experiential effect. As adolescents move through the teenage years, experiences related to their beginning to engage in delinquency would likely show up as a negative influence because those experience often teach students that punishment risks are lower than they anticipated. Third, self-reported drug use was a negative and statistically significant predictor of punishment perceptions in four of six models. As with the age covariate, it is likely that self-reported delinquency is tapping into the experiential effect.

Finally, race exerted a significant effect in the first two models, but had a less consistent impact in the last four models. Black students served as the reference category, so the negative coefficients observed for the other racial groups reveals that Black students reported significantly higher levels of perceptions of punishment risk for the two least serious offenses: disrupting class and disrespecting a teacher. Although we cannot be certain, this pattern of findings may reflect teacher discretion in punishment outcomes. If, given the same offense, teachers are more likely—even due to implicit biases (Okonofua & Eberhardt, 2015; Price & Wolfers, 2010)—to punish minorities harsher than non-minorities (e.g., Whites), then we may expect to find Black students reporting the highest average punishment risk for the least serious outcomes. And indeed, there is precedent for this finding in the criminal justice literature. For instance, research has shown that racial inequalities in sentencing outcomes are most prominent for the least serious offenses where judges have the most discretion (Spohn & Cederblom, 1991; Steffensmeier, Ulmer, & Kramer, 1998; but see also Warren, Chiricos, & Bales, 2012).

Supplemental & Sensitivity Analyses

We are sensitive to the criticism that a pattern of null findings for the calibration between objective punishment levels and students' perceptions might arise if, for instance, there is an interaction between student offending status and perception formation or if measurement error exists for the punishment risk measures. We conducted several supplementary tests that were performed to assess these sorts of concerns.

Prior work, and theoretical arguments, suggest active offenders may have a better grasp of the risks of punishment compared to the average citizen (see Apel & Nagin, 2017). Thus, we re-estimated each of the six models presented in Table 2 (all covariates were included but are not presented for parsimony), this time specifiying a random variance component for the effect of self-reported delinquency and including a multiplicative term for *self-reported delinquency*$_{ij}$ × *objective punishment risk*$_j$. The estimated effects for the three variables of interest are presented in Table 3. As can be seen in

Table 3 Results from interaction between self-reported delinquency and principal reports of punishment risk

	Disrupt class		Disrespect teacher		Cut class		Fighting		Alcohol/Drug use		Bring weapon to school	
	b	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error	*b*	Standard error
School-level predictors												
Disrupt class	−0.033	[0.034]										
Disrespect teacher			0.011	[0.031]								
Cut class					−0.003	[0.034]						
Fighting							0.019	[0.059]				
Alcohol/Drug use									−0.007	[0.027]		
Bring weapon to school											−0.084*	[0.042]
Self-reported delinquency	0.005	[0.004]	0.002	[0.003]	−0.011**	[0.003]	−0.012**	[0.003]	−0.024**	[0.003]	−0.028**	[0.004]
School-level predictor X Self-reported Delinq.	0.003	[0.002]	0.001	[0.002]	−0.002	[0.002]	0.002	[0.002]	0.004	[0.002]	0.004	[0.003]

*$p < 0.05$, ** $p < 0.01$, ***$p < 0.001$; Standard errors in brackets; Black students served as the reference category for the race/ethnic variables; All covariates are included but are not presented for parsimony

the bottom row of the table, none of the interactions emerged as statistically significant. This test failed to provide support for the argument that offenders have stronger calibration than non-offenders.

The self-reported delinquency scale captures a range of behaviors, only some of which may occur at school. It could be that calibration is higher for students who engage in delinquent behaviors at school compared to students who primarily offend outside of school. In order to investigate this possibility, we singled out the behaviors from the self-reported delinquency scale that overlapped with the behaviors of focus in the six statistical models. Four items overlapped: students were asked to report how often they assaulted a teacher (similar to disrespecting a teacher), got into a serious fight, used drugs, and carried a weapon at school. We re-estimated the four models that aligned with these behaviors, this time including an interaction term between the objective punishment risk and the student's self-reported invovlement in the behavior of focus. These tests produced no statistically significant interactions.

As for concerns over measurement quality, it is possible that the measurement strategies employed here led to restricted variation that attenuated the observed calibration. Braga and Apel (2016) recently noted that a low-end estimate for the reliability for certain measures (they were focused on the clearance rate) of objective punishment risk might be around 0.11. Building on this, they noted, "…we would expect the relationship between [objective punishment risk] and [perceptions of punishment risk] to be heavily attenuated because of measurement error—it would be about $1/9^{th}$ its true size" (Braga & Apel, 2016, p. 819). Taking this at face value, one might calculate an unattenuated/corrected estimate of the calibration by multiplying the coefficients we present by 9. Focusing only on the two coefficients that were positive from our models in Table 2, we see that the unattenuated calibration coefficient might be as high as 0.126 (for fighting: 0.014*9) and as low as 0.036 (for disrespecting a teacher: 0.004*9).

In a similar vein, we conducted additional analyses to explore how measurement error on perceptions of punishment might have affected our substantive conclusions. Specifically, we were concerned that the ordinal coding of our punishment variables might have biased the estimates if students had a hard time differentiating between different levels of punishment. Additionally, the ordering of the categories may not accurately capture perceived increases in severity for all respondents. Some, for example, may feel that being sent to the principal's office is worse than a notification sent to parents. In an effort to circumvent these potential measurement issues, we recoded all of the outcome variables into a binary indicator of punishment using the following scheme: 0 = *nothing*, *discipline by teacher*, *sent to principal's office*, or *parents are notified*; 1 = *detention*, *suspension or expulsion*, or *police are called.* Punishments coded as 1 are likely to be considered by all students as being more serious than those coded as 0. We followed this same coding scheme for the objective punishment risk variable so that the coding schemes matched. Finally, we re-estimated the models presented in Table 2 under a multi-level logistic regression framework. The substantive story was unchanged; and only one of the coefficient estimates attained statistical significance: fighting ($b = -0.28$, SE = 0.13, $P = 0.04$).

We also performed a sensitivity test to determine whether missing values affected our substantive conclusions. To test this, we used standard routines for multiple imputation to estimate the missing values for the principal reports of the objective risk measures (Allison, 2002). We imputed these missing values using all the covariates as predictors and by creating 10 imputed datasets. We then re-estimated all of the equations from Table 2, averaging across the 10 imputed files. The results from these models did not substantively differ from those shown in Table 2.

Finally, we performed a supplemental analysis to determine whether a "global" measure of objective punishment risk was predictive of a "global" measure of student perceptions. Specifically, we averaged the objective punishment risk measures across the six domains and we did the same for the student-level perceptions. We then estimated a multi-level model to determine whether this global objective risk measure predicted the global punishment perceptions measure. The results from this analysis were: $b = -0.08$, $\text{SE} = 0.03$, $P < 0.05$. The effect was statistically significant, but the direction of association was opposite of expectation.

Discussion

Policy based on general deterrence usually assumes a close calibration between objective punishment risks and individuals' perceptions of those punishment risks (Pickett & Roche, 2016a). The goal of this paper was to test this fundamental assumption. Relatively few studies have directly estimated that calibration (Apel et al., 2009; Kleck & Barnes, 2014; Kleck et al., 2005; Lochner, 2007), which has come to be known as the "dirty little secret" of deterrence theory (Paternoster, 2010, p. 804). Of the prior studies that have analyzed it, findings have revealed a weak and substantively insignificant relationship between objective punishment and perceptions of punishment. But a notable critique of prior studies (viz. Kleck & Barnes, 2014; Kleck et al., 2005; Lochner, 2007) is that measures of objective punishment were too far removed from the individual citizen (Apel, 2013; Nagin, 2013b; Pogarsky, 2009). Specifically, Kleck's studies and Lochner's analysis relied on county-level measures of objective punishment risk. Apel and colleagues (Apel et al., 2009) argued that the school-level may provide a better context for analyzing the calibration between objective punishment risks and perceptions of punishment risk.

For three-fourths of the year, American youth spend roughly 8 hours a day, 5 days a week in school. By middle and high school—the age of the sample in the present study—students can expect to spend about 1000 h a year in school (Farbman, Davis, Goldberg, & Rowland, 2015). Even large schools surround students with a relatively stable environment. By implication, students are exposed to a stable sanctioning environment for most of the waking hours of their day. Punishment will usually be meted out quickly, and students will be back in class to discuss the event. All the delays that come with the criminal justice system are relatively nonexistent in schools. This is all to say that if a calibration between objective punishment and perceptions of punishment does exist, it should emerge in a school-level analysis.

Yet, the current study found few substantively (or statistically) meaningful relationships between objective punishment levels and perceptions of punishment levels. We tested for an association between objective punishment and perceptions of punishment in six different statistical models. After including relevant controls, none of these six models revealed a substantive or statistically significant relationship.

Our models revealed that the average perception of risk increased monotonically from the least serious offense to the most serious offense (see the pattern of intercept values in Table 2). These results suggest students (citizens) are aware of which behaviors (crimes) are punished most severely relative to other behaviors (crimes), but that they cannot identify specific punishment levels with any precision (Thomas et al., 2017). This is consistent with work from behavioral economics that shows individuals' valuation of goods often seems arbitrary when observed one good at a time (Ariely et al., 2003). But when valuations are made relative to other goods, the process appears orderly and takes on a certain level of coherent arbitrariness (see also Pogarsky & Loughran, 2016). Thus, it may be the case that, as Ariely and colleagues (Ariely et al., 2003) suggest, relative perceptions are more informative than independently assessed perceptions. Individuals might be poor judges of the risks of punishment for behaviors considered independently, but they are in-tune enough to know that one behavior carries larger risk than some other behavior (Thomas et al., 2017).

Behavioral economists and psychologists have shown that humans use convenient, cognitively inexpensive information to form their beliefs about everyday risks (Kahneman, 2011). Given the difficulty of obtaining precise and accurate measures of objective punishment levels (see, for example, the arguments spelled out by Braga & Apel, 2016), it seems safe to conclude that accurate information about objective punishment risks are even more difficult to come by for the average citizen. In short, objective punishment levels are not convenient to locate, making them expensive to use when forming perceptions. This all suggests that people use other sources of information (i.e., heuristics) to form their perceptions of punishment risk (see Sampson & Raudenbush, 2004 for additional evidence that suggests individuals' perceptions are linked to more convenient indicators that may only loosely relate to the actual information that is being requested).

But let us be clear: our results and the theoretical points made above should not be taken as an indication that individual-level perceptions of punishment are impervious to change. Quite the contrary, as Pogarsky and Loughran (2016, p. 783) summarize, "… there *is* substantial evidence that individuals change their beliefs in response to acquiring new information regardless of from where it is learned." In other words, the lack of a calibration found in this study should not be interpreted as argument that individuals' beliefs are fixed and unmanipulable. It does, however, raise questions about the sources that affect variation in punishment perceptions and it suggests that the sources of variation may lie at the individual level (see generally Piquero, Paternoster, Pogarsky, & Loughran, 2011).

This leads us to speculate about where and how people actually form their perceptions of punishment risks. For instance, we have solid evidence to suggest life-experiences—such as being arrested—and other individual-level factors affect the formation of punishment perceptions (Lochner, 2007; Piquero et al., 2011). This literature has shown that experiencing punishment may lead to temporary adjustments in perceptions of punishment, and that those adjustments work differently as a function of one's unique punishment history. One more arrest for someone who already has amassed 10 will do little to affect his/her perceptions of punishment. One arrest for someone who has none, is expected to lead to bigger shifts (see, generally, Anwar & Loughran, 2011). But these adjustments to perceptions of punishment risk may only be temporary (Piquero & Pogarsky, 2002). And it is important to keep in mind that when an individual experiences an arrest, that becomes specific and not general deterrence.

All of this supports crime prevention theories (e.g. Eck & Guerrette, 2012), which it can be argued are agnostic about the calibration between *general* objective punishment risks and perceptions of those punishment risks. Indeed, the crime prevention approach is only concerned with the offender's perceptions about the probability of punishment in a specific situation. Prevention (i.e., specific deterrence) might work, but it perhaps is a context-specific effect that is impacted by temporary changes in an offender's perceptions. In other words, it could be the case that humans have a "baseline" perception of punishment risk that they carry with them from place-to-place. But if perceptions are dynamic—and they most certainly are (Kahneman, 2011)—then citizens' perceptions can be influenced by acute, localized circumstances such as a concentration of police on a city block or in a neighborhood. There may be no calibration, on average, between the prevailing levels of police strength and citizens' perceptions of police levels (Kleck & Barnes, 2014). But this does not undermine the possibility that citizens are sensitive to acute changes such as when a certain "spot" is "hot" (Groff et al., 2015). There is no expectation that the change in perceptions is lasting over time or across place, meaning an acute "shock" might shift one's perceptions temporarily. After a time (or after the person has moved locations), the "baseline" level may then take over again.

Also, as we noted earlier with our discussion of Fig. 1, the results of this analysis say nothing about whether deterrence "works" in the sense that people's beliefs and perceptions about punishment risks cause them to refrain from criminal acts. The lack of a calibration does not mean perceptions are unimportant. It simply means that, on average, perceptions are not linked with objective risk levels.

In light of these points, though, there are certain aspects of the present study that might inspire caution among readers. First, previous research (Pratt, Cullen, Blevins, Daigle, & Madensen, 2006) and reviews (Nagin, 1998, 2013b; Paternoster, 2010) have indicated severity of punishment is a weaker predictor of behavior than certainty of punishment. Because severity is a weaker predictor, one might assume that the calibration between objective punishment severity and perceptions of that severity is necessarily low. But that need not be the case. It could be that the calibration is high, but punishment severity is simply a weak predictor of behavior. Given that our research interests were centered only on the calibration, the strength of punishment severity as a predictor of crime does not strictly concern us here.

Second, the data came from one county in Ohio. This may restrict the observed variation in objective punishment. However, as Fig. 2 demonstrates, this concern is ot sufficient to explain the pattern of null results.

Third, the measures of objective and perceived punishment risk, albeit similar to ones used in previous research (Apel et al., 2009), may include a non-trivial amount of measurement error. It is entirely possible that principals are somewhat removed from the reality of their schools. Especially in large schools, minor infractions—which make up most of the deviance committed by students—are unlikely to be brought to the attention of a principal. A principal could easily have an idea of how punishment is meted out that contrasts with reality. We do not discount this possibility, but we see little reason to think such a problem would consistently wipe out an otherwise large positive calibration.

Fourth, one could reasonably argue that the school context is still too far removed from the individual to pick up on any meaningful level of objective risk for punishment.

We built on the arguments laid out by Apel and colleagues (Apel et al., 2009) that the school is nearer a closed system than is the county, but it is possible that the problems that plague county-level measures also arise with school-level measures. In other words, it may be a poor assumption that schools are more effective risk communicators. We were not able to account for specific steps schools took to communicate risks (for example, through a code of conduct).

Of course, there are many other social systems that are even closer to the "closed" end of the spectrum. For instance, one could study the calibration of punishment risk in a peer group or a family unit. And, if one conducted such a study, the calibration could reasonably be expected to exceed what we see in the present study. We submit, however, that the school level is more applicable to a scholarly audience because this represents one of the institutions that might be amenable to changes in policy if it can be shown that student perceptions are closely calibrated to actual risk levels. This is far less likely to hold—at least with any fidelity—for policies aimed at increasing punishment risks among units like peer groups or families. Scholars are more comfortable hypothesizing (and making recommendations) about evidence-based policies for school disciplinary tactics than about recommending peers/families increase punishment risks for their friends/children.

Related to this, one could note that school sanctions and criminal sanctions are different, which may make the school a poor proxy for the criminal justice system. However, the difficulty in measuring calibration in the criminal justice system led us to schools in the first place. In our case, the differences described by Apel et al. (2009) were precisely what we were counting on to give us the best chance of finding calibration.

Finally, critics might argue that calibration studies are unnecessary because punishment probability is temporally and even spatially specific (Jacobs, 2010). Thus, people may not know the prevailing level of risk at any given time or location, but those same people are expected to respond to shifts in policy by updating their perceptions even if their initial perceptions were wrong (see, generally, Pogarsky & Loughran, 2016). We agree that this may help explain the pattern of findings we observed in this study. But we do not believe it fully explains why our calibration estimates were so consistently close to zero. The reason is that in order to see calibrations estimates close to zero—if citizens pick up on shifts in policies but perhaps not stagnant risk levels—any updates to one's perceptions must be temporary. Otherwise these policy shifts would, over time, push the calibration toward 1.00 (rather than toward zero) as more and more policies are added, amended, and repealed. There is some theoretical precedent here that we can point to. Specifically, Ariely and colleagues (Ariely et al., 2003, p. 101) argued, "We would predict, therefore, that one should find short-term deterrence effects in narrowly focused studies that examine the impact of policy changes, but little or no deterrence effects in cross-sectional studies. This is, indeed, the observed pattern. Interrupted time series studies have measured sizable reactions in criminal behavior to sudden, well-publicized, increases in deterrence [Ross, 1973; Sherman, 1990], but these effects tend to diminish over time." Based on these points, we believe it would be fruitful for future work to estimate the decay rate that might exist for the calibration between changes in policies and changes in citizens' perceptions of risk. Knowing that piece of information might help tie together our findings with arguments set forward by those who advocate for general deterrence-based policies.

Conclusion

We are not advocating for an "anti-deterrence" campaign nor do we believe general deterrence is a failed theory or a bad approach to policy. We raise this point for two reasons. First, recall our study says nothing about the well established perceptions➔offending behaviors relationship. Thus, deterrence can work as long as the deterrence-based policy actually affects perceptions. Second, there are numerous examples that can be cited to support the fact that deterrence works, even in massively open systems such as between countries (think the cold war era policies of nuclear deterrence). And we all know what happens when the police disappear like during times of crisis or police strikes (Andenaes, 1974). But these examples reflect scenarios where the information about punishment risks is convenient and cognitively inexpensive. In other words, these scenarios reflect big changes in objective risk. What our study tests—and ultimately draws into question—is whether respondents are able to pick up on more subtle, everyday changes in risk levels. In light of this study and previous work, we conclude that the pervasive general deterrence idea—that changes in objective risk levels will alter perceptions of risk—is incorrect. This conclusion is not very surprising and has, indeed, been anticipated by for quite some time. Specifically, Lattimore and colleagues (Lattimore, Baker, & Witte, 1992, p. 377) noted that, "Most subjects are considerably less sensitive to changes in mid-range probability than is proposed by the expected utility model…".

But citizens' perceptions of punishment risks can be changed and a close calibration between those perceptions and the objective levels can be realized in an open-system if changes in punishment risk are communicated to the public (Kennedy, 2009; Ross, 1973). If we are to "re-calibrate" citizens' perceptions to align more closely with reality, then citizens need to be informed about those realities. Without communication of actual punishment risks, deterrence mechanisms are simply drowned out by the white noise of information that comes with everyday life. This means that policy proposals based on deterrence should include specific noisemaking mechanisms if they are to be successful in deterring criminal activity.

References

Allison, P. D. (2002). *Missing data*. Thousand Oaks: Sage.

Andenaes, J. (1974). *Punishment and deterrence*. Ann Arbor: University of Michigan Press.

Anwar, S., & Loughran, T. A. (2011). Testing a bayesian learning theory of deterrence among serious juvenile offenders. *Criminology, 49*, 667–698.

Apel, R. (2013). Sanctions, perceptions, and crime: Implications for criminal deterrence. *Journal of Quantitative Criminology, 29*(1), 67–101.

Apel, R., & Nagin, D. S. (2017). Perceptual deterrence. In W. Bernasco, H. Elffers, & J-L. van Gelder (Eds.), *Oxford handbook of offender decision making* (pp. 121–140). New York: Oxford University Press.

Apel, R., Pogarsky, G., & Bates, L. (2009). The sanctions-perceptions link in a model of school-based deterrence. *Journal of Quantitative Criminology, 25*(2), 201–226.

Ariely, D., Lowenstein, G., & Prelec, D. (2003). Coherent arbitrariness: Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118*, 73–105.

Ball, J. C. (1955). The deterrence concept in criminology and law. *Journal of Criminal Law Criminology and Police Science, 46*, 347–354.

Beccaria, C. (1995). On crimes and punishments. In R. Bellamy, R. B. Davies, & V. Cox (Eds.), *Beccaria: 'On crimes and punishments' and other writings* (pp. 9–114). Cambridge: Cambridge University Press.

Braga, A. A., & Apel, R. (2016). And we wonder why criminology is sometimes considered irrelevant in real-world policy conversations. *Criminology & Public Policy, 15*, 813–829.

Braga, A. A., & Weisburd, D. L. (2012). The effects of focused deterrence strategies on crime: A systematic review and meta-analysis of the empirical evidence. *Journal of Research in Crime and Delinquency, 49*(3), 323–358.

Cook, P. J. (1980). Research in criminal deterrence: Laying the groundwork for the second decade. *Crime and Justice, 2*, 211–268.

Eck, J. E., & Guerrette, R. T. (2012). Place-based crime prevention: Theory, evidence, and policy. In B. C. Welsh, & D. P. Farrington (Eds.), *The Oxford handbook of crime prevention* (pp. 354–383). New York: Oxford University Press.

Farbman, D., Davis, J., Goldberg, D., & Rowland, J. (2015). Learning time in america: Trends to reform the American school calendar. National Center on Time & Learning.

Geerken, M. R., & Gove, W. R. (1975). Deterrence: Some theoretical considerations. *Law and Society Review, 9*(3), 497–513.

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime*. Stanford: Stanford University Press.

Grasmick, H. G., & Bryjak, G. J. (1980). The deterrent effect of perceived severity of punishment. *Social Forces, 59*(2), 471–491.

Groff, E. R., Ratcliffe, J. H., Haberman, C. P., Sorg, E. T., Joyce, N. M., & Taylor, R. B. (2015). Does what police do at hot spots matter? The Philadelphia policing tactics experiment. *Criminology, 53*, 23–53.

Horney, J., & Marshall, I. H. (1992). Risk perceptions among serious offenders: The role of crime and punishment. *Criminology, 30*, 575–594.

Jacobs, B. A. (2010). Deterrence and deterrability. *Criminology, 48*(2), 417–441.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kennedy, D. M. (2009). *Deterrence and crime prevention: Reconsidering the prospect of sanction*. Routledge Studies in Crime and Economics (Vol. 2). Routledge.

Kleck, G. (2016). Objective risks and individual perceptions of those risks. *Criminology & Public Policy, 15*, 767–775.

Kleck, G., & Barnes, J. C. (2014). Do more police lead to more crime deterrence? *Crime & Delinquency, 60*, 716–738. https://doi.org/10.1177/0011128710382263.

Kleck, G., Sever, B., Li, S., & Gertz, M. (2005). The missing link in general deterrence research. *Criminology, 43*, 623–660.

Lab, S.P., & Clark, R.D. (2006). Controlling victimization in schools: Effective discipline and control strategies in a county in Ohio, 1994. ICPSR02587-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR02587.v1

Lattimore, P. K., Baker, J. R., & Witte, A. D. (1992). The influence of probability on risky choice: A parametric examination. *Journal of Economic Behavior & Organization, 17*(3), 377–400.

Lochner, L. (2007). Individual perceptions of the criminal justice system. *The American Economic Review, 97*(1), 444–460.

Maimon, D., Antonaccio, O., & French, M. T. (2012). Severe sanctions, easy choice? Investigating the role of school sanctions in preventing adolescent violent offending. *Criminology, 50*, 495–524.

Nagin, D. S. (1998). Criminal deterrence research at the outset of the twenty-first century. *Crime and Justice, 23*, 1–42.

Nagin, D. S. (2013a). Deterrence: A review of the evidence by a criminologist for economists. *Annual Review of Economics, 5*, 83–105.

Nagin, D. S. (2013b). Deterrence in the twenty-first century. *Crime and Justice, 42*, 199–263.

Nagin, D. S. (2016). What we've got here is failure to communicate. *Criminology & Public Policy, 15*(3), 753–765.

Nagin, D. S., Solow, R. M., & Lum, C. (2015). Deterrence, criminal opportunities, and police. *Criminology, 53*, 74–100. https://doi.org/10.1111/1745-9125.12057.

Okonofua, J. A., & Eberhardt, J. L. (2015). Two strikes: Race and the disciplining of young students. *Psychological Science, 26*, 617–624.

Paternoster, R. (2010). How much do we really know about criminal deterrence? *The Journal of Criminal Law and Criminology, 100*, 765–824.

Paternoster, R., Saltzman, L. E., Waldo, G. P., & Chiricos, T. G. (1983). Perceived risk and social control: Do sanctions really deter? *Law and Society Review, 17*, 457–479.

Pickett, J. T., Mancini, C., Mears, D. P., & Gertz, M. (2015). Public (mis)understanding of crime policy: The effects of criminal justice experience and media reliance. *Criminal Justice Policy Review, 26*(5), 500–522.

Pickett, J. T., & Roche, S. P. (2016a). Arrested development: Misguided directions in deterrence theory and policy. *Criminology & Public Policy, 15*(3), 727–751.

Pickett, J. T., & Roche, S. P. (2016b). A few clarifying comments on Pickett and Roche (2016). *Criminology & Public Policy, 15*(3), 831–836.

Piquero, A. R., Paternoster, R., Pogarsky, G., & Loughran, T. (2011). Elaborating the individual difference component in deterrence theory. *Annual Review of Law and Social Science, 7*, 335–360.

Piquero, A. R., & Pogarsky, G. (2002). Beyond Stafford and Warr's reconceptualization of deterrence: Personal and vicarious experiences, impulsivity, and offending behavior. *Journal of Research in Crime and Delinquency, 39*(2), 153–186.

Pogarsky, G. (2009). Deterrence and decision making: Research questions and theoretical refinements. In M. D. Krohn, A. J. Lizotte, & G. P. Hall (Eds.), *Handbook on crime and deviance* (pp. 241–258). Springer.

Pogarsky, G., & Loughran, T. A. (2016). The policy-to-perceptions link in deterrence. *Criminology & Public Policy, 15*, 777–790.

Pogarsky, G., Roche, S. P., & Pickett, J. T. (2017). Heuristics and biases, rational choice, and sanction perceptions. *Criminology, 55*, 85–111.

Pratt, T. C., Cullen, F. T., Blevins, K. R., Daigle, L. E., & Madensen, T. D. (2006). The empirical status of deterrence theory: A meta-analysis. In F. T. Cullen, J. P. Wright, & K. R. Blevins (Eds.), *Taking stock: The status of criminological theory—Advances in criminological theory* (Vol. 15, pp. 367–395). New Jersey: Transaction Publishers.

Price, J., & Wolfers, J. (2010). Racial discrimination among NBA referees. *The Quarterly Journal of Economics*, *125*(4), 1859–1887.

Rosenfeld, R., & Fornango, R. (2017). The relationship between crime and stop, question, and frisk rates in New York city neighborhoods. *Justice Quarterly, 34*(6), 931–951. https://doi.org/10.1080/07418825.2016.1275748.

Ross, H. L. (1973). Law, science, and accidents: The British road safety act of 1967. *The Journal of Legal Studies, 2*(1), 1–78.

Sampson, R. J., & Raudenbush, S. W. (2004). Seeing disorder: Neighborhood stigma and the social construction of "broken windows". *Social Psychology Quarterly, 67*(4), 319–342.

Sherman, L. W. (1990). Police crackdowns: Initial and residual deterrence. *Crime and Justice, 12*, 1–48.

Sherman, L. W., & Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review, 49*, 261–272.

Sherman, L. W., Schmidt, J. D., Rogan, D. P., & Smith, D. A. (1992). The variable effects of arrest on criminal careers: The Milwaukee domestic violence experiment. *Journal of Criminal Law and Criminology, 83*(1), 137–169.

Spohn, C., & Cederblom, J. (1991). Race and disparities in sentencing: A test of the liberation hypothesis. *Justice Quarterly, 8*(3), 305–327.

Stafford, M. C., & Warr, M. (1993). A reconceptualization of general and specific deterrence. *Journal of Research in Crime and Delinquency, 30*(2), 123–135.

Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology, 36*(4), 763–798.

Tanner-Smith, E. E., Fisher, B. W., Addington, L. A., & Gardella, J. H. (2018). Adding security, but subtracting safety? Exploring schools' use of multiple visible security measures. *American Journal of Criminal Justice, 42*, 102–119.

Thomas, K. J., Hamilton, B. C., & Loughran, T. A. (2017). Testing the transitivity of reported risk perceptions: Evidence of coherent arbitrariness. *Criminology, 56*(1), 59–86.

Tittle, C. R. (1977). Sanction fear and the maintenance of social order. *Social Forces, 55*(3), 579–596.

Warren, P., Chiricos, T., & Bales, W. (2012). The imprisonment penalty for young black and Hispanic males: A crime-specific analysis. *Journal of Research in Crime and Delinquency, 49*(1), 56–80.

Welch, K., & Payne, A. A. (2010). Racial threat and punitive school discipline. *Social Problems, 57*(1), 25–48.

Zhang, G. (2018). The effects of a school policing program on crime, discipline, and disorder: A quasi-experimental evaluation. *American Journal of Criminal Justice, 43*, 1–18. https://doi.org/10.1007/s12103-018-9440-z.

Timothy Nixon is a crime analyst at the New York City Police Department and a doctoral candidate in the School of Criminal Justice at the University of Cincinnati.

J.C. Barnes is an associate professor in the School of Criminal Justice at the University of Cincinnati.