Cooperation and Punishment in Regulating Labor Standards: Evidence from the Gap Inc Supply Chain

Matthew Amengual
Greg Distelhorst

August 18, 2020

Multinational firms with global supply chains manage the labor compliance of their supplier firms. Does this private regulation cause compliance to improve? And what approaches to private regulation are effective? This study analyzes supply chain regulation under a largely cooperative approach and under an approach that incorporates the threat of penalties. Drawing on data from over one thousand factories supplying the multinational retailer Gap Inc over 2010–2019, it uses a regression discontinuity design to estimate the causal effects of assigning compliance ratings under these two approaches to supply chain regulation. Under the cooperative approach, we estimate precise, near-zero effects of compliance ratings on future social compliance. However, when the buyer incorporated penalties in the form of threats from the sourcing department to discontinue the business relationship with noncompliant suppliers, a failing grade caused factory compliance to improve by 0.8 standard deviations and reduced the probability of future failure by 22 percentage points. These effects are validated using independent labor compliance data from the ILO/IFC Better Work program. We also test hypotheses about the mediating effect of long-term supplier relationships and find that private regulation had a larger impact on longer-term suppliers, but again only in the presence of penalties. Notwithstanding an emerging consensus about the need for cooperation and committed commercial relationships in supply chain regulation, this study affirms the importance of incentives to enhance sustainability.

Keywords: private regulation, supply chain, compliance, social responsibility, labor.

Matthew Amengual: Associate Professor, Saïd Business School, University of Oxford, Park End Street, OX1 1HP, United Kingdom. www.mattamengual.net, matthew.amengual@sbs.ox.ac.uk,

Greg Distelhorst: Assistant Professor, University of Toronto, Centre for Industrial Relations and Human Resources and Rotman School of Management, 121 St George St, Toronto ON M5S 2E8, Canada. www.gregdistelhorst.com, g.distelhorst@utoronto.ca,

We thank Danny Tobin for research assistance, managers at Gap Inc for contributing their time and data to this project, and Sarosh Kuruvilla for his support. We gratefully acknowledge financial support from the MIT Good Companies Good Jobs Initiative. Conference and workshop participants at Brown University, UC-Berkeley, the London School of Economics, Kings College London, University of Manchester, University of Oxford, University of Toronto, University of Warwick, Alliance for Research on Corporate Sustainability, European Group for Organizational Studies, International Political Economy Society, International Studies Association, Labor and Employment Relations Association, and Strategy and the Business Environment offered generous feedback. We also thank Jennifer Bair, Tim Bartley, Ben Cashore, Alice Evans, Akshay Mangla, Layna Mosley, Ben Richardson, Mari Sako, Jodi Short, and Mike Toffel for helpful feedback on this research.

Global supply chains pose risks to the reputation and financial performance of multinational enterprises. Recent media reports link factories and farms in global supply chains to environmental damage, industrial accidents, wage theft, and even modern slavery.[1] These abuses invite activist campaigns that target multinational enterprises, threaten their brands and reputations, and impose financial costs (Baron 2001, King and Soule 2007, Bartley and Child 2011).

To manage these risks, many multinationals, including over 90% of Fortune's top 250 global corporations, have adopted supplier codes of conduct,[2] and a majority of all publicly listed corporations in the food, textile, and wood products industries manage social responsibility in their supply chains (Thorlakson, de Zegher and Lambin 2018). Firms are now expected to manage not only the operational performance of their suppliers but also their *social* performance.

A broad literature explains how buyers manage supplier relationships to enhance economic performance (Williamson 1991, Uzzi 1997, Dyer and Singh 1998), but scholarship on how firms manage suppliers' *social* performance is more recent and leaves many unanswered questions. Improving supplier social performance involves many of the challenges identified in the literature on supplier governance, such as fostering inter-organizational learning and avoiding opportunism. Yet social performance standards differ because they were adopted to maintain legitimacy in the face of stakeholder pressure. Labor and environmental compliance have a more uncertain relationship to value creation and capture than the traditional objectives of price, speed and quality. Indeed, many buyers have demonstrated ambivalence about whether suppliers should prioritize social performance, especially when doing so might harm other performance metrics (Locke 2013, Bird, Short and Toffel 2019).

This study addresses two questions about private regulation of social standards in supply chains. First, does private regulatory activity cause compliance to improve? Given that buyers adopted social standards for their suppliers to maintain legitimacy, core elements of private regulatory programs may serve to mollify stakeholders while having no effect on outcomes. Indeed, this question has been at the center of a longstanding tension between private regulation optimists, who note its potential to raise standards in emerging markets (Fung, O'Rourke and Sabel 2001, Elliott and Freeman 2003, Lee and Tang 2017, Bird, Short and Toffel 2019), and more skeptical studies pointing to widespread violations even in supply chains subject to private regulation for many years (Barrientos and Smith 2007, Locke 2013, Vogel 2010, Anner 2012, Bartley 2018). Despite nearly two decades of research, we still have a limited understanding of whether and how much buyers' standard practices of private regulation cause suppliers to improve compliance.

Second, is private regulation effective under a largely cooperative approach? If not, does

[1] Amazon Deforestation, Once Tamed, Comes Roaring Back. *The New York Times*. https://tinyurl.com/zqs8atw. More Than 300 Killed in Pakistani Factory Fires. *The New York Times*. https://tinyurl.com/ybsnf2ko. Building Collapse in Bangladesh Leaves Scores Dead. *The New York Times*. https://tinyurl.com/yaln4pgo. Human Rights Watch. "Thailand: Forced Labor, Trafficking Persist in Fishing Fleets," January 23, 2018. Worker Rights Consortium "Stealing from the Poor: Wage Theft in the Haitian Apparel Industry" 2013.

[2] KPMG International. 2008. "KPMG International Survey of Corporate Responsibility Reporting 2008."

introducing penalties make it effective? These questions speak to debates in the supplier governance literature between approaches that emphasize the role of relationships and those that emphasize the role of incentives. On the one hand, previous research establishes that buyer–supplier relationships are socially embedded (Uzzi 1997) and that cooperation enables trust and knowledge-sharing (Dyer and Chu 2003, Gulati and Nickerson 2008, Janowicz-Panjaitan and Noorderhaven 2009, Short, Toffel and Hugill 2020). This literature's emphasis on cooperative approaches to managing supplier performance for value creation and capture, especially within longer-term relationships (Elfenbein and Zenger 2014), is echoed by research on labor and environmental compliance. Both the specialized literature on supply chain labor standards (Locke, Amengual and Mangla 2009, Lund-Thomsen and Lindgreen 2014, Short, Toffel and Hugill 2020) and a more general literature on regulatory compliance (Bardach and Kagan 1982, Short and Toffel 2010) argue that collaboration is key to improving compliance. However, other theories stress the importance of incentives and the calculative nature of relationships based on the expectation of future gains (Williamson 1991, Baker, Gibbons and Murphy 2002, Susarla, Holzhacker and Krishnan 2020). These theories suggest that buyer–supplier relationships benefit from aligned incentives, yet there has been little empirical research on the use of incentives for improving supplier social performance.

Against the backdrop of these competing theoretical views, this study estimates the causal effects of supply chain regulation first under a largely cooperative approach and then under an approach augmented by penalties. Evidence comes from ten years of compliance audit results, sourcing records, and administrative data from over one thousand factories in the supply chain of Gap Inc, a large multinational apparel retailer. To estimate the causal effect of private regulatory activities, the analysis exploits a sharp discontinuity in the assignment of labor compliance ratings. Because Gap's approach to supply chain regulation changed during the period we study (2010–2019), the actions triggered by compliance ratings also changed. In the earlier period of the study, the approach was largely cooperative. A failing compliance rating communicated buyer displeasure and led social responsibility managers to provide more support to remediate violations. Beginning in mid-2016, a failing compliance rating additionally triggered the threat of commercial penalties. The buyer informed failing factories that the inability to pass compliance audits would harm the business relationship, and serially failing factories began being terminated at higher rates. We study the efficacy of private regulation in each of these periods, estimating the causal effects of failing ratings on future compliance under the earlier cooperative approach and the later approach that included penalties.

In addition to the theoretical stakes discussed below, this research has implications for management practice and public policy. Activists have long pushed buyers to privately regulate their suppliers in hope of improving labor and environmental practices in global supply chains. Policymakers have similarly worked to expand buyers' scope of responsibility for the actions of their suppliers, most recently in legislation including the UK Modern Slavery Act, the California Transparency in Supply Chains Act, and the French Law on Duty of Care. Yet there remains widespread uncertainty over which approaches to private regulation, if any, can improve social

compliance. This study offers evidence that, under certain conditions, buyers *can* improve the social compliance of their suppliers. These findings can assist the wide range of actors—including activists, policymakers, and managers—searching for ways to make corporate supply chain responsibility programs more effective.

Private Regulation of Global Supply Chains

Buyers in global supply chains increasingly manage not only their suppliers' business performance, but also their suppliers' social and environmental impacts. This private (i.e. non-governmental) regulation of global supply chains emerged from private politics, in which "interest and activist groups attempt to influence economic activity directly without reliance on public institutions" (Baron 2001, 7). Social movements concerned with worker rights, deforestation, conflict minerals, marine life, and other issues have targeted large firms in advanced economies with the goal of influencing behavior in hundreds or even thousands of their supplier firms, often in emerging markets. These "proxy targeting" campaigns (Walker, Martin and McCarthy 2008) pressed multinational enterprises to bring their suppliers' practices in line with international standards of socially responsible production. Such activism prompted many multinational firms to develop programs to regulate the social and environmental performance of their supplier factories, farms, or forests (Bartley 2003; 2007, Cashore, Auld and Newsom 2004). Contractual requirements for social and environmental performance in buyer–supplier relations are now commonplace. This private regulation of labor and environmental practices expanded the scope of supplier activities that buyers sought to manage and attracted scholarly interest in strategic management (Vogel 2010, Short, Toffel and Hugill 2016), operations (Lee and Tang 2017, Jira and Toffel 2013, Kalkanci, Ang and Plambeck 2016, Liu et al. 2019, Kalkanci and Plambeck 2019), and organizations (Bird, Short and Toffel 2019), as well as economics (Harrison and Scorse 2010, Tanaka 2017, Boudreau 2020), sociology (Bartley 2007, Seidman 2007), and political science (Anner 2012, Locke 2013).

Despite its extensive diffusion, private regulation of global supply chains remains controversial. Suppliers are asked to comply with social and environmental standards while simultaneously satisfying performance demands in price, quality, and delivery that may conflict with compliance. Because private regulation was adopted to maintain legitimacy in the face of stakeholder pressure, buyers have been more tolerant of violations of social and environmental standards than underperformance in traditional business metrics. A large and growing literature on compliance has repeatedly shown that most suppliers fail to meet buyers' social performance standards (Frenkel 2001, Barrientos and Smith 2007, Locke, Qin and Brause 2007, Anner 2012, Toffel, Short and Ouellet 2015, Bird, Short and Toffel 2019). Although most studies focus on the apparel and footwear industries, similar patterns have been established in electronics (Raj-Reichert 2013, Distelhorst et al. 2015, Yang and Gallagher 2017) and agricultural products (Riisgaard 2009, Coslovsky and

Locke 2013, Dietz, Grabs and Estrella Chong 2019).[3] These findings raise questions about the efficacy of private regulation of supply chains and whether certain approaches are more effective than others.

The Cooperative Approach to Supply Chain Regulation

One approach to improving supplier compliance emphasizes cooperation between buyers and suppliers. The cooperative approach has theoretical roots in scholarship on supplier governance and regulation. Scholars have long observed that buyer–supplier interactions are embedded in social relationships (Uzzi 1997) and that such relationships enable information transfer (Sako 2004) and increase value capture (Dyer and Singh 1998). A cooperative approach builds on these social relationships, using reciprocity and information sharing to improve supplier social performance (Locke, Amengual and Mangla 2009). Such an approach was highlighted in the earliest scholarship on managing supplier social performance (Frenkel and Scott 2002) and has gained widespread acceptance in research and practice. Lund-Thomsen and Lindgreen (2014, p.19) argue that "a coalition of academics, consultants, leading retailers, and NGOs has advocated a new, cooperation-based paradigm to rectify the shortcomings" of punitive approaches to supply chain regulation. Short, Toffel and Hugill (2020) also note that, "a consensus has begun to emerge among academics and practitioners that suppliers are more likely to improve with a less punitive, more cooperative approach."

The cooperative approach responds to one view of labor and environmental violations in global supply chains. As regulatory scholars observe, violations often result from "ignorance, incompetence, inattention, and internal conflict" within organizations (Ayres and Braithwaite 1992, p.82). Compliance with social standards often requires substantial changes in organizational processes and systems that are difficult for managers to carry out and sustain. The cooperative approach envisions buyers sharing knowledge to enable the adoption of new management practices and exhibiting patience as suppliers undergo complex transitions (Jiang 2009, Locke 2013, Gereffi and Lee 2016). Under a cooperative approach, monitoring for compliance seeks to help suppliers identify problems and find solutions. Buyers act more as consultants that help suppliers than as police that seek to uncover and penalize violations (Ayres and Braithwaite 1992). Research suggests that buyers are well-qualified to play this role, with an ability to understand and troubleshoot production problems that lead to compliance challenges (Amengual 2010). Effective private regulation therefore uses, "joint problem solving, information sharing, and the diffusion of best practices" to improve supplier compliance (Locke, Amengual and Mangla 2009).

A growing body of evidence suggests that opportunities to learn and develop new organizational capabilities can improve compliance in global supply chains. Suppliers that adopt certified management systems show greater improvement in compliance, in part due to their ability to identify problems and learn from their failings (Bird, Short and Toffel 2019). Buyer-introduced

[3]One exception to this trend comes from the toy manufacturing industry, where a series of projects in China found evidence of substantial improvement over time (Egels-Zandén 2007; 2014).

capabilities around new compensation systems (Lollo and O'Rourke 2020), worker participation institutions (Boudreau 2020), and work organization (Distelhorst, Hainmueller and Locke 2017) have all been shown to improve supplier compliance or working conditions. Supply chain auditors that enjoy more cooperative relationships with suppliers appear to be more effective in improving supplier compliance (Short, Toffel and Hugill 2020). These findings align with evidence that firms in emerging markets can improve a range of performance outcomes by adopting new management practices with support from external consultants (Bloom et al. 2013).

The cooperative approach does not involve threats to penalize suppliers because research suggests threats have adverse effects on learning and trust, thereby inhibiting improved compliance. Indeed, surveys of buyers suggest that only a small minority use penalties, such as the terminating commercial relationships, to promote compliance with sustainability standards (Porteous, Rammohan and Lee 2015). Researchers have distinguished the use of "voice" and "exit" in buyer–supplier relationships, with the former supporting greater innovation and capability development (Helper 1990). While a minimal threat of exit is necessary for voice, excessive reliance on exit undermines voice (Hirschman 1970). Coercive practices by buyers, most notably the threat of ending commercial relationships, are understood to have a broadly negative effect on buyer–supplier relations (Maloni and Benton 2000). When suppliers perceive the demands of buyers to be threatening, they are less likely to enter into knowledge-sharing relationships with buyers needed to improve compliance (Soundararajan and Brammer 2018). The possible negative effects of penalties suggest that "carrots" are preferable incentives to "sticks" in managing supplier sustainability (Porteous, Rammohan and Lee 2015).

Congruently, scholars of regulation have noted the corrosive effects of strong penalties on cooperative efforts to improve compliance (Bardach and Kagan 1982). When the targets of regulation view threatened penalties as overly burdensome or arbitrary, they may respond indignantly and resist (Kagan and Scholz 1984). Research on regulated organizations shows that personnel take varying views of regulators; some interpret regulators as allies, who help them improve their organizations, while others view regulators as threats or obstacles (Gray and Silbey 2014). When the regulator is perceived to be an ally, individuals within organizations charged with making changes can draw upon regulators for advice and information. By contrast, threatening and police-like behavior by regulators can damage the basis for cooperation and engender opposition (Kagan and Scholz 1984). In addition to these adverse effects on learning and partnership, punitive approaches can also reduce the intrinsic motivations that firms have to comply (Short and Toffel 2010).

Research also suggests that the cooperative approach to regulation may be more effective when buyers have long-term relationships with suppliers. Long-term relationships and repeated exchange between firms have a variety of positive effects, including the cultivation of social relationships, goodwill, and trust (Dore 1983, Gulati 1995, Gulati and Nickerson 2008). When transactions are repeated and embedded in social relationships, buyers and suppliers tend to engage in more cooperation, information exchange, and joint problem solving (Uzzi 1997, McMillan and Woodruff

1999, Dyer and Chu 2003, Sako 2004), increasing the likelihood of sharing information that can help improve compliance (Jiang 2009). Both buyer and supplier benefit from the relational capital generated through repeat exchange (Elfenbein and Zenger 2014), reducing fear that either party will take advantage of the other. Trusting relationships are especially important for knowledge sharing among boundary spanners, such as social compliance managers who visit suppliers (Janowicz-Panjaitan and Noorderhaven 2009). This trust could in turn enable the transfer of compliance-enhancing capabilities from buyers to suppliers. The combination of long-term commercial relationships with a cooperative approach to supply chain regulation may therefore be most effective in promoting supplier compliance with sustainability standards.

In sum, the cooperative perspective suggests that buyers will improve the compliance of their suppliers when they focus supplier attention on unacceptably low levels of compliance and jointly plan steps for remediation, even without threatening to penalize suppliers. Scholarship on supply chain governance additionally suggests that suppliers engaged in longer-term relationships should improve more under this cooperative approach to regulation.

H1. Cooperative Approach. Buyers notifying suppliers their performance falls below expectations and jointly developing plans to remediate violations will cause improved compliance.

H2. Cooperative Approach in Long-Term Relationships. Buyers notifying suppliers their performance falls below expectations and jointly developing plans to remediate violations will cause greater improvement in compliance among suppliers that have longer relationships with the buyer.

Penalties in Supply Chain Regulation

Notwithstanding these benefits of cooperative buyer–supplier relationships, other perspectives on supplier governance and regulation suggest the necessity of penalties to raise supplier compliance. This argument arises from a different view of the drivers of regulatory violations. Whereas the cooperative perspective points to a lack of knowledge and capabilities, suppliers might be more calculative and require incentives to comply with buyer demands (Williamson 1991). Violations can result from the decisions of "amoral calculators" that correctly judge that complying would cost more than violating (Kagan and Scholz 1984, p. 67). Suppliers operating by this logic may act opportunistically (Wathne and Heide 2000)—initially promising to comply with social standards and then reneging on that promise after the commercial relationship is established. By this logic, increasing the costs imposed on offenders will generally reduce the supply of violations (Becker 1968). Although our understanding of supply chain relationships and regulatory compliance has been enriched by the perspectives discussed in the preceding section, the expected penalties associated with offending remain important in many contemporary models of regulatory compliance (Basu, Chau and Kanbur 2010, Kelly 2010, Ji and Weil 2015).

In the absence of penalties, suppliers may infer that sustainability demands are subordinate to the many other demands that buyers make. Although suppliers contractually commit to complying with social and sustainability standards, they may believe that this requirement is a

mere formality, meant for avoiding criticism from external stakeholders but not a priority of the buyer. Indeed, studies have shown that many buyers often undermine their own sustainability standards by placing competing demands on suppliers. International sportswear brand Nike Inc determined that its own purchasing practices led to overtime violations in factories, undermining its requirement that suppliers respect overtime limits (Locke 2013). Even as buyers ask suppliers to develop new management systems to maintain compliance, a "price squeeze" ensures suppliers have little leeway to adopt any practice that might increase costs (Anner 2018). Conflicts between efficiency demands and sustainability demands are exacerbated by an unforgiving competitive environment in many global supply chains. Suppliers in many industries face intense competition and have correspondingly narrow profit margins. They face strong pressure from buyers to keep costs low, maintain product quality, and meet shipping deadlines (Frederick and Gereffi 2011). In light of these pressures and conflicting demands from buyers, suppliers could reasonably interpret sustainability standards as ceremonial and subordinate to other performance demands (Kuruvilla et al. 2020). Introducing penalties for noncompliance is one way to elevate the importance of labor and environmental standards, communicating to suppliers that they are similarly important to established performance expectations around price, quality, and delivery.

Long-term commercial relationships may also moderate the effectiveness of penalties—especially threats to end the buyer–supplier relationship—in stimulating compliance. First, suppliers likely capture greater value from their longer-term commercial relationships. Suppliers tend to capture more value from transactions with existing buyers (Elfenbein and Zenger 2014), and evidence suggests that the value of existing relationships increases with relationship duration (Macchiavello and Morjaria 2015). Buyers engaged in longer-term relationships with fewer suppliers provide higher markups to those suppliers (Cajal Grossi, Macchiavello and Noguera 2019). Because suppliers derive greater benefit from long-term commercial relationships, they have greater incentive to preserve these relationships.

In addition to their greater value to suppliers, longer-term relationships may also be associated with greater supplier lock-in (i.e. higher costs of switching buyers). Long-term commercial relationships allow for greater asset specificity—in which converting current assets to serve some new buyer would result in loss of value—in the absence of vertical integration (Joskow 1988, Williamson 1991). In industries like apparel, where physical assets are highly flexible, asset specificity can arise through organizational processes, such as integration into a buyer's design and planning systems (Abernathy et al. 1999). If long-term relationships are associated with greater supplier asset specificity, ending a long-term commercial relationships will impose greater switching costs than ending a short-term relationship (even if both relationships were equivalent from a value-capture perspective). Dependence on current buyers generated by asset specificity can in turn make suppliers more responsive to buyer demands for sustainability (Delmas and Montiel 2009).

These perspectives suggest private regulation may be effective when buyers penalize noncompliance. Rather than undermining information transfer, penalties may encourage learning by providing incentives for suppliers to take the advice of compliance managers seriously. In addition,

the higher value and switching costs associated with long-term commercial relationships may make suppliers more responsive to threats to end the commercial relationship.

H3. Penalties. Notifying suppliers their performance falls below expectations, jointly developing plans to remediate violations, *and threatening penalties for failing to improve* will cause improved compliance.

H4. Penalties in Long-Term Relationships. Notifying suppliers their performance falls below expectations, jointly developing plans to remediate violations, *and threatening penalties for failing to improve* will cause greater improvement in compliance among suppliers that have longer relationships with the buyer.

Research Design

We study the effects of private regulation in over one thousand suppliers of Gap Inc, an international apparel retailer whose brands include Gap, Banana Republic, Old Navy, and Athleta.[4] We begin by describing Gap Inc and its compliance program. Then we describe the regression discontinuity design and our approach to estimating the effects of key elements of Gap Inc's program. In the following empirical sections we report the results, probe the robustness of these analyses, and test hypotheses about heterogeneous effects by the length of supplier relationship.

Gap Inc is a large retailer; in 2018, it had 3,666 stores, 135,000 employees, and USD $16.6 billion in worldwide sales. Like many leading apparel retailers, Gap Inc has been targeted by activist campaigns for the labor conditions in its supplier factories. These campaigns seek to leverage Gap Inc's position as a major apparel buyer to improve working conditions in its worldwide network of supplier factories. Its suppliers are located primarily in emerging markets that dominate the global apparel trade, such as China, India, and Vietnam. These countries generally have weak regulatory institutions of labor standards enforcement and poor respect for freedom of association.

Gap Inc's Supplier Responsibility Program

To manage labor and environmental issues in its supply chain, Gap Inc has adopted an internal supplier responsibility program. Suppliers agree to a code of conduct that imposes constraints on their workplace practices, including maximum overtime hours, the content of employment contracts, and workplace safety. The code generally stipulates that suppliers comply with local legal requirements and the code's standards, whichever is more stringent. Gap Inc's supplier responsibility department audits supplier practices and implements programs to improve compliance. Social auditors visit each factory annually for a 1–2 day inspection, depending on factory size. They

[4]Gap Inc provided access to its data and personnel for this study. Gap Inc did not fund the research or the researchers, nor did it exercise any control over the conclusions of the study. The company was given an opportunity to review the research output for disclosure of confidential information as well as an opportunity to appear anonymously in publication if desired.

inspect the physical plant for compliance with basic health and safety standards, such as adequate fire exits and ventilation. They also review documents, including worker contracts and payroll records, comparing them to legal standards. Finally, auditors interview managers and between 5 and 50 workers. Most audits are conducted by Gap Inc's supplier responsibility staff, but in a subset of factories they are conducted by the independent Better Work program, which is jointly run by the International Labor Organization and the International Finance Corporation.[5]

The social auditing system defines more than 700 categories of violations, each of which is associated with a severity level from 'low' to 'highest.' Table 1 summarizes the most frequently detected labor violations in each severity level. The most common violations are low-severity and deal with health and safety compliance, such as protective equipment and emergency preparedness. Audits detect more severe violations, such as excessive overtime hours and insufficient rest days, somewhat less frequently. In a small minority of factories, audits detect the most serious violations, such as verbal abuse or obstructed emergency exits.

Buyers use social audits to assess the compliance of their supplier factories. However, social audits are not well suited for detecting violations of certain standards, such as the rights to form or join a union (Anner 2012, Distelhorst et al. 2015). Notwithstanding this limitation, these inspection results offer some of the best available data on a range of labor standards—such as wages, overtime, contracts, and benefits—in developing countries and are therefore commonly used in research (Locke, Qin and Brause 2007, Kortelainen 2008, Short, Toffel and Hugill 2016, Locke 2013, Bird, Short and Toffel 2019, Liu et al. 2019, Short, Toffel and Hugill 2020).

Each labor audit produces a corrective action plan and a compliance rating. The corrective action plan details all the violations detected, outlines steps the supplier should take to correct them, and provides timelines for this process. Supplier responsibility staff then provide technical support to factories as they implement the plan and monitor their progress. For example, supplier responsibility staff might explain how to improve emergency evacuation routes or how to re-organize production to reduce excessive overtime. These are typical cooperative remediation steps, documented in previous research in other buyers (Locke, Amengual and Mangla 2009).

Each audit also distills the high-dimensional compliance data into a simple categorical rating of overall compliance. The categorical rating in turn prompts responses that derive from the category, rather than the complex set of circumstances in each case, a common phenomenon in organizations (Ashforth and Humphrey 1997). In the Gap Inc system, categorical compliance ratings first communicate the level of (dis)satisfaction with the factory's social performance. The lowest rating, failure, is intended to communicate both to the supplier and to internal audiences

[5]Better Work audits factories in the garment industry and seeks to improve labor conditions. In 2018 it engaged over 1,400 factories across Bangladesh, Cambodia, Haiti, Indonesia, Jordan, Nicaragua, and Vietnam. URL: http://www.betterwork.org. Where Better Work operates, buyers often rely on Better Work social audits rather than their own internal audits. In certain countries, such as Cambodia, all garment exporters must participate in Better Work. In others, such as Indonesia, factories opt in to Better Work, often at the encouragement of their buyers.

Table 1: Most frequently-detected violations in labor audits, by severity

Low-severity violations	
43%	Exit routes / emergency routes
36%	Personal protective equipment
35%	Machine/Equipment safety
29%	First aid and medical
26%	Emergency procedures and evacuation drills
Medium-severity violations	
17%	One day off in 7
13%	Worker schedules
12%	Payment of benefits
10%	Reasonable emergency leave period
8%	Employment contracts availability
High-severity violations	
18%	Payment of benefits
15%	Overtime hours
10%	Overtime and incentive rates
9%	Minimum wage requirement
5%	Payment for leave not taken
Highest-severity violations	
5%	Documentation transparency
5%	Exit routes / emergency routes
4%	Verbal abuse
3%	Access to facility, workers, and records
2%	Unauthorized subcontracting

that labor practices have fallen below acceptable standards.[6] Compliance ratings also trigger different responses within the supplier responsibility team. Throughout the period of our study, supplier responsibility managers prioritized failing factories for additional engagement and increased monitoring of progress remediating violations.

Gap Inc's social compliance ratings are determined by a formula. Each factory begins with a full score, from which points are deducted for each violation, according to their severity and whether the violation was detected in previous audits. Before 2015 violations had two levels of severity, and suppliers could earn additional points for achieving sustainability certifications such as SA8000. In 2015, a re-scaled audit scoring system created four levels of severity (see Table 1), stopped adding points for certifications, and changed the point values for repeat violations. Notwithstanding these changes to the scoring formula, the large majority of violation definitions themselves remained unchanged.

To assign the compliance rating, auditors enter all detected violations in a system that computes a numeric score associated with the audit. This computing system resides in the regional

[6]Gap Inc uses colors in its social compliance ratings. The highest performers are rated green, middling performers are rated yellow, and the lowest-performing factories are rated red. For clarity, we refer to red ratings as failing throughout.

offices—the auditor does not calculate the score and rating until after departing from the factory. The compliance rating is entirely determined by whether the score is greater or less than a numeric threshold. Factories scoring below the failing threshold fail, and those above this threshold pass. Gap Inc set the failing threshold at a level broadly related to the risks associated with the supplier, but locally the threshold is arbitrary.[7] After the rating is determined, Gap Inc's staff communicates it to the factory management, which is not permitted to lobby for changes in the rating. The rating then places the factory in a particular category, and that categorization triggers a set of actions.

Cooperation and Penalties in the Supplier Responsibility Program

Gap Inc offers an instructive setting to study private regulation because it used different approaches to regulating supplier compliance during our study period.[8] Prior to 2016, a failing audit rating communicated that the factory's level of labor compliance was substandard and prompted supplier responsibility staff to increase monitoring and support for implementing corrective action plans. Engagement with failing factories was constructive and cooperative. Failing factories were prioritized for more frequent communication and more attention was paid to designing corrective action plans. Compliance staff assisted these factories to learn how to address the underlying causes of violations, and followed up to ensure that the actions were taken. Compliance staff also encouraged these suppliers to take ownership over the process, seeking to engage their intrinsic motivation to improve. Only rarely did egregious violations lead to termination of a buyer–supplier relationship. Analyzing the probability of termination from the supply chain, we see little difference between factories that passed and failed social compliance audits over 2010–2015 (Figure 1).

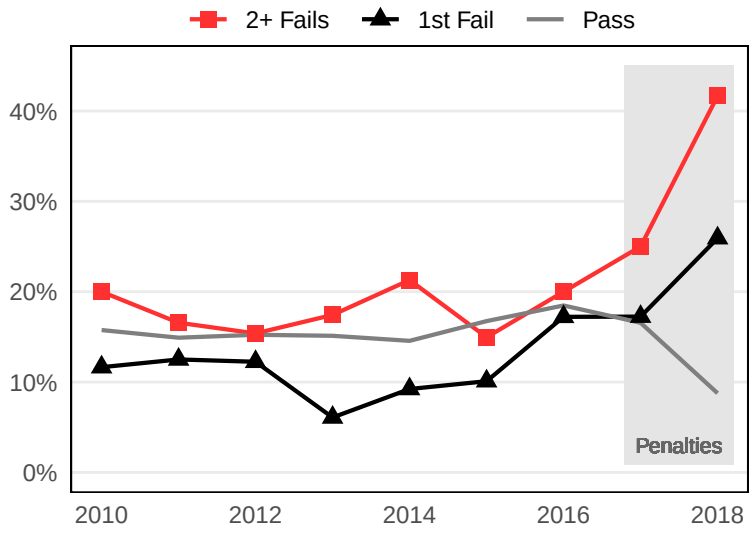
Gap shifted its approach in 2016 by incorporating penalties into its supplier responsibility system. New managers determined that to improve compliance they needed to continue supporting factory remediation while also penalizing those that repeatedly failed compliance audits.[9] When suppliers failed an audit, the relevant sourcing and social responsibility managers received an automatic notification: "Please be advised that the following factories in your region each have a newly submitted assessment with [a failing compliance rating]." These managers contacted the failing factories and instructed them to improve sufficiently to pass subsequent audits or else

[7]Compliance ratings are similar to exam grades in school. Exams will have cutoffs in meaningful ranges, but locally the difference between a student that scores one point above or one point below any given threshold is minimal.

[8]To understand the compliance rating system and the changing relationship between sourcing and compliance, we conducted 22 interviews with Gap Inc managers in both its corporate headquarters in San Francisco and Asia office in Hong Kong.

[9]This change was prompted by the combination of the Rana Plaza disaster several years earlier and a new management team. The Rana Plaza disaster, in which over one thousand people were killed in the collapse of an export factory in Bangladesh, shockingly illustrated the risks workers faced in the global garment industry. Activist campaigns increased pressure on apparel firms to improve labor conditions in their suppliers. This resulted in increased public scrutiny and increased awareness of supply chain labor issues among senior managers at Gap. When new leaders later took over the supplier responsibility program, they sought to change the approach to supply chain compliance.

Figure 1: Factory terminations by labor compliance audit result



Notes. Conditional probability of factory termination within 12 months by audit rating. Factories passing, failing for the first time, and failing in two or more consecutive audits plotted separately. Results from the first two months of 2019 pooled into the 2018 data.

risk termination of the commercial relationship. To make the threat of penalties credible, Gap Inc sourcing managers who maintained commercial relationships with suppliers assumed a more prominent role in managing labor compliance. Senior sourcing managers began attending meetings with factories that failed audits, emphasizing the commercial consequences if suppliers did not bring their labor practices into compliance. Lower-level sourcing staff in more frequent communication with suppliers also reinforced this message in both informal conversations and formal quarterly reviews. In interviews, social responsibility managers noted the motivating power of these messages from sourcing personnel.

The introduction of penalties did not end efforts to support compliance by transferring knowledge and engaging in joint problem-solving. Sustainability managers continued to work with failing factories on corrective action plans and to provide guidance on implementation. Factories that failed typically received at least one opportunity to remediate violations and reach compliance. Gap managers described an informal "two strikes" rule: suppliers that failed one audit would receive at least one more chance to achieve compliance in the following audit. This new approach to supply chain labor compliance, which began in mid-2016, is evident in the changing pattern of supplier terminations over this period. Figure 1 shows that termination of failing factories increased markedly in 2017 and 2018, especially for those that failed two audits in a row.

Table 2: Consequences of failed labor compliance audits

	2010–2016	2016–2019
Failing grade indicates that supplier's employment practices fall short of buyer's standards.	✓	✓
Supplier responsibility personnel provide technical support for remediating violations.	✓	✓
Sourcing personnel communicate possibility of commercial penalties for failing to improve.		✓

Regression Discontinuity Design and Data

This study uses a regression discontinuity design to estimate the causal effects of labor compliance ratings under the two different approaches to supply chain regulation described above. As discussed above, passing or failing a labor compliance audit is decided by whether a factory reaches a numeric threshold in scores generated by rules related to the severity of violations and their recurrence. We therefore collected the results of labor compliance audits of Gap supplier factories from 2010 to 2019. We study repeated compliance audits at the same factory: a "previous" audit that determines whether a factory previously failed and a second "outcome" audit that shows their labor compliance in the following audit (our dependent variable). Due to a rescaling of the compliance rating system in 2015, our main analyses cover three periods: under the old compliance rating system (2010–2014), under the current compliance scoring system but *before* the the use of penalties (2015 to mid-2016), and *after* the incorporation of penalties (mid-2016 to 2019). Over 2010–2019, we observe before-and-after audit pairs in 1,366 unique factories across 33 countries.

Regression discontinuity designs can recover unbiased estimates of causal effects if the potential outcomes of the dependent variable (here, compliance scores in the outcome audit) are a continuous function of the running variable (here, the previous audit compliance score) across the threshold that determines treatment status (Cattaneo, Idrobo and Titiunik 2018). By modeling these functions and computing the difference in predicted values at the threshold, our analysis estimates a local average treatment effect of a failing rating—and the attendant regulatory actions summarized in Table 2—on compliance.[10] One detail of Gap's supply chain program posed a threat to this design. Roughly one year after introducing penalties, the interval between audits for some failing factories was shortened from one year to approximately six months. The change was prompted by the view that six months offered sufficient time to remediate even major violations,

[10]Regression discontinuity designs estimate *local* causal effects, meaning the effect we estimate only pertains to factories precisely at the threshold of failure. The regression discontinuity estimate is not informative of factories whose scores fall farther from that threshold. We might imagine that factories that fail with scores farther from the threshold of failure could either: (a) feel more pressure to improve, resulting in larger effects of a failing, or (b) view achieving compliance as impossible and put less effort into improvement. It is difficult to say whether the causal effect of failure would be greater or smaller away from this threshold.

and that the uncertainty generated by an unresolved failed audit hampered planning by both Gap sourcing staff and their supplier factories. However, if these quick audits were timed to coincide with the completion of corrective action plans needed to pass labor audits, they might violate the assumption of as-if random variation around the numeric threshold of failure. To address this threat to identification, we limit the audits that determine pass/fail status after the 2016 introduction of penalties to only the *first* audit conducted in this period. This reduces our statistical power but offers improved causal identification. We note that the findings reported below are not dependent on restricting our analysis to the sample of first audits. Analyzing a sample of *all* audits after introduction of penalties yields more precise and stronger estimated effects. Moreover, none of these issues affect the subsample of independent audits conducted by the ILO/IFC Better Work program used to test the robustness of our findings.

An important assumption of the regression discontinuity design is that neither factories or auditors strategically adjust scores to slightly exceed or undercut the threshold of failure. In other words, while factory managers may put more or less effort into compliance overall, they cannot precisely control their scores to put in the minimum effort necessary to just barely pass. Such precise manipulation would bias our estimates if it was correlated with potential outcomes in the following audit.

Precise score manipulation is unlikely for several reasons. First, although managerial decisions strongly influence compliance with labor standards, events outside of management control make it impossible to precisely control compliance levels on the day(s) of the audit. For example, unexpected power outages or delays in inputs can trigger excessive overtime in factories that would have otherwise complied with overtime limits. Management systems within factories can also break-down or over-perform, affecting compliance. Even in a factory with management systems to prevent emergency exits from becoming blocked, a worker could make a bad decision on the day of the audit and place material to block an exit. Conversely, a factory with no systematic safeguards might comply simply because workers had no need to pile up material in front of the exit on the day of the audit. Thus, while management systems affect the probability of violations, they do not exert complete control over the portfolio of violations that auditors can detect. These minor shocks to compliance levels make it nearly impossible for management to precisely control compliance to land just above the threshold of failure in the Gap Inc audit.

Second, even if factories could control their exact level of compliance, factory management would also need to know which violations would be detected by auditors to precisely manipulate scores. Factory managers, however, do not have this knowledge. The audit process is designed to penetrate attempts by management to hide true levels of compliance by cross-checking documents, inspecting visible manifestations of compliance, and interviewing workers. In addition, some parts of the audit processes involve random checks. For example, auditors usually select a sample of workers to interview effectively at random (e.g. by walking up and down the production line and choosing a subset). What auditors learn depends in part on the varying experiences of those workers and variation in their willingness to disclose issues. For issues like remuneration and

overtime, auditors sample paystubs and management cannot know in advance which employees' records will be audited. Therefore, even if a factory could choose to have the minimum number of violations to be just above the threshold of failure (which it cannot), managers would also have to know exactly which payroll records auditors would review to have precise control over their scores.

Third, calculating the score that determines the compliance rating is complex. Auditors can assess over 700 different violations which vary in their point value. Even within a particular labor standard (e.g. overtime limits), violations differ in their severity and point value. Gap Inc's rating system also applies multipliers for repeat offenses. To calculate the score, an individual needs to know not only the point values of each violation, but also the factory's audit history to assess repeat offenses. Given the complexity of calculating scores, even the auditors do not know the score until they return to Gap Inc offices and enter the detected violations into the computer system. Factories have no opportunity to "lobby" for changes in their scores after the rating is assigned.

To examine the assumption that factories that just-passed and just-failed compliance audits were on average very similar to one another, we collected administrative records of factory covariates relevant to social compliance, including total workers, location, previous audit scores, units purchased from the factory, and many others. To test hypotheses about the heterogeneous effects according to the length of the buyer–supplier relationship, we also gathered information about the the duration of each factory's commercial relationship to Gap Inc. Descriptive statistics for supplier factories under the old and current ratings systems appear in Tables 3 and 4.[11] We observe more factory covariates in the later period than under the older audit scoring system.

If scores near the threshold are as-if randomly distributed, we expect supplier covariates to exhibit little difference on either side of the threshold. We therefore conduct regression discontinuity estimates of these indicators for each period in our study. We detect no serious threats to inference in the 2010–2014 compliance rating data (Table 5).[12] We note that Vietnamese factories are slightly overrepresented just above the threshold of failure. Because each validation exercise generates a large table, we report results of the 2015–2016 period in the Appendix (Table A1). We note that in this period, we detect one imbalance in pretreatment covariates—the count of workers in failing factories was lower ($p = 0.04$). Our analyses find a very weak, *negative* association between total workers and compliance scores; increasing the workforce by 50% is associated with a small 0.7 point (0.039 standard deviation) reduction in compliance score. Because the correlation is weak and negative, we expect this difference would only generate a very small upward-bias in the estimated effect of failing.

[11] A few features of the summary statistics require explanation. In some cases the length of the commercial relationship is *negative* at the time of the audit. Labor audits may occur before the factory is formally approved as a supplier. Also, some audits take place years apart rather than annually, likely due to the deactivation and later reactivation of a supplier. Finally, we standardized units shipped from the factory to avoid revealing sensitive corporate information.

[12] Throughout our regression discontinuity analyses, we use first order polynomials, triangular kernel weights, and algorithmically-selected bandwidths, following recommended procedures in Cattaneo, Idrobo and Titiunik (2018).

Table 3: Descriptive statistics: old rating system (2010–2014)

	median	mean	sd	min	max	N
Previous audits						
Audit score	5	1.5	11.9	-114	14	2,366
Low-severity violations	16	19.8	15.3	0	110	2,366
Medium-severity	2	3.0	2.8	0	21	2,366
High-severity	2	2.2	2.2	0	15	2,366
Highest-severity	0	0.6	1.0	0	8	2,366
Better Work audit?	0	0.1	0.2	0	1	2,366
Days since last audit	335	330.8	128.6	91	1218	1,586
Days until outcome audit	336	338.4	139.3	91	1882	2,366
Outcome audits						
Audit score	6	3.6	11.6	-57	35	2,366
Low-severity violations	21	26.1	18.0	0	115	2,366
Medium-severity	3	4.0	3.3	0	24	2,366
High-severity	2	2.9	2.6	0	19	2,366
Highest-severity	0	0.9	1.2	0	8	2,366
Better Work audit?	0	0.1	0.2	0	1	2,366
Factory characteristics						
Pre-previous audit score	6	3.1	12.8	-114	33	1,586
Relationship (years)	5	5.3	4.2	-3	18	2,277
ln(units shipped) (std)	0	-0.0	1.0	-1	1	1,133
Workers	708	1237.7	1735.6	1	20800	2,362
Female workers (%)	67	61.6	24.6	0	100	1,352
Manufacturer?	1	0.8	0.4	0	1	2,366
Factory in China	0	0.3	0.5	0	1	2,366
India	0	0.2	0.4	0	1	2,366
Indonesia	0	0.0	0.2	0	1	2,366
Vietnam	0	0.1	0.3	0	1	2,366
other country	0	0.3	0.5	0	1	2,366

Notes. Descriptive statistics for data under the old compliance rating system (942 total factories, unit of analysis is one audit). Audit scores zeroed at the threshold of failure; positive scores indicate the factory passed the audit. In this audit-level data factories may appear repeatedly, depending on how many total audits they experienced. For consistency with the current scoring system, we mapped violations in the old system on to the four levels of severity used today. The old system only recognized two severity levels.

Table 4: Descriptive statistics: current rating system (2015–2019)

	median	mean	sd	min	max	N
Previous audits						
Audit score	8	4.4	15.7	-64	22	1,266
Low-severity violations	10	11.2	5.9	0	45	1,254
Medium-severity	2	2.3	1.7	0	12	1,254
High-severity	1	1.0	1.1	0	6	1,254
Highest-severity	0	0.3	0.6	0	3	1,254
Better work audit?	0	0.2	0.4	0	1	1,266
Days since last audit	344	328.1	87.6	91	749	709
Days until outcome audit	350	336.5	99.0	91	1091	1,266
Outcome audits						
Audit score	12	9.9	15.7	-64	36	1,266
Low-severity violations	9	10.3	5.8	0	44	1,112
Medium-severity	2	2.2	1.7	0	14	1,112
High-severity	1	0.9	1.0	0	7	1,112
Highest-severity	0	0.2	0.5	0	3	1,112
Better work audit?	0	0.2	0.4	0	1	1,266
Factory characteristics						
Pre-previous audit score	12	7.5	18.7	-64	36	709
Relationship (years)	6	6.9	5.3	-2	19	1,244
ln(units shipped) (std)	1	0.1	1.0	-1	2	1,266
Workers	807	1351.6	1695.2	1	15500	1,237
Female workers (%)	66	62.0	23.6	0	100	1,137
Manufacturer?	1	0.8	0.4	0	1	1,266
Americas	0	0.0	0.2	0	1	1,266
Mediterranean	0	0.0	0.2	0	1	1,266
North Asia	0	0.3	0.4	0	1	1,266
South Asia	0	0.3	0.5	0	1	1,266
Southeast Asia	0	0.3	0.5	0	1	1,266
Factory in China	0	0.3	0.4	0	1	1,266
India	0	0.2	0.4	0	1	1,266
Indonesia	0	0.1	0.3	0	1	1,266
Vietnam	0	0.2	0.4	0	1	1,266
other country	0	0.2	0.4	0	1	1,266

Notes. Descriptive statistics for data in the current compliance rating system (702 total factories, unit of analysis is an audit). Audit scores zeroed at the threshold of failure; positive scores indicate the factory passed the audit. Because this is audit-level data, factories may appear repeatedly, depending on how many audits they experienced.

Table 5: Covariate balance, cooperative period (2010–2014, old rating system)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	0.01	1.00	[4.7, -4.6]	7.8	1,127
Medium-severity	-0.64	0.12	[0.2, -1.5]	11.2	1,641
High-severity	-0.52	0.11	[0.1, -1.2]	10.7	1,566
Highest-severity	-0.00	0.98	[0.3, -0.3]	12.9	1,953
Better Work audit?	0.02	0.61	[0.1, -0.0]	12.9	1,954
Days since last audit	-35.50	0.10	[6.9, -77.9]	9.8	972
Days until outcome audit	16.48	0.43	[57.3, -24.3]	7.2	1,025
Better Work audit?	0.02	0.61	[0.1, -0.0]	12.9	1,954
Pre-previous audit score	0.80	0.61	[3.9, -2.3]	11.5	1,160
Relationship (years)	1.10	0.19	[2.7, -0.5]	7.5	1,064
ln(units shipped) (std)	0.12	0.60	[0.6, -0.3]	9.4	653
Workers	160.75	0.55	[685.7, -364.2]	11.2	1,634
Female workers (%)	2.54	0.59	[11.7, -6.6]	10.9	955
Manufacturer?	-0.04	0.50	[0.1, -0.2]	8.6	1,260
Factory in China	0.01	0.84	[0.1, -0.1]	12.7	1,918
India	-0.01	0.89	[0.1, -0.1]	10.5	1,492
Indonesia	0.01	0.75	[0.1, -0.1]	8.4	1,214
Vietnam	-0.11	0.04	[-0.0, -0.2]	7.8	1,127
other country	0.10	0.17	[0.3, -0.0]	9.0	1,297

Notes. Tests of continuity of pre-treatment covariates at the threshold of failure in the old ratings system (2010-2014). Positive values indicate that factories that passed have higher values than those that failed. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018).

Table 6 reports covariate balance after introduction of penalties. In covariates that might pose threats to inference—like factory workforce ($p = 0.70$), total units shipped to Gap ($p = 0.80$), the length of the buyer–supplier relationship ($p = 0.68$), and audits by the independent Better Work program ($p = 1.00$)—we observe good balance in passing and failing factories at the threshold of failure. Across two dozen statistical tests, only one rises to conventional levels of statistical significance: days elapsed between the previous and outcome audits. As discussed above, we understood that Gap reduced the period between audits for failing factories and therefore expected this imbalance. In observational analyses of audit data, we find that longer time between audits is generally associated with *higher* scores, consistent with a model in which factories benefit from having more time to bring themselves into compliance. We therefore do not believe the shorter audit cycle conferred any strong advantage to the failing factories. Notwithstanding this observation, in a later robustness check we exclude these "quick audits" from the estimation sample.

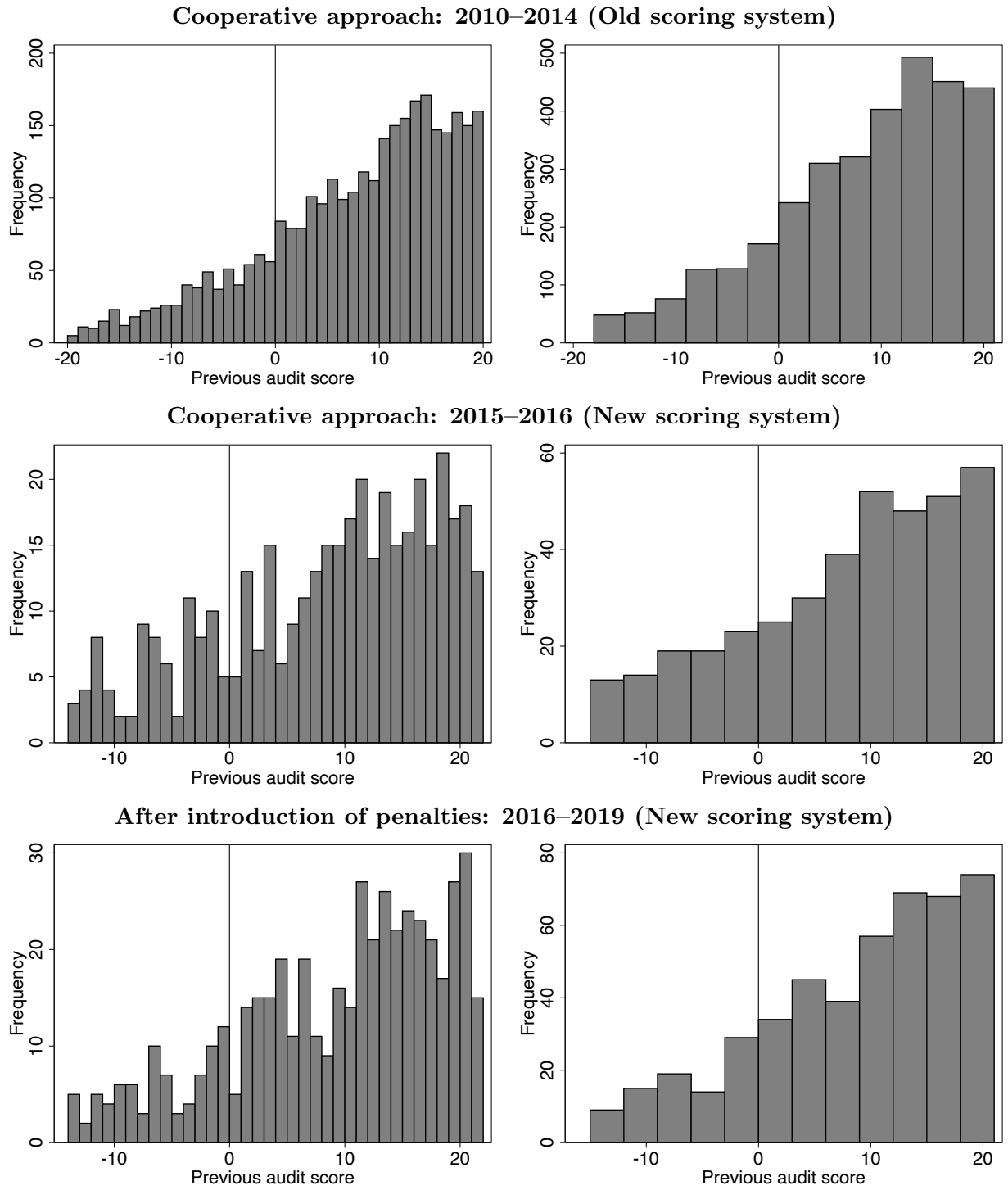
In addition to covariate balance, we also examine the distribution of audit scores around the threshold of failure. If factories were able to precisely foretell their audit scores and put in minimum efforts to pass, we expect to see a "bulge" of audits just above the threshold of failure. Figure 2 plots histograms of the audit scores near the threshold in three time periods—during the cooperative approach under the old scoring system, during the cooperative approach under the new scoring system, and after the introduction of penalties. Observational density varies smoothly across the threshold in all three time periods. Formal tests of continuity in observation density using local polynomial density estimation fail to reject the null hypothesis of no imbalance, yielding p-values of 0.48, 0.98, and 0.70, respectively.

Table 6: Covariate balance, after introduction of penalties (2016–2019)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	-0.09	0.96	[3.1, -3.3]	13.5	272
Medium-severity	0.29	0.65	[1.5, -0.9]	13.5	302
High-severity	-0.48	0.20	[0.3, -1.2]	10.7	218
Highest-severity	0.16	0.31	[0.5, -0.1]	13.8	302
Better work audit?	0.00	0.98	[0.3, -0.3]	10.3	202
Days since last audit	-17.61	0.59	[47.1, -82.3]	14.8	253
Days until outcome audit	-62.02	0.03	[-5.6, -118.4]	17.6	409
Pre-previous audit score	2.13	0.83	[21.7, -17.4]	10.3	156
Outcome audit by ILO	-0.00	1.00	[0.3, -0.3]	11.8	252
ln(units shipped) (std)	0.08	0.80	[0.7, -0.5]	15.6	360
Relationship (years)	-0.70	0.68	[2.6, -4.0]	13.0	267
Workers	-168.65	0.70	[686.4, -1023.7]	14.6	316
Female workers (%)	2.01	0.79	[16.8, -12.8]	14.9	287
Manufacturer?	-0.11	0.41	[0.2, -0.4]	8.9	180
Americas	-0.11	0.14	[0.0, -0.2]	7.8	165
Mediterranean	0.04	0.57	[0.2, -0.1]	14.1	306
North Asia	0.15	0.38	[0.5, -0.2]	8.0	165
South Asia	-0.25	0.15	[0.1, -0.6]	10.4	202
Southeast Asia	0.16	0.25	[0.4, -0.1]	13.0	275
Factory in China	0.15	0.38	[0.5, -0.2]	7.9	165
India	-0.23	0.19	[0.1, -0.6]	9.9	202
Indonesia	0.09	0.36	[0.3, -0.1]	10.8	220
Vietnam	0.06	0.62	[0.3, -0.2]	10.4	202
other country	-0.07	0.67	[0.2, -0.4]	9.3	180

Notes. Tests of continuity of pre-treatment covariates at the threshold between failing and passing audit scores. Positive values indicate that passing factories have higher values than failing ones. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018). We make no adjustment for multiple comparisons.

Figure 2: Compliance audit score density around the threshold of failure



Notes. Histograms of audit scores that determine pass/fail status in each period of analysis. Vertical lines show the pass/fail threshold. Left-hand plots shows integer-wide bins and right-hand plots shows bins that are three points wide. Plots zoomed to highlight the area around the threshold relevant to regression discontinuity estimations.

Results

Cooperative Approach to Supply Chain Regulation

Our first hypothesis asked whether notifying suppliers their performance was substandard and developing remediation plans (in the absence of penalties), caused compliance to improve (H1). Under this cooperative approach to supply chain regulation, the estimated effect of a failed compliance audit on future compliance is near-zero. Figure 3 visualizes this null finding, showing no difference in average outcome compliance scores at the threshold of failure (0 on the horizontal axis). Detailed results from four alternative algorithmic approaches to selecting the bandwidth around the threshold appear in Table 7. The left-hand panel reports the bandwidth around the threshold and the observations within the bandwidth; the right-hand panel reports the estimated effect magnitude, p-value, and 95% confidence interval. Each bandwidth selection technique shows a precisely estimated null effect of failing a social responsibility audit on subsequent compliance scores. The widest confidence interval across our four estimations spans 6.3 points, less than one half of a standard deviation in outcome compliance scores.

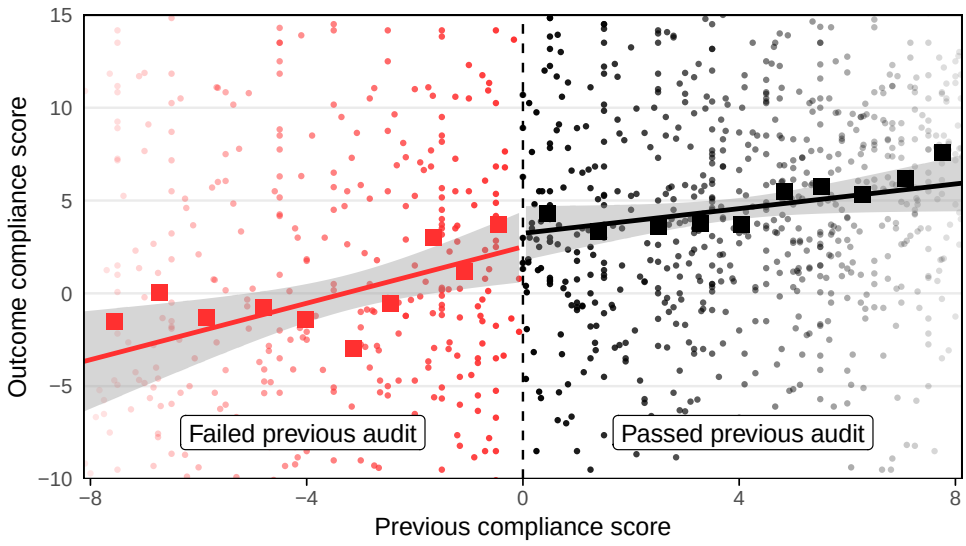
Under the updated compliance rating system in place since 2015, regression discontinuity models again estimate near-zero effects of failing labor compliance audits in the absence of penalties, although these estimates have wider confidence intervals due to smaller samples (Table 8). When the buyer used a predominantly cooperative approach, we detect no evidence that failing a compliance audit prompted factories to improve.

Table 7: Effects during cooperative period (2010-2014)

Bandwidth properties				Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
MSE symmetrical	8.1	8.1	1,164	-1.17	0.41	[-4.0, 1.6]
MSE separate	11.3	6.4	1,079	-1.26	0.34	[-3.8, 1.3]
CER symmetrical	5.6	5.6	800	-0.53	0.74	[-3.7, 2.6]
CER separate	7.8	4.4	758	-1.70	0.25	[-4.6, 1.2]

Notes. Regression discontinuity estimates of the effect of failing ratings on subsequent audit scores in the old rating system (2010–2014). Reports four alternative data-driven approaches to selecting the bandwidth. The first approach selects a bandwidth that minimizes the expected mean squared error (MSE) of the point estimator. The second seeks to minimize the coverage error (CER) of the confidence interval around the point estimate. For each, we report both symmetrical bandwidths and separately chosen bandwidths for distributions above and below the threshold. Details of each procedure in Cattaneo, Idrobo, and Titiunik (2018).

Figure 3: Effect during cooperative period (2010–2014)



Notes. Regression discontinuity estimate of the effect of failing the previous labor compliance audit on the next compliance score *before* introducing penalties (old audit scoring system, 2010–2014). Estimate uses MSE-minimizing symmetric bandwidths and triangular kernel weights. Squares show binned means of equally-sized subgroups around the discontinuity. Points show individual observations, using fading to indicate the kernel weight assigned to each observation. Details including alternative bandwidth selection techniques reported in Table 7.

Table 8: Effects during cooperative period (2015–2016)

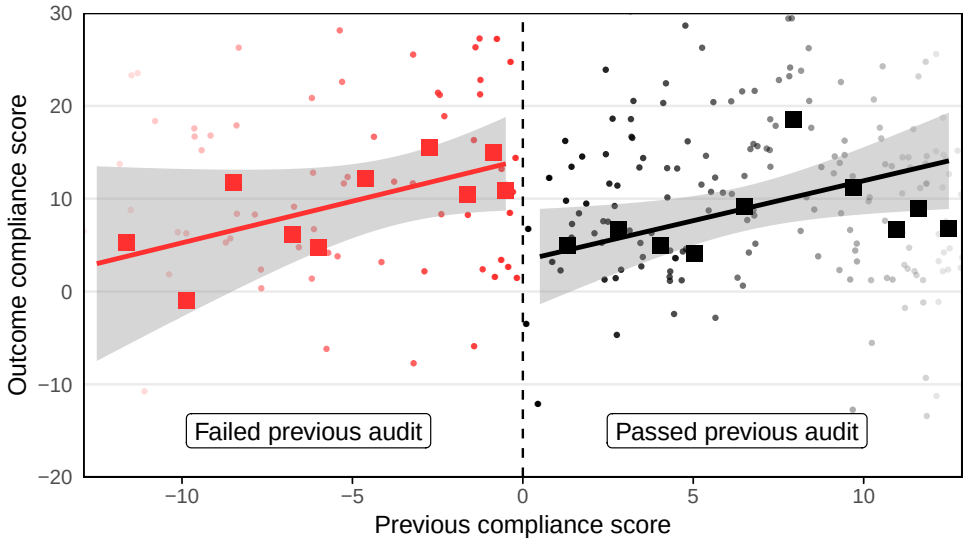
	Bandwidth properties			Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
MSE symmetrical	10.4	10.4	172	-2.23	0.80	[-19.6, 15.1]
MSE separate	13.1	11.7	225	-0.11	0.99	[-15.7, 15.5]
CER symmetrical	7.5	7.5	116	2.00	0.85	[-19.1, 23.1]
CER separate	9.4	8.4	140	1.17	0.90	[-17.1, 19.4]

Notes. Regression discontinuity estimates of the effect of failing ratings on subsequent audit score. Details of each bandwidth selection procedure in Cattaneo, Idrobo, and Titiunik (2018).

Supply Chain Regulation with Penalties

Does the combination of notifying suppliers of substandard performance, developing plans for remediation, and threatening penalties for failure to improve lead to improved compliance (H3)? After the introduction of threats to terminate the buyer–supplier relationship, failing a previous audit caused compliance to improve by 11.6 points $[2.8, 20.4]$ (Figure 4 and Table 9). The local average treatment effect is an increase in mean compliance score of 0.8 standard deviations, causing factories to be 22 percentage points more likely to pass their next audit. The larger squares in Figure 4 show mean compliance scores for equally-sized groups of observations above and below the threshold, illustrating a non-parametric approach to the same analysis. Mean compliance in each of the first four subgroups of failing factories below the threshold is greater than any of the first four groups of passing factories above the threshold. When failing a compliance audit led to both remediation support and threats to discontinue the buyer–supplier relationship, it caused supplier factories to improve.[13]

Figure 4: Effect after introduction of penalties (2016–2019)



Notes. Regression discontinuity estimate of the effect of failing the previous labor compliance audit on the next compliance score *after* the introduction of penalties for failing audits (2016–2019). Estimate uses MSE-minimizing symmetric bandwidths and triangular kernel weights. Squares show binned means of equally-sized subgroups around the discontinuity. Points show individual observations, using fading to indicate their kernel weights. Details including alternative bandwidth selection techniques reported in Table 9.

[13]Simpler analytic approaches reach similar conclusions to those from the regression discontinuity design. Examining cross-tabulations under the current rating system, during the largely cooperative period 64% of failing factories passed their next audit. After the introduction of penalties 85% did.

Table 9: Effects after introduction of penalties (2016–2019)

Bandwidth properties				Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
MSE symmetrical	12.6	12.6	275	11.61	0.01	[2.8, 20.4]
MSE separate	10.5	11.5	220	12.27	0.01	[2.9, 21.6]
CER symmetrical	9.0	9.0	180	11.71	0.02	[2.1, 21.3]
CER separate	7.6	8.3	165	10.63	0.04	[0.7, 20.6]

Notes. Regression discontinuity estimates of the effect of failing ratings on subsequent audit score. Details of each bandwidth selection procedure in Cattaneo, Idrobo, and Titiunik (2018).

Robustness: Excluding quick audits for failing factories

We noted above that Gap subjected failing factories to shorter audit cycles after the introduction of penalties starting in the second half of 2017. Table 6 estimates that, at the threshold, failing factories were next audited two months (62 days) sooner than factories that passed after the introduction of penalties. This does not threaten the validity of regression discontinuity identification assumptions—note that Table 6 estimates no difference in the time since the *last* audit for factories that just-failed and just-passed. However, if we believed it was difficult to sustain compliance for longer periods of time, this shorter audit cycle could allow failing factories to exhibit greater improvement. We therefore re-estimate our effects after excluding any failing factories audited less than nine months (270 days) since their previous audit. In this more restricted sample, we detect no difference in the days until the outcome audit across passing and failing factories ($p = 0.69$, see Appendix Table A2). Table 10 shows results after eliminating the quick follow-up failing factories from the sample are very similar to the full sample, ranging from 9.7 to 12.9 points, with p-values ranging from 0.03 to 0.07.

Table 10: Effects with penalties, excluding quick follow-up (2016–2019)

Bandwidth properties				Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
MSE symmetrical	13.8	13.8	260	9.73	0.07	[-0.7, 20.2]
MSE separate	10.0	11.3	183	12.92	0.03	[1.1, 24.8]
CER symmetrical	9.9	9.9	169	12.19	0.03	[0.9, 23.5]
CER separate	7.2	8.1	136	11.24	0.07	[-0.9, 23.3]

Notes. Regression discontinuity estimates of the effect of failing ratings on subsequent audit score, imposing a minimum of 9 months (270 days) between audits for failing factories. Details of each bandwidth selection procedure in Cattaneo, Idrobo, and Titiunik (2018).

Robustness: Independent compliance audits by ILO/IFC

Despite features of Gap Inc audit system that make it amenable to the regression discontinuity design, we cannot fully rule out a bias in these in-house audits. We therefore replicate the analyses above using an independent source of labor compliance information: audits conducted by the ILO/IFC Better Work program. Better Work is a regulatory program managed by the International Labour Organization and the World Bank's International Finance Corporation in Bangladesh, Cambodia, Haiti, Indonesia, Jordan, Lesotho, Nicaragua, and Vietnam (Rossi, Luinstra and Pickles 2014). Its labor audits are conducted independently and have several helpful features. Better Work is overseen by advisory committees that include representatives from trade unions, employer associations, chambers of commerce, and governments. Under this governance structure, Better Work auditors operate independently from the retailers, such as Gap Inc, that source from these factories. Discussions with officials at Better Work indicate that the program staff who conduct audits are generally unaware of the details of the ratings systems used by the buyers that participate in the program, creating additional separation between the audit process and the assignment of failing scores. In addition, Better Work separates audits from factory improvement work; the ILO officials who evaluate a factory are not involved in helping that factory remediate violations. All audits are unannounced and conducted by two ILO officials. They occur annually and their timing is unrelated to previous audit scores. These features—organizational independence, program design to reduce auditor conflict of interest, and consistent audit timing—address possible concerns about analyses of in-house audits.

The ILO/IFC analyses before and after the introduction of penalties use a similar regression discontinuity design, with the cautionary note that a smaller sample size leaves us with less statistical precision. Observation density around the threshold during both time periods is balanced ($p = 0.72$ before and $p = 0.99$ after, see Appendix Figure A2). Covariates around the threshold are also largely balanced. We detect covariate imbalances in 4 out of 31 statistical tests across both periods (Appendix Tables A3 and A4). In the cooperative period, failing factories had shorter commercial relationships and shipped fewer units. After introduction of penalties, failing factories had a different distribution of violations, with more medium-severity violations but fewer high-severity violations.

Before the introduction of penalties, failing audits did not cause any change in compliance measured by the ILO/IFC auditors (Table 11). After the introduction of penalties, failing caused subsequent compliance scores to increase by 29 points on average (Table 12). Due to fewer observations, these estimates are imprecise, with confidence intervals that span nearly 50 points, but all are significant at conventional levels.

Table 11: Effects during cooperative period (2010–2014), ILO/IFC Better Work audits only

Bandwidth properties				Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
MSE symmetrical	8.8	8.8	72	-4.30	0.42	[-14.8, 6.2]
MSE separate	9.5	7.3	67	-4.29	0.42	[-14.8, 6.2]
CER symmetrical	7.1	7.1	57	-2.47	0.64	[-12.9, 8.0]
CER separate	7.7	5.9	54	-2.95	0.59	[-13.7, 7.8]

Notes. Regression discontinuity estimates of the effect of failing ratings on subsequent audit score *before* the introduction of penalties (2010–2014). Exclusively uses audits conducted by the ILO/IFC Better Work program. Details of each bandwidth selection procedure in Cattaneo, Idrobo, and Titiunik (2018).

Table 12: Effects after introduction of penalties, ILO/IFC Better Work audits only

Bandwidth properties				Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
MSE symmetrical	11.9	11.9	38	29.31	0.02	[5.5, 53.1]
MSE separate	15.6	12.5	40	25.67	0.03	[2.0, 49.3]
CER symmetrical	9.2	9.2	26	29.77	0.01	[6.4, 53.2]
CER separate	12.1	9.7	32	27.54	0.02	[4.1, 51.0]

Notes. Regression discontinuity estimates of the effect of failing ratings on subsequent audit score after the introduction of penalties, examining only the audits conducted by the International Labor Organization / International Finance Corporation Better Work program. Details of each bandwidth selection procedure in Cattaneo, Idrobo, and Titiunik (2018).

Increasing compliance through improvement or selection?

Average compliance can improve through both supplier improvement and supplier terminations. The preceding analysis suggested that—after the introduction of penalties for failing audits—failing ratings prompted factories to improve social compliance. However, the same effect could also be generated through selective termination near the threshold of failure. If Gap managers had private information about which factories were capable of sustaining compliance, they could use that information to terminate factories in ways that bias our estimated effects of failing an audit. Specifically, if Gap Inc terminated the factories that managers knew were unable to improve and did so *more often when factories failed their first audit*, this could create higher expected potential

outcomes just below the threshold of failure.[14]

Table 13 reports analyses after the introduction of penalties (2016–2019) that investigate the possibility of selective termination of failing factories below the threshold. If Gap selectively terminated after the first audit at higher rates below the threshold, we expect that failing should (at the threshold) increase the probability of termination, but we find no evidence of this. This finding aligns with evidence from our interviews; Gap Inc did not immediately terminate after the first failed audit. Instead factories usually received an additional audit round to provide opportunity for improvement.

Table 13: Failing at threshold does not increase probability of termination (after introduction of penalties, 2016–2019)

	Bandwidth properties			Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
MSE symmetrical	8.6	8.6	233	0.02	0.89	[-0.2, 0.23]
MSE separate	11.8	6.7	217	0.06	0.58	[-0.2, 0.28]
CER symmetrical	6.1	6.1	151	-0.04	0.72	[-0.3, 0.18]
CER separate	8.4	4.7	151	0.08	0.48	[-0.1, 0.29]

Notes. Regression discontuity estimates of the effect of failing ratings on the probability of not having a subsequent audit. Because audits are on a 12-month cadence, the sample is limited to audits conducted at least 365 days prior to the final audit in our data (March 6, 2019). Details of the bandwidth selection procedures are described in Cattaneo, Idrobo, and Titiunik (2018).

Longer-term commercial relationships and supply chain regulation

Two hypotheses anticipated varying effects of supply chain regulation depending on the length of the commercial relationship between buyer and supplier. We next test these hypotheses, both during the cooperative period and after the introduction of penalties. In each period of analysis, we divide factories into shorter- and longer-term commercial relationships based on the median relationship length in that period. We chose the median as the dividing line in the absence of strong theoretical priors about what constitutes "long" or "short" relationships. This approach also equalizes the sizes of the subsamples, generating estimates of similar precision for each group. Analyses of observation density (Figures A3 and A4) and covariate balance (Tables A5–A8) for each of the four subsamples appear in the appendix.

Did the near-zero effect of failing ratings in the cooperative period mask heterogeneous effects by relationship length? Table 14 reports separate effects within longer- and shorter-term suppliers (based on the median relationship length of 4.5 years) in the cooperative period. It again

[14]Note that if supplier responsibility staff instead used that private information to selectively terminate factories near the threshold, but did so equally no matter whether a factory passed or failed its first audit, this behavior would not necessarily bias our estimates.

shows two relatively precise near-zero effects, with no significant difference based on the length of the commercial relationship. We find no evidence that long-term commercial relationships make the purely cooperative approach to supply chain regulation more effective (H2).

Table 14: Effects by relationship length, cooperative period (2010–2014)

Bandwidth properties				Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
Supplier for under 4.5 years						
MSE symmetrical	9.0	9.0	621	-0.38	0.85	[-4.2, 3.5]
MSE separate	12.9	10.1	720	0.23	0.89	[-3.1, 3.6]
CER symmetrical	6.4	6.4	427	0.38	0.86	[-3.8, 4.6]
CER separate	9.1	7.1	519	-0.44	0.81	[-4.1, 3.2]
Supplier for over 4.5 years						
MSE symmetrical	9.1	9.1	654	-1.59	0.36	[-5.0, 1.8]
MSE separate	13.7	7.9	627	-1.81	0.24	[-4.8, 1.2]
CER symmetrical	6.6	6.6	467	-1.69	0.38	[-5.5, 2.1]
CER separate	9.9	5.7	472	-2.19	0.19	[-5.5, 1.1]

Notes. Regression discontinuity estimates of the effect of failing ratings on subsequent audit score (2010–2014). Effects reported by equally sized factory subgroups based on shorter and longer commercial relationships with the buyer.

However, we find a different pattern after the introduction of penalties. Table 15 reports effects estimated in two equally-sized groups of factories divided by the median commercial relationship length (6 years) after the introduction of penalties. Among longer-term suppliers, failing caused improved compliance ranging from 20.5 to 28.4 points depending on the bandwidth selection procedure. However, the effects among factories with shorter commercial relationships are near-zero. This offers support for Hypothesis 4; long-term supplier relationships are associated with greater responsiveness to private regulation augmented with the threat of commercial penalties.[15]

[15]The difference in effect magnitudes across long- and short-term relationships in the MSE symmetrical bandwidth estimate is $25.91 - 3.80 = 22.11$ points (Table 15). Assuming independence of these two subsamples, the standard error of this difference estimate is $\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} = \sqrt{7.47^2 + 8.34^2} = 11.2$ points (two-tailed test $p = .05$).

Table 15: Effects by relationship length after introduction of penalties

Bandwidth properties				Regression discontinuity est.		
Procedure	lower	upper	N	Est.	p-val.	95% CI
Supplier for under 6 years						
MSE symmetrical	8.8	8.8	84	3.80	0.61	[-10.8, 18.4]
MSE separate	9.7	10.9	100	5.97	0.39	[-7.8, 19.7]
CER symmetrical	6.6	6.6	73	-1.64	0.82	[-16.0, 12.7]
CER separate	7.2	8.1	80	2.63	0.71	[-11.0, 16.3]
Supplier for 6+ years						
MSE symmetrical	7.3	7.3	73	25.91	0.00	[9.6, 42.3]
MSE separate	16.7	5.6	90	28.38	0.00	[10.4, 46.3]
CER symmetrical	5.4	5.4	50	20.53	0.03	[2.3, 38.8]
CER separate	12.4	4.2	64	27.03	0.01	[7.6, 46.5]

Notes. This table estimates separate regression discontinuity effects among factories that have longer and shorter commercial relationships to the buyer.

Discussion

These findings offer contributions to scholarship on buyer–supplier governance, sustainability in global supply chains, and private politics. Scholarship on buyer–supplier relations repeatedly highlights the social and cooperative aspects of high-quality supplier relationships, with attendant benefits for value creation and capture (Uzzi 1997, Dyer and Chu 2003, Gulati and Nickerson 2008, Janowicz-Panjaitan and Noorderhaven 2009). Our study shifts the focus from traditional metrics of supplier performance—price, quality, and delivery—to a more recent addition: supplier *social* performance. The evidence suggests that punitive threats help buyers to shape the social performance of their suppliers. Under a more cooperative approach to governing supplier social performance, the analysis detects no evidence that notifying suppliers they are performing below standards and supporting remediation leads to improvement.

Why might the threat of penalties stimulate improvements in supplier social performance? Many factors suggest to suppliers that social performance standards are subordinate to more established metrics of performance. Buyers adopted supplier responsibility programs under duress from activist campaigns, in part to seek legitimacy and protect against consumer backlash, raising questions about their commitment to social performance demands (Bartley 2003; 2007). Many buyers demonstrate ambivalence about these social standards in their economic behavior; they simultaneously instruct suppliers to comply and engage in purchasing practices that undermine compliance (Locke 2013, Anner 2018). In the context of many buyers sending mixed signals, penalties create clarity and elevate the importance of social performance among the many competing performance demands on suppliers. Although previous research suggests that penalties can undermine both information transfer and intrinsic motivation for compliance, they may prompt

increased compliance with social standards. Cultivating long-term supplier relationships remains important; the largest effects on compliance are among long-term suppliers. However, in the absence of penalties, neither short- nor long-term suppliers appear responsive to failing compliance ratings.

In the literature on supply chain sustainability, there has been considerable uncertainty about the causal effects of private regulation. Locke (2013) writes that "it is not at all clear that global buyers have the ability or leverage (let alone the credibility) to pressure [their] suppliers...to improve working conditions" (p. 32). Despite roughly two decades of scholarship (Frenkel 2001, Barrientos and Smith 2007, Locke 2013, Toffel, Short and Ouellet 2015, Porteous, Rammohan and Lee 2015, Lee and Tang 2017, Caro et al. 2018, Bartley 2018, Dietz, Grabs and Estrella Chong 2019, Liu et al. 2019, Bird, Short and Toffel 2019), evidence on whether the actions of global buyers can *cause* improved labor compliance is limited. Harrison and Scorse (2010) used a combination of sectoral and geographic variation in the presence of reputation-sensitive buyers to show that activist campaigns raised wages in Indonesia, but could not examine the effects of private regulatory efforts themselves. Boudreau (2020) experimentally manipulated a private intervention to promote worker safety committees in Bangladeshi exporters, offering strong evidence on the efficacy of such committees. However, committee formation only appears in more specialized private regulatory programs. Most supply chain responsibility programs instead revolve around some combination of audits, ratings, corrective action plans, and threatened penalties. We contribute to this literature by estimating the causal effects of these core elements of supply chain regulation.

This study of Gap Inc's supplier responsibility program also offers an empirical contribution to the study of private politics. Previous research shows how social movements target firms and influence their behavior and financial performance (Eesley and Lenox 2006, King and Soule 2007, Lenox and Eesley 2009, Ingram, Yue and Rao 2010, McDonnell and King 2013, McDonnell, King and Soule 2015). Researchers have extended the study of social movement impact to non-targeted peer firms in the same industry (Yue, Rao and Ingram 2013, Briscoe, Gupta and Anner 2015, Soule, Swaminathan and Tihanyi 2014). Contemporary social movements go further by seeking to transform practices in the networks of commercial partners, in part by proxy-targeting firms that have large supply chains (Baron 2001, Schurman and Munro 2009). By studying the impact of private regulation on supplier firms, this study reveals that, under certain conditions, corporate responses to these social movements can have an impact on the social performance of these commercial partners.

There are important limitations to what this research can accomplish. Although the regression discontinuity approach has strong internal validity in its estimation of causal effects, the external validity of these findings depends on the typicality of the buyer–supplier relationships in Gap Inc's supply chain. We follow a tradition in the study of buyer–supplier relations by learning from the supply chain of a single buyer (Elfenbein and Zenger 2014, Aoki and Wilhelm 2017), in part due to the substantial challenges accessing sensitive corporate data. However, this approach also necessitates considering the ways in which the buyer and suppliers are typical or distinctive. First, Gap Inc is an extremely large apparel firm, which makes it a major buyer for

many of its suppliers. It often purchases more than one-third of estimated supplier output, similar to other large apparel firms discussed in Locke (2013). Losing this particular buyer would likely impose significant costs on these suppliers. Many smaller buyers also engage in private regulation (Amengual, Distelhorst and Tobin 2020), and suppliers may be less responsive to the prospect of losing these smaller customers. Second, stable sourcing relationships and the expectation of future business may also be necessary conditions for the effects we observe. Half of Gap Inc's suppliers have been in a commercial relationship for more than six years, and failing ratings had the clearest effect on this subgroup (Table 15). Not all supply chains are characterized by such stable relations. A recent study of the apparel industry in Bangladesh, one of the world's leading garment exporters, found that the average length of a buyer–supplier relationship was just two years (Cajal Grossi, Macchiavello and Noguera 2019). Gap Inc's comparatively long commercial relationships may increase suppliers' expectation of future business. When commercial relationships are less stable, suppliers may place lower value on the future of the commercial relationship, anticipating it may end even if they invest in improving social compliance.

Comparing effects across our two time periods also depends on the assumption that the key change between these periods was the introduction of penalties. We probed the validity of this assumption in several ways, but cannot eliminate all uncertainty. First and most importantly, our interviews and many informal conversations with Gap Inc social compliance and sourcing managers were what allowed us to identify the change in approach that defines these two periods. Those interviews affirmed that the only major change in the consequences of audit ratings was the introduction of threatened penalties from the sourcing department, and offered concrete examples of the efficacy of sourcing managers threatening to end the commercial relationship.

Second, we considered whether slow-moving changes in background conditions might have influenced the potency of audit ratings over time. Our largest concern here was consolidation of the supplier base. Gap Inc, along with many other large apparel firms, has reduced its count of suppliers over the past decade, focusing on maintaining a smaller number of longer-term supplier relationships. This consolidation tends to increase both the average length of the buyer–supplier relationship and the total output purchased by the buyer, both of which are known to increase the value of the buyer–supplier relationship to suppliers. Note that we already examined whether longer-term relationships were associated with larger effects of audit ratings during the cooperative period, and found they were not (Table 14). In supplemental analyses, we additionally examined heterogenous effects by sourcing volume prior to introduction of penalties. Again, during the cooperative period we find null effects even among high-volume suppliers, suggesting that neither higher sourcing volume nor longer-term relationships associated with consolidation explain our finding.

Finally, we considered whether changes in the institutional environment that occurred after the Rana Plaza tragedy in Bangladesh could account for the positive effect in the period with penalties. The Rana Plaza factory collapse occurred in 2013, and Gap Inc introduced penalties several years later in mid-2016. Rana Plaza's impact on the garment industry was primarily in

Bangladesh, where new factory safety regimes were established that included Gap Inc suppliers (Boudreau 2020). To ensure that this change did not explain our result, we additionally replicated our results after excluding all factories located in Bangladesh. Results for all supplemental analyses are available on request. Overall, our interview research and these supplemental analyses offer confidence that the central difference between the periods was the introduction of penalties. However, it is impossible to fully rule out all the unobservable shifts that take place when comparing two periods of time.

In addition to the private regulatory practices analyzed in this study, buyers have other tools to influence the social compliance of their suppliers. Our study has less to say about improvements prompted by supplier selection or interventions to transform suppliers' management practices (Bloom et al. 2013). Many factories with poor labor conditions were likely screened out by Gap Inc's social audits. They never entered the supply chain, and we therefore do not observe their data. If certain social practices are disqualifying for exporting to multinational enterprises, supplier selection may generate incentives for labor upgrading in order to gain access to export markets (Malesky and Mosley 2018). Other buyer-led management interventions have also been shown to improve factory labor standards. Recent research suggests that interventions by multinational enterprises to improve worker voice (Boudreau 2020), upgrade manufacturing management practices (Distelhorst, Hainmueller and Locke 2017), and optimize the wage and bonus system (Lollo and O'Rourke 2020) offer alternative pathways to improving working conditions in global supply chains.

Table 16: How much do failing factories improve?

	Failed audit	Next audit	diff	std err
Total violations	21.6	16.3	-5.2	0.85
Counts by severity				
Low-severity violations	13.5	9.8	-3.6	0.66
Medium-severity	3.3	2.4	-0.9	0.20
High-severity	1.6	1.1	-0.5	0.14
Highest-severity	0.8	0.2	-0.5	0.07

Notes. Change in violations after receiving a failing rating, across all failing audits after introduction of penalties. In t-tests for differences in means assuming unequal variances, $p < .001$ for all comparisons. $N = 130$.

Finally we caution that although we find that private regulation can improve labor compliance of suppliers, stronger actions are likely needed to bring suppliers into *full* compliance with international standards. Table 16 shows average audit-to-audit improvement in all failing factories since the introduction of penalties. Failing factories markedly improved in their subsequent audits, reducing highest-severity violations by 75% (from 0.8 to 0.2 on average). Subsequent audits of failing factories still detected an average of 16 violations and more than one high-severity violation. It is possible that this is merely one stop on the way to even higher levels of compliance, but it

seems more likely that suppliers will stop improving once they achieve a rating that removes the threat of commercial penalties. Future improvement may depend on supply chain buyers adopting increasingly stringent definitions of social compliance "failure."

This study's main finding—that private regulatory activity improved compliance most when long-term suppliers were exposed to some threat of penalties—suggests ways to improve supply chain responsibility programs. The cooperative approach to supply chain regulation encouraged buyers to commit to long-term relationships with their suppliers, rather than threatening to terminate the commercial relationship if suppliers fell below certain performance standards. This allowed sourcing departments and supplier sustainability departments to operate in separate silos. Sourcing could continue to purchase from factories that met commercial needs while sustainability departments worked to improve factory compliance through monitoring and cooperative problem-solving. However, if commercial penalties make private regulation more potent, improving supplier social performance may instead require dismantling these silos and forging stronger organizational linkages between sourcing and supplier responsibility *within buyers*, as we observed inside Gap Inc. This form of internal alignment may help supplier responsibility programs to improve labor and environmental compliance in global supply chains.

References

Abernathy, Frederick H, John T Dunlop, Janice H Hammond and David Weil. 1999. *A stitch in time: Lean retailing and the transformation of manufacturing–lessons from the apparel and textile industries*. Oxford University Press.

Amengual, Matthew. 2010. "Complementary Labor Regulation: The Uncoordinated Combination of State and Private Regulators in the Dominican Republic." *World Development* 38(3):405–414.

Amengual, Matthew, Greg Distelhorst and Danny Tobin. 2020. "Global Purchasing as Labor Regulation: The Missing Middle." *Industrial and Labor Relations Review* 73(4):817–840.

Anner, Mark. 2012. "Corporate Social Responsibility and Freedom of Association Rights The Precarious Quest for Legitimacy and Control in Global Supply Chains." *Politics & Society* 40(4):609–644.

Anner, Mark. 2018. "CSR Participation Committees, Wildcat Strikes and the Sourcing Squeeze in Global Supply Chains." *British Journal of Industrial Relations* 56(1):75–98.

Aoki, Katsuki and Miriam Wilhelm. 2017. "The role of ambidexterity in managing buyer–supplier relationships: The Toyota case." *Organization Science* 28(6):1080–1097.

Ashforth, Blake E and Ronald H Humphrey. 1997. "The ubiquity and potency of labeling in organizations." *Organization Science* 8(1):43–58.

Ayres, Ian and John Braithwaite. 1992. *Responsive regulation: Transcending the deregulation debate*. Oxford University Press, USA.

Baker, George, Robert Gibbons and Kevin J Murphy. 2002. "Relational Contracts and the Theory of the Firm." *The Quarterly Journal of Economics* 117(1):39–84.

Bardach, Eugene and Robert Kagan. 1982. *Going by the Book: The Problem of Regulatory Unreasonableness (1982)*. Temple University Press.

Baron, David P. 2001. "Private politics, corporate social responsibility, and integrated strategy." *Journal of Economics & Management Strategy* 10(1):7–45.

Barrientos, S. and S. Smith. 2007. "Do Workers Benefit from Ethical Trade? Assessing Codes of Labour Practice in Global Production Systems." *Third World Quarterly* 28(4):713–729.

Bartley, Tim. 2003. "Certifying forests and factories: States, social movements, and the rise of private regulation in the apparel and forest products fields." *Politics & Society* 31(3):433–464.

Bartley, Tim. 2007. "Institutional Emergence in an Era of Globalization: The Rise of Transnational Private Regulation of Labor and Environmental Conditions." *American Journal of Sociology* 113(2):297–351.

Bartley, Tim. 2018. *Rules without rights: Land, labor, and private authority in the global economy*. Oxford University Press.

Bartley, Tim and Curtis Child. 2011. "Movements, markets and fields: The effects of anti-sweatshop campaigns on US firms, 1993-2000." *Social Forces* 90(2):425–451.

Basu, Arnab K, Nancy H Chau and Ravi Kanbur. 2010. "Turning a blind eye: Costly enforcement, credible commitment and minimum wage laws." *The Economic Journal* 120(543):244–269.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76(2):169–217.

Bird, Yanhua, Jodi L Short and Michael W Toffel. 2019. "Coupling Labor Codes of Conduct and Supplier Labor Practices: The Role of Internal Structural Conditions." *Organization Science* .

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie and John Roberts. 2013. "Does Management Matter? Evidence from India." *The Quarterly Journal of Economics* 128(1):1–51.

Boudreau, Laura. 2020. "Multinational enforcement of labor law: Experimental evidence from Bangladesh's apparel sector." *Working Paper* .
URL: *https://goo.gl/b7CRTH*

Briscoe, Forrest, Abhinav Gupta and Mark S Anner. 2015. "Social activism and practice diffusion: How activist tactics affect non-targeted organizations." *Administrative Science Quarterly* 60(2):300–332.

Cajal Grossi, Julia, Rocco Macchiavello and Guillermo Noguera. 2019. "International buyers' sourcing and suppliers' markups in Bangladeshi garments.".

Caro, Felipe, Prashant Chintapalli, Kumar Rajaram and Chris S Tang. 2018. "Improving Supplier Compliance Through Joint and Shared Audits with Collective Penalty." *Manufacturing & Service Operations Management* 20(2):363–380.

Cashore, Benjamin William, Graeme Auld and Deanna Newsom. 2004. *Governing through markets: Forest certification and the emergence of non-state authority*. Yale University Press.

Cattaneo, Matias D, Nicolas Idrobo and Rocıo Titiunik. 2018. *A Practical Introduction to Regression Discontinuity Designs: Volume 1*. Cambridge University Press.

Coslovsky, Salo V and Richard Locke. 2013. "Parallel Paths to Enforcement: Private Compliance, Public Regulation, and Labor Standards in the Brazilian Sugar Sector." *Politics & Society* 41(4):497–526.

Delmas, Magali and Ivan Montiel. 2009. "Greening the supply chain: when is customer pressure effective?" *Journal of Economics & Management Strategy* 18(1):171–201.

Dietz, Thomas, Janina Grabs and Andrea Estrella Chong. 2019. "Mainstreamed voluntary sustainability standards and their effectiveness: Evidence from the Honduran coffee sector." *Regulation & Governance* .

Distelhorst, Greg, Jens Hainmueller and Richard M Locke. 2017. "Does lean improve labor standards? Management and social performance in the Nike supply chain." *Management Science* 63(3):707–728.

Distelhorst, Greg, Richard M Locke, Timea Pal and Hiram M Samel. 2015. "Production Goes Global, Compliance Stays Local: Private regulation in the global electronics industry." *Regulation & Governance* 9:224–242.

Dore, Ronald. 1983. "Goodwill and the spirit of market capitalism." *The British journal of sociology* 34(4):459–482.

Dyer, Jeffrey H and Harbir Singh. 1998. "The relational view: Cooperative strategy and sources of interorganizational competitive advantage." *Academy of management review* 23(4):660–679.

Dyer, Jeffrey H and Wujin Chu. 2003. "The role of trustworthiness in reducing transaction costs and improving performance: Empirical evidence from the United States, Japan, and Korea." *Organization science* 14(1):57–68.

Eesley, Charles and Michael J Lenox. 2006. "Firm responses to secondary stakeholder action." *Strategic Management Journal* 27(8):765–781.

Egels-Zandén, Niklas. 2007. "Suppliers' Compliance with MNCs' Codes of Conduct: Behind the Scenes at Chinese Toy Suppliers." *Journal of Business Ethics* 75(1):45–62.

Egels-Zandén, Niklas. 2014. "Revisiting supplier compliance with MNC codes of conduct: Recoupling policy and practice at Chinese toy suppliers." *Journal of Business Ethics* 119(1):59–75.

Elfenbein, Daniel W and Todd R Zenger. 2014. "What is a relationship worth? Repeated exchange and the development and deployment of relational capital." *Organization Science* 25(1):222–244.

Elliott, Kimberly A. and Richard B. Freeman. 2003. *Can Labor Standards Improve Under Globalization?* Peterson Institute.

Frederick, Stacey and Gary Gereffi. 2011. "Upgrading and restructuring in the global apparel value chain: why China and Asia are outperforming Mexico and Central America." *International Journal of Technological Learning, Innovation and Development* 4(1-3):67–95.

Frenkel, Stephen J. 2001. "Globalization, athletic footwear commodity chains and employment

relations in China." *Organization studies* 22(4):531–562.
Frenkel, Stephen J and Duncan Scott. 2002. "Compliance, collaboration, and codes of labor practice: The Adidas connection." *California Management Review* 45(1):29–49.
Fung, Archon, Dara O'Rourke and Charles F. Sabel. 2001. *Can We Put an End to Sweatshops? A New Democracy Forum on Raising Global Labor Standards*. Beacon Press.
Gereffi, Gary and Joonkoo Lee. 2016. "Economic and social upgrading in global value chains and industrial clusters: Why governance matters." *Journal of business ethics* 133(1):25–38.
Gray, Garry C and Susan S Silbey. 2014. "Governing inside the organization: Interpreting regulation and compliance." *American Journal of Sociology* 120(1):96–145.
Gulati, Ranjay. 1995. "Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances." *Academy of management journal* 38(1):85–112.
Gulati, Ranjay and Jack A Nickerson. 2008. "Interorganizational trust, governance choice, and exchange performance." *Organization Science* 19(5):688–708.
Harrison, Ann and Jason Scorse. 2010. "Multinationals and Anti-Sweatshop Activism." *The American Economic Review* 100(1):247–273.
Helper, Susan. 1990. "Comparative supplier relations in the US and Japanese auto industries: an exit/voice approach." *Business and Economic history* pp. 153–162.
Hirschman, A.O. 1970. *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard Univ Pr.
Ingram, Paul, Lori Qingyuan Yue and Hayagreeva Rao. 2010. "Trouble in store: Probes, protests, and store openings by Wal-Mart, 1998–2007." *American Journal of Sociology* 116(1):53–92.
Janowicz-Panjaitan, Martyna and Niels G Noorderhaven. 2009. "Trust, calculation, and interorganizational learning of tacit knowledge: An organizational roles perspective." *Organization Studies* 30(10):1021–1044.
Ji, MinWoong and David Weil. 2015. "The impact of franchising on labor standards compliance." *ILR Review* 68(5):977–1006.
Jiang, Bin. 2009. "Implementing supplier codes of conduct in global supply chains: Process explanations from theoretic and empirical perspectives." *Journal of business ethics* 85(1):77–92.
Jira, Chonnikarn and Michael W Toffel. 2013. "Engaging supply chains in climate change." *Manufacturing & Service Operations Management* 15(4):559–577.
Joskow, Paul L. 1988. "Asset specificity and the structure of vertical relationships: empirical evidence." *Journal of Law, Economics, & Organization* 4(1):95–117.
Kagan, Robert A and John T Scholz. 1984. The "criminology of the corporation" and regulatory enforcement strategies. In *Enforcing Regulation*, ed. K Hawkins and J.M. Thomas. Springer pp. 67–95.
Kalkanci, Basak and Erica L Plambeck. 2019. "Reveal the supplier list? A trade-off in capacity vs. responsibility." *Manufacturing & Service Operations Management*. .
Kalkanci, Basak, Erjie Ang and Erica L Plambeck. 2016. Strategic disclosure of social and environmental impacts in a supply chain. In *Environmentally responsible supply chains*. Springer pp. 223–239.
Kelly, Erin L. 2010. "Failure to update: An institutional perspective on noncompliance with the Family and Medical Leave Act." *Law & Society Review* 44(1):33–66.
King, Brayden G and Sarah A Soule. 2007. "Social movements as extra-institutional entrepreneurs: The effect of protests on stock price returns." *Administrative Science Quarterly* 52(3):413–442.
Kortelainen, Ketty. 2008. "Global supply chains and social requirements: case studies of labour condition auditing in the People's Republic of China." *Business Strategy and the Environment* 17(7):431–443.
Kuruvilla, Sarosh, Mingwei Liu, Chunyun Li and Wansi Chen. 2020. "Field opacity and practice

outcomes decoupling: private regulation of labor standards in global supply chains." *Industrial and Labor Relations Review* .

Lee, Hau L and Christopher S Tang. 2017. "Socially and environmentally responsible value chain innovations: New operations management research opportunities." *Management Science* 64(3):983–996.

Lenox, Michael J and Charles E Eesley. 2009. "Private environmental activism and the selection and response of firm targets." *Journal of Economics & Management Strategy* 18(1):45–73.

Liu, Xiaojin, Anant Mishra, Susan Goldstein and Kingshuk K Sinha. 2019. "Toward improving factory working conditions in developing countries: An empirical analysis of Bangladesh Ready-Made Garment factories." *Manufacturing & Service Operations Management* 21(2):251–477.

Locke, Richard M. 2013. *The Promise and Limits of Private Power: Promoting Labor Standards in a Global Economy*. Cambridge University Press.

Locke, Richard M., Fei Qin and Alberto Brause. 2007. "Does Monitoring Improve Labor Standards? Lessons from Nike." *Industrial and Labor Relations Review* 61(1):3–31.

Locke, Richard M., Matthew Amengual and Akshay Mangla. 2009. "Virtue out of Necessity? Compliance, Commitment, and the Improvement of Labor Conditions in Global Supply Chains." *Politics & Society* 37(3):319.

Lollo, Niklas and Dara O'Rourke. 2020. "Factory benefits to paying workers more: The critical role of compensation systems in apparel manufacturing." *PloS one* 15(2):e0227510.

Lund-Thomsen, Peter and Adam Lindgreen. 2014. "Corporate social responsibility in global value chains: Where are we now and where are we going?" *Journal of Business Ethics* 123(1):11–22.

Macchiavello, Rocco and Ameet Morjaria. 2015. "The value of relationships: evidence from a supply shock to Kenyan rose exports." *American Economic Review* 105(9):2911–45.

Malesky, Edmund J and Layna Mosley. 2018. "Chains of Love? Global Production and the Firm-Level Diffusion of Labor Standards." *American Journal of Political Science* 62(3):712–728.

Maloni, Michael and Wilhelm C Benton. 2000. "Power influences in the supply chain." *Journal of business logistics* 21(1):49–74.

McDonnell, Mary-Hunter, Brayden G King and Sarah A Soule. 2015. "A dynamic process model of private politics: Activist targeting and corporate receptivity to social challenges." *American Sociological Review* 80(3):654–678.

McDonnell, Mary-Hunter and Brayden King. 2013. "Keeping up appearances: Reputational threat and impression management after social movement boycotts." *Administrative Science Quarterly* 58(3):387–419.

McMillan, John and Christopher Woodruff. 1999. "Interfirm relationships and informal credit in Vietnam." *The Quarterly Journal of Economics* 114(4):1285–1320.

Porteous, Angharad, Sonali Rammohan and Hau Lee. 2015. "Carrots or sticks? Improving social and environmental compliance at suppliers through incentives and penalties." *Production and Operations Management* 24(9):1402–1413.

Raj-Reichert, Gale. 2013. "Safeguarding labour in distant factories: Health and safety governance in an electronics global production network." *Geoforum* 44:23–31.

Riisgaard, L. 2009. "Global Value Chains, Labor Organization and Private Social Standards: Lessons from East African Cut Flower Industries." *World Development* 37(2):326–340.

Rossi, Arianna, Amy Luinstra and John Pickles. 2014. *Towards Better Work: Understanding labour in apparel global value chains*. Springer.

Sako, Mari. 2004. "Supplier development at Honda, Nissan and Toyota: comparative case studies of organizational capability enhancement." *Industrial and Corporate Change* 13(2):281–308.

Schurman, Rachel and William Munro. 2009. "Targeting capital: A cultural economy approach to understanding the efficacy of two anti–genetic engineering movements." *American Journal of*

Sociology 115(1):155–202.

Seidman, Gay. 2007. *Beyond the Boycott: Labor Rights, Human Rights, and Transnational Activism*. Russell Sage Foundation Publications.

Short, Jodi L and Michael W Toffel. 2010. "Making self-regulation more than merely symbolic: The critical role of the legal environment." *Administrative Science Quarterly* 55(3):361–396.

Short, Jodi L, Michael W Toffel and Andrea R Hugill. 2016. "Monitoring global supply chains." *Strategic Management Journal* 37(9):1878–1897.

Short, Jodi L, Michael W Toffel and Andrea R Hugill. 2020. "Improving Working Conditions in Global Supply Chains: The Role of Institutional Environments and Monitoring Program Design." *Industrial and Labor Relations Review* 73(4):873–912.

Soule, Sarah A, Anand Swaminathan and Laszlo Tihanyi. 2014. "The diffusion of foreign divestment from Burma." *Strategic Management Journal* 35(7):1032–1052.

Soundararajan, Vivek and Stephen Brammer. 2018. "Developing country sub-supplier responses to social sustainability requirements of intermediaries: Exploring the influence of framing on fairness perceptions and reciprocity." *Journal of Operations Management* .

Susarla, Anjana, Martin Holzhacker and Ranjani Krishnan. 2020. "Calculative Trust and Interfirm Contracts." *Management Science* .

Tanaka, Mari. 2017. "Exporting Sweatshops? Evidence from Myanmar." *Review of Economics and Statistics* pp. 1–44.

Thorlakson, Tannis, Joann F de Zegher and Eric F Lambin. 2018. "Companies' contribution to sustainability through global supply chains." *Proceedings of the National Academy of Sciences* pp. 2072–2077.

Toffel, Michael W, Jodi L Short and Melissa Ouellet. 2015. "Codes in context: How states, markets, and civil society shape adherence to global labor standards." *Regulation & Governance* 9(3):205–223.

Uzzi, Brian. 1997. "Social structure and competition in interfirm networks: The paradox of embeddedness." *Administrative science quarterly* pp. 35–67.

Vogel, David. 2010. "The private regulation of global corporate conduct: achievements and limitations." *Business & Society* 49(1):68–87.

Walker, Edward T, Andrew W Martin and John D McCarthy. 2008. "Confronting the state, the corporation, and the academy: The influence of institutional targets on social movement repertoires." *American Journal of Sociology* 114(1):35–76.

Wathne, Kenneth H and Jan B Heide. 2000. "Opportunism in interfirm relationships: Forms, outcomes, and solutions." *Journal of marketing* 64(4):36–51.

Williamson, Oliver E. 1991. "Comparative economic organization: The analysis of discrete structural alternatives." *Administrative science quarterly* pp. 269–296.

Yang, Yujeong and Mary Gallagher. 2017. "Moving In and Moving Up? Labor Conditions and China's Changing Development Model." *Public Administration and Development* 37(3):160–175.

Yue, Lori Qingyuan, Hayagreeva Rao and Paul Ingram. 2013. "Information spillovers from protests against corporations: A tale of Walmart and Target." *Administrative Science Quarterly* 58(4):669–701.

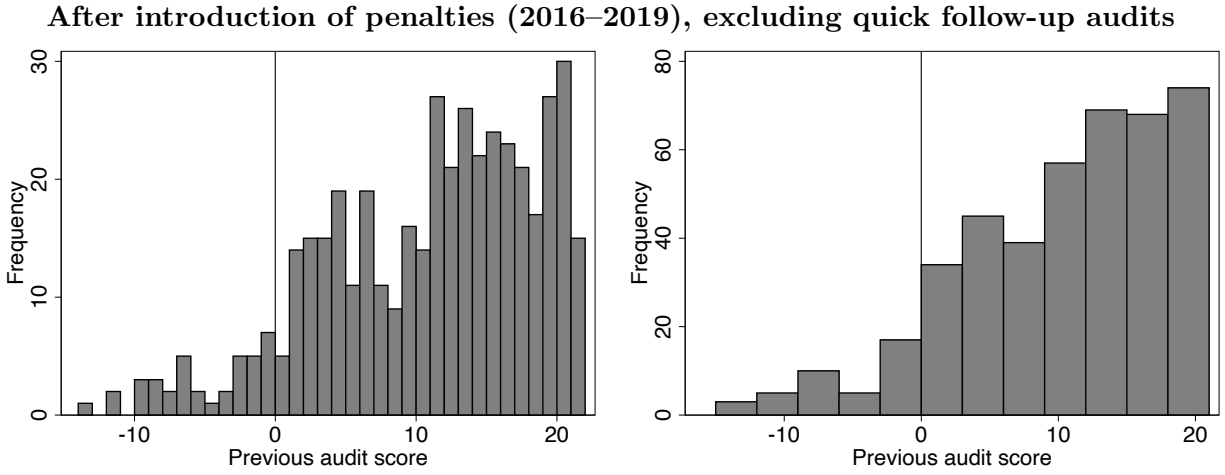
Appendix

Table A1: Covariate balance, cooperative period (2015–2016)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	-0.69	0.71	[2.9, -4.3]	15.0	280
Medium-severity	0.17	0.77	[1.3, -1.0]	16.0	299
High-severity	-0.08	0.81	[0.6, -0.8]	12.2	219
Highest-severity	0.09	0.61	[0.5, -0.3]	17.6	340
Better work audit?	-0.14	0.25	[0.1, -0.4]	9.8	172
Days since last audit	-75.65	0.44	[116.7, -268.0]	13.0	21
Days until outcome audit	-7.94	0.89	[101.7, -117.6]	10.2	172
Pre-previous audit score	-22.03	0.28	[17.8, -61.9]	8.5	12
Outcome audit by ILO	-0.15	0.42	[0.2, -0.5]	9.7	172
ln(units shipped) (std)	-0.28	0.47	[0.5, -1.0]	11.2	193
Relationship (years)	-1.65	0.38	[2.0, -5.3]	12.2	219
Workers	-1530.26	0.04	[-44.6, -3015.9]	7.9	138
Female workers (%)	12.61	0.14	[29.2, -4.0]	12.9	218
Manufacturer?	-0.14	0.33	[0.1, -0.4]	12.5	239
Americas	-0.04	0.54	[0.1, -0.2]	9.3	155
Mediterranean	-0.00	0.99	[0.1, -0.1]	7.2	116
North Asia	0.26	0.17	[0.6, -0.1]	9.0	155
South Asia	-0.30	0.10	[0.1, -0.7]	10.0	172
Southeast Asia	0.09	0.64	[0.5, -0.3]	12.0	221
Factory in China	0.26	0.17	[0.6, -0.1]	9.0	155
India	-0.26	0.13	[0.1, -0.6]	10.2	172
Indonesia	0.01	0.89	[0.1, -0.1]	11.2	193
Vietnam	0.00	0.99	[0.3, -0.3]	15.3	282
other country	-0.02	0.89	[0.3, -0.3]	10.4	172

Notes. Tests of continuity of pre-treatment covariates at the threshold between failing and passing audit scores. Positive values indicate that passing factories have higher values than failing ones. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018). We make no adjustment for multiple comparisons.

Figure A1: Compliance audit score density, after introduction of penalties (2016–2019) – Excluding Quick Follow-ups



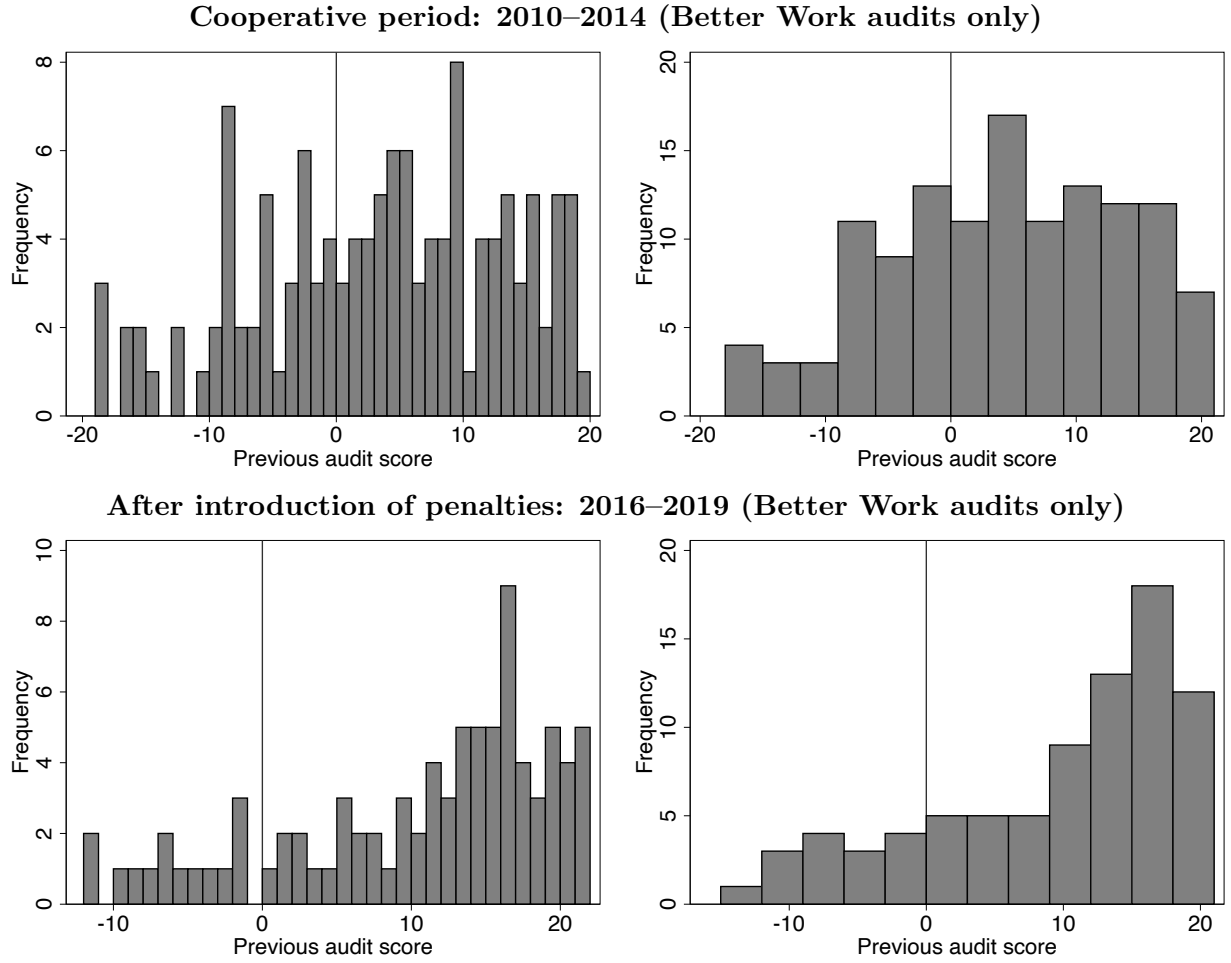
Notes. Histograms of audit scores, with a vertical line showing the threshold between failing and passing audit scores. Quick follow-up audits with fewer than 270 days from the previous audit are excluded. Left-hand plot shows all scores (each is an integer) and right-hand plot shows bins that are three points wide. A manipulation test using local polynomial density estimation fails to reject the null of no sorting around the threshold ($p = 0.77$).

Table A2: Covariate balance after introduction of penalties (2016–2019) – excluding quick follow-ups on failing factories

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	-1.39	0.46	[2.3, -5.1]	11.6	210
Medium-severity	-0.09	0.88	[1.1, -1.3]	12.7	231
High-severity	-0.46	0.36	[0.5, -1.4]	11.4	181
Highest-severity	0.31	0.15	[0.7, -0.1]	13.0	231
Better work audit?	-0.02	0.89	[0.3, -0.3]	9.6	169
Days since last audit	-51.65	0.10	[9.9, -113.2]	12.7	183
Days until outcome audit	9.97	0.69	[59.1, -39.1]	13.9	260
Pre-previous audit score	2.01	0.87	[26.7, -22.7]	9.3	116
Outcome audit by ILO	0.03	0.89	[0.4, -0.3]	11.0	183
ln(units shipped) (std)	0.32	0.44	[1.1, -0.5]	11.1	183
Relationship (years)	0.28	0.90	[4.6, -4.1]	11.4	179
Workers	-73.25	0.90	[1055.3, -1201.8]	10.4	162
Female workers (%)	5.30	0.59	[24.5, -13.9]	10.9	160
Manufacturer?	-0.16	0.39	[0.2, -0.5]	7.4	125
Americas	-0.12	0.12	[0.0, -0.3]	7.5	125
Mediterranean	0.09	0.46	[0.3, -0.1]	12.7	233
North Asia	0.10	0.55	[0.4, -0.2]	10.3	169
South Asia	-0.15	0.45	[0.2, -0.6]	11.6	212
Southeast Asia	0.02	0.89	[0.3, -0.3]	12.0	212
Factory in China	0.10	0.55	[0.4, -0.2]	10.3	169
India	-0.03	0.91	[0.4, -0.5]	8.3	138
Indonesia	0.09	0.45	[0.3, -0.2]	11.6	212
Vietnam	-0.16	0.15	[0.1, -0.4]	8.4	138
other country	0.07	0.68	[0.4, -0.3]	13.2	233

Notes. Tests of continuity of pre-treatment covariates at the threshold between failing and passing audit scores. Positive values indicate that passing factories have higher values than failing ones. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018). We make no adjustment for multiple comparisons.

Figure A2: Compliance audit score density around the threshold of failure: Better Work audits



Notes. Histograms of audit scores, with a vertical line showing the threshold between failing and passing audit scores. This plot shows the first audit after the introduction of penalties only if both this audit and the following (outcome) audit were conducted by the ILO Better Work program. Left-hand plot shows all scores (each is an integer) and right-hand plot shows bins that are three points wide. A manipulation test using local polynomial density estimation fails to reject the null of no sorting around the threshold ($p = 0.72$ in the cooperative period and $p = 0.99$ after introduction of penalties).

Table A3: Covariate balance, cooperative period (2010–2014), ILO/IFC Better Work audits only

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	-2.86	0.63	[8.7, -14.4]	7.5	59
Medium-severity	-1.55	0.43	[2.3, -5.4]	6.9	55
High-severity	-1.67	0.16	[0.6, -4.0]	9.1	73
Highest-severity	-0.93	0.24	[0.6, -2.5]	6.3	53
Days since last audit	-84.87	0.10	[15.0, -184.7]	8.9	51
Days until outcome audit	8.99	0.92	[188.0, -170.1]	7.0	55
Pre-previous audit score	-5.63	0.17	[2.5, -13.7]	7.4	39
Relationship (years)	-6.41	0.00	[-2.6, -10.2]	5.3	41
ln(units shipped) (std)	-1.56	0.05	[-0.0, -3.1]	9.1	34
Workers	-55.21	0.93	[1109.7, -1220.1]	7.3	58
Female workers (%)	-16.62	0.55	[38.1, -71.3]	6.1	33
Manufacturer?	0.15	0.56	[0.6, -0.4]	13.9	99
Factory in Vietnam	-0.07	0.43	[0.1, -0.2]	7.8	61
other country	0.07	0.43	[0.2, -0.1]	7.8	61

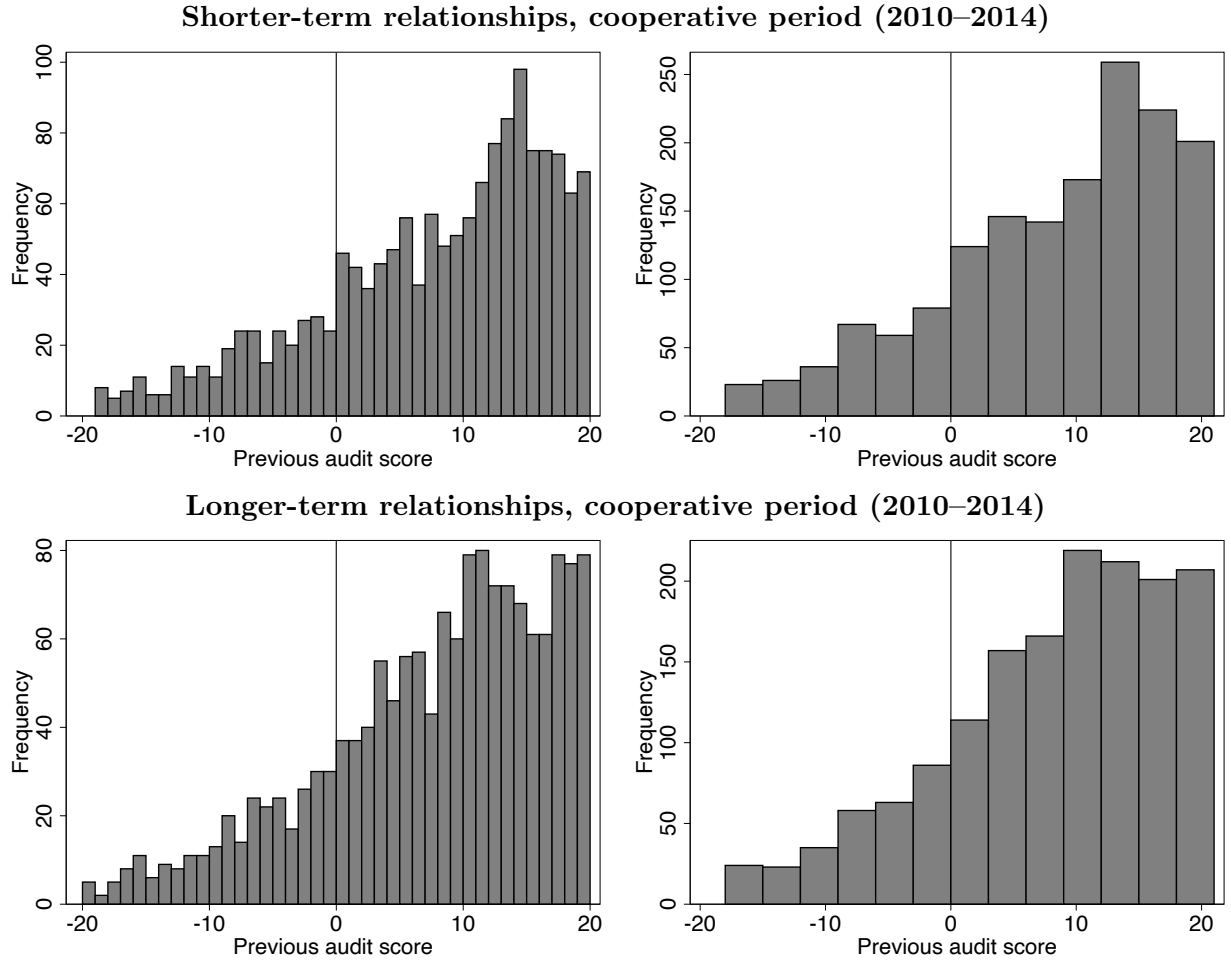
Notes. Tests of continuity of pre-treatment covariates at the threshold of failure in the old ratings system (2010-2014). Positive values indicate that factories that passed have higher values than those that failed. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018).

Table A4: Covariate balance, after introduction of penalties (2016–2019), ILO/IFC Better Work audits only)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	-3.84	0.26	[2.8, -10.5]	13.5	40
Medium-severity	-2.95	0.08	[0.4, -6.3]	10.1	29
High-severity	3.11	0.07	[6.5, -0.2]	9.3	25
Highest-severity	-0.57	0.25	[0.4, -1.5]	10.2	29
Days since last audit	-64.10	0.55	[144.6, -272.8]	10.9	24
Days until outcome audit	-9.31	0.70	[38.4, -57.1]	9.0	26
Pre-previous audit score	14.53	0.35	[44.9, -15.9]	10.9	24
ln(units shipped) (std)	0.41	0.55	[1.7, -0.9]	10.3	30
Relationship (years)	0.08	0.99	[9.1, -9.0]	12.2	38
Workers	-1074.53	0.59	[2823.5, -4972.6]	12.9	40
Female workers (%)	-2.53	0.85	[22.8, -27.9]	10.5	29
Manufacturer?	0.04	0.65	[0.2, -0.1]	6.3	17
Factory in Bangladesh	0.19	0.65	[1.0, -0.6]	11.4	32
Cambodia	0.40	0.31	[1.2, -0.4]	10.2	30
Indonesia	0.12	0.75	[0.9, -0.6]	12.2	38
Vietnam	-0.24	0.56	[0.6, -1.1]	12.0	38
other country	-0.50	0.14	[0.2, -1.1]	7.6	24

Notes. Tests of continuity of pre-treatment covariates at the threshold between failing and passing audit scores for the Better Work audits, after the introduction of penalties. Positive values indicate that passing factories have higher values than failing ones. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018). We make no adjustment for multiple comparisons.

Figure A3: Audit score density by relationship length, cooperative period (2010–2014)



Notes. Histograms of audit scores that serve as running variable in the regression discontinuity analysis. Manipulation tests using local polynomial density estimation fail to reject the null that the density of the running variable is continuous at the threshold ($p = 0.34$ for shorter-term relationships, $p = 0.64$ for longer-term relationships).

Table A5: Covariate balance in cooperative period (2010–2014) – shorter sourcing relationship (under 4.5 years)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	-2.67	0.33	[2.7, -8.1]	11.1	754
Medium-severity	-0.82	0.09	[0.1, -1.8]	13.9	1,000
High-severity	-0.66	0.13	[0.2, -1.5]	8.1	554
Highest-severity	0.06	0.78	[0.5, -0.3]	12.4	855
Better Work audit?	0.08	0.15	[0.2, -0.0]	13.0	919
Days since last audit	-76.19	0.04	[-2.3, -150.1]	6.2	242
Days until outcome audit	8.06	0.75	[57.9, -41.8]	9.6	657
Better Work audit?	0.08	0.15	[0.2, -0.0]	13.0	919
Pre-previous audit score	2.61	0.44	[9.2, -4.0]	8.7	339
Relationship (years)	-0.31	0.34	[0.3, -0.9]	11.6	801
ln(units shipped) (std)	0.02	0.95	[0.7, -0.7]	9.8	300
Workers	0.59	1.00	[472.5, -471.3]	12.4	853
Female workers (%)	3.82	0.63	[19.3, -11.6]	10.4	381
Manufacturer?	-0.01	0.88	[0.1, -0.2]	9.1	621
Factory in China	-0.02	0.83	[0.2, -0.2]	13.6	975
India	0.05	0.65	[0.2, -0.2]	10.1	683
Indonesia	-0.01	0.72	[0.1, -0.1]	8.0	552
Vietnam	-0.17	0.04	[-0.0, -0.3]	8.2	557
other country	0.11	0.27	[0.3, -0.1]	8.8	607

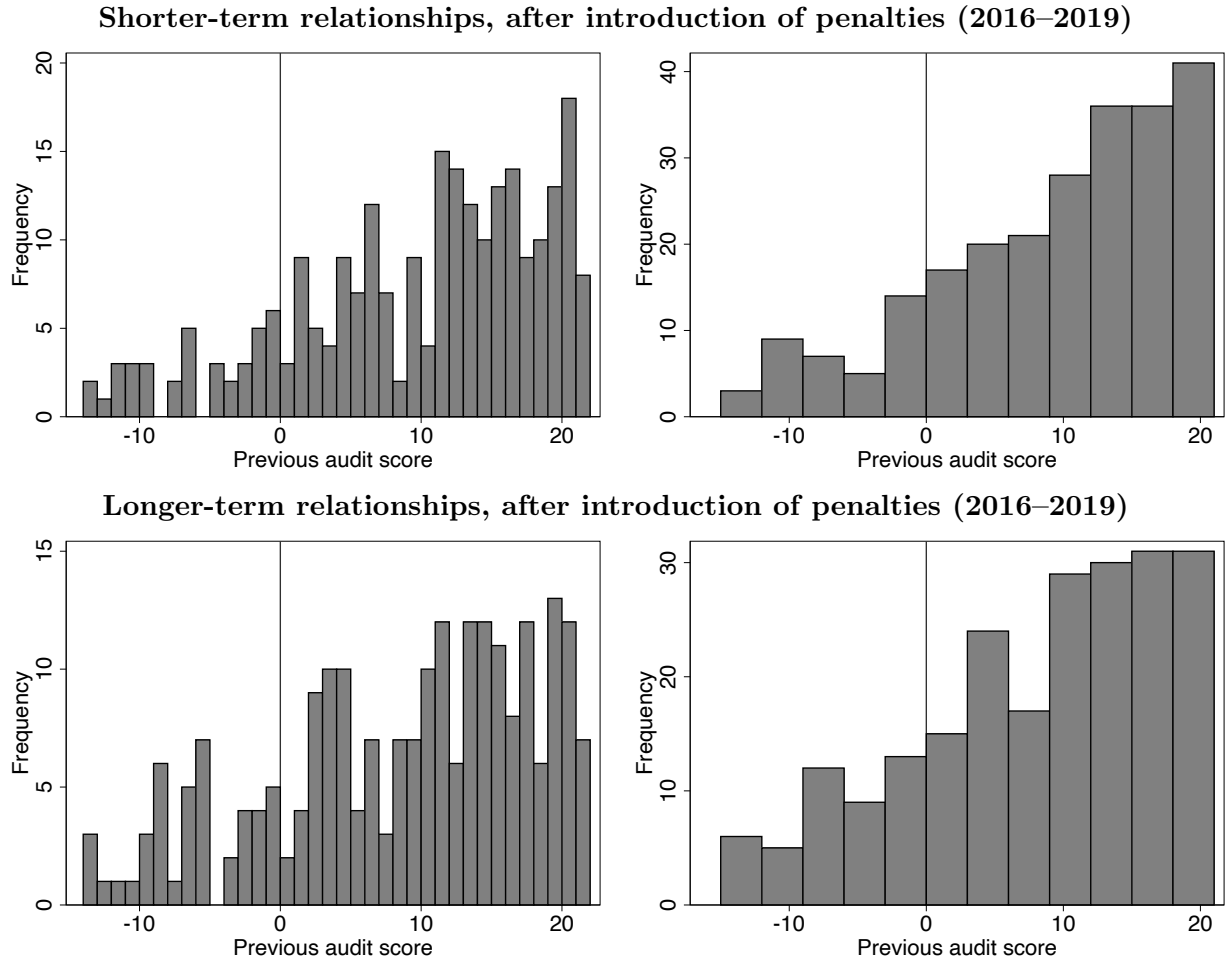
Notes. Tests of continuity of pre-treatment covariates at the threshold of failure in the old ratings system (2010-2014). Positive values indicate that factories that passed have higher values than those that failed. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018).

Table A6: Covariate balance in cooperative period (2010–2014) – longer sourcing relationship (4.5+ years)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	3.76	0.27	[10.4, -2.9]	8.5	605
Medium-severity	-0.50	0.48	[0.9, -1.9]	10.0	722
High-severity	-0.36	0.50	[0.7, -1.4]	8.4	604
Highest-severity	-0.03	0.91	[0.5, -0.5]	10.7	784
Better Work audit?	-0.04	0.33	[0.0, -0.1]	13.0	986
Days since last audit	-27.97	0.36	[31.5, -87.5]	10.3	608
Days until outcome audit	8.91	0.70	[54.9, -37.1]	13.4	1,016
Better Work audit?	-0.04	0.33	[0.0, -0.1]	13.0	986
Pre-previous audit score	-0.19	0.90	[2.9, -3.2]	10.2	599
Relationship (years)	0.94	0.22	[2.4, -0.6]	8.3	585
ln(units shipped) (std)	-0.06	0.81	[0.4, -0.5]	7.6	281
Workers	353.00	0.40	[1172.6, -466.6]	12.6	951
Female workers (%)	2.25	0.70	[13.8, -9.3]	11.8	576
Manufacturer?	-0.06	0.45	[0.1, -0.2]	9.1	646
Factory in China	0.02	0.82	[0.2, -0.2]	11.9	889
India	-0.08	0.41	[0.1, -0.3]	13.4	1,011
Indonesia	0.04	0.51	[0.2, -0.1]	7.1	503
Vietnam	-0.01	0.82	[0.1, -0.1]	9.3	667
other country	0.07	0.49	[0.3, -0.1]	13.8	1,050

Notes. Tests of continuity of pre-treatment covariates at the threshold of failure in the old ratings system (2010-2014). Positive values indicate that factories that passed have higher values than those that failed. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018).

Figure A4: Audit score density by relationship length, after introduction of penalties (2016–2019)



Notes. Histograms of audit scores that serve as running variable in the regression discontinuity analysis. Manipulation tests using local polynomial density estimation fail to reject the null of no difference in density in the shorter-term relationship data ($p = 0.48$). However, in the longer-term relationship subsample, the audit density is slightly higher above the threshold ($p = 0.03$).

Table A7: Covariate balance after introduction of penalties (2016–2019) – shorter sourcing relationship (under 6 years)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	0.63	0.83	[6.3, -5.1]	11.7	120
Medium-severity	0.06	0.94	[1.7, -1.5]	13.3	135
High-severity	-0.30	0.59	[0.8, -1.4]	15.7	174
Highest-severity	0.08	0.73	[0.5, -0.4]	11.1	102
Better work audit?	0.02	0.92	[0.3, -0.3]	14.0	150
Days since last audit	51.40	0.17	[124.1, -21.3]	8.1	48
Days until outcome audit	-15.74	0.70	[64.2, -95.7]	12.6	136
Pre-previous audit score	7.33	0.53	[30.1, -15.4]	9.0	50
Outcome audit by ILO	0.10	0.64	[0.5, -0.3]	12.1	121
ln(units shipped) (std)	0.31	0.45	[1.1, -0.5]	16.9	190
Relationship (years)	0.15	0.86	[1.8, -1.5]	17.6	199
Workers	1014.56	0.09	[2191.2, -162.0]	8.0	78
Female workers (%)	-10.81	0.35	[11.8, -33.4]	16.1	150
Manufacturer?	-0.11	0.48	[0.2, -0.4]	16.5	190
Americas	-0.18	0.16	[0.1, -0.4]	8.6	84
Mediterranean	0.08	0.48	[0.3, -0.1]	18.3	199
North Asia	0.08	0.74	[0.6, -0.4]	8.8	84
South Asia	-0.15	0.45	[0.2, -0.5]	17.3	190
Southeast Asia	0.27	0.16	[0.6, -0.1]	17.0	190
Factory in China	0.08	0.74	[0.6, -0.4]	8.8	84
India	-0.17	0.37	[0.2, -0.6]	13.5	136
Indonesia	0.38	0.03	[0.7, 0.0]	9.2	84
Vietnam	-0.20	0.24	[0.1, -0.5]	8.3	82
other country	-0.09	0.71	[0.4, -0.6]	9.1	84

Notes. Tests of continuity of pre-treatment covariates at the threshold between failing and passing audit scores. Positive values indicate that passing factories have higher values than failing ones. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018). We make no adjustment for multiple comparisons.

Table A8: Covariate balance after introduction of penalties (2016–2019) – longer sourcing relationship (6+ years)

	Regression discontinuity est.			MSE-optimal	Eff.
	Est.	p-val.	95% CI	bandwidth	obs.
Low-severity violations	0.48	0.83	[4.9, -3.9]	12.3	123
Medium-severity	-0.45	0.50	[0.9, -1.8]	12.5	130
High-severity	-0.35	0.42	[0.5, -1.2]	10.2	99
Highest-severity	0.21	0.43	[0.7, -0.3]	14.7	159
Better work audit?	-0.11	0.55	[0.2, -0.5]	10.2	100
Days since last audit	-54.34	0.33	[55.0, -163.7]	12.6	124
Days until outcome audit	-106.43	0.05	[-2.1, -210.8]	13.7	146
Pre-previous audit score	-0.46	0.97	[27.3, -28.2]	11.8	118
Outcome audit by ILO	-0.06	0.77	[0.4, -0.5]	12.5	124
ln(units shipped) (std)	-0.30	0.48	[0.5, -1.1]	9.5	100
Relationship (years)	-1.16	0.50	[2.2, -4.5]	12.2	124
Workers	-789.59	0.38	[959.9, -2539.1]	8.7	90
Female workers (%)	11.52	0.27	[32.1, -9.1]	11.4	100
Manufacturer?	-0.12	0.51	[0.2, -0.5]	8.3	77
Americas	0.04	0.46	[0.2, -0.1]	13.7	146
Mediterranean	0.03	0.76	[0.2, -0.1]	14.5	160
North Asia	0.37	0.07	[0.8, -0.0]	9.6	100
South Asia	-0.35	0.19	[0.2, -0.9]	8.8	90
Southeast Asia	-0.07	0.68	[0.3, -0.4]	10.9	111
Factory in China	0.37	0.07	[0.8, -0.0]	9.5	90
India	-0.26	0.35	[0.3, -0.8]	9.5	100
Indonesia	-0.12	0.07	[0.0, -0.2]	11.5	111
Vietnam	-0.05	0.73	[0.2, -0.3]	14.8	160
other country	-0.04	0.86	[0.4, -0.4]	12.1	124

Notes. Tests of continuity of pre-treatment covariates at the threshold between failing and passing audit scores. Positive values indicate that passing factories have higher values than failing ones. Estimations use first-order polynomials and algorithmically selected symmetrical bandwidths that minimize mean squared error. Estimated p-values and confidence intervals come from robust estimators described in Cattaneo, Idrobo, and Titiunik (2018). We make no adjustment for multiple comparisons.