8-4-2020

Is a Fine Still a Price? Replication as Robustness in Empirical Legal Studies

Cherie Metcalf
Queen's University (Canada) - Faculty of Law, metcalfc@queensu.ca

Emily Satterthwaite
University of Toronto - Faculty of Law, emily.satterthwaite@utoronto.ca

Brock V. Stoddard
Appalachian State University - Department of Economics, brvstod@gmail.com

Shahar Dillbary
University of Alabama - School of Law, sdillbary@law.ua.edu

ALABAMALAW
THE UNIVERSITY OF ALABAMA

Is a Fine Still a Price? Replication as Robustness in Empirical Legal Studies

Cherie Metcalf
Emily A. Satterthwaite
J. Shahar Dillbary
Brock Stoddard

63 INTERNATIONAL REVIEW OF LAW AND ECONOMICS 1
(2020)

Is a Fine Still a Price?
Replication as Robustness in Empirical Legal Studies

Cherie Metcalf, Emily A. Satterthwaite, J. Shahar Dillbary & Brock Stoddard[1][2]

Abstract

Can fines lead to more of an undesirable behavior, rather than deterring it? This was the surprising finding in Uri Gneezy and Aldo Rustichini, "A Fine is a Price" published in the *Journal of Legal Studies* in 2000. In this field experiment at Israeli daycares, the introduction of fines caused an increase in late pick-ups by parents. The study has been frequently cited, especially for its suggestion that a fine can act as a price for non-compliance that "crowds-out" social norms and motivations for individuals.

In this article, we conduct two related studies to explore the robustness of Gneezy & Rustichini's findings and the relevance of their suggested explanations. We seek to replicate their results using experimental surveys administered on MTurk, an increasingly common methodology in empirical legal studies, psychology and economics. While not an exact replication, it allows us to control aspects of the experimental design that are difficult to replicate in the field. We are also able to directly investigate whether fines persistently change the way respondents perceive the consequences (signaling, completing the contract) or relevant social motivations (crowding-out), as suggested by Gneezy & Rustichini. In the first study we translate the original daycare field setting into a vignette-based experimental survey. Our second study similarly investigates the effect of introducing a fine on income tax reporting compliance – an example suggested in the original study. In both studies, respondents are randomized into experimental conditions exposing them to one of two alternate fines or a social norm-based measure. We solicit multiple compliance measures for respondents along with measures of the importance of the alternate explanations to their decisions.

Our survey results do not replicate the original findings. In both our daycare and tax studies, the introduction of fines causes respondents to reduce non-compliant behaviour. Respondents expect others to behave similarly. Fines do not cause respondents to adjust their concerns about an incomplete contract consistently with Gneezy & Rustichini's theory. They also show very little evidence of fines crowding out social motivations, despite being responsive to our social treatments. The effects of fines on outcome behaviours and respondents' reasons are transitory. Once the fines are removed, our respondents return to their baseline behaviours. The survey results are consistent with our intuitions (and standard rational-choice theory) that fines deter. A survey is not a field experiment; however, our results suggest that more research is required to understand when and how any "fine is a price" effect may arise.

Keywords: fines; deterrence; tax compliance; social norms; motivational-crowding; crowding-out; experimental survey; replication

---

[1] Queen's University Faculty of Law, University of Toronto Faculty of Law, University of Alabama School of Law, and Appalachian State University, Department of Economics respectively. Please address correspondence to Cherie Metcalf – metcalfc@queensu.ca.

[2] Forthcoming in *International Review of Law & Economics* 63 (2020) 105906, https://doi.org/10.1016/j.irle.2020.105906. Thanks to participants at the 2019 PELS Replication Conference at Claremont McKenna College, especially our commentator Paige Skiba, participants at the 2019 SIOE, CLEA and CELS Conferences and to James MacKinnon and Robin Boadway for helpful comments and discussion of this project; thanks to Suzanne Rath and Kelly Zhang for excellent research assistance. Financial support from the *Social Sciences & Humanities Research Council of Canada* (Queen's SIG – Explore Grant (Metcalf)) is gratefully acknowledged.

I. Introduction

Uri Gneezy and Aldo Rustichini published "A Fine is a Price" in the *Journal of Legal Studies* in 2000.[3] The authors used a randomized field experiment at Israeli daycares to test whether "when negative consequences are imposed on a behaviour, they will produce a reduction of that particular response."[4] Instead of validating this standard assumption in psychology, economics, and legal studies, the authors found that introducing a fine *increased* late behaviour in their experiment.

Gneezy & Rustichini offered two possible explanations – both of which involved the fine transforming parents' perceptions of the context. The first explanation was that the fine provided information about the consequences of coming late that had been unspecified in the contract between parents and the daycare. While parents may have previously feared severe consequences from coming late, introduction of the fine made it clear this was not a threat.

Gneezy and Rustichini offered another powerful explanation: before the fine existed, a social norm restrained parents from coming late. Once the fine was introduced, it transformed the parents' perceptions of the extra care into a commodity. They were now free to consume it at the "price" of the fine, without the guilt or shame felt when transgressing the social norm. Parents' internal, moral motivations were "crowded out" by the market signal sent by the fine. Gneezy & Rustichini's experiment has been widely accepted as support for this latter, intuitively appealing theory. The paper has been cited more than 2000 times including in popular non-fiction for general audiences.[5]

Gneezy and Rustichini's social norms explanation fit well with emerging developments in economic theory that expanded standard approaches to individual preferences. Theorists developed models for individual preferences that reflected altruism and other-regarding behaviour, drawing on empirical support from standard economic games.[6] Another strand of emerging theory developed models of preferences dependent on reference points, providing a related theory of how a shift in behaviour might occur.[7] Scholars also provided

[3] (2000) Vol. XXIX *Journal of Legal Studies* 1.
[4] *Ibid* at 2.
[5] See e.g. Camerer, C., G. Loewenstein, and M.Rabin, *Advances in Behavioral Economics* (Princeton University Press, 2003), and popular works such as Dan Ariely's *Predictably Irrational* (Harper Collins, 2008).
[6] See e.g. Elizabeth Hoffman, Kevin McCabe and Vernon L. Smith, "Social Distance and Other-Regarding Behavior in Dictator Games" (1996) 86(3) *AER* 653; Gary Charness & Matthew Rabin, "Understanding Social Preferences with Simple Tests" (2002) 117(3) *QJE* 817; Ernst Fehr & Urs Fischbacher, "Why Social Preferences Matter – the Impact of non-Selfish Motives on Competition, Cooperation and Incentives" (2002) 112(478) *The Economic Journal* C1 (overview).
[7] See e.g. Botond Koszegi and Matthew Rabin, "A Model of Reference-Dependent Preferences" (2006) 121 *Q. J. Econ*. 1133; Robert Sugden, "Reference-Dependent Subjective Expected Utility" (2003) 111 *J. Econ. Th'y* 172. For a recent review of related theory and empirical literature, see Uri Gneezy, Lorenz Goette Charles Sprenger & Fliroan Zimmermann, "The Limits of Expectations-Based Reference Dependence" (2017) 15 (4) *J. European Econ Ass'n* 861.

models and empirical evidence consistent with individuals behaving differently in contexts perceived as social as opposed to market exchanges.[8]

Despite these developments, questions remain about the extent to which social norms influence choices in economic environments and whether fines "crowd out" the effects of norms, as Gneezy and Rustichini suggested.[9] Most relevant for our project, there has been some criticism of Gneezy and Ristichini's original field study[10] and the authors themselves suggested potential benefits from replication, even undertaking their own follow-up study.[11]

Given the significant influence of the original paper and the importance of the prospective behavioral insight it uncovers, we believe it is a good candidate for a form of replication study. However, in the case of field studies like Gneezy and Rustichini's original, faithfully reproducing the full set of underlying conditions embedded in the social and temporal context for the field experiment is challenging. This would make it difficult to view any null result as a true failure of replication. For empirical work that is inherently contingent on complex background factors such as legal institutions, culture and political conditions, we suggest that a "many studies"[12] approach as a way to test robustness of the underlying mechanism may be more appropriate than strict replication. Empirical work and results like this are perhaps best conceptualized as an ongoing project that inherently involves the collaboration of a community of researchers. The creation of a body of related work will ground confidence in the insights and an understanding of their limits.

With this approach in mind, we conduct a number of related studies to explore the robustness of Gneezy & Rustichini's original results. We transform the original daycare field setting into a vignette-based experimental survey administered on Amazon's Mechanical Turk (MTurk) platform. This approach allows us to control aspects of the experimental design that are difficult to replicate in the field, and to directly investigate the alternative explanations Gneezy and Rustichini offered for their original results. We also test for evidence of the underlying mechanisms they identify in a tax compliance setting – an alternate context where Gneezy and Rusitchini suggested their "fine is a price" effect could appear. Survey-based research in empirical legal studies has grown, particularly as

---

[8] For a paper pre-dating Gneezy & Rustichini's study, see e.g. George A. Akerlof, "Labor Contracts as a Partial Gift Exchange" (1982) 97(4) *QJE* 543, see also e.g. Torgler, B. "Tax Morale, Rule Governed Behavior & Trust" (2003) 14 *Constitutional Political Economy* 119 online: https://doi.org/10.1023/A:102364362228.

[9] See e.g. Fellner, Gerlinde and Sausgruber, Rupert and Traxler, Christian (2013) "Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information" (2013) 11(3) *J. Eur. Econ Ass'n* 634 (field study, enforcement of TV license fees); Tor Helge Holmås, Egil Kjerstad, Hilde Lurås & Odd Rune Straume, "Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay" (2010) 75(2) *J. Econ Behav. & Org.* 261.

[10] Ariel Rubenstein has provided several critiques of the field study, see e.g. http://arielrubinstein.tau.ac.il/papers/behavioral-economics.pdf.

[11] See a reply to Rubenstein http://arielrubinstein.tau.ac.il/papers/WC05/GR.pdf.

[12] See Krin Irvine, David Hoffman and Tess Wilkinson-Ryan, "Law and Psychology Grows Up, Goes Online and Replicates" (2018) 15 *JELS* pp. 320-355, generally and at 345 (contextual sensitivity and replication).

accessibility through online platforms such as MTurk has expanded.[13] We seek to replicate Gneezy & Rustichini's famous result with this increasingly common approach. While our approach is thus not a direct opportunity to confirm or refute the original field trial findings, we hope to provide evidence relevant to their generality and contribute to the emerging literature on the use of MTurk in empirical legal studies, psychology and economics.

Our survey results do not replicate the original findings. In our daycare and tax studies, the introduction of fines causes a reduction in anticipated non-compliant behaviour by our survey respondents. Moreover, subjects anticipate that other respondents will be similarly affected. Fines do not cause subjects to adjust their concerns about an incomplete contract in a manner that is consistent with Gneezy & Rustichini's theory. They also show very little evidence of fines crowding out any social norms. Moreover, the effect of fines on outcome behaviours and the reasons behind the outcomes of our respondents is transitory. Once the fines are removed, they return to their baseline behaviours. These survey results are consistent with our intuition (and standard rational-choice theory) that fines deter.

The remainder of the paper is organized as follows. In Section II we introduce our primary replication of Gneezy and Rustichini's daycare study. Section II includes discussion of our experimental structure and survey instrument, survey implementation and data collection, and results. In Section III we extend the "Fine is a Price" replication to the tax compliance context, discussing these same features for our tax study. In Section IV we use regression analysis to investigate the sensitivity of our results to individual characteristics. Section V concludes.

II. Replication of the Original Study: Robustness to Alternate Empirical Strategies

Our first study replicates Gneezy & Rustichini's (G&Rs) original research as faithfully as possible, while translating it to an experimental survey.[14] We retain G&Rs context of parents dropping off children at daycare centres. We test for the impact of introducing a fine, compared with an initial setting without any explicit consequence for late pickups specified, and then test for any change in behaviour when the fine is removed. This mirrors

[13] Examples of experimental survey-based research abound in many legal contexts, see e.g. David A Hoffman & Tess Wilkinson-Ryan, "The Psychology of Contract Precautions" (2013) 80 *U Chicago L. Rev.* 395 (contract); Janice Nadler & Shari Seidman Diamond, "Eminent Domain & the Psychology of Property Rights: Proposed Use, Subjective Attachment & Taker Identity" (2008) 5 (4) *JELS* 713; Cherie Metcalf, "Property Law Culture: Public Law, Private Preferences & the Psychology of Expropriation" (2014) 39 *Queen's L.J.* 685 (property); W. Jonathan Cardi, Randall Penfield & Albert Yoon, "Does Tort Law Deter Individuals?: A Behavioural Science Study" (2012) 9 *JELS* 567 (torts), Jerome Olsen, Mattias Kasper, Janina Enachescu, Serkan Benk & Tamer Budak, "Emotions & Tax Complaince among Small Business Owners: An Experimental Survey" (2018) 56 *Int. Rev. Law & Econ* 42 (tax); for numerous additional examples, see Irvine *et al* (2018), *ibid* at 321 (notes 1-6), Adriana Robertson & Albert Yoon, "You Get What You Pay For: An Empirical Examination of the Use of MTurk in Legal Scholarship" (2019) 72 *Vand. L. Rev.*1633 at 1640-41 (notes 25-40).

[14] Data and supporting materials for both our daycare and tax studies are available to researchers, see Cherie Metcalf, Emily Satterthwaite, Shahar Dillbary & Brock Stoddard, " Is a Fine Still a Price: Replication as Robustness in Empirical Legal Studies – Data and Supporting Materials 2019" https://doi.org/10.5683/SP2/9ZNNS4. The studies were subject to oversight and approval by Queen's General Research Ethics Board.

the structure of the original experiment and allows us to see how the fine may shift the perceptions of our respondents.

A. Initial Conditions

We use details from the original study regarding the daycare and important elements of the setting that were emphasized by the authors in their original study to construct our vignettes (nature of the centres, number of children, similarity of centres, ownership structure, who receives the fine, how it is administered, the nature of the notice posted, size of fine, how long the fine is in place, etc.).[15]

As other authors have noted, replications must be sensitive to the need to update the details of vignette studies, so that they will continue to faithfully test the underlying mechanisms.[16] In order to make sure that details of the G&R original would not be too far from those found in a current US context, we reviewed a number of sources to get a sense of details such as how daycares would typically be described, usual operating hours, fees, numbers of children, etc.[17]

Respondents in our study were all exposed to the same initial description of the daycare setting, as below[18]:

The Daycare

> The daycare provides care to children from age 1-4. It is regulated and a maximum of 35 children can attend. It is overseen by a Director who is qualified as an Early Childhood Educator. The Director is at the daycare from when it opens at 7:30 until 3:00 pm each day. After 3:00 pm, daycare teachers supervise the children until they are picked up.
>
> Parents sign a contract when their children begin attending the daycare. It states that the daycare operates between 7:30 am and 5:30 pm each weekday.
>
> If any parent is late, the daycare teachers take turns waiting with the children after closing time until all the children are picked up. The teachers are required to record the names of any parents who arrive late, and the time that they arrived. Parents almost never come later than 6:00pm.

[15] See Gneezy & Rustichini (2000), at 3-5 for description of details, at 16 for wording of the notice.
[16] See e.g. Irvine *et al* (2018), at 344-345.
[17] For background facts about childcare arrangements in the US, see e.g.: https://www.census.gov/topics/families/child-care/about.html (includes facts from Survey of Income and Program Participation); US – terminology, qualifications for supervisor / workers – US BLS: https://www.bls.gov/ooh/personal-care-and-service/childcare-workers.htm#tab-8 (supervisors); https://www.bls.gov/ooh/personal-care-and-service/childcare-workers.htm (workers); qualifications – ECE (US) – e.g. https://www.leg.state.nv.us/Session/77th2013/Exhibits/Assembly/HHS/AHHS370G.pdf . Similar sources were consulted for Canada to get a sense of comparability / commonality.
[18] Respondents were already aware that they would be answering questions about daycare policy, from the brief description of the HIT in MTurk, and the information and consent page for the study which was the first screen in the Qualtrics online survey.

> Imagine that you are a parent, and that your child attends this daycare. Please answer a few questions for us.

Some of the key elements of these initial conditions which we included to mirror the setting as described in G&R include: the nature and size of the daycare, its status as a regulated center, the presence of a "Director", as distinct from daycare "teachers", the time between when the Director left and closing, the inclusion of operating hours in a contract signed by parents, and the procedure for late pickups.

One issue we faced in converting the field trial to an experimental survey was consideration of what additional information actual daycare parents would likely know, that survey participants might not. This is reflected in our decision to include the line, "Parents almost never come later than 6:00 pm" in our initial conditions.[19] The typical outer boundary for late pick-ups seemed like information that parents would likely acquire over time, or have some contextually informed estimate of as part of their decision-making. Without specifying some information to provide guidance about how parents behave at daycares, we were concerned that our MTurk participants would differ from field trial parents in a way that was not related to testing the underlying hypothesis. Considering just this detail led us to generally question the extent to which relevant contextual information was embedded in the field trial, but not specified or discussed in the original study.

Once our respondents had acquired the baseline information through the initial conditions screen, we solicited measures for variables to test the main results from the G&R study; the frequency with which parents came late in a typical week, and the length of time parents would arrive after closing. Again, this involved decisions about how to best convert the field trial into a survey replication. G&R's original results reported measures of the "number of late coming parents" in the control and treatment groups each week over their study period.[20] In our one-shot survey of individual respondents, we used a qualitative question about late frequency in a typical week, followed by a numerical measure of lateness in minutes on any single day, on a continuous scale from 0-60.[21]

An advantage of using a survey vignette approach is that it allowed us to include additional measures for the alternate explanations of G&R's original result. The social norms story - parents initially show social restraint in taking advantage of the generosity of teachers who watch their children when they are late - is perhaps most commonly associated with the study. However, a second explanation offered in the paper was that the introduction of a fine helped fill a gap in the incomplete contract. Without any specific contractual term for late behaviour, G&R thought that "parents could form any belief" on the consequences of

[19] This mirrors G&R's description of their study setting – they indicated that parents rarely came later than 30 minutes past the formal closing time, Gneezy &Rustichini (2000), at 4.
[20] See e.g. *ibid* at 6-8. There is some limited discussion of the "values of delay", at 7, which we have interpreted to mean the temporal length of delay associated with each "late coming parent" observation.
[21] The frequency of lateness asked how often respondents thought they might be late in a typical week, ranging from "never" to "always". The question measuring late time was "What is the maximum amount of time you would want to be late on any day?"

coming late – ranging from mild to severe.[22] Parents' initial hesitation about being late could be driven by fear of unknown, potentially serious consequences. In our survey, we directly solicited measures of the relative importance of these two types of explanation for respondent late behaviour. We used an 11 point numeric scale to assess the contribution of concerns that being late would "affect my child's ability to keep attending the daycare" vs. feeling "bad about the teacher having to wait for me."[23] We also directly ask respondents if they "would feel guilty or ashamed about being late" and how many other parents they thought would be late on a typical day.[24]

Together with the late measures, these responses following our initial description of the daycare setting establish our baseline observations – corresponding to G&R's observations of their control and experimental groups prior to introduction of the fine.

B. Experimental Conditions

i) Gneezy & Rustichini – Replication Fine Condition

With the baseline established, we introduced the "G&R fine" condition. Respondents saw the following information in a single screen of the survey:

> **The Late Pick-up Announcement**
>
> Recently the Director posted the notice below at the daycare, on a notice board that all parents check daily for important news and information.
>
> **Announcement: Fine for Coming Late**
>
> As you all know, the official closing time of the daycare center is 5:30 pm every day. Since some parents have been coming late, we have decided to impose a fine on parents who come late to pick up their children (this is approved under our regulations).

[22] Gneezy &Rustichini (2000), at 10-13. The severe consequence example was being excluded / kicked out of the daycare.

[23] The scale ranged from 0-10, with qualitative descriptors at each end ("not at all" to "extremely" important). An additional question about the importance of the respondent's child "feeling bad if I am late to pick them up" was also included. This was partly to give an "air of reality" to the set and provide a relative benchmark for the contractual and social norms explanations against what many people might consider the main reason for parents to be on time. The questions appeared together in a table with a slider to select the number within the range (0-10) on their own page within the survey.

[24] The first question is a 7-point qualitative scale, from "strongly agree (1)" to "strongly disagree" (7). The latter is a multiple-choice question, with options from 0-5 late parents, and an additional category for "another number" of late parents in which respondents could indicate their own figure. This last option was treated as a categorical indicator for more than 5 late parents in our analysis. Respondents were reminded that there were 35 children at the center. This structure for beliefs about others late behavior was intended to strike a balance between providing enough context that survey respondents could respond in a manner comparable to a parent in a field trial and allowing individuals freedom to make their own assessments. The numbers of late parents in Gneezy &Rustichini (2000) was used to help determine the range for the multiple-choice options.

As of next week, a fine of $15 will be charged every time a child is picked up after 5:40 pm.

This fine will be calculated monthly, and it is to be paid together with the regular monthly payment.

This late fine announcement again mirrors the original field trial conditions as closely as possible.[25] One issue was the choice of value for the fine. In the original study the fine was "10 NIS" – to provide context, G&R compared the size of the fine to various other fines (e.g. parking ticket, red light infraction, etc.) However, the extent to which the prohibited behaviour would be detected varied considerably across their comparators – and we would have no easy way of assessing the similarity of detection for any analogous fines in our context. The fine size comparators we settled on in this pilot were G&R's comparison of the fine to the hourly cost of a baby-sitter, and their description of the fine as "small but not insignificant". Our fine is roughly in the neighbourhood of the hourly price for a baby-sitter / hourly wage for a daycare teacher.[26] We chose $15 as a "round" figure that would be easy for respondents to think about, as with the original 10 NIS.

With the G&R Fine condition in place, we then revisit the questions on late measures, and the additional reasons, feelings and beliefs about late behaviour. Answers to these questions provide our analog to G&R's observations of their experimental group during the time period with the fine in place.

Finally, we provide information to tell respondents that the fine has been cancelled:

Another Change

After the late policy with the fine had been in place for two months, a new notice was posted on the announcement board:

Announcement: Fine for Coming Late - *Cancelled*

The policy imposing fines for late pick-up of children is cancelled, effective immediately.

Respondents are then asked to revisit their answers to the same questions, with responses this time serving as our analog to the observations in the field study period following removal of the fine.

ii) Robustness: An Alternate Fine Condition

In order to assess the potential sensitivity of the "fine is a price" effect, we also created a new experimental condition that adopts a slightly more common structure for a late fine. G&R's fine was unusual in being both quite modest in size, and a flat fee that did not vary at all with the length of time parents were late. The structure of the fine in our alternate

[25] The original announcement (translated into English) is at 16 in Gneezy &Rustichini (2000).
[26] There is considerable variation in typical rates for babysitters or daycare teachers across US states / by area / qualifications, but these provided a general guide to the magnitude.

condition was based on the desire to have a fine that would increase with the amount of delay in pick up, while being similar to the G&R fine in size at low levels of late behaviour, and not so large that (as G&R suggest) it would become self-evident that a fine would deter. These criteria combined with an exploratory investigation of actual daycare fine structures led us to adopt the following description for the Alternate Fine (AF):

> **As of next week, a fine of $15 will be charged every time a child is picked up after 5:40 pm. An additional $5 will be charged for each additional 10 minutes thereafter.**

All other aspects of the fine condition remained unchanged. The rest of the survey structure was identical to the G&R Fine condition above.

iii) Robustness: Influence of Social Norms?

The powerful intuitive appeal of G&Rs original result draws on our acceptance that there would be significant social constraints operating in an environment like the daycares of their field study. In our survey, we introduce a further experimental condition (SN) that appeals directly to these presumptive social norms. This provides us with a way to see if respondents are sensitive to social norms in the survey environment (within-subjects) and use a between-subjects comparison to assess the relative impact of the fine conditions vs. social norms in determining late behaviour.

In this treatment we adopt a slightly different sequencing than that in the G&R and ALT F conditions. In our first "social norms" treatment, we use a structure similar to the announcement of the late fine – but it is simply a social "reminder":

> **The Late Pick-up Announcement**
>
> Recently the Director posted the notice below at the daycare, on a notice board that all parents check daily for important news and information.
>
> **Announcement: Coming Late**
>
> As you all know, the official closing time of the daycare center is 5:30 pm every day. Since some parents have been coming late, we have decided to remind all parents about closing time.
>
> **Remember, when you are late teachers must wait! Please be considerate and be on time.**

This social reminder allows us to test whether our MTurk subjects are responsive to social norms and compare the incentive effects on behaviour with those from the fines. However, the social reminder lacks an analog to the explicit removal of the monetary fines in G&R, or our Alternate Fine conditions.

As a secondary experimental condition in our "social norms" treatment, we went the other direction and escalated the social norms penalty by adding a public dimension – the "Late List":

Another Change

> After the announcement asking parents to remember closing time and be considerate had been posted for two months, a new notice was posted on the announcement board:
>
> **Announcement: Coming Late - Late List**
>
> As you all know, the official closing time of the daycare center is 5:30 pm every day. Since some parents have been coming late, even though we have reminded everyone of the hours and asked for your cooperation, we have decided to post a list of parents who come late to pick up their children (this is approved under our Regulations).
>
> **As of next week, every time a child is picked up after 5:40 pm we will post it in a "Late List" on this announcement board.**
>
> This "Late List" will be compiled monthly, and it will be posted at the time of the regular monthly payment.

This additional condition allows us to assess how a social sanction in the form of "naming and shaming" might work in comparison with the fine conditions. While this condition goes somewhat beyond a simple replication of the original, it provides additional information about the relative influence of social vs. economic incentive effects. The public disclosure of "late" behaviour has been shown to affect individuals in other contexts.[27]

The order of our primary and secondary treatments within our three experimental conditions was not varied from its presentation above. While in some settings one might randomize the order of treatments, in general the ordering within our experimental conditions is driven by the structure of the original field trial and our desire to replicate it as faithfully as possible.

C. Survey Administration – MTurk

We administered our survey using the Mechanical Turk (MTurk) platform. This Amazon platform allows "Requesters" to post "Human Intelligence Tasks" (HITs) for the workers who have signed up to provide services through the site.[28] Our survey was hosted and programmed within Qualtrics survey software, using a link to make the survey available to workers through a HIT on MTurk.

Our sample of respondents was solicited through MTurk by posting the HIT for our surveys. We used MTurk's qualifications feature to restrict access to the survey to MTurk

[27] For example, see recent work on the influence of the public "Six Month List" on speed and timing with which federal judges in the U.S. decide pending cases and motions; Miguel de Figueiredo, Alexandra Lahav and Peter Siegelman, "The Six Month List and the Unintended Consequences of Judicial Accountability" (2020) 105 (2) *Cornell L. Rev* 363 (SSRN-id2989777.pdf), Jonathan Petkun, "Can (and Should) Judges Be Shamed? Evidence from the 'Six-Month List'" available at: https://jbpetkun.github.io/pages/working_papers/CJRA_working_draft_20190524.pdf.

[28] See https://www.mturk.com/ for Amazon's own description of MTurk.

workers with US addresses, who had an approval rating for their work on MTurk of 90% or better, and who had completed at least 50 HITs on MTurk.[29] We evenly randomized assignment of workers to one of our experimental conditions and used the "create new qualifications" feature in MTurk to prevent MTurk workers from taking more than one version of the survey. We assigned a random number code to workers on completion of the survey within Qualtrics that workers entered manually in the MTurk HIT window in order to receive credit for competing the assignment within MTurk. This code was used to match survey responses in Qualtrics with worker IDs for those who had completed our HIT and prevented duplicate survey taking.[30] Only workers who were eligible for our survey were able to see the HIT on MTurk. An attention check was used to screen inattentive workers out of the experiment at the outset, and a second attention check was used following a section on basic demographic questions.[31] Participants had to complete the survey fully to receive compensation through MTurk, however they could quit at any time during the survey.[32] Data from partial responses was discarded.

Although we are replicating a field experiment that involved only parents, we did not similarly restrict our MTurk sample. Gneezy and Rustichini suggest that the "fine is a price" effect is quite general. While an actual daycare parent would likely experience social pressure around being late differently than our MTurk respondents, G&R suggest the "fine is a price" effect arises from introducing a fine in a setting people *perceive* as governed by social norms in the baseline. Their incomplete contracts explanation is driven by how introduction of a fine shifts perceived consequences, which does not appear to depend on being an actual parent. By allowing MTurk workers to select into our HIT based on its description and pay, we are able to test the generality of the "fine is a price" effect through these perceptual channels. We did use pre-experiment questions to elicit basic demographic and income data, and used post-experiment questions to determine respondents' status as parents and experience with daycare as potential controls.

---

[29] While restricting the HIT to these highly qualified MTurkers may skew the results somewhat compared to making the task generally accessible, we are hoping to attract workers who will be attentive enough to the task to be influenced by considering the experimental conditions.

[30] Our survey was administered anonymously, so we were not able to use any IP address features to determine the location of workers or verify whether they had completed surveys more than once. However, our alternate check using the randomly-generated code and their MTurk worker ID should have prevented any duplicate survey takers.

[31] The initial attention check was similar to that used in Irvine *et al* (2018), at fn. 57. We used a question to set out the following text:
This study seeks to understand how people process the questions that are being asked to them. There are many aspects of a person's behaviour that are related to the way they answer questions.
One aspect is their ability to stay engaged throughout a survey and a person's willingness to read the directions fully. To make sure you are currently paying attention, we would like you to answer "none of the above" to the following question. Which of the following adjectives would you use to describe yourself?
This was followed by a multiple-choice question with a series of attributes (e.g. leader, bold, shy, trustworthy, etc.) and the last choice, "none of the above". Respondents who chose anything but "none of the above" were told they were ineligible for the survey and asked to return the HIT. The secondary attention check required respondents to type the word "survey" into a text box. Almost none of the respondents failed this check, and we simply discarded data from those who did.

[32] An alternate non-research task was available for any workers who wanted to earn the compensation without completing the survey, to preserve the voluntariness of participation.

A significant issue in using MTurk to solicit respondents is determining an appropriate level of pay. Recent work has shown that workers on MTurk are often underpaid – earning far less than the hourly minimum wage in the US for their work.[33] Moreover, their pay has also been shown to relate to their attentiveness and the quality of their work.[34] We used estimates for the time to complete our survey generated within Qualtrics, and actual completion times from a small pool of test subjects (not MTurkers, and not included in our data) to gauge how long the survey would take. We anticipated that our pay of $1.50 per survey would translate into an hourly wage of approximately $US 7-9 per hour. We anticipated this "above market" wage would allow us to gather new samples of respondents quickly and secure attentive performance in the survey task.[35]

Once we account for incomplete, inattentive or invalid responses, we obtained 1200 individual participants for our daycare survey through MTurk[36]

D. Data and Results

i) Is a Fine (Still) a Price?

The first question is whether our study reproduces the key outcomes from G&R's field trial. We focus on our outcome measures related to respondents' late behavior to test the following replication hypotheses based on G&R's original results:

> **H_0**: The introduction of a small fine against the no fine baseline should *increase* late behaviour.
>
> **H_1**: The effect of introducing the fine should be sticky. Once respondents perceptions of the context have changed, the removal of the fine should not have an effect.

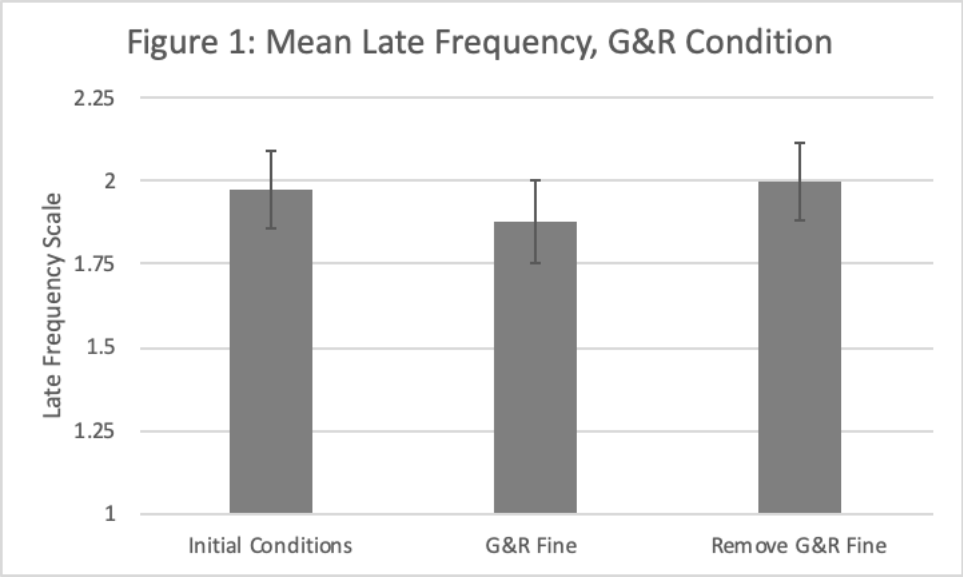
We use two main approaches to test for effects of our treatments: changes in means as average treatment effects and changes in the full distribution of responses.

a) Equivalence of Means as Treatment Effects

We first consider a within-subjects evaluation of the average treatment effects for both introduction and removal of the fine, looking first at the G&R fine condition and then the

[33] See e.g. Kotaro Hara *et al.* "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk" (manuscript) available at: https://www.researchgate.net/publication/321902451_A_Data-Driven_Analysis_of_Workers'_Earnings_on_Amazon_Mechanical_Turk.
[34] See e.g. Adriana Robertson & Albert Yoon (2019).
[35] In order to further enhance the attentiveness of our survey participants, workers were given only 30 minutes to complete the HIT on MTurk.
[36] The survey was launched in six discrete batches of HIT assignments: March 25 (45 responses, 9:39-10:52 am PDT); March 27 (150 responses, 12:21-1:41 PDT); March 28 (150 responses, 1:47-2:57 PDT); March 29 (150 responses, 8:16-9:33 PDT); June12 (350 responses, 11:41-1:47 PDT); June14 (350 responses, 8:06 PDT-9:05 PDT).

alternate fine (Alt F) condition as a robustness check. We consider the effects on both late frequency and maximum time respondents would want to be late on any day.



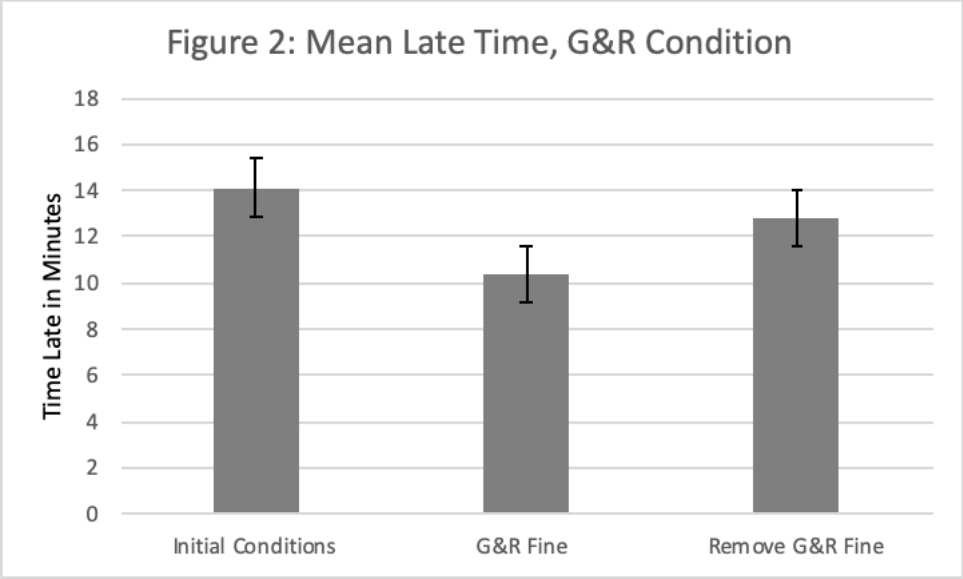
Notes to Figure 1: Late Frequency measured by 7-point qualitative scale, 1 = "never" late to 7= "always" late. Initial Conditions is the mean Late Frequency following initial description of the daycare setting. G&R Fine is mean late frequency after G&R replication fine introduced. Remove G&R Fine is the mean following removal of the G&R fine. Error bars are 95% confidence intervals.

As can be seen in Figure 1, introduction of the G&R replication fine condition produced a lower mean late frequency. With our qualitative responses scaled with "1" assigned to the category "Never" late, this drop reflects respondents' beliefs they would be late *less* often with the fine in place. Once the fine was cancelled, they returned to their baseline estimate of late frequency. However, these differences in average late frequency are small and not statistically significant.[37]

We observe a similar pattern in our measure of late time, but here there are statistically significant effects. Figure 2 illustrates the drop in the mean for the time in minutes that respondents indicate they would be willing to be late on any given day. The drop of roughly 4 minutes from the initial conditions once the G&R fine is introduced is statistically significant (P<.0000), as is the rebound when the fine is removed (P<.0078), leaving the final mean for our late time measure indistinguishable from the initial conditions (P<.1543).[38]

[37] The tests for equivalence of means are as follows for Late Frequency: H_0 $\mu_{IC}=\mu_{G\&R}$, $F_{(1,1199)} = 1.20$ (P< 0.2735); H_0 $\mu_{G\&R}=\mu_{FRemov}$, $F_{(1,1199)} = 1.89$ (P< 0.1690); H_0 $\mu_{IC}=\mu_{FRemov}$, $F_{(1,1199)} = 0.07$ (P< 0.7874).

[38] The respective test statistics are as follows for equivalence of means for Late Time: H_0 $\mu_{IC}=\mu_{G\&R}$, $F_{(1,1199)} = 16.81$; H_0 $\mu_{G\&R}=\mu_{FRemov}$, $F_{(1,1199)} = 7.10$; H_0 $\mu_{IC}=\mu_{FRemov}$, $F_{(1,1199)} = 2.03$.



Notes to Figure 2: Initial Conditions is mean time respondents are late following initial description of the daycare setting, before introduction of any fine treatments. G&R Fine is mean Late Time after G&R replication fine treatment. Remove G&R Fine is the mean Late Time when fine removed. Error bars are 95% confidence intervals.

The means for our key outcome variables in our Alt F fine condition produce a similar pattern, but with greater statistical significance. Introduction of the late fine produces a drop in the mean for late frequency (P<0.0129), while removal of the Alt F late fine creates a rebound in the mean for late frequency that again leaves it statistically indistinguishable from the initial conditions (P<0.0906).[39] The same pattern repeats for our measure for the amount of time that respondents would want to be late. When the Alt F fine is introduced, the drop in mean time respondents would be late of just under 3 minutes is strongly significant (P<0.0010). Once the fine is removed, respondents mean late time increases (P<0.0181) and becomes indistinguishable from their no-fine baseline estimates (P<0.3570).[40]

The comparisons of means for our key late measures indicates a failure to replicate results consistent with $\mathbf{H_0}$ and $\mathbf{H_1}$ from Gneezy and Rustichini. We obtain similar, consistent results from both our fine conditions. Introduction of a fine causes respondents to *decrease* their anticipated late behaviour. The change is *not* sticky – removal of the fine produces a rebound in respondents' late behaviour that is generally a statistically significant departure from behaviour with the fine, but indistinguishable from baseline behaviour.

b) Distributions of Responses as Treatment Effects

[39] The means for Late Frequency in the alternate fine condition are: 1.96, 1.76, and 1.89 respectively for initial conditions, introduction of the Alternate Fine, and Alternate Fine removal. Test statistics for equivalence of means for Late Frequency are: $H_0\ \mu_{IC}=\mu_{AF}\ \ F_{(1,1208)} = 6.20$ (P< 0.0129); $H_0\ \mu_{AF}=\mu_{FRemov}$, $F_{(1,1208)} = 2.87$ (P< 0.0906); $H_0\ \mu_{IC}=\mu_{FRemov}$, $F_{(1,1208)} = 0.83$ (P< 0.3627).

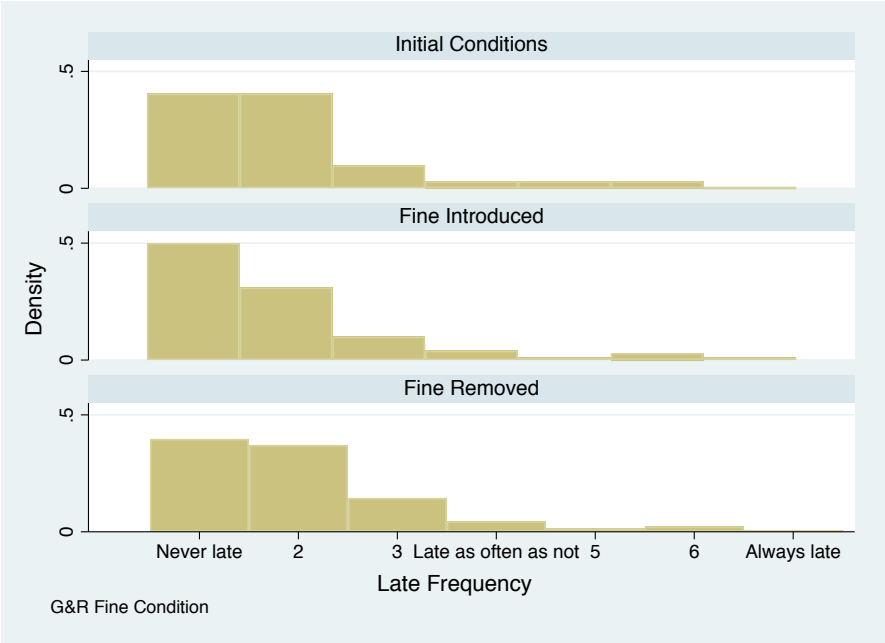
[40] The means for Late Time in the alternate fine condition are: 12.28, 9.40, and 11.47 respectively for initial conditions, introduction of the Alternate Fine, and Alternate Fine removal. Test statistics for equivalence of means for Late Time are: $H_0\ \mu_{IC}=\mu_{AF}\ \ F_{(1,1208)} = 10.90$; $H_0\ \mu_{AF}=\mu_{FRemov}$, $F_{(1,1208)} = 5.60$; $H_0\ \mu_{IC}=\mu_{FRemov}$, $F_{(1,1208)} = 0.85$.

In addition to examining changes in the mean as an average treatment effect, we also examine patterns in the full distribution of responses. The implication of **H0** is that we expect to see movement into the larger values of the distribution of our late measures when we introduce the fine. **H1** implies that when we remove the fine, the distribution of responses should not change. We include plots for the response distributions for Late Frequency and Late Time for the G&R condition below. Those for the Alt F condition are also examined as a robustness check. We use the Kolmogorov-Smirnov (K-S) test to determine whether there are any statistically significant differences in the distributions as treatments are varied within each condition.

Figure 3 below provides the plot for the distribution of Late Frequency in our G&R condition by treatment. The distributions again reveal basic inconsistency with **H0** and **H1** from G&R. Introduction of the fine causes responses to move into lower value categories – which corresponds with *reduced* late behaviour (1=never late). When the fine is removed, responses shift back into higher valued categories (increasing late behaviour) and become indistinguishable from the initial distribution. G&R's hypotheses would require a shift in responses toward the other end of the distribution when the fine was introduced, that remained stable when it was removed. The inconsistent patterns we identify instead are strongly statistically significant.[41]

[41] The Kolmogorov-Smirnov test statistic values for G& R Late Frequency distributions are as follows: **H0**: values in initial conditions (IC) < values in fine introduction (FI) D=.0075 (P<0.978); FI<IC D=-.0950 (P<0.027); **H1**: values in FI<values when fine removed (FR) D=0.1025 (P<0.015); values FR<FI D=-.0125 (P<0.939). The initial and final distributions are indistinguishable: Combined K-S D=0.0425 (P<0.863).

Figure 3: Distribution of Late Frequency by Treatment, G&R Condition



Notes to Figure 3: Vertical axis is density measure of response distribution, Horizontal axis is Late Frequency Scale. Initial Conditions panel is distribution prior to introduction of any fine treatments, Fine Introduced is distribution of late frequency after G&R replication fine introduced, Fine Removed panel is distribution of late frequency after cancelling G&R replication fine.

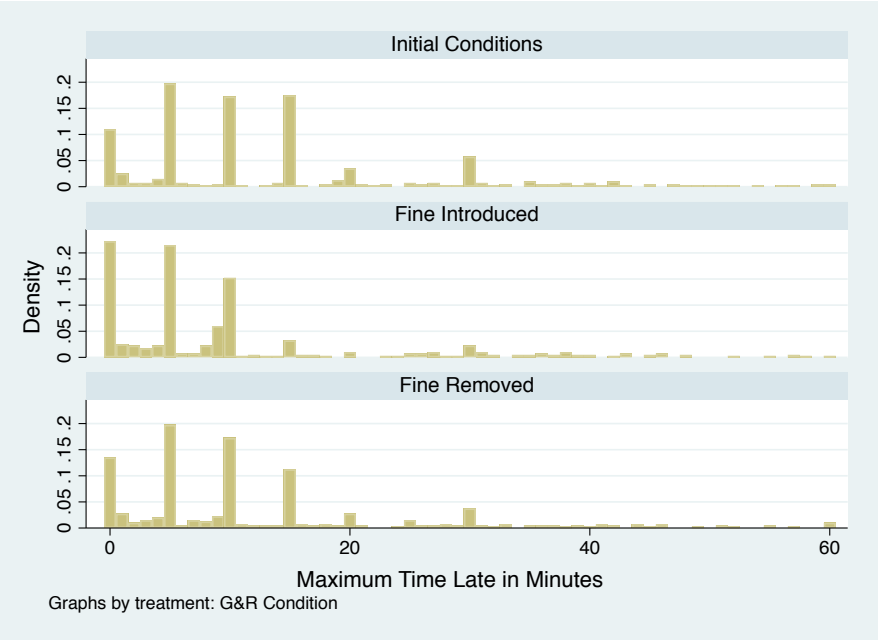
We obtain similar results when we examine the distribution of responses for our Late Time measure. The effect of introducing the fine can be seen in Figure 4, reducing late behaviour by shifting respondents into lower values of late time. Once the fine is removed, responses shift back toward higher values. The changes, again inconsistent with both $\mathbf{H_0}$ and $\mathbf{H_1}$, are strongly significant.[42]

The distributional results for our Alt F fine condition mirror those above for late frequency and late time in the G&R condition. The outcome measure distributions change (statistically significantly), but not in ways consistent with $\mathbf{H_0}$ and $\mathbf{H_1}$.[43]

[42] The Kolmogorov-Smirnov test statistic values for G&R Late Time distributions are as follows: **H0**: IC<FI D=.0000 (P<1.000); FI<IC D=-.2400 (P<0.000); **H1**: FI<FR D=0.11625 (P<0.000); values FR<FI D=-.0025 (P<0.998). The initial and final distributions are marginally different under the combined K-S D=0.0900 (P<0.078) (this is driven by significant test stat for FR<IC D=-.0900 (P<0.039). There is some slight persistence in effects from the fine even after it is removed – but it lowers the late behaviour, rather than increasing it relative to the baseline as in. G&R's story.

[43] The Kolmogorov-Smirnov test statistic values for Alternate Fine condition Late Frequency distributions are as follows: **H0**: IC<FI D=.0000 (P<1.000); FI<IC D=-.1464 (P<0.000); **H1**: FI<FR D=0.1315 (P<0.001); values FR<FI D=-.0074 (P<0.978). The initial and final distributions are indistinguishable: Combined K-S D=0.0174 (P<1.000). The corresponding K-S stats for ALT F Late Time distributions are as follows: **H0**: IC<FI D=.0149 (P<0.915); FI<IC D=-.1911 (P<0.000); **H1**: FI<FR D=0.1439 (P<0.000);

Figure 4: Distribution of Late Time by Treatment, G&R Condition



Notes to Figure 4: Vertical axis is density measure of response distribution, Horizontal axis is maximum time in minutes respondents are willing to be late. Initial Conditions panel is distribution prior to introduction of any fine treatments, Fine Introduced is distribution of Late Time after G&R replication fine introduced, Fine Removed panel is distribution of Late Time after treatment removing G&R replication fine.

ii) Are MTurkers Sensitive to Social Norms?

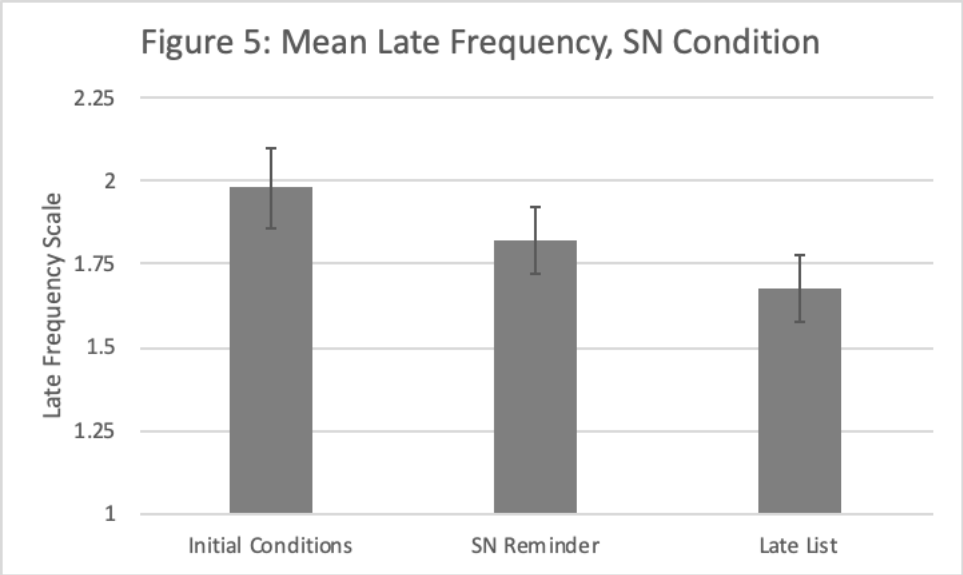
The existence of perceived social constraints on behaviour is important for the social norms story of G&R. We use our within-subjects late measures for our Social Norm Condition to assess whether they exert an influence on individuals' behaviour within our survey context. This can also help us test for the possibility that our results are a reflection of MTurkers being insensitive to social norms in the daycare vignette setting (e.g. in contrast to actual parents) and overly sensitive to monetary incentives.

a) Equivalence of Means as Treatment Effects

As seen in Figure 5 below, when the Social Reminder treatment is introduced, it produces a drop in late frequency. This is a statistically significant change from the initial conditions ($P<0.0588$). The Late List treatment causes mean late frequency to drop again and is marginally statistically distinguishable from the simple Social Reminder's effect ($P<0.0695$). The Late List does, however, produce a strongly significant difference in respondents estimates of how frequently they would be late compared with their initial responses ($P<0.0002$).[44]

values FR<FI D=-.0050 (P<0.990). The initial and final distributions are indistinguishable: Combined K-S D=0.0620 (P<0.420).

[44] The means for Late Frequency in the Social Norm condition are: 1.98, 1.82, and 1.68 respectively for initial conditions, introduction of the Social Reminder, and Late List. Test statistics for equivalence of

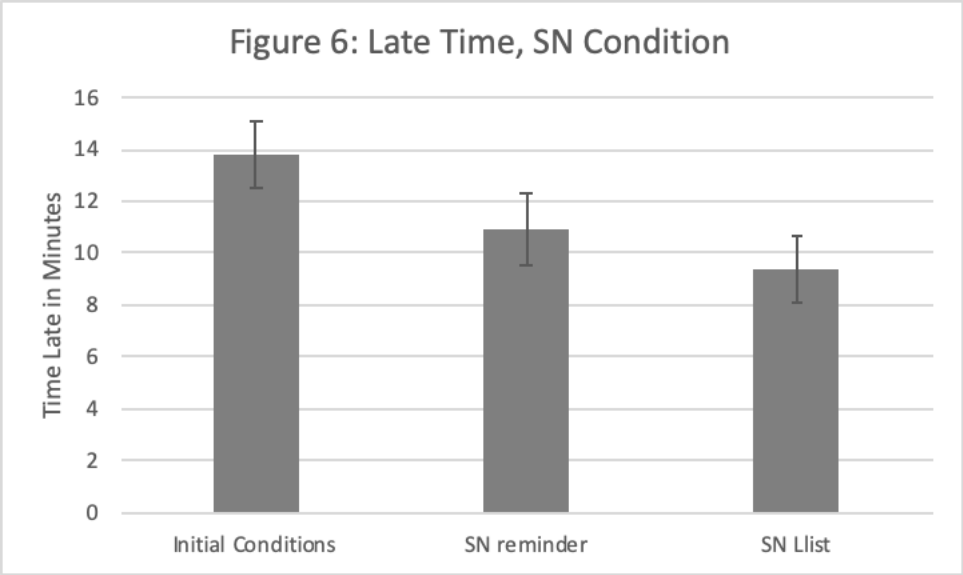


Notes to Figure 5: Late Frequency measured by 7-point qualitative scale, 1 = "never" late to 7= "always" late. Initial Conditions is the mean Late Frequency following initial description of the daycare setting, before introduction of any social norms treatments. SN Reminder is mean Late Frequency after Social Reminder is introduced. Late List is the mean Late Frequency after Late List treatment introduced. Error bars are 95% confidence intervals.

A similar pattern is seen in Figure 6 for the estimated mean Late Time in our Social Norms treatments. The Social Reminder causes respondents to reduce their estimates of the maximum time they want to be late, causing the mean to drop significantly from initial conditions (P>0.0022). The Late List causes the time to drop again, although not statistically differently compared with the simple reminder (P>0.1253). Again, there is a highly significant difference between respondents' initial baseline responses and the final mean late time in the Late List treatment (P>0.0000).[45]

means for Late Frequency are: $H_0\ \mu_{IC}=\mu_{SR}$ $F_{(1,1190)} = 3.58$; $H_0\ \mu_{SR}=\mu_{LList}$, $F_{(1,1190)} = 3.30$; $H_0\ \mu_{IC}=\mu_{LList}$, $F_{(1,1190)} = 13.72$.

[45] The means for Late Time in the Social Norm condition are: 13.8, 10.9, and 9.41 respectively for initial conditions, introduction of the Social Reminder, and Late List. Test statistics for equivalence of means for Late Time are: $H_0\ \mu_{IC}=\mu_{SR}$ $F_{(1,1190)} = 9.40$;$H_0\ \mu_{SR}=\mu_{LList}$, $F_{(1,1190)} = 2.35$; $H_0\ \mu_{IC}=\mu_{LList}$, $F_{(1,1190)} = 21.34$.

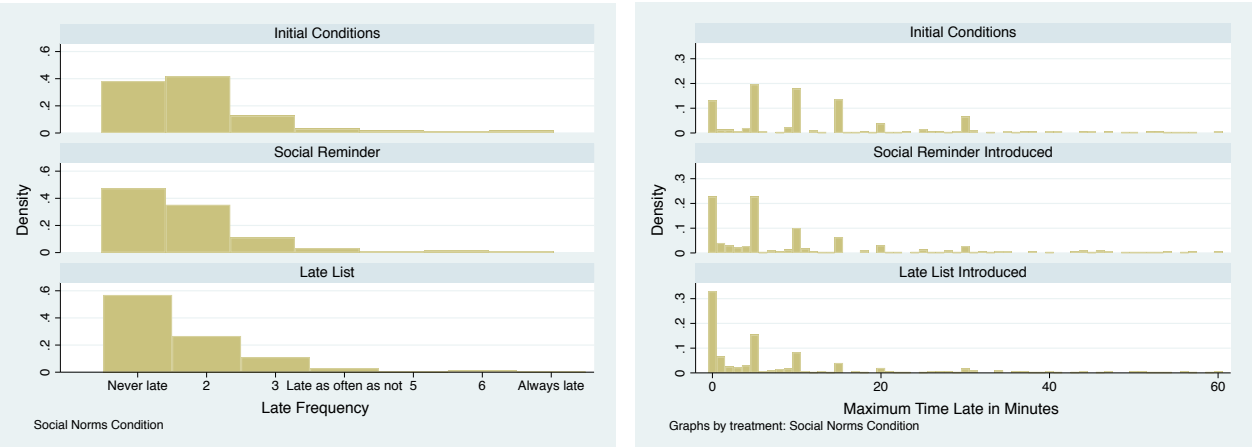


Notes to Figure 6: Late Time is maximum time in minutes respondents late on any day. Initial Conditions is the mean following initial description of the daycare setting, before introduction of any social norms treatments. SN Reminder is mean Late Time after Social Reminder treatment. Late List is the mean after Late List announced. Error bars are 95% confidence intervals.

b) Distribution of Responses as Treatment Effects

We confirm that our MTurk respondents are sensitive to social norm-based influences on their behaviour by examining patterns in the distribution of responses for our Social Norms condition. Distributions for Late Frequency and Late Time across treatments within this condition are set out below in Figure 7

Figure 7: Late Frequency & Time Distributions by Treatment, Social Norm Condition



Notes to Figure 7: Vertical axis is density measure for distribution of responses. Initial Conditions panel is distribution of Late Frequency and Late Time following description of the daycare setting. Social Reminder Panel shows distribution of responses for Late Frequency and Time following introduction of the Social Reminder treatment, and Late List shows the distribution following treatment announcing the Late List.

The initial distributions themselves appear to reflect the presence of a latent social norm similar to what G&R suggest for the parents in their field study. Responses are not normally distributed around a norm of being late as often as not or being on average later than the "grace" pick-up time of 5:40 pm. *Most* respondents intend to be quite timely in their pick-ups in the initial conditions, although there is quite a long tail, revealing a degree of heterogeneity amongst respondents. The Social Reminder that simply stresses that teachers must wait if parents are late shifts the distribution of responses for both Late Frequency[46] and Time[47] toward lower values. The announcement of the public Late List policy continues to similarly shift the distributions toward lower values and more on-time behaviour.[48]

Our Social Norms condition reassures us that our survey respondents have a similar anticipated response to being late in the daycare vignette context as G&R theorized for their field trial parents. Our MTurkers appear to be responsive to social prompts in our treatment in a way that supports use of the platform as a legitimate testing ground for replication of G&R's results.

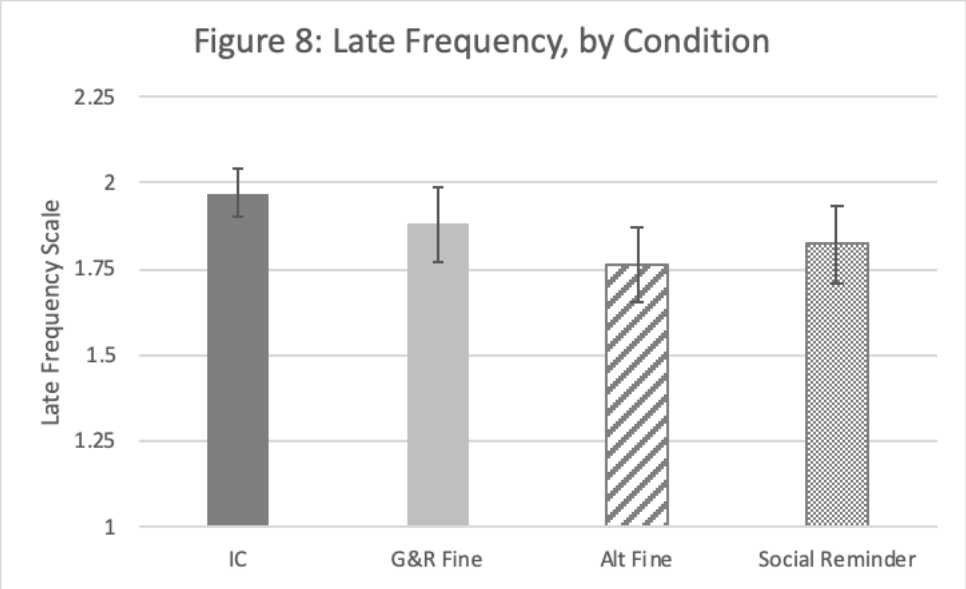
iii) Exploring the Relative Effect of Fines vs. Social Norms

G&R's results and their explanations suggest that our fine conditions and social norm conditions should work in opposite directions in shifting respondents from their initial conditions. Imposing fines should increase late behaviour relative to the no-fine baseline, while reinforcing the social obligations should decrease it.[49] This is the consequence of fines crowding out social norms, and if sufficiently low, incentivizing late behaviour. We use between-subjects comparison of means below to assess the relative impact of our experimental conditions on the key variables of interest – Late Frequency and Late Time.[50]

As seen in Figure 8 below, the average Late Frequency is lower in all our experimental treatments than in the Initial Conditions, although with the exception of the Alt F fine, the differences are not statistically significant. With our scale (1=never late), all treatments work to reduce respondent's anticipated late behaviour. The fines and social reminder thus work in the same direction. There are no statistically significant differences between

[46] The Kolmogorov-Smirnov test statistic values for the Social Norm condition Late Frequency distributions are as follows: **H0**: IC<Reminder D=.0000 (P<1.000); Reminder<IC D=-.0907 (P<0.038).
[47] The Kolmogorov-Smirnov test statistic values for the Social Norm condition Late Time distributions are as follows: **H0**: IC<Reminder D=.0000 (P<1.000); Reminder<IC D=-.0907 (P<0.038).
[48] The K-S test statistics for Late Frequency in the Late List treatment are: **H1(sticky)**: Reminder**<**LList D=0.0000 (P<0.1.000); LList<Reminder D=-.0957 (P<0.026). The initial and final late frequency distributions are strongly statistically distinguishable: Combined K-S D=0.1864 (P<0.0000). The K-S test statistics for Late Time are: **H1(sticky)**: Reminder**<**LList D=0.0000 (P<0.1.000); LList<Reminder D=-.0957 (P<0.026). The initial and final late time distributions are strongly statistically distinguishable: Combined K-S D=0.1864 (P<0.0000).
[49] If social norms are the latent restraint in the initial conditions, then making them more salient with the reminder / late list should not cause lateness to increase, even it if did not produce a drop in lateness.
[50] The mean for Initial Conditions is pooled across conditions for this comparison. In this section we focus only on our main treatment, introduction of our experimental treatments (fine, reminder) rather than secondary removal. We have already established with our within subjects results that the stickiness G&R expected is not replicated.

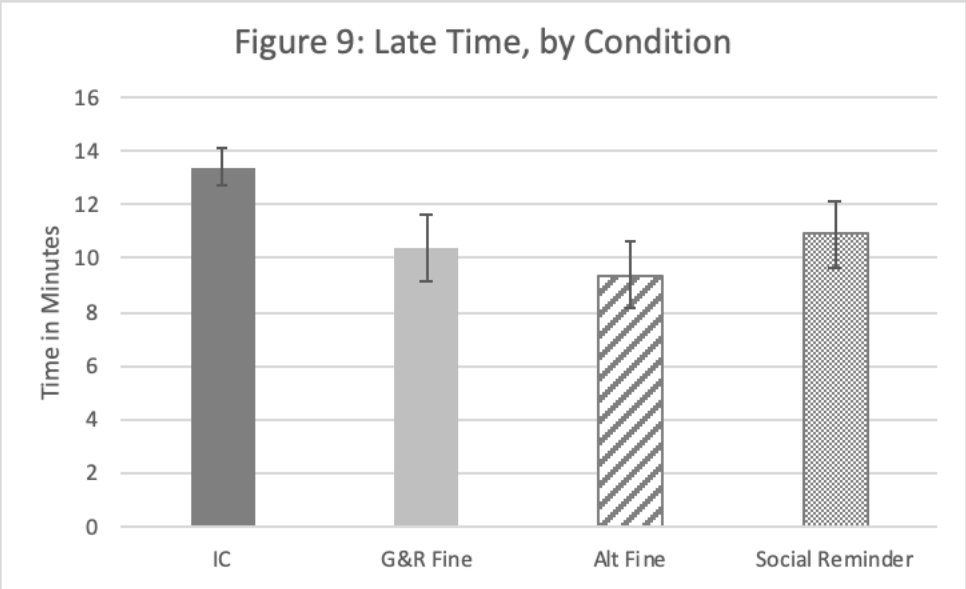
the mean effects for any of our treatments, although the Alt F fine produces the lowest point estimate of the mean effect. [51]



Notes to Figure 8: IC is the mean Late Frequency for all respondents in the initial conditions, before introduction of any treatment imposing consequences for late behaviour. G&R Fine is mean Late Frequency among respondents exposed to the G&R replication fine condition. Alt Fine is mean Late Frequency for respondents exposed to the Alt F fine condition. Social Reminder is mean Late Frequency for respondents exposed to the Social Reminder condition. All means are for initial treatment introducing the measures. Error bars are 95% confidence intervals.

The pattern for Late Time across conditions is similar. In Figure 9 below, we see that all experimental treatments work to reduce Late Time relative to the Initial Conditions, statistically significantly. Once again, all the treatments work in the same direction, rather than the introduction of either of our fines producing an effect in the opposite direction of the social norm reinforcement. There are no statistically significant differences between the main effects of the experimental conditions on Late Time.[52]

[51] We use a simple OLS regression of individual observations of Late measures on dummies for each experimental condition and use F-tests for significance of the treatment parameter estimates. Comparing G&R to Alt F and SN respectively, we do not reject equivalence with P<.1418, P<.4944 and P<.4340.
[52] Using F-tests: G&R=Alt F (P<.2725), G&R=SN (P<.5911) and Alt F=SN (P<.1025).



Notes to Figure 9: IC is the mean Late Time for all respondents in the initial conditions, before introduction of any treatment imposing consequences for late behaviour. G&R Fine is mean Late Time among respondents exposed to the G&R replication fine. Alt Fine is mean Late Time for respondents exposed to the Alt F fine condition. Social Reminder is mean Late Time for respondents exposed to the Social Reminder condition. All means are for initial treatment within our conditions introducing fines / social norm measures. Error bars are 95% confidence intervals.

This simple comparison illustrates a basic conclusion: respondents behave differently than in Gneezy and Rustichini's field trial – despite evidence that social considerations are a factor in respondents decision-making. In our replication, social and financial incentives work in the same direction in influencing late behaviour. Fines and social norms both appear to increase the costs of being late for our respondents, deterring their intended late behaviour.

iv) How do Fines Change Perceptions?

The second major focus of our replication study is to directly investigate how our experimental conditions relate to changes in the perceptions of our respondents. Does the introduction of a fine cause changes in our subjects' reasons for their decisions in the way that Gneezy and Rustichini suggest? G&R's two alternate hypotheses are set out below:

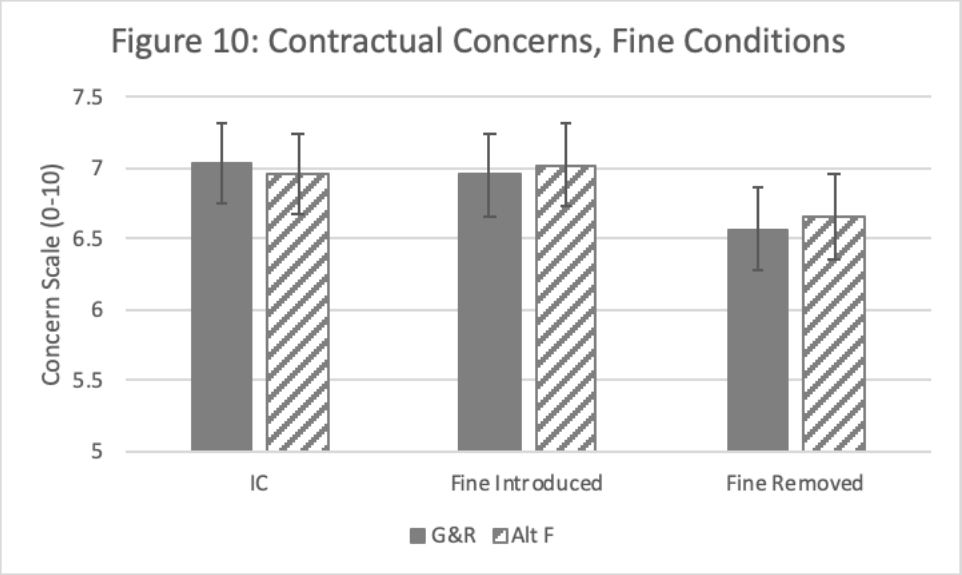
> **H2**: The introduction of a (small) fine will decrease fears of severe contractual consequences relative to the incompletely specified baseline. This effect should be sticky.
>
> **H3**: The introduction of a (small) fine will reduce social concerns, by crowding out social norms with the market signal. This effect should be sticky.

Translating these into hypotheses specific to our measures, **H2** implies we should see reduced concern about being able to attend daycare in future with introduction of the fine conditions. We expect that if **H3** is operative we will see decreases in concerns over making

teachers wait and less guilt about being late when fines are introduced. In both cases, if G&R's hypotheses hold, we expect to see respondents' beliefs about other parents' late behaviour change when the fines are introduced; they should expect **more** late behaviour from others (**H4**).

a) Fines & Contractual Concerns

We explore the impact of our fine conditions on our measure of contractual concern in Figure 10 below. With our scale (10=extremely concerned), a drop in value reflects a reduction in concern.



Notes to Figure 10: IC is mean level of contracutual concerns for respondents in fine treatments in the initial conditions, before introduction of any late fine, scaled from 0 to 10 ("extremely concerned"), means illustrated from mid-point. Fine Introduced is mean contractual concern after introducing G&R replication and Alt F fine treatments. Fine removed illustrates the mean level following announcement removing the fines. Error bars are 95% confidence intevals.

The results reveal some interesting patterns. The introduction of the fine does not result in any significant change in respondents' own concerns about consequences in terms of being able to continue attending the daycare – the "severe" contractual consequence for being late that G&R believe parents may fear in the baseline before the fine completes the contract. Instead, *removing* the fine causes a drop in the mean importance of this consideration compared to both the initial conditions and concerns with the fine in place. The drop from introduction of the G&R fine when it is removed is marginally significant (P<.0747), and the difference between the initial conditions and the final level of contractual concern is also significant in the G&R condition (P<.0250). The pattern is similar for Alt Fine, with the drop between the fine introduction and removal being marginally statistically significant (P<.0830). When we directly test whether our fines are different from each other, we find that they are statistically indistinguishable. Our results fail to match G&R's hypothesis, **H2**, both in terms of producing results in the expected direction when our fines are introduced and in generating a sticky effect from the fines once they are removed.

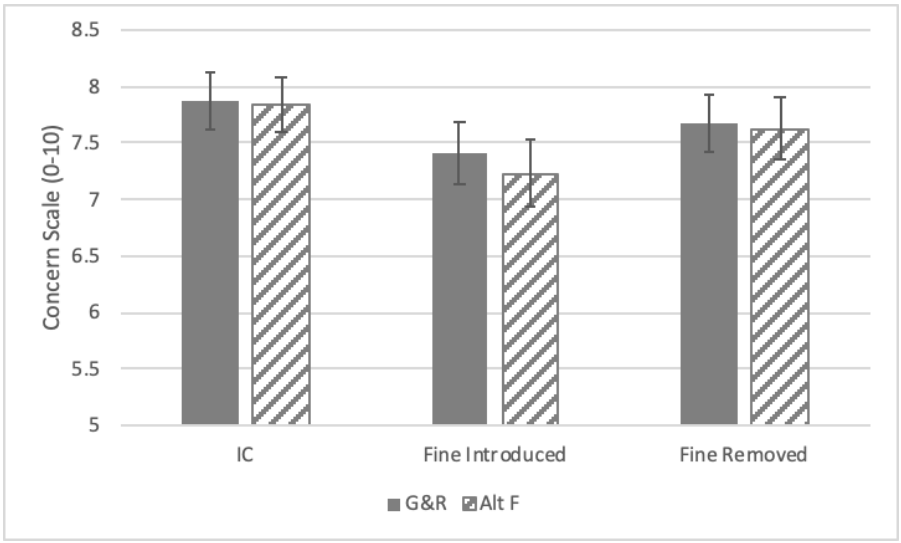
We also check these fine treatment effects by examining the full distribution of scaled responses for our question on contractual concerns. We check for equivalence in distributions across treatments within both of our G&R and Alt F fine conditions. The results are consistent with the main average treatment effects discussed above. For the G&R condition, there is no difference in distributions between IC and introduction of the fine. The removal of the fine produces a marginally significant shift in the distribution toward lower values (K-S $P<0.066$), leaving the final distribution of contractual concerns also marginally lower in values than the initial distribution (K-S $P<0.077$). There are no statistically significant changes in the distributions with the Alt F fine treatments.

Overall, our results do not support an incomplete contracts explanation for the behaviour of respondents in response to introduction of fines. There is no substantial change in concerns about severe consequences of the sort G&R suggest a small fine might produce – despite the fact our respondents have a relatively high level of concern initially (7/10). Instead, it is removal of the fine that appears to signal future lenience / reveal the type of regulator to our respondents, to the extent there is any change at all.

b) Fines & Social Norms

Figure 11 below presents our means across treatments for social reasons in our two fine treatments. Respondents indicated how important it was to their decision that they would "feel bad" about making the teachers wait (10 = extremely important).

Figure 11: Social Concerns, by Fine Treatments



Notes to Figure 11: IC is mean level of social concern for teachers initially expressed by respondents, prior to our late fine treatments, scaled from 0-10, illustrated from mid-point. Fine Introduced is mean level of social concern after introduction of the G&R (replication) and Alt F fine treatments respectively. Fine removed is mean levels following the announcement removing the fines. Error bars are 95% confidence intervals.

Here we see some support for G&R's prediction, **H3**, about how fines might interact with social norms. In both the G&R and Alt F fine conditions, the introduction of the fines causes our respondents to *reduce* their expressed level of social concern for the teachers. The initial drop in concern when the fine is introduced is strongly significant in both the G&R ($P<0.0120$) and Alt Fine ($P<0.0014$) conditions. However, this "crowding out" effect is not sticky in the way G&R suggest for our survey respondents. When the fine is removed, they adjust their concerns back upward. This shift from the level with the fine in place is not statistically significant in G&R condition ($P<0.1523$), but is in Alt F ($P<0.0385$). In both fine conditions, the final level of concern is statistically indistinguishable from baseline responses (G&R, $P<0.2663$; Alt F, $P<0.2697$).

The response in relation to social concerns provides the most support for G&R among our measures. It does move in the anticipated direction, with the fine "crowding out" respondents' concerns for workers, but only while the fine is in place. There does not appear to be any sticky transformation in the way our respondents thought about the problem. It should also be noted that even though there was a reduction in concern for workers with the fine in place, the level of concern remained relatively high at more than 7/10 on average. This is roughly equal to the highest mean measures for contractual concerns.

Again, when we examine the full distribution of responses as a check on these effects of our treatment, the results are consistent. Introduction of the fine moves the distribution of social concerns toward lower values (G&R, K-S $P<0.056$; Alt F, K-S $P<0.019$). There is some "stickiness" of the distribution in the G&R condition; we cannot reject that the distributions are indistinguishable with the fine in place or once it is removed (Combined K-S, $P<.758$). However, in the Alt F fine condition, we can reject stickiness of the fine (Combined K-S, $P<0.096$). However, in both fine conditions, the initial and final distributions of social concerns are indistinguishable.

Although some responses related to social concerns are consistent with G&R's explanation, overall, these effects are neither substantial enough in size or persistent enough to provide strong support for their social norms story.[53]

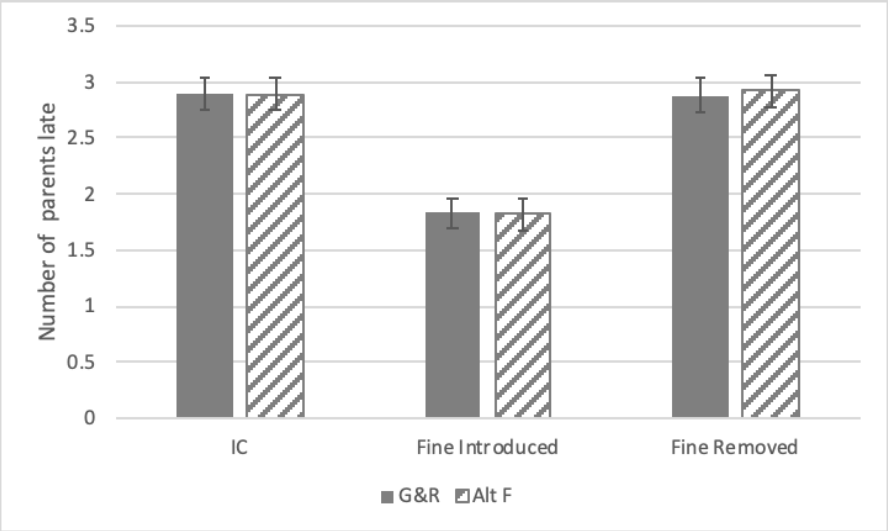
c) Fines & Expectations about Others' Behaviour

A corollary of both of G&R's explanations for their results is that introducing a fine also changes parents' expectations about how other parents will behave. We examine results from our question regarding how many parents our respondents think would be late to test for this effect in our fine conditions. This also serves as a useful check on how respondents themselves understand the fine's incentive effects – as presumably they will expect other

[53] We also examine respondents' answers to the question about whether they would feel guilty or ashamed about being late as a check on how the fine treatments may have affected social / moral concerns in line with G&R's theory. The G&R and Alt F fine conditions both produced a small drop in mean guilt with the fine in place, but only in Alt F was it significant ($P<0.0363$). In neither condition was the effect persistent or distinguishable from initial conditions. There were no significant changes in the distributions of responses for guilt across treatments for either condition.

parents to be influenced by the fine in a way that reflects their own reactions. Results for means are shown in Figure 12 below.

Figure 12: Late Parents, by Fine Treatments



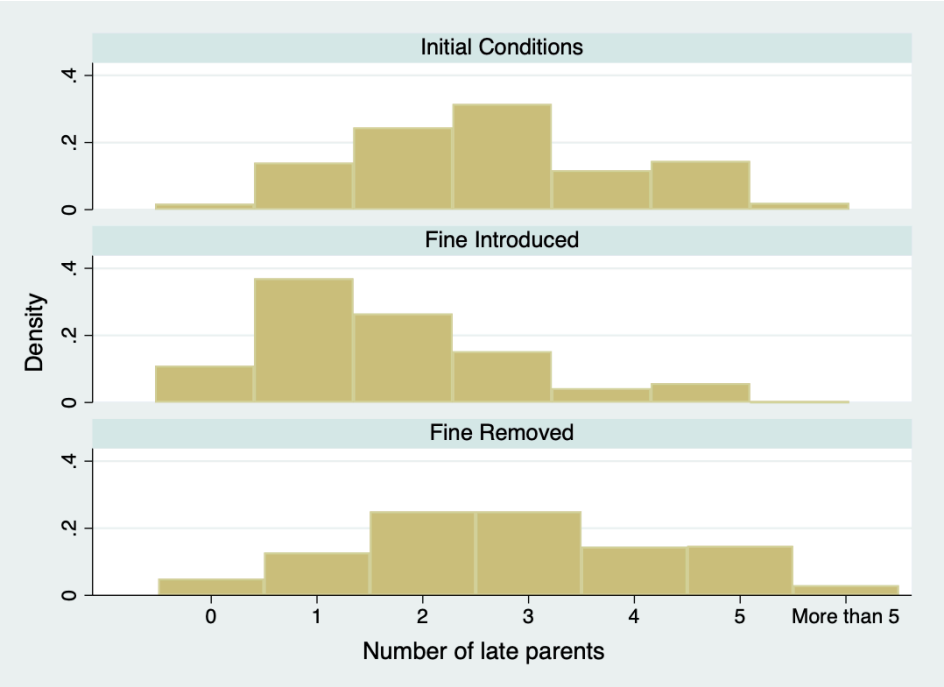
Notes to Figure 12: IC is mean number of parents respondents expect to be late initially, prior to late fine treatments. Fine Introduced is the mean number late following introduction of the G&R (replication) and Alt F fine treatments respectively. Fine removed illustrates the mean number of parents expected to arrive late following the announcement removing the fines. Error bars are 95% confiodence intervals.

Introduction of the fine produces a substantial and strongly significant shift in respondents' ideas about how many others will be late. Respondents anticipate *fewer* late parents with the fine in place. The mean number of expected late parents drops from approximately 3 down to 1-2 once the fine is put in place, then rebounds back to almost 3 on average when the fine is removed. Both transitions are highly significant (G&R, Alt F: P<0.000). This effect works in the opposite direction from that G&R expected in their explanation of their field experiment results. They thought that the "fine is a price" effect would cause parents to anticipate that others would also now be more inclined to be late, which would decrease any consequences of their own late behavior (individual parents "less responsible"). Instead, our respondents anticipate a general drop in late behaviour. This is in line with the standard theory about the incentive effects of fines – they deter – but only when fines are actually in place.

Once again, we confirm consistency of our results by looking to the full distributions of responses for the number of parents coming late. In both the G&R and Alt F conditions, introduction of the fine significantly shifts the distribution to one with lower values. Removal of the fine shifts it back to one with higher values, that is indistinguishable from

the initial distribution.[54] This pattern is apparent in the plot for the G&R replication fine condition below, in Figure 13.[55]

Figure 13: Distribution Others Late, G&R Condition



Notes to Figure 13: Vertical axis is density measure of response distribution, Horizontal axis is number of parents that respondents expect to be late. Initial Conditions panel is distribution prior to introduction of any fine treatments, Fine Introduced is distribution after G&R replication fine introduced, Fine Removed panel is distribution after removing G&R replication fine.

v) Conclusions – Daycare Study

Having set out and examined the various hypotheses drawn from G&R's field trial, we can only conclude that our study fails to replicate the original findings or support the suggested explanations. We produce significant effects for outcome measures with the introduction of our fines, but they are generally modest in size and in the opposite direction to G&R's striking results. Our fine treatments do not produce any sticky changes in the way our respondents appear to think about their decisions or (generally) their reasons. Instead, they appear to respond to the introduction and removal of the fine in a way that is in line with a fine's incentive effects in standard rational choice theory. This is despite the fact that our respondents *are* sensitive to social norms – evident in their reluctance to be late initially even in the hypothetical vignette, and their responsiveness to our social condition's reminder and late list treatments. Our results do confirm that social norms and social controls can influence anticipated behaviour in a vignette context like G&R's field study.

[54] K-S statistics generate the corresponding respective P-values: IC<F, P<0.000; F<FR P<1.000; IC≠FR, P<.906).
[55] The Alt F plot is virtually identical.

However, these effects do not appear to be crowded out by the introduction of fines in any significant way.

II. Replication: Extending the "Fine is a Price" Mechanism to the Tax Context

In line with emerging best practices, we use a second study to test the "fine is a price" effect.[56] We take our cue from the original article, in which Gneezy & Rustichini suggest that "the announcement of a government that tax evasions are going to be more severely pursued" may generate similar perceptual shifts as those experienced by the daycare parents.[57]

The possibility that fines for tax evasion may have unintended consequences reflects a central debate in the tax compliance literature that was well underway when G&R were conducting their research in the late 1990s: to what extent, or under what conditions, might sanctions (such as fines) backfire?[58]

In the same way that the counterintuitive result of a "fine is a price" challenges the straightforward application of regulatory tools, the "crowding out" literature challenges the neoclassical model of tax compliance that has "dominated the academic research in economics and in part also in psychology."[59] Developed independently by Allingham and Sandmo (1972) and Srinivasan (1973) using Becker's (1968) rational expected utility model of crime framework,[60] the main parameters of the neoclassical model are the income of the taxpayer, the tax rate applied to reported income, the probability of audit, and the magnitude of the penalty assessed on unreported income.[61] Individual taxpayers maximize their expected net utility, which involves a comparison of the probability-weighted payoff of the safe option (income less taxes paid) and that of the risky option (income less taxes and penalties paid in the event of audit/detection).[62] The comparative statics of the neoclassical model are unambiguously positive for the key parameter in which G&R are interested: the fine/penalty rate.[63] Increasing the fine/penalty rate on unreported income should result in an increase in voluntary compliance.[64]

---

[56] See Irvine *et al* (2018) at 346-7.
[57] See Gneezy & Rustichini (2000), at 16.
[58] See Leandra Leaderman, Does Enforcement Reduce Voluntary Tax Compliance? (2018) *8 BYU L. Rev.* (forthcoming), available at: http://ssrn.com/abstract=3222803.
[59] See Erich Kirchler, *The Economic Psychology of Tax Behaviour* (Cambridge University Press: Cambridge, UK, 2007), at 105 and 107 (noting that "[s]ince its publication, the vast majority of economic studies on tax behaviour refer to the Allingham and Sandmo model").
[60] See G. Becker "Crime and punishment: An economic approach" (1968) 76(2) *Journal of Political Economy* 169.
[61] See Michael G. Allingham & Agnar Sandmo, Income Tax Evasion: A Theoretical Analysis, 3 J. Pub. Econ. 323 (1972) and T.N. Srinivasan, "Tax Evasion: A Model" (1973) 2 *J. Pub. Econ.* 339.
[62] *Ibid* at 325.
[63] In the model, the penalty rate on unreported income is constrained to be higher than the tax rate on reported income.
[64] *Ibid* at 330 (stating that "unambiguous results can be derived for the two parameters of the model which are of interest for policy purposes in this field, viz. the penalty rate and the probability of detection [the audit rate]").

However, the results of empirical studies frequently diverge from the predictions of the neoclassical model, so other models focus on a broader range of factors such as social norms,[65] trust in government,[66] reciprocity,[67] justifications for audits,[68] and attitudes towards the tax system.[69] For example, Feld and Frey theorize that a "psychological tax contract" mediates the relationship between the government and the taxpayer and theorize the concept of "crowding-out": "…[e]xternal interventions undermine intrinsic motivation when they are perceived to be intrusive by the individuals concerned."[70] Lederman's recent review of this literature explores the contributions of scholars who "argue that enforcement ultimately will have the perverse effect of reducing voluntary tax compliance."[71]

Empirical studies on the effect of fines have yielded conflicting results, further fueling the debate about the limits of the neoclassical model. On the one hand, there are numerous studies concluding that fines strongly increase compliance.[72] On the other hand, many

---

[65] See Kent W. Smith and Karyl A. Kinsey, "Understanding Taxpaying Behavior: A Conceptual Framework with Implications for Research" (1987) 21(4) *Law & Society Rev*. 639.

[66] See Erich Kirchler, Erik Hoelzl, and Ingrid Wahl, "Enforced versus voluntary tax compliance: the slippery slope framework" (2008) 29 *J Econ. Psychology*210.

[67] See Kent W. Smith, "Reciprocity and Fairness: Positive Incentives for Tax Compliance", in Slemrod, J. (Ed.), *Why People Pay Taxes: Tax Compliance and Enforcement* (University of Michigan Press: Ann Arbor, MI, 1992), at 223-250.

[68] See Sebastian Beer, Matthias Kasper, Erich Kirchler, and Brian Erard, Audit Impact Study, Taxpayer Advocate Service Annual Report to Congress (2015), at 75, available at https://taxpayeradvocate.irs.gov/Media/Default/Documents/2015ARC/ARC15_Volume2_3-AuditImpact.pdf (noting that "On the one hand, audits may increase a taxpayer's trust in the state, and therefore, serve to reinforce the social norm of voluntary compliance. On the other hand, audits could be perceived as unjustified measures, thereby undermining one's willingness to comply voluntarily)."

[69] Steven M. Sheffrin and Robert K. Triest, "Can brute deterrence backfire? Perceptions and attitudes in taxpayer compliance" in: Slemrod, J. (Ed.) (1992), at 193–218.

[70] Lars P. Feld & Bruno Frey, "Tax Compliance as the Result of a Psychological Tax Contract: The Role of Incentives & Responsive Regulation" (2007) 29(1) *Law & Pol'y* 102, at 105-6 (noting that the implications of the psychological tax contract theory are that sanctions should be deployed with caution: "[t]axpayers' reward from that contract must be understood in a broad sense going beyond pure exchanges of goods and services for the payment of a tax price….As deterrence and tax morale interact, it would be counterproductive solely to rely on punishment or monetary (non-authentic) rewards, because tax morale can be undermined"). See also Bruno S. Frey and Reto Jegen, Motivation Crowding Theory: A Survey of Empirical Evidence, Journal of Economic Surveys, Vol. 15, No. 5, pp. 589-611 (2001).

[71] See Lederman (2018) at 627 (citing Benno Torgler's statement that "[w]hen monitoring and penalties for noncompliance are intensified, individuals notice that extrinsic motivation has increased, which . . . crowds out their intrinsic motivation to comply with taxes;" also citing Dan M. Kahan, "The Logic of Reciprocity: Trust, Collective Action, and Law" (2003) 102 *Mich. L Rev.* 71, at 83 and Marjorie E. Kornhauser, "Normative and Cognitive Aspects of Tax Compliance: Literature Review and Recommendations for the IRS Regarding Individual Taxpayers", in 2 *National Taxpayer Advocate 2007 Annual Report to Congress* 138, 15).

[72] See, e.g., N. Friedland, S. Maital, and A. Rutenberg, "A simulation study of income tax evasion" (1978) 10(1) *Journal of Public Economics* 10(1) 107 (conducting a laboratory experiment in Israel, finding that penalties increased compliance and their effect was stronger than the positive effect of audit rates on compliance); J. Alm, Sanchez, I., and deJuan, A, "Economic and noneconomic factors in tax compliance" (1995) 48(1) *Kyklos* 3 (using experimental data from laboratories in both the U.S. and Spain, similarly found that fines and audit rates were positively associated with compliance, and that the fine had the stronger effect); C.G. Park, and Hyun, J. K., "Examining the determinants of tax compliance by experimental data: A case of Korea" (2003) 25(8) *Journal of Policy Modeling* 673 (finding specifically that

studies from a wide variety of countries have found that higher penalties have a weak or statistically unobservable effect.[73] The possibility that enforcement will cause crowding-out under some conditions is broadly consistent with the one of the two mechanisms that G&R speculate may apply: imposing a fine may shift the psychological context of the tax compliance choice from being dominated by a norm of compliance for social reasons to being driven by a "rational choice" calculation in which the private net benefits of compliance take center stage. In addition, lower compliance in response to a fine could also be consistent with G&R's other proposed mechanism: the fine may provide information about an otherwise uncertain tax enforcement context and, if it is set too low, it may signal that severe consequences are unlikely. In this way, a sanction could function as a "price" for illegally underreporting income.

---

the elasticity of fines with respect to compliance was higher than that of audit rates); Friedland, N., "A note on tax evasion as a function of the quality of information about the magnitude and credibility of threatened fines: Some preliminary research" (1982) 12(1) *Journal of Applied Social Psychology* 54 (failing to replicate the 1978 finding that the effect of fines is stronger than audit rates, but confirming a positive effect of fines); Boris Maciejovsky et al., "Misperception of Chance and Loss Repair: On the Dynamics of Tax Compliance" (2007) 28 *J. Econ. Psychol.* 678, at 684 (stating, on the basis of results from a laboratory experiment in Vienna that "[e]ffect sizes suggest that audits have a stronger impact on compliance than fines"); Lederman (2018), at 672 (summarizing, at the conclusion of her literature review of tax penalty studies, "[o]verall, these studies suggest that audit threats are much more effective than sanction threats at increasing compliance").

[73]See Barak Ariel, "Deterrence and Moral Persuasion Effects on Corporate Tax Compliance: Findings from a Randomized Controlled Trial" (2012) 50 *American J. of Crim.* 27, at 43–52 (finding no statistically significant effect on corporate taxpayers in Israel of receiving a letter "inform[ing] taxpayers that filing a false report will result in harsh sanctions.); Fjeldstad, O.-H., and Semboja, J. (2001) "Why people pay taxes: The case of the development levy in Tanzania" (2001) 29(12) *World Development* 2,059 (using data from Tanzania, finding evidence that compliance varied positively with ability to pay, the perceived probability of being prosecuted, and the number of tax evaders known personally to the participant); Das-Gupta, A., Lahiri, R., and Mookherjee, D., "Income tax compliance in India: An empirical analysis" (1995) 23(12) *World Development* 2051 (analyzing aggregate administrative data from India, finding that both revenues collected and compliance were impacted by marginal tax rates and exemption limits but that the effects of traditional enforcement tools, such as searches, penalties, and prosecution, were weak by comparison); James Alm, Betty R. Jackson, and Michael McKee, "Estimating the Determinants of Taxpayer Compliance with Experimental Data" (1992) 45 *National Tax Journal* 107, at 110 (finding that "[c]ompliance increases with an increase in the fine rate; however, the coefficient on FINERATE is so small that the fine rate elasticity is virtually zero, and the coefficient is also not highly significant"); Webley, P., Robben, H. S. J., Elffers, H., and Hessing, D. J., *Tax evasion: An experimental approach* (Cambridge: Cambridge University Press, 1991) (reporting experimental results in which fines and audit rates were varied; the severity of the fine had no significant influence on the frequency of underreporting). See also Paul Webley, "Audit Probabilities and Tax Evasion in a Business Simulation", (1987) 25 *Econ. Letters* 267; Brooks, N., and Doob, A. N., "Tax evasion: Searching for a theory of compliant behavior" In M. L. Friedland (ed.), *Securing compliance: Seven case studies*(1990) (conducting a mail survey in Canada and finding that individual taxpayers' perceptions of sanctions/fines are less important to evasion decisions than their perceptions of the likelihood of being audited and apprehended); Kirchler (2007) at 113 (summarizing Weck-Hannemann, H., and Pommerehne,W. (1989). Einkommenssteuerhinterziehun in der Schweiz: Eine empirische Analyse. *Schweizerische Zeitschrift fu¨r Volkswirtschaft und Statistik*, 125(4), 515–56, which in a lab simulation of taxpaying in Switzerland found no significant effect of punishment on voluntary compliance, although the authors conjectured this could be due to awareness of Swiss regulations providing that the parliament rather than the tax agency had authority to set and vary penalty rates).

On this basis, we believe that exploring the tax context as part of a larger effort to replicate G&R as part of a "many studies" approach is justified. However, there are several challenges to translating G&R's field experiment in the context of daycare pickups to a vignette survey in the context of taxpaying. These challenges, and our responses to them, are discussed in the next section.

B. Is Tax Compliance a Sensible Alternative Context for Replication?

In contrast to empirical studies of parents' daycare pickup choices, researchers have studied the relationship of fines to voluntary tax compliance for decades. And, while their findings have cast doubt on the consistent strength of the deterrence effect of penalties, to the best of our knowledge none have found that penalties *decrease* voluntary compliance. Lederman's recent review of the literature relating to crowding-out states: "[w]ith respect to monetary sanctions, studies sometimes find a positive effect, but they generally do not find as strong an effect of fines as they do of audit threats, or they do not get statistically significant results."[74] This raises the following question: what is the utility of replicating the G&R study in a survey vignette context if there are already dozens of papers that examine the effect of fines on tax compliance, and to date none of them have found the (negative) association that would replicate G&R?

One answer is that, to the best of our knowledge, there are no experiments that take the approach required by a translation of G&R to the tax setting. Specifically, none feature an initial condition in which there are *no* penalties whatsoever for underreporting, which is the appropriate analog for the G&R study.[75] Rather, all experimental tax compliance studies of which we are aware vary the magnitude of the fine across two or more positive values (small and large, for instance). Designing the experiment to start from a no-fine baseline is, to the best of our knowledge, novel.

A second challenge relates to the realism of the no-fine baseline. If it is the case that most self-assessment-based tax regimes *do* feature penalties, would having a no-fine vignette be "counterproductively rigid"?[76] We believe that the answer is no, for two reasons. One is that the key moment in the experiment is the introduction of a fine where none had existed before.[77] Simply increasing a pre-existing (but perhaps low) fine would obscure this moment and miss the point of the sanctions treatment. Additionally, there are real-world

[74] See Lederman (2018), at 670.
[75] One of the reasons for the absence of such studies may be that the standard neoclassical model doesn't easily accommodate a very low or zero-penalty condition. An interior (non-corner) solution for the Allignham & Sandmo model requires that the penalty rate be greater than tax rate on reported income. See Allingham & Sandmo (1972), at 324-5 (equation (3)).
[76] Irvine *et al* (2018), at 22 (going on to state: "[s]ocial, political, and legal change over time make rote replication of law and psychology experiments not just challenging but almost irrelevant….'Replication' of law and psychology demands that researchers focus on the underlying mechanism of action and work to translate those insights to a modern vernacular.")
[77] See Gneezy &Rustichini (2000), at 4 (emphasizing that "[t]here is no mention of what happens if parents come late to pick up their children. In particular, before the beginning of the study, there was no fine for coming late.")

examples of no-fine tax enforcement regimes, including the high-profile General Anti-Avoidance Rule in Canada. [78]

A third challenge relates to the structural differences between the information asymmetries of the two settings. In the daycare setting, one of G&R's interpretations of their result relates to a "partially specified contract" about which parents update their beliefs over time.[79] The daycare center/director (in an abstract sense, the regulator or social planner) has private information, *not* the parent (the individual making the compliance decision). Moreover, repeat play is required to arrive at the observed equilibrium.[80] The neoclassical model of tax compliance, by contrast, posits that the informed party is the individual taxpayer (making the compliance decision about her known quantity of income), rather than the tax agency (the regulator). The taxpayer is assumed to have complete information about all the parameters of the model, including the audit probability, penalty rate, and so on. This knowledge allows the taxpayer to optimize her reporting choices in each period; multi-period updating on the part of the tax agency is unnecessary. G&R's "partially specified contract" interpretation thus would seem inapposite to the neoclassical model of tax compliance.[81]

One response is that the neoclassical model may not include factors that are likely to be important to taxpayers' decision-making.[82] In line with theories of crowding-out, taking these factors seriously makes translating G&R's possible mechanisms more plausible, as well as more interesting. Specifically, the median taxpayer seems unlikely to have perfect information about the values of the parameters in the neoclassical model. It is more plausible that such a taxpayer is uninformed (or has poor or approximate information) about the probability of being audited. Likewise, taxpayers may be aware only vaguely of

---

[78] See Canada's federal *Income Tax Act*, subsections 245(1) through (5). The "general anti-avoidance rule" can be used as a last resort by the Crown or a provincial tax authority to challenge aggressive tax planning. Since its adoption in 1988, the federal GAAR imposes no penalties on a taxpayer who loses her GAAR case in court—the taxpayer owes unpaid taxes and interest only. Most provincial GAARs mirror this no-penalty posture. However, because of differences in the federal-provincial tax agency relationship as compared to other provinces, in 2009 Quebec responded to concerns and scandals about aggressive tax avoidance by introducing penalties into its GAAR. See Canada, Department of Finance, *Federal Administration of Provincial Taxes: New Directions*, (Ottawa: Department of Finance, 2000) ("[t] he existing TCAs [tax collection agreements, which are in place for all provinces except Quebec] place a strong emphasis on tax harmonization and simplicity for taxpayers and employers"). Our experimental setup can be seen as reflecting this policy experiment.

[79] See Gneezy &Rustichini (2000), at 10 ("[i]n the initial period, parents who are not facing a fine can refer only to a partially specified contract to anticipate any consequences of their delay. As the initial weeks go by, parents acquire some additional information. For instance, they learn that the penalty is not severe for the mild level of late arrivals that is actually taking place").

[80] *Ibid* at 11 (discussing the result as "[t]o be precise, a ''sequential equilibrium" and explaining that concept).

[81] See Emily A. Satterthwaite, "Can Audits Increase Tax Evasion?" (2016) 20 Florida Tax Review 1, 4-5 (noting that "[t]he lived experience of being audited" is not part of the basic neoclassical model; it "is merely the probabilistic manifestation of a known random variable: the audit rate….In the standard model of tax compliance, the experience of a past audit should be wholly irrelevant to a taxpayer's post-audit compliance behavior").

[82] See James Andreoni, Brian Erard and Jonathan Feinstein, "Tax Compliance" (1998) 36:2 *Journal of Economic Literature* 818, at 831-34.

the magnitude of penalties for noncompliance. Last, there is evidence that many taxpayers may be ignorant of their marginal tax rate.[83] Thinking about taxpayer information constraints more realistically creates space to translate from the daycare to the tax context.

A fourth, and perhaps most formidable, challenge in translating from daycare to tax is that the former context lacks a feature that is fundamental to the latter: audits. In daycare centers, non-compliance is observed with certainty: the daycare teachers are required to record the time that each child is picked up, each and every day. In a self-assessment income tax system, by contrast, fines cannot be imposed unless there has been an audit that determines that the taxpayer has unreported income. Typically, only an audit would allow a taxpayer's underreporting to be discovered by the tax agency. Accordingly, one of the explanations offered for the result that audits are more effective deterrents than fines is that fines are of second-order importance for taxpayers: they only matter if they are actually imposed, so it is the probability of detection prior to the fine—that is, the chance of audit—that is the more salient parameter.[84]

Given that there is no analog to audits in the daycare context, we faced the question of whether the survey vignettes should remain silent about audits, or go ahead and specify the audit policy. Arguably, mentioning any sanctions (including an audit) taints the sanction-free no-fine baseline conditions. However, participants' background beliefs about the probability of audit may not be distributed normally, and it is plausible that the background beliefs themselves might be *influenced by* the fine treatment. For instance, the absence of fines in the initial condition might lead participants to infer that the tax agency lacks resources for frequent or effective audits. Not specifying the audit rate might compound this inference. Similarly, experiencing the fine treatment might cause participants to infer that the audit rate had also increased. In the end, the risk of introducing endogeneity into the experimental design by not specifying the audit rate outweighed our concern about tainting the no-fine baseline. Thus, in all experimental conditions, the audit rate is clearly specified.

C. Design of the Tax Compliance Survey

i) Vignette Scenarios

Using vignette scenarios rather than a tax reporting simulation game in which participants' payoffs are linked to their reporting decisions is common in the tax compliance literature.[85] One of its key advantages is helping to investigate the two possible mechanisms for the

[83] See David Weisbach, "Is Knowledge of the Tax Law Socially Desirable?" (2013) 15:1 *American Law and Economics Review* 187, at 188 (stating with respect to the U.S. income tax schedule: "[m]arginal rates in the phase-out ranges are hard to compute and are not equal to the rate stated in the tax tables. Taxpayers, therefore, may not know what their marginal tax rate is").

[84] See Lederman (2018) at 627 (stating "[t]hus, a large fine may be little deterrent if there is little likelihood it will be imposed," citing Alm, Jackson & McKee (1993) as stating: "[s]ince the probability of detection is small, large responses to changes in the fine rate would require extreme degrees of risk aversion").

[85] See Olsen *et al* (2018) at 44 ("[s]cenario studies are widely used in business ethics research, as they allow assessing complex research questions in real-world environments").

"fine is a price" effect discussed by G&R.[86] Accordingly, the vignette scenarios have a common setup that mirrors, to the greatest extent possible, the structural details from the original daycare field experiment and as well as those from the daycare vignettes.

A design choice that did not present itself in the daycare setting is whether to specify details about the taxpaying jurisdiction in which the vignettes take place. Should participants assume that they are in their own legal environment (the U.S.), or should the vignette specify that they are in fictionalized jurisdiction, about which specific details are provided?[87] One advantage of designing the survey vignette using a fictional jurisdiction relates to the high salience of penalties for income tax evasion in many jurisdictions (including the U.S.). The idea of a tax law in their own jurisdiction without penalties may strike participants as odd or unrealistic, and they may find it difficult to envision themselves making compliance decisions as instructed by the vignette. Participants would likely face less difficulty envisioning their behaviour under specified conditions when the vignette describes then as being in a fictionalized country. Use of a fictionalized country may make the exercise less realistic, potentially compromising external validity. However, in our replication setting, the importance of the no-fine baseline and maintaining this structural consistency led us to place less weight on any drawback the fictional country might have for external validity. Following the naming convention used in prior experimental research, we situate all participants as paying taxes in a fictional country named Varosia.

a) Initial Conditions / No-fine Baseline

In all of our tax study experimental conditions, participants were exposed to the same initial conditions, in which there is no fine for underreporting. They were told that in Varosia there is a legal obligation to pay tax at a flat rate of 30 percent on all income, some of which is reported by the payer to the government. Participants were informed that the tax agency does not check all returns, but that each taxpayer has a 10 percent chance of being audited, and an audit will always find the full extent of unreported income. However, in the initial conditions, no fine was mentioned. Further, the participant was told that she has earned $1,000 of "extra" income, such as from consulting work on the side in addition to her wages, and this income was not reported by the payer to the tax agency. She was then

[86]There is substantial discussion in the literature about the reliability of surveys of taxpayers about tax compliance, but being able to ask "why" questions and linking responses to detailed demographic data is identified as a key advantage. See Andreoni *et al.* (1998), at 836 ("Surveys provide an alternative source of information about noncompliance. The main advantage of survey data is that they often include many socioeconomic, demographic, and attitudinal variables that are not available with tax return and audit data, allowing researchers to investigate a rich set of hypotheses about the factors associated with noncompliance. The major disadvantage of survey data is that they are based on self-reports, which often provide very inaccurate information. In general, survey results substantially overstate the degree of compliance"); Henk Elffers, Russell Weigel and Dick J. Hessing, "The Consequences of Different Strategies for Measuring Tax Evasion Behavior" (1987) 8(3) *J. Economic Psychology*, 311 (linking tax audit results with survey responses for several hundred Dutch taxpayers; finding little to no correlation between assessed evasion and evasion reported on the survey).

[87] See Olsen *et al.* (2018), at 44 ("…Following Kogler *et al.* (2013) and Wahl *et al.* (2010), we used scenarios that described the tax system of a fictitious country named Varosia"). See Metcalf *et al.* "Is a Fine Still a Price? Replication as Robustness in Empirical Legal Studies – Data and Supporting Materials, 2019" https://doi.org/10.5683/SP2/9ZNNS4 for the complete tax study survey.

asked to answer some questions about her reporting decision that were designed to be analogous to the daycare outcome questions about frequency and extent of lateness:

- How likely are you to report all of your extra income? (7-point qualitative, coded from 1 ("extremely likely") to 7 ("extremely unlikely"))
- How much of your extra income do you think you would report on your tax return? (7 point qualitative scale coded as 1 ("All") to 7 ("none")?
- What is the amount of your extra income you think you would report? (slider that allows participants to specify a dollar amount from $0 to $1,000)

As in our daycare study, we included a subsequent set of questions focused on participants' reasons for their choices in the questions above, using the same scale and structure as in the daycare study.[88] To distinguish between fines functioning to resolve incomplete information about enforcement vs. fines crowding out social norms around paying taxes we used two main questions. Respondents indicated the importance to their decision of: a) possible consequences with the Tax Agency, b) "feeling bad" about not reporting when other pay their full share.[89] Participants were also asked whether they agreed (scale from 1 "strongly agree" to 7 "strongly disagree") that they would "feel guilty or ashamed about underreporting".

Finally, we again check to see how our treatments affect respondents' expectations about other taxpayers' behavior. Respondents were asked to estimate "the percentage of other taxpayers in Varosia who have extra income like yours (i.e., not reported by the payer to the Varosia Tax Agency) who will not report it fully" (slider for percentage from 0 to 100). They were also asked to specify the amount of income underreported for the sub-group of those not reporting fully using the same qualitative scale as in their own earlier reporting amount question. These responses established our baseline observations that correspond to G&R's observations prior to the introduction of the fine.

b) Experimental Conditions

<u>G&R – Tax Replication Fine Condition</u>

With the baseline established, participants who were randomly assigned to this experimental condition advanced to a screen with the heading "The Fine Announcement" in which the "G&R Fine" treatment was introduced. Here, we sought to translate G&R's fine for lateness into an equivalent flat, relatively low fine for underreporting any income through the following language: "Effective immediately, a fine of $200 will be charged every time that an audit finds that you did not report all your income. In addition to the fine, you will also pay the tax owed on the unreported income." All other aspects of the

[88] "How important are each of the reasons below to your decision to report your extra income?" (11-point scale from 0 "not at all important" to 10 "extremely important").

[89] We also ask an additional question that focuses on respondents' self-image and possible guilt about non-compliance, rather than concern about others being imposed on, by allowing them to indicate the importance of "Not reporting fully would make me feel bad about myself."

setup were the same, and participants were told that they will "continue to have a 10 percent chance of being audited."

Participants then revisited the questions above on the likelihood, proportion and amount of underreporting, and their reasons, feelings, and beliefs about underreporting.[90]

Next, participants were exposed to the secondary treatment canceling the fine. They advanced to a screen with the heading "Another Change" and were told that after the fine having been in place for one tax reporting year, the fine was "cancelled, effective immediately." They were asked to revisit their answers to the above set of questions. Their responses serve as the analog to the observations in the daycare field study following removal of the fine.

Following these questions but before concluding, participants answered a set of "tax demographic questions" that provided data on their tax behaviour for potential controls.[91]

Robustness: An Alternate Tax Fine Condition

To assess the potential sensitivity of any "fine is a price" effect in the context of tax-reporting, we again included an experimental condition (Alt F) featuring an alternate fine. As in the daycare study, we chose a fine that is more realistic in scaling with the amount of unreported income, but that remains somewhat similar in magnitude to the G&R fine at low levels of under-reporting. Again, we sought to avoid a fine so large that it would become expropriative and thus likely to be a self-evident deterrent. We thus set the alternate fine at $200 plus 20 percent of unreported income.[92]

Participants who were randomly assigned to this experimental condition advanced (after the initial conditions) to a screen with the heading "The Fine Announcement" in which the Alt F tax fine treatment was introduced. Participants were exposed to a screen identical to that of the G&R fine treatment above, but with the following language: "Effective immediately, a fine of $200 plus 20% of any underreported income will be charged every time that an audit finds that you did not report all your income." Participants revisited their answers to the questions, and then advanced to the secondary treatment in which the fine was cancelled. They answered the questions a final time before providing responses to the tax demographic questions.

Robustness: Influence of Social Norms

[90] The same questions that followed the initial conditions are asked again, verbatim, but with the addition of "reminder" language. In this case, each question was prefaced with the phrase, "Now that the fine has been announced,…"

[91] E.g. whether they had filed tax returns, had ever underreported their income, had been audited in the past, or had been required to pay a penalty to the tax authority.

[92] In the U.S., many penalty rates are set statutorily at 20 percent of unreported income. The expected maximum value of this additional fine component for our participants (earning $1000 with a 10% audit rate) was the same as our flat fine example.

As in our daycare replication, we tested an additional "Social Norms" condition designed to assess directly whether social motivations for taxpayer compliance are operating in our survey setting, and to compare them with the impact of a fine. This allowed us to directly investigate the potential "crowding out" of social norms by fines in a tax compliance setting.

Following the structure of our daycare vignette, for participants randomly assigned to this experimental condition, following exposure to the initial conditions, we introduced a simple "Social Reminder" treatment designed to appeal to social reasons for tax compliance. It stated "…Since some people have not been reporting all of their income, the Varosia Tax Agency has decided to remind everyone to report all of their income. When you underreport, the Varosian government needs to rely more on those who fully pay their share so it can provide public services! Please be considerate of your fellow Varosians and report all of your income." Participants then answered the set of questions above.

Next, similar to the daycare vignette, we introduced a secondary Social Norms treatment involving public disclosure of non-compliance – this time through an "Underreporting List:"[93] It stated: "Effective immediately, every time that an audit finds that you underreported your income, the name of the taxpayer will be published in a list in a full-page advertisement in the Varosian Daily News, both in print and online. In addition to being listed on the Underreporting List, you will remain liable for tax owed on any unreported income. This list will be compiled annually, and it will be published in advance of the next year's tax return filing deadline."

Participants revisited their answers to the questions, and then provided answers to the tax demographic questions.

Administration of Survey

As with our daycare study, we administer our tax vignettes via an online Qualtrics survey hosted on MTurk. We use similarly structured language to recruit MTurk workers to our tax HIT as we used in the daycare survey. The HIT provides workers with the basic information that they are participating in a study regarding tax compliance behaviour that will ask them to imagine how they would respond. We use the same approach as in the daycare study with regard to pay, worker qualifications, prevention of duplicate surveys,

[93] This experimental condition is likely more realistic in the tax context than it is in the daycare context. Numerous jurisdictions have experimented with "tax shaming." See, e.g., James Alm et al., "Culture, Compliance and Confidentiality: A Study of Taxpayer Behavior in the United States and Italy" University Ca' Foscari of Venice, Dept. of Economics Research Paper Series No. 36, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2888276; Ricardo Perez-Truglia Ugo Troiano, "Shaming Tax Delinquents: Theory and Evidence from a Field Experiment in the United States", NBER Working Paper 21264 http://www.nber.org/papers/w21264 (June 2015); Giorgio Coricelli et al., "Cheating, emotions, and rationality: an experiment on tax evasion" (2010) 13 *Exp Econ* 226, DOI 10.1007/s10683-010-9237-5.

attention checks, etc. Our sample for the tax study includes 648 participants randomized evenly across our three experimental conditions.[94]

D. Data & Results in the Tax Compliance Setting

i) Is a Fine a Price that Incentivizes Tax Cheating?

We first consider the effects of our fine conditions on our key outcome variables for tax reporting compliance. In line with G&R's study and our daycare study, we are looking to see if the introduction of fines against our no-fine baseline produces effects in the tax setting that match the G&R replication hypotheses:

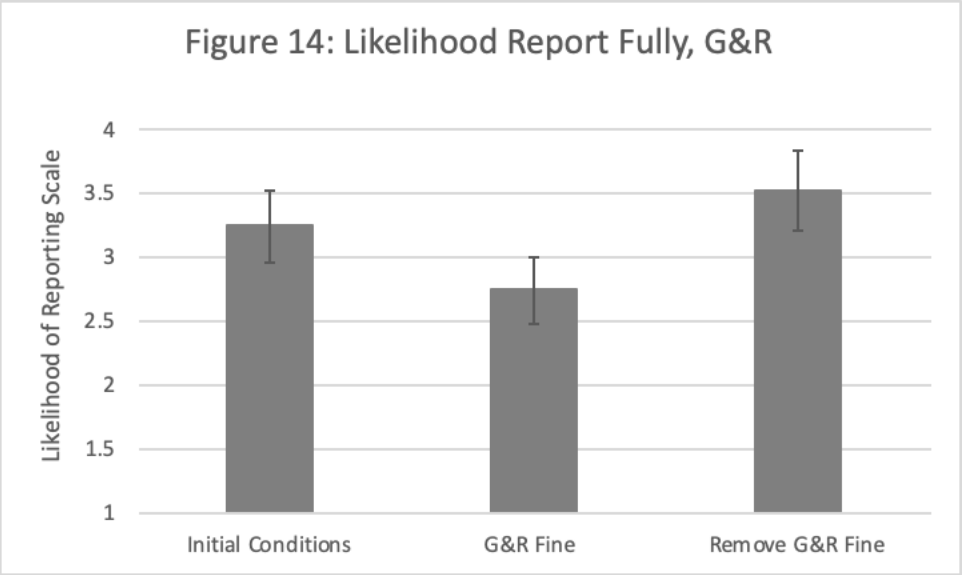
> **H_0**: The introduction of a small fine against the no fine baseline should *increase* underreporting of income in the tax setting.
>
> **H_1**: The effect of introducing the fine should be sticky. Once respondents' perceptions of the context have changed, the removal of the fine should not have an effect.

We use the same two main approaches to test for effects of our treatments: changes in means as average treatment effects and changes in the full distribution of responses as alternate treatment effects.

a) Equivalence of Means as Treatment Effects

We again begin by examining the within-subjects evaluation average treatment effects for both introduction and removal of the fine, looking first at the G&R fine condition and then the Alt F fine condition as a robustness check. We consider the effects on our key tax compliance outcome measures: how likely respondents would be to fully report their income, and how much of their income they would report.

[94] Additional details for our tax study, including summary statistics for participant characteristics, survey, and data are available for researchers in our supplementary materials, see Metcalf *et al* (2019) https://doi.org/10.5683/SP2/9ZNNS4.

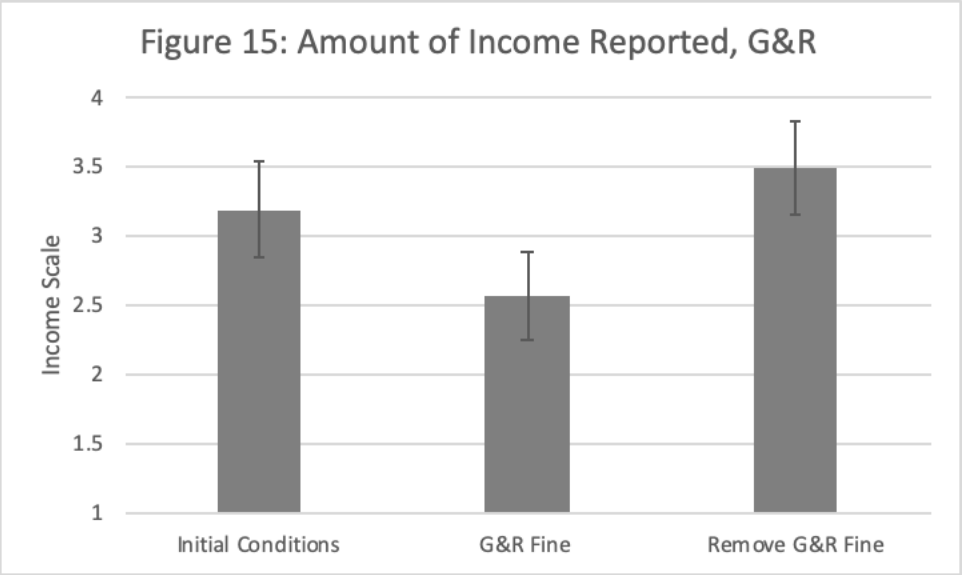


Notes to Figure 14: Likelihood report income fully measured by 7-point qualitative scale, 1 = "extremely likely" to 7= "extremely unlikely". Initial Conditions is the mean likelihood of fully reporting prior to any fine treatments. G&R Fine is mean after introduction of the G&R replication tax fine treatment. Remove G&R Fine is the mean after the G&R tax fine is cancelled. Error bars are 95% confidence intervals.

As can be seen in Figure 14, and similar to our daycare study results, introduction of the G&R tax fine produced a lower mean for our income reporting compliance measure. With our responses scaled so "1" is assigned to the category for "extremely likely" to fully report income, this drop in the mean implies respondents increase their tax compliance when the fine is introduced. On average they are *less* likely to underreport income with the fine. Mirroring our daycare study results, when the fine is removed, underreporting changes again, rebounding to a level that is actually above the initial conditions mean.[95] The effects of both introducing and removing the fine on income reporting compliance are highly statistically significant.[96]

[95] The initial and final mean compliance measures are not statistically distinguishable: H_0 $\mu_{IC}=\mu_{FRemov}$, $F_{(1,647)} = 1.80$ ($P< 0.1804$).

[96] The tests for equivalence of means are as follows for Probability of Reporting: H_0 $\mu_{IC}=\mu_{G\&R}$, $F_{(1,647)} = 6.59$ ($P< 0.0105$); H_0 $\mu_{G\&R}=\mu_{FRemov}$, $F_{(1,647)} = 14.15$ ($P< 0.0002$).



Notes to Figure 15: Amount of extra income reported as measured by 7-point qualitative scale, 1 = "all" to 7= "none". Initial Conditions is the mean response for amount of income reported prior to any fine treatments. G&R Fine is mean income reported once G&R replication tax fine introduced. Remove G&R Fine is the mean after the fine is cancelled. Error bars are 95% confidence intervals.

We observe a similar pattern in our compliance measure of the amount of income respondents report. Our scale codes with 1=all extra income reported, so once again the drop in the mean reflects an *increase* in tax reporting (i.e., compliance) with the G&R tax fine in place. Once again, when the fine is removed, respondents' reporting relapses back toward the no-fine baseline, with the final mean at a lower compliance level. The transitions with introduction and removal of the fine are highly significant, while the initial and final compliance means are indistinguishable.[97] When we examine the alternative measure for income reporting (dollar amount respondents would report), we find a similar pattern. The mean amount increases with introduction of the G&R fine from $617 to $683 (P<.0891). When the fine is removed, it falls back to $550 (P<.0007). This leaves the average amount reported at a marginally significant lower level than initial conditions (P<.0959).

The means for these key outcome variables in our Alt F tax fine condition tell the same story, but generally with increased significance levels. Introduction of the Alt F underreporting fine produces a drop in the mean for likelihood of fully reporting income (P<0.0000) which means *increasing* compliance with our scale, while removal of the Alt F tax fine creates a rebound in the mean (P<0.0000) that again leaves it statistically indistinguishable from the initial conditions (P<.5816).[98] The pattern repeats in the Alt F condition for the amount of income reported. When the underreporting fine is introduced, the drop in the mean for income reporting category indicates *more* income reported and is

[97] The tests for equivalence of means are as follows for amount of income reported: H0 μIC = μG&R, F(1,647) = 4.67 (P < 0.0311); H0 μG&R = μFRemov, F(1,647) = 12.12 (P < 0.0005); initial and final mean for amount of income reported: H_0 $\mu_{IC} = \mu_{FRemov}$, $F_{(1,647)} = 1.60$ (P < 0.2067).

[98] The means for Likelihood of Fully Reporting Income in the alternate fine condition are: 3.44, 2.46, and 3.55 respectively for initial conditions, introduction of the Alternate Fine, and Alternate Fine removal.

highly significant (P<0.0000) and the rebound when the fine is removed equally so (P<0.0000). Initial and final means are indistinguishable (P<0.5919).[99] Our alternative measure for the dollar amount of income reported is consistent. The mean amount increases with introduction of the Alt F tax fine from $576 to $744 (P<.0000). When the fine is removed, it falls to $571 (P<.0000). In the Alt F condition, the initial and final mean dollar amounts reported are indistinguishable (P<.8850).

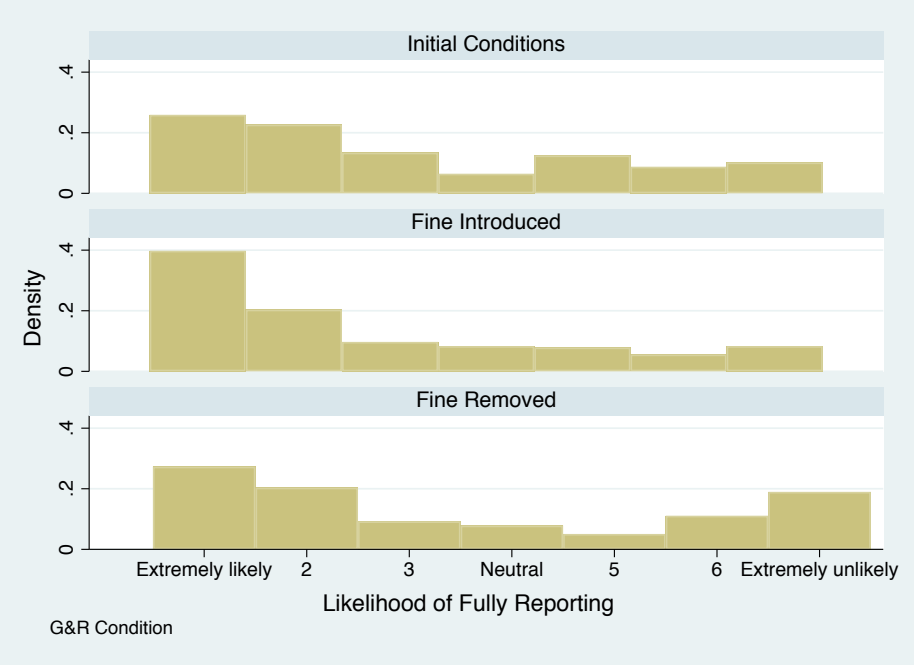
As in our daycare study, the comparisons of means for our key income reporting measures fail to produce results consistent with **H0** and **H1** from Gneezy and Rustichini. We continue to obtain similar, consistent results from both fine conditions that are also consistent with those in our daycare study. Introduction of a fine causes respondents to *decrease* their anticipated non-compliance behavior. The change is *not* sticky – removal of the fines leads to highly significant rebounds in our respondents' anticipated tax underreporting. The changes in tax compliance are quantitatively larger than for outcomes in the daycare setting. They reflect movements around the mean of our compliance scales themselves, rather than the small changes around highly compliant behavior we saw in the daycare setting. However, if the fine is a price for tax cheating, in our vignettes it has the standard anticipated effect of reducing behaviour when the costs increase, even by a relatively small amount.

b) Distributions of Responses as Treatment Effects

Once again in our tax context, we examine patterns in the full distribution of responses as evidence for treatment effects of our fines. Recall that **H0** implies that we should see movement into the higher scale values of the distribution of our tax compliance measures when fines are introduced. **H1** implies that this change should be sticky, the distribution of responses should not change when the fine is removed. We examine the distributions for Likelihood of Fully Reporting Income and Amount of Income reported for both our qualitative and dollar measures, but do not present them all visually as they are extremely similar. We check all our distributional results for both the G&R and Alt F tax fines, again using Kolmogorov-Smirnov tests to determine whether there are any statistically significant differences in the distributions as treatments are varied within each condition.

Figure 16: Distributions of Responses - Income Reporting, G&R Condition

[99] The respective means for the amount of income reported in Alt F condition are: 3.4, 2.42, and 3.55 for IC, AF and AF Removed.



Notes to Figure 16: Vertical axis is density measure of response distribution, Horizontal axis is likelihood of fully reporting extra income, qualitative scale from 1="extremely likely" to 7="extremely unlikely". Initial Conditions panel is distribution prior to introduction of any tax fine treatments, Fine Introduced is distribution for likelihood of reporting extra income after G&R replication tax fine introduced, Fine Removed panel is distribution after the fine is cancelled.

The general pattern for changes in distributions of our tax compliance outcomes can be seen in Figure 16 above. It shows the distributions for Likelihood of Fully Reporting Income across treatments in the G&R condition. The inconsistency with **H0** and **H1** from G&R is apparent. Introduction of the fine causes responses to move into lower value categories, which corresponds with *reduced* tax cheating. When the fine is removed, responses shift back into higher valued categories (increasing tax underreporting) and become indistinguishable from the initial distribution. The effect of the fine lasts only as long as it is in place. No sticky holdover effect appears in the distribution of responses. As in our Alt F fine treatment in the daycare study, it appears that there may even be a shift into stronger non-compliance than in the no-fine baseline for some respondents with *removal* of the fine. All our main distributional results are strongly statistically significant.[100]

We obtain similar results when we examine the distribution of responses for the amount of income reported. The effect of introducing the fine is to increase the amount of income respondents report by shifting respondents into lower values of our categorical measures (1=report all income). Once the fine is removed, responses shift back toward higher values and less income reported. The changes, again inconsistent with both **H0** and **H1** from G&R,

[100] The Kolmogorov-Smirnov test statistic values for G& R Late Frequency distributions are as follows: **H0**: values in initial conditions (IC) < values in fine introduction (FI) D=.0000 (P<1.000); FI<IC D=-.1389 (P<0.016); **H1**: values in FI<values when fine removed (FR) D=0.1620 (P<0.003); values FR<FI D=-.0000 (P<1.000). The initial and final distributions are indistinguishable: Combined K-S D=0.1111 (P<0.139).

are strongly significant.[101] Introduction of the G&R tax fine in our dollar amount reporting compliance measure also shifts the values upward, but not statistically significantly (K-S P<.157). Removal of the fine does cause a statistically significant drop in the distribution of dollar amounts reported into lower values (K-S P<0.001). The initial and final distributions are also indistinguishable for this income reporting compliance outcome measure (K-S P<.139).

The distributional results for our Alt F fine condition mirror those above for all income reporting compliance measures discussed for the G&R condition, with higher statistical significance. The outcome measure distributions change (statistically significantly), but not in ways consistent with the G&R replication hypotheses **H0** and **H1**.

As in our daycare vignette study, we reliably find statistically significant effects from our fine treatments in both our fine conditions. However, when we introduce fines, effects are in the wrong direction for consistency with G&R's results and hypotheses. Moreover, the effects from the fines are only apparent when the fines are in place; there is no evidence of the persistence beyond that suggested by G&R.

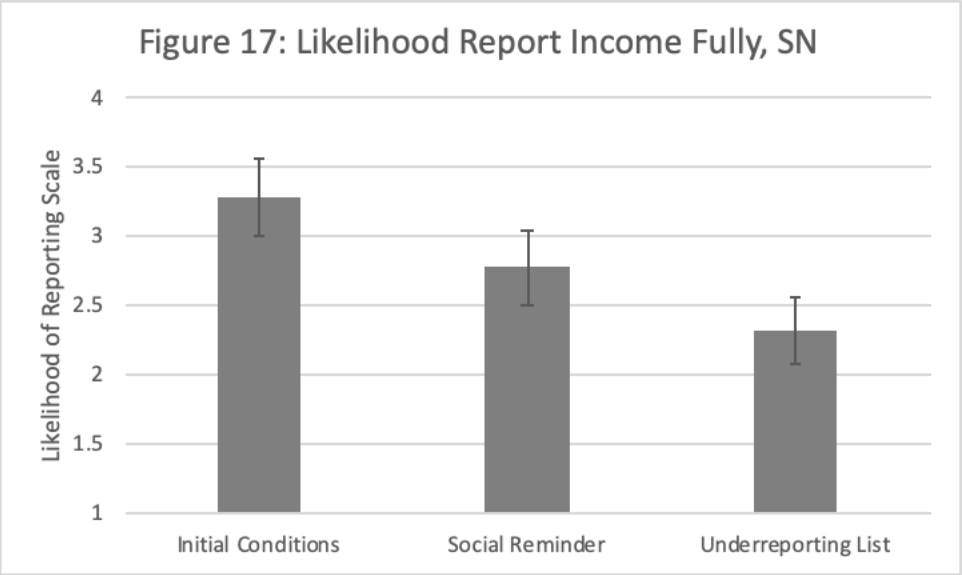
ii) Do MTurkers Respond to Social Norms around Tax Compliance?

As discussed above, particularly given the structure of the traditional Allingham & Sandmo model of tax compliance, the existence of background social norms that generate tax morale for voluntary compliance is critical for the kind of crowding out effect Gneezy and Rustichini suggest. We check here to see if our MTurk workers are receptive to social norms around tax compliance. As in the daycare study, we use our within-subjects tax reporting results for our Social Norm Condition as a test.

a) Equivalence of Means as Treatment Effects - Tax Norms

As seen in Figure 17 below, when the Social Reminder treatment is introduced, it produces a significant drop in mean compliance category – increasing the likelihood of fully reporting (P<.0085). Introduction of the Underreporting List treatment – which publicizes non-compliance when discovered – causes a further drop in the mean. With our scale, this again implies an increased likelihood on average that respondents report income fully in this tax shaming treatment (P<0.0111). There is a highly significant difference between initial and final mean compliance when our social norm-based treatments are introduced (P<0.000).

[101] The Kolmogorov-Smirnov test statistic values for G&R Qualitative Amount of Income Reported distributions are as follows: **H0**: IC<FI D=.0000 (P<1.000); FI<IC D=-.1157 (P<0.055); **H1**: FI<FR D=0.2037 (P<0.000); values FR<FI D=-.0000 (P<1.000). The initial and final distributions are indistinguishable under the combined K-S D=0.0880 (P<0.374).



Notes to Figure 14: Likelihood report income fully measured by 7-point qualitative scale, 1 = "extremely likely" to 7= "extremely unlikely". Initial Conditions is the mean prior to any social norm-based treatments. Social Reminder is mean likelihood of reporting income after social tax reminder treatment. Underreporting List is mean once publication of Underreporting List is announced. Error bars are 95% confidence intervals.

A similar pattern appears in our metrics for the amount of income reported.[102] The Social Reminder increases the mean amount reported relative to initial conditions ($P<0.0116$). The Underreporting List treatment further enhances compliance by continuing to increase the amount respondents choose to report ($P<0.0458$). Again, the cumulative impact of our social norm-based tax treatments is to enhance mean reporting compliance in a highly significant way relative to participants' baseline responses in the initial conditions ($P<0.000$). The dollar amount income reporting measure behaves similarly. The reminder causes mean income reported to increase from \$589 to \$672 ($P<0.0318$). The introduction of the Underreporting List causes a further increase in mean income reported to \$765 ($P<0.0087$). This is a quantitatively large and highly significant change from the initial conditions ($P<0.0000$).

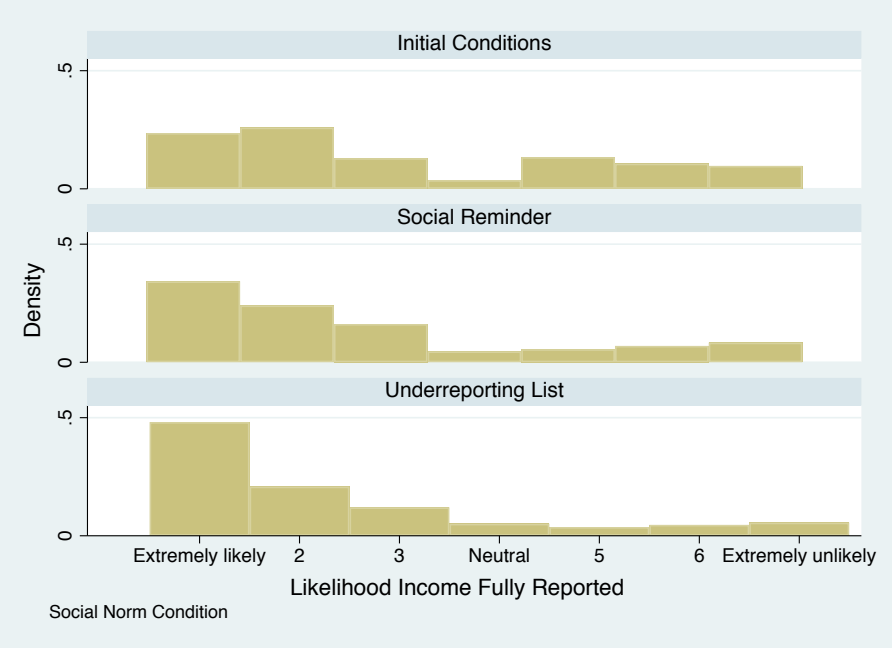
b) Distribution of Responses as Treatment Effects – Tax Norms

We again use the full distribution of responses to our tax compliance measures to confirm that our MTurk respondents are sensitive to social norm-based influences on their tax reporting behaviour. Distributions for our measure of the Likelihood Income is Fully Reported across treatments in the Social Norm Condition is shown below in Figure 18.

One thing that is immediately apparent from the Initial Conditions plot is that there is no predominant latent norm of full compliance among our MTurkers in the tax setting. In contrast with the highly compliant responses in our daycare study, the initial income reporting distribution is pushed into two fat tails of those inclined to comply vs. those who will not.

[102] We do not present these visually due to the strong similarity.

Figure 18: Likelihood Income Reported Fully, Social Norm Condition



Notes to Figure 18: Vertical axis is density measure of response distribution, Horizontal axis is likelihood of fully reporting extra income, qualitative scale from 1="extremely likely" to 7="extremely unlikely". Initial Conditions panel is distribution prior to introduction of any social norms treatments, Social Reminder is distribution for likelihood of reporting extra income after the social tax reminder is introduced, Underreporting List panel shows distribution after announcement of the public underreporting list.

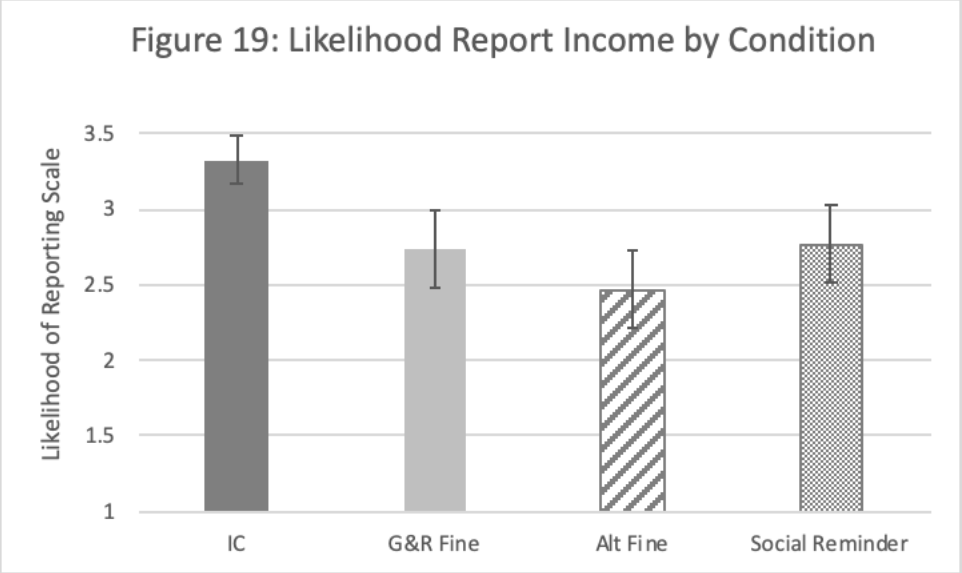
Despite this lack of a strong latent compliance norm compared with our daycare setting, respondents remain sensitive to the social norm treatments. Our Social (Tax) Reminder stresses that underreporting will force government to rely more on those who fully pay their share and make it harder for Varosia to provide public services. This reminder has a significant effect, shifting respondents out of the non-compliant categories for reporting income fully (K-S P<.027). Introduction of the Underreporting List, publishing names of those caught underreporting their income in a national news outlet, causes a further reduction in non-compliant income reporting. Responses shift significantly into higher probability compliance categories (K-S P<0.016).

The effect on our compliance measures for the amount of income respondents report is similar. The Tax Reminder shifts the distribution into lower value income reporting categories, consistent with reporting more income (K-S P<0.055). The Underreporting list produces a further shift into lower value categories, enhancing compliant reporting of more of total income (K-S P<0.006). Our dollar reporting distributions behave similarly, with the reminder shifting amounts into higher values of income reported (K-S P<0.069), and the Underreporting List further increasing values in the distribution of dollar amounts reported (K-S P<0.016).

Our Social Norms condition reassures us that MTurkers remain sensitive to social norms around tax compliance behaviour in the way that G&R and others tax morale scholars suggest. A simple reminder produces statistically significant enhancements in tax compliance behaviour in our sample.[103] Again, this reassures us that our MTurk workers are not a fundamentally inappropriate sample to test G&R's replication hypotheses in a tax reporting context.

iii) Relative Effect of Fines vs. Appeals to Tax Morale / Social Norms

G&R's theory and the scholarship on how tax enforcement may crowd out voluntary compliance suggest that fines and social norms are at odds in incentivizing tax enforcement. The introduction of fines may undermine voluntary compliance by crowding out pro-reporting social norms. If this account is operative in our vignette scenario, we should see our fine conditions and social norm conditions working in opposite directions. The introduction of (small) fines should increase noncompliant reporting relative to the baseline, while reinforcing the social obligations in the tax context should improve compliance. As in our daycare study, we use between-subjects comparison of means below to test the relative impact of fines vs. social norms in our tax vignettes.[104]



Notes to Figure 19: IC is the mean likelihood of fully reporting extra income for all respondents in the initial conditions, before introduction of any treatment imposing consequences for underreporting. Scale ranges from 1= "extremely likely" to 7= "extremely unlikely". G&R Fine is mean likelihood of reporting fully for respondents exposed to the G&R tax fine. Alt Fine is mean likelihood of reporting fully for respondents exposed to the Alt F tax fine. Social Reminder is mean for respondents exposed to the tax reminder treatment. Error bars are 95% confidence intervals

[103] This is consistent with results from real world tax reminder experiments, see e.g. Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. (2014) "The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance", NBER Research Working Paper 20007.

[104] The mean for Initial Conditions is pooled across conditions for this comparison. In this section we focus only on our main treatment, introduction of our experimental treatments (fine, reminder) rather than secondary removal. We have already established with our within subjects results that the stickiness G&R expected is not replicated.

A glance at Figure 19 above shows that the average compliance category is significantly smaller in all experimental conditions relative to the initial conditions. As in our daycare study, these movements *enhance* income reporting compliance (1=report fully). There is no evidence that introducing fines works to increase underreporting relative to social norm-based controls on taxpayer behavior. The fines are not statistically distinguishable from each other in their effects. There is also no significant difference between the G&R fine and the effect of the Tax Reminder (P<.8614). There is a marginally significant difference between the Alt F tax fine and the Tax Reminder (P<0.0978). However, the main conclusion is that we do not see evidence that fines and social controls work in opposite directions in our tax vignette setting.

The results for the amount of income reported across treatments replicates the pattern above. All experimental treatments work similarly to *increase* the mean compliance in the category amount of income reported with high statistical significance. There is no statistically significant difference between the experimental conditions themselves – all produce indistinguishable effects on the categorical measure of income reported.[105] Results for the dollar amount of income reported across conditions are similar. All conditions produce significantly higher mean dollar values for income reported than initial conditions; fines and the tax reminder continue to work in the same direction. There is a marginally significantly higher mean income reported under the Alt F tax fine relative to both the G&R fine (P<.0945) and the Tax Reminder (P<0.0474). There is no difference between the G&R Fine and the Tax Reminder in terms of mean income reported (P<.7545).[106]

As in our daycare study, our between-subjects review of the treatment effects across conditions confirms that our respondents behave differently than the parents in Gneezy and Rustichini's field trial. Once again, this is not because our MTurkers are insensitive to social norms around tax compliance. In our tax study, fines and social controls work in the same direction for our respondents, increasing the cost of non-compliant tax reporting.

iv) How do Fines Change Perceptions in the Tax Reporting Context?

The final component of our tax study returns to trying to find evidence for which (if any) of G&R's explanations for their results may provide a basis for our survey respondents' behaviour. We translate G&R's two alternate hypotheses for our tax context below:

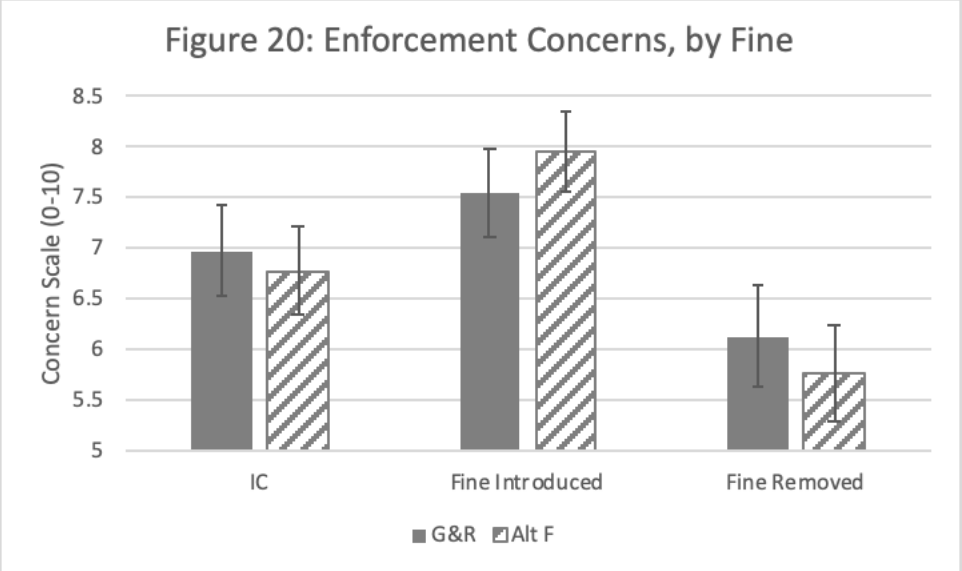
> **H2**: The introduction of a (small) fine will decrease fears of severe consequences from the Tax Agency in its enforcement efforts. This effect should be sticky.
>
> **H3**: The introduction of a (small) fine will reduce social concerns about underreporting income, by crowding out social norms / tax morale sustained through voluntary compliance with a financial signal of the value of tax compliance. This effect should be sticky.

[105] Using F-tests: G&R=Alt F (P<.2182), G&R=SN (P<.9826) and Alt F=SN (P<.2265).
[106] For comparison, the mean amounts of income reported are: $594 (IC); $683 (G&R); $744 (Alt F); $672 (Tx Reminder).

In terms of our measures, **H2** implies we should see *reduced* concern about possible enforcement consequences with the Tax Agency when the fines are introduced. If **H3** is true we should see *decreases* in respondents' concerns that underreporting will leave a higher burden on other taxpayers in order to fund public services. As in our daycare vignette, if G&R's hypotheses hold, respondents should adjust their beliefs about others when our fines are introduced to anticipate *increased* non-compliance in income reporting (**H4**).

a) Fines & Tax Authority Enforcement Fears

We explore the impact of our fine conditions on our measure of respondent's concerns about consequences with the tax enforcement authority in Figure 20 below. With our scale (10=extremely concerned), higher values reflect greater concern.



Notes to Figure 20: IC is mean level of concern over tax authority enforcement for respondents in fine treatments in the initial conditions, prior to introduction of any underreporting fine. Scaled from 0 to 10 (extremely concerned); means illustrated from mid-point. Fine Introduced is mean enforcement concern after introduction of the G&R and Alt F tax fine treatments respectively. Fine removed is mean level following announcement cancelling the underreporting fines. Error bars are 95% confidence intervals.

The results in our tax study are broadly consistent with those in our daycare study – but reveal more sensitivity in respondents' fears about enforcement consequences than we saw in respondents' contractual concerns re the daycare. Introduction of the tax fines produces significant increases in mean concern about enforcement consequences (P<.0669, G&R; P<.0001, Alt F). When the fines are removed, another statistically significant change is produced, with a substantial drop in mean enforcement concern (P<.0000, G&R; P<.0000, Alt F). This leaves the final mean level of enforcement concern significantly below the initial conditions (P<.0108, G&R; P<.0022, Alt F).

Although we choose fines that were not particularly severe and that had a low expected value when combined with the explicit audit rate of 10%, they appear to have a fairly significant effect on our participants' fears about tax enforcement. The fines do *not* assuage concerns by providing certainty and ruling out latent extreme fears (e.g. imprisonment in a tax setting). As in our daycare study, the effect of the underreporting fine lasts only while it is in place. Removal of the fine appears to signal leniency to our respondents and causes them to substantially revise and reduce their fears of Tax Agency enforcement consequences.

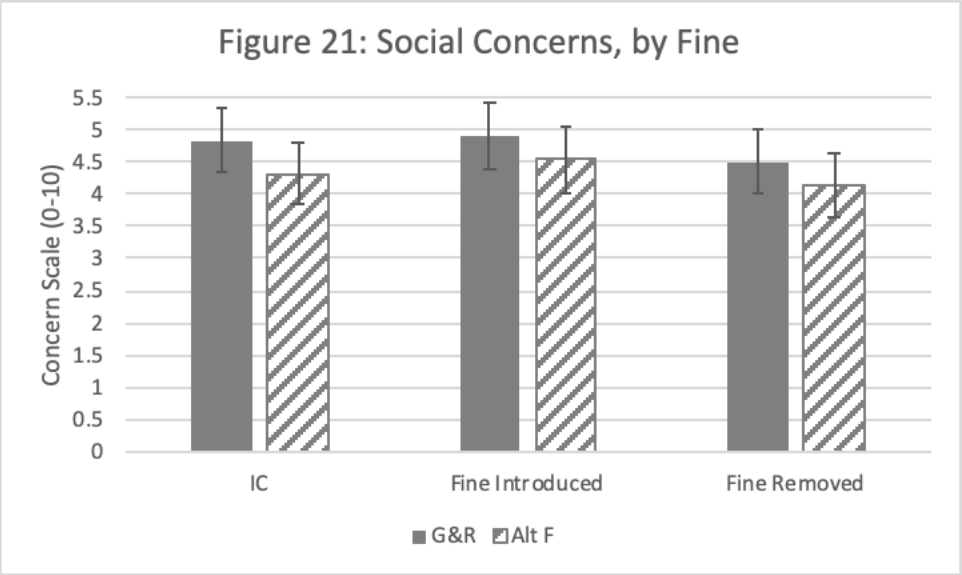
As in the daycare setting, our tax study results fail to match G&R's hypothesis, $\mathbf{H_2}$, in the direction and persistence of effects when the fines are introduced, then removed.

We confirm these effects of our fines on enforcement concern by examining the distributions of responses in both fine conditions. The results are consistent with the main average treatment effects discussed above. For the G&R condition, introduction of the underreporting fine increases the values in the distribution of responses relative to IC (K-S $P<0.086$). Removal of the fine shifts the distribution into lower values for enforcement concern (K-S $P< .0000$), which leaves the final distribution with significantly lower values than the IC (K-S $P< 0.069$). The pattern for changes in the distribution for the Alt F tax fine is the same, but with higher significance levels.

Overall, our results from the tax study provide support for interaction effects between the use of fines and the enforcement concerns of taxpayers. However, they are not consistent with a translated incomplete contracts / information account similar to that suggested by G&R to explain their field trial results. The effects run the "wrong" direction, with even our relatively small fines increasing enforcement fears that potentially contribute to the deterrent effect we see in our tax compliance outcome measures. G&R's semi-permanent signalling of enforcer type by the introduction of a (small) fine against a no-fine baseline alternative is also not consistent with our tax study results. Instead, our respondents seem to treat all changes in the tax fine policy as signals about enforcement.

b) Fines & Social Norms in the Tax Setting

Figure 21 below presents our means across treatments for social reasons in our two tax fine conditions (G&R, Alt F). Respondents indicated how important it was to their decision that they would "feel bad" about not reporting fully when others paid their full share (10 = extremely important). Do fines undermine tax morale and the social incentives to comply by reducing this concern?



Notes to Figure 21: IC is mean level of social concern around underreporting in the initial conditions, prior to introduction of any underreporting fine treatments. Scaled from 0-10 (extremely concerned). Fine Introduced illustrates the mean level of social concern following introduction of the G&R and Alt F tax fine treatments respectively. Fine removed illustrates the mean after underreporting fines cabncelled. Error bars are 95% confidence intervals.

In contrast to our results in the daycare study, in our tax study we see no support for prediction, **H3**, that fines may crowd out social concerns and damage tax morale. In both the G&R and Alt Fine conditions, there is no statistically significant effect from either the introduction or the removal of the tax fines. The level of concern expressed because others are paying while the respondent "free rides" does not vary systematically with the different tax fine treatments. Interestingly, the average value of social concerns related to the tax compliance decision is amongst the lowest for any of our measures. This suggests weak evidence of "tax morale" among our respondents as a motivation for voluntary compliance.

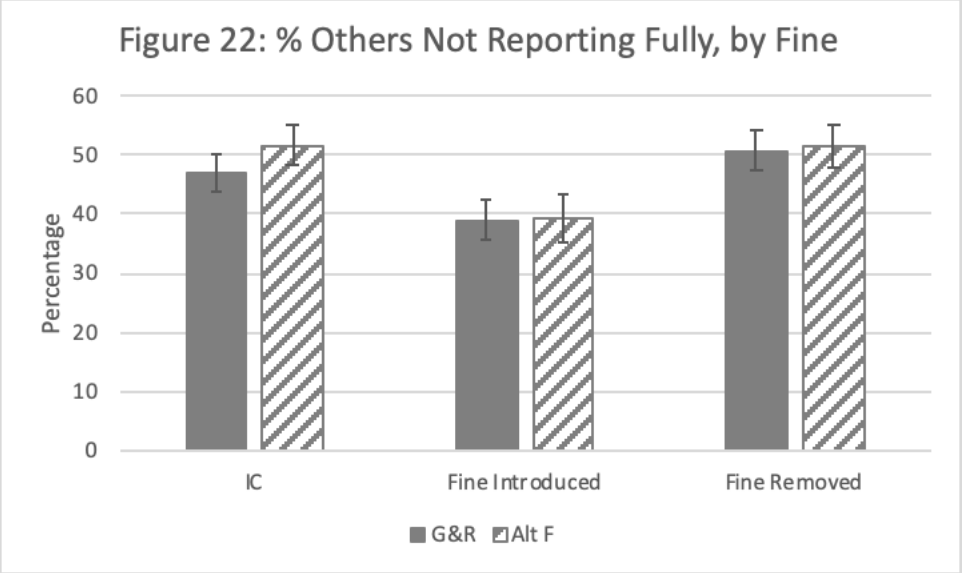
Again, when we examine the full distribution of responses as a check on these effects of our treatment, the results are consistent. There are no significant effects on the distribution of responses with either introduction or removal of the fine, in either the G&R or the Alt F tax fine conditions.

The complete non-responsiveness of our measure of social tax concerns with our fine treatments is at odds with G&R's claims about the broad applicability of their social norm / crowding out theory for fines. Interestingly, the combination of the Social Tax Reminder and the Underreporting List do produce shifts in the mean level of concern about free riding on other taxpayers through non-compliance. Again, this suggests that there is a degree of sensitivity to social controls in our MTurk sample and that social norms can contribute to tax compliance. However, they do not appear to interact with fines in the way G&R suggest.[107]

[107] Our results for the measure of guilt about not reporting fully are generally consistent with the analysis for social concerns above. There are no significant effects on the measure in the G&R condition. There is one significant effect in the Alt Fine condition – when the fine is removed the mean for guilt about not

c) Fines & Expectations about Others' Tax Compliance Behaviour

The final test of G&R's theories in our tax study is to examine how the introduction of fines affect respondents' beliefs about how other taxpayers will behave. Recall that when fines are introduced, according to G&R, respondents should expect others to increase non-compliance. We test this by looking at results from our question asking respondents to predict the percentage of others *not* reporting fully, and the amount of income those others would report.



Notes to Figure 22: IC is mean percentage of others initially expected to report extra income fully by respondents, prior to underreporting fine treatments. Fine Introduced illustrates the mean percentage of others fully reporting following introduction of the G&R and Alt F tax fine treatments respectively. Fine removed illustrates the means after announcement cancelling underreporting fines. Error bars are 95% confidence intervals.

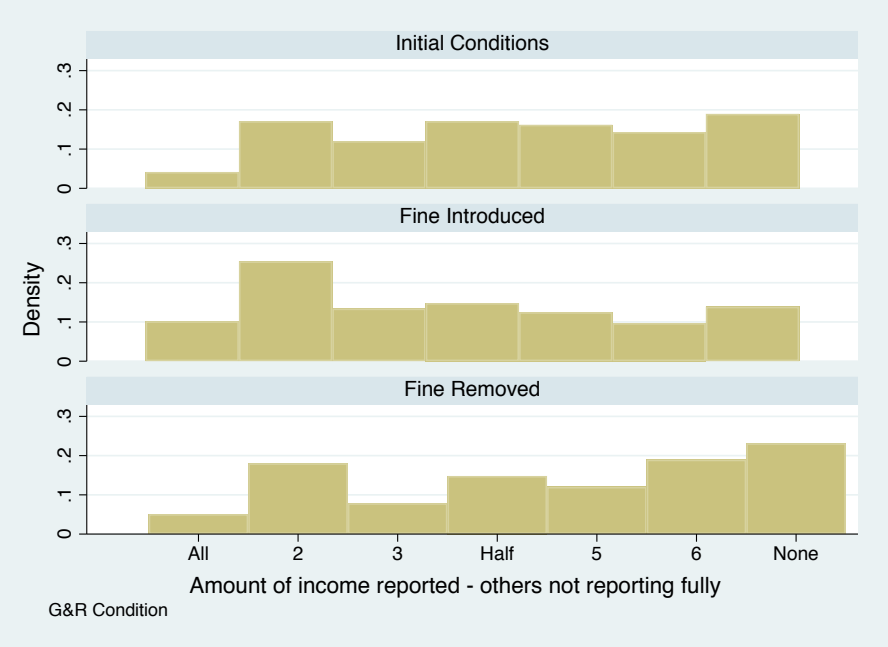
A quick glance at Figure 22 above is enough to address this issue. It is clear that respondents do adjust their expectation about others behaviour when the fine is introduced – but expect it will *reduce* the number of others who do not comply by fully reporting their income. Again, as in the daycare study, this is a transient effect that depends on the fine remaining in place. Once it is removed, expectations about non-compliance rise back to their baseline levels. The two fine conditions work virtually identically. Results for the full distribution of responses confirm this same story.[108]

reporting fully falls significantly relative to when the fine was in place (P<0.0278). This is however not supportive of G&R's hypotheses, since it is the removal of the fine that causes guilt to be assuaged, not its introduction.

[108] In the G&R Condition, the IC distribution is slightly lower valued that the final distribution (K-S P< 0.069), although with a combined K-S test we do not reject equivalence of the initial and final distributions (K-S P<0.139). In any case, this effect does not offer any support for G&R's theories.

When we look at how much income respondents expect taxpayers not reporting fully to disclose, again we do not see support for G&R. Figure 23 provides the full distribution for the G&R Condition to illustrate. The introduction of the fine condition shifts responses to lower values. With the fine in place, respondents expect others will report *more* income – even if not reporting fully. Once the fine is removed, the distribution shifts into the higher values of the scale – respondents expect that non-compliers will try to hide more of their income. Similar results obtain for comparison of means across treatments within each fine condition for this question, as well as if we examine the pattern of changes in the distributions for the Alt F condition.

Figure 23: Distribution Income Reported by Others, G&R Condition



Notes to Figure 23: Vertical axis is density measure of response distribution, Horizontal axis is amount of extra income reported by others who don't report fully, qualitative scale from 1="all" to 7="none". Initial Conditions panel is distribution prior to introduction of fine treatments. Fine Introduced panel is distribution following introduction of G&R replication tax fine, Fine removed panel is distribution after tax fine is cancelled.

Just as in the daycare study, our respondents *do* change their expectations in response to the introduction of the fines. It is just not in a way that supports G&R's account in "A Fine is a Price". Instead, it appears to reflect expectations that the fines will operate in the standard way, as a cost that will generally deter the targeted behavior.

v) Conclusions – Tax Study

The results of the tax study are generally very consistent with those from the daycare vignette replication. The impact of fines on our key outcome measures is to consistently incentivize *more* compliant behaviour by respondents, not increase rule-breaking. There does not appear to be any movement in the reasons behind our respondents' decisions that reflects the causal changes from introduction of a fine in G&R's hypotheses. Fines increase

concerns about enforcement – but only when they are in place. Fines have no significant effect on either respondents' social norm-based concerns for other taxpayers, or their own guilt about non-compliance. Our respondents are concerned most about tax enforcement consequences, and the introduction of fines enhances that concern and deters intended non-compliance. Although there is evidence of social norms having an impact that can enhance taxpayer compliance, introduction of fines does not appear to work at cross-purposes with these social incentives in our survey. Overall, our results again fail to replicate outcomes consistent with G&R.

IV. Sensitivity of Results to Individual Characteristics

In this section, we use regression analysis to check whether our results on the effect of fines are sensitive to any particular individual characteristics of our sample respondents.[109] We address two main questions in this analysis:

- Are there individuals who are especially sensitive to the fines (i.e., they adjust their behaviour more when fines are introduced)?
- Are there particular individuals in our sample for whom "a fine is a price" in the sense of the original study (i.e., the fine increases the outcome measures it is supposed to deter)?

These two questions address concerns related to the generalizability of our main results. Strong associations between particular individual characteristics and respondents' reactions to our treatments would suggest caution in generalizing our results, and possibly a need for further causal investigation. We check general sensitivity of our results to sample characteristics in our first set of regressions. The second question tracks a similar concern about the generalizability of our qualitative result - that fines deter – which relies on average treatment effects or aggregate movement in responses for our full sample. It is clear from our distributional results above that there is underlying heterogeneity in individuals' responses. Here we select on whether or not individuals react to fines in the way that Gneezy and Rustichini's original study predicts and ask whether this distinction appears to be driven in a systematic way by individual respondent characteristics.

We include three broad sets of individual characteristics as controls. The first set are fairly standard demographic controls, including age, gender, race, education, employment status and income. This data was collected prior to exposure to the vignettes. Secondly, for both studies we collected "contextual" individual controls – characteristics that track participant experience with "real world" analogs to our vignettes. These data were collected after the completion of the experimental survey. For the daycare study, these controls include: parental status, status as a daycare user, and if so frequency with which the subject picked up their child.[110] For the tax study, contextual controls include: status as an individual who

[109] An overview of the characteristics of our sample respondents for the Daycare and Tax studies can be found in our supplementary materials; see Metcalf *et al* (2019).
[110] We also collected some additional information, such as the number of children for parents, but ultimately chose to focus on the smaller set of "primary" controls above for which we had better data.

earns non-third-party-reported income[111], whether or not respondents had previously under-reported their income[112], and whether respondents had been audited or fined.[113] The final individual control relates to the extent to which respondents might be other-regarding or otherwise sensitive to social controls. This is a factor that some, such as Kornhauser, Lu and Tontrup, suggest may help explain Gneezy and Rustichini's original result.[114] For the daycare study, we use the initial condition's scale of each respondent's concern about making the daycare teachers wait as a way to capture this characteristic. For the tax study, we use the World Values Survey "trust" question as our measure.[115]

A. Sensitivity of Response to Fines: Individual Characteristics

To investigate the sensitivity of responses to individual characteristics, we estimate linear OLS regressions of the following form:

$$|\Delta Y| = X\beta + \epsilon$$

Where: $|\Delta Y|$ is the absolute value of the change in our outcome measure of interest (e.g. late time, likelihood of reporting income, etc.), X is a matrix of individual characteristics (demographic, contextual, pro-sociality) for all participants, and ϵ is an error term. Illustrative results are included in the Appendix.

For our daycare study, Table 1 presents results using the change in time late as our dependent variable. We provide results for regressions using respondents in only the G&R condition, only the Alt F condition, and finally for a pooled regression for both fine conditions combined. The most consistently significant variable is the pro-sociality variable that measures each respondent's initial concern for others (the daycare workers).

---

[111] This is self-reported income that was not also reported by the payer. We used a question prior to this that asked about whether respondents earned income reported by the payer in order to make this clear, to get more accurate information about this measure, more relevant to sensitivity to the tax study scenario conditions.

[112] This is a self-reported measure and was not restricted to those who said they earned income that they had to self-report. Respondents were provided with the opportunity to choose a category, "prefer not to answer" in the original question. We included only those who admitted to under-reporting in our dummy.

[113] As with the daycare study, we collected some additional information that was ultimately not used in our regressions. For example, we asked whether respondents were tax filers – over 95% of the sample had filed previously – making this unhelpful as a control.

[114] Kornhauser, Lewis A. and Lu, Yijia and Tontrup, Stephan, "Testing a Fine Is a Price in the Lab" (October 1, 2019). NYU Law and Economics Research Paper No. 19-39, Available at SSRN: https://ssrn.com/abstract=3477534 or https://doi.org/10.2139/ssrn.3477534, forthcoming in 63 *Int. Rev. Law & Econ.* (September 2020) (pro-social individuals more susceptible to crowding out effect from fines).

[115] Kornhauser, Lu & Tontrup (2020) use the Social Value Orientation (SVO) measure of DeCremer & Dijk (2002), a measure commonly used in laboratory experiments, but which does not translate well to our experimental survey vignette setting. We use alternatives more suitable to inclusion in a survey, see e.g. Yating Chuang & Laura Schechter, "Stability of Survey and Experimental Measures of Risk, Time and Social Preferences: A Review and Some New Results" (2015) 117 *Journal of Development Economics* 151. Note, for our survey measures of pro-sociality, we have a categorical scale from 0-10 in the daycare, where 10 is the highest concern for others; we use an indicator to code WV survey trust question, with 1="trusting of others".

The higher that initial concern, the less sensitive respondents are to introduction of the fine.[116] However, the size of the effect is not large. Our contextual controls are generally not significant and neither are our demographic controls.

Results for our Tax study using changes in the likelihood of reporting income fully as our dependent variable are shown in Table 2.[117] In general, our demographic controls are not significant. Those in higher income categories are less sensitive to the fine when choosing *how much* income to report in some specifications, but this control is not consistently significant. In terms of contextual controls, those who admit to previously underreporting income appear more responsive to introduction of the fines, but this is only a weakly significant effect in some specifications. Those who have been fined for under reporting adjust less in response to introduction of the fines in some specifications – perhaps because they were already more sensitive to the risk of being caught under-reporting. As with our daycare study, the most consistently significant control over all our specifications is our measure of pro-sociality. Those who are "trusting" (pro-social) are consistently *less* responsive to introduction of the fines.

Overall, our individual regressions do not suggest that our results are specific to a particular demographic cohort. Individual characteristics generally do not correlate with systematic differences in sensitivity to fines in our vignettes. The one exception is the relative pro-sociality of individuals.

B. A "Fine is a Price" that Incentivizes Misbehaving for a Few?

Our second series of regressions investigates whether any individual characteristics systematically predict the probability an individual behaves as a "fine is a price" individual, as in Gneezy & Rustichini's original.

We estimate a series of probit regressions:

$$P(y_i = 1|x) = 1 - F(-x_i'\beta|x)$$

Where $y_i = \begin{cases} 1, & \Delta y_i^* > 0 \\ 0, & \Delta y_i^* \leq 0 \end{cases}$

With an underlying model $\Delta Y^* = X\beta + \epsilon$

The variable Δy_i^* is the true change in behaviour for individual i and x_i is our vector of individual controls (demographic, contextual, pro-sociality) and ϵ is an error term. We code for y_i using the reported changes in the relevant outcome measures (late time, likelihood of reporting income fully, etc.).[118] The percentage of individuals who appear to increase the target outcome behaviour after fines are introduced is not large: in our daycare

[116] Note that this result runs contrary to the findings in Kornhauser, Lu & Tontrup (2020).
[117] We also performed regressions for the amount reported (categorical and dollar values) as alternative outcome measures. Except as noted, the results are generally similar.
[118] Recall that our outcome variables are scaled so that 1 is the highest compliance category. Increases in the value of the outcome measure represent increased non-compliance on our scale.

study the raw percentage of "fine is a price" individuals ranges from 12-15%; however, when we eliminate individuals who only increase their late behaviour up to the end of the new "grace period" the proportion falls to 9-10%.[119] In the tax study the proportion of "fine is a price" individuals is at most 11% with respect to likelihood of reporting income fully. The percentage is considerably smaller for the amount of income reported. These proportions mean that our regressions are identifying off only a small number of individuals. The results should thus be viewed with appropriate caution.

i) Daycare Study

Results for our daycare study using increases in the time late to measure "fine is a price" (FiP) behaviour can be found in Table 3 of the Appendix. We again provide regression results for the G&R condition alone, the Alt Fine condition alone and for a pooled regression for both fines. The most consistent demographic control is educational attainment, significant in all specifications. Respondents in educational categories other than "high school" are more likely to show "fine is a price" behaviour, however the increase in probability with additional education is not large (≈2%). Status other than full-time employment is significant in our Alt Fine and combined regressions and decreases the likelihood of FiP responses. Being in a high-income category is similarly significant and also decreases the probability of a FiP response. In the combined regression, age is a significant demographic control. Increased age reduced the probability of FiP behaviour – although with a small effect size (<1%).

Among our contextual controls, none were consistently significant across our regression specifications. In the G&R condition, parental status and being both a daycare user and the consistent pick-up parent were significant. In both cases, the effect of these characteristics was to increase the probability that a respondent showed FiP behaviour. However, beyond these variables in the G&R condition, none of our contextual controls were significant.

Our measure of pro-sociality was highly significant across all specifications. Those who have higher initial measures of concern for others (the daycare workers) are *less likely* to show the FiP response as a reaction to introduction of a fine.

ii) Tax Study

Results for our tax study using the likelihood that a respondent reports their income fully as our outcome measure are in Table 4 of the Appendix. In general, few controls were consistently significant. In the Alt Fine condition, those not in full-time employment were less likely to show FiP responses. Those who were more educated (beyond high school) were more likely to be FiP responses. No other demographic variables were significant. Among our contextual controls, in the Alt Fine condition, those who earn non-third-party-

[119] This adjustment rules out a possible "confound" in Gneezy & Rustichini's original design, which is also possible in our daycare replication vignette: some individuals may view the announcement of a grace period in combination with the fine as the announcement of a "new" closing time of 5:40. Increasing late behaviour up to this limit then may not be a "true" increase in late behaviour that reflects any "fine is a price" effect.

reported income (i.e., earn income that is not required to be reported by the payer to the government) were less likely to give FiP responses, while those who admitted to previously underreporting income were more likely to do so. In the combined regression for both fines, the only significant control was being audited, which increased the likelihood of FiP behaviour. However, previously under-reporting one's income also increased the probability and was only marginally insignificant ($P<0.108$). Our pro-sociality measure – which was consistently significant in determining sensitivity to the fines in the tax setting – was not significant at all in explaining who is an FiP individual in our sample.

We perform similar regressions using the categorical and dollar measures of the amount of income reported. Here, we are capturing FiP behaviour with respect to the amount of income reported – a decision in which changes take place only if respondents have decided not to report fully. In general, the results are similar to those above, except as noted below. For these outcome measures, the share of individuals who display FiP behaviour becomes even smaller, further reducing the weight that should be placed on the results. Again, we generally did not have controls that were significant across all specifications.

For our categorical income measure, a number of our demographic controls were significant in some but not all specifications. Age was significant in the G&R condition and combined regression and reduced the probability of FiP behaviour. Medium income (above the low category) and status other than full-time employment were also significant in G&R and combined regressions and increased the probability of FiP behaviour. This was a change in sign for employment status at this secondary decision stage, compared with deciding whether or not to report fully. In the Alt Fine regression, being female was associated with a lower probability of FiP responses. Higher education was significant in the Alt Fine and combined regressions and increased FiP responses.

In terms of contextual controls, being fined previously was significant in the Alt Fine and combined specifications. It increased the likelihood that individuals reported less income after the fine was introduced and generated the largest effect size in adjustments to the value of income reported. For dollar value of income reported, having previously under-reported income was the only significant contextual control (more likely FiP), but only in the combined regression.

There was no significant influence for our measure of pro-sociality (trust) in explaining who was likely to show FiP responses when we used our categorical measure for the amount of income reported. However, when we used the dollar amount reported, pro-sociality did become significant and decreased the likelihood of a FiP response. To the extent we have significant results, they do not support Gneezy & Rustichini's theory that the fines crowding out social norms explains FiP behaviour in our respondents. Those with stronger pro-sociality are not more susceptible to responding to fines this way.

iii) Conclusion – Individual characteristics & "fine is a price" behaviour

Overall, although we do find some significant results for our controls, we do not find any strong indication that particular controls are consistently driving "fine is a price" behaviour.

To the extent we have any individual variables that appear to help explain this response, they are ones we would anticipate being relevant from a rational choice perspective. In our daycare setting, full time employees who use daycare and pick up children routinely were more likely to show FiP behaviour. Those with higher sensitivity to social considerations (concern for teachers) were less likely to be in this group. It thus does not appear that "crowding out" of social norms is the mechanism for our FiP respondents in the daycare setting.

Perhaps concerningly, in our tax study it appears that those who admitted to previously under-reporting their income, and those who have been audited or fined, were more likely to show FiP effects in their hypothetical tax behaviour. This suggests some potential influence for the "information" story of G&R's original, rather than the social norms explanation. Respondents who were admitted tax cheaters were more likely to engage in hypothetical tax cheating when exposed to the fine information in our vignettes.[120] The extent to which individuals were pro-social generally did not help predict the hypothetical tax behaviour.

V. Conclusion

The results from our direct "replication" in the daycare survey do not produce similar results to Gneezy &Rustichini's daycare field experiment. Instead, we generally obtain results that are more consistent with a standard understanding of the way that fines operate – as a negative consequence that should (other things equal) result in *less* of the target behaviour. Our tax study, which tests for replication effects in a context suggested by Gneezy & Rustichini, also does not reproduce their key results. Instead, results are consistent with those in our daycare study. The introduction of fines deters anticipated under-reporting behaviour in our respondents. The effect of fines is tied to their applicability. There does not appear to be any regime shift between market-based and social norm-based decision-making evident in our results.

What explains the difference between G&R's results and ours? One possibility is that the effect identified in G&R is not as universal as they articulated in their original paper. There may have been particular features of that environment that produced the surprising results of their study. While the "fine is a price" effect they describe cannot be ruled out through additional studies such as ours, we may hope to try to better understand the particular circumstances needed to generate it. Our regressions do not suggest any universally-relevant individual characteristics that would help us predict the behaviour. To the extent that we have significant controls, they are largely what we would anticipate being relevant from a rational choice perspective. While social norms and concern for others do appear to motivate our respondents, the strength of this characteristic is not associated with

[120] This was a rational decision: in both our G&R and Alt F fine conditions, the expected value of not reporting the extra income was positive – e.g. if none of the income was reported.

systematic "crowding out" of the sort Gneezy & Rustichini proposed as an explanation for their striking results. Instead, characteristics more likely to relate to the economic impact of the fine itself seem more relevant. For fines to deter, they have to be large enough relative to the benefits of non-compliance for an individual.

Another possible explanation for the divergence between our results and Gneezy & Rustichini's original may lie in the translation of their field experiment to our context. Our survey is a one-shot projection of anticipated behaviour, which likely differs from a real-world, dynamic choice experience. Our survey participants do not actually make teachers wait, fail to report income, face any actual fines or social consequences for their choices. The unincentivized nature of our experimental survey vignette may limit its external validity as a check on field trial results. Our survey respondents may be choosing in a way that they think is a rational projection of the fine's impact on their behaviour. In a field setting, additional considerations (e.g., work constraints) may become more salient at the moment of decision. Fines that are imposed only later, or are not apparent as separate fines, may lost the deterrent effect they have in our survey – in which they are undoubtedly highly salient to our respondents. While we believe the survey tells us something important about how respondents *think* they would behave, and perhaps how they *want* to behave, differences between this *ex ante* decision environment and real life may lead them to behave differently.

While consulting parents about daycare conditions for this project, we received a copy of a parent newsletter for a local daycare. In red lettering at the bottom was the following message:

> We close at 5:30pm. Our Late Pick-Up Penalty is in place to offset the additional staffing costs, and to discourage this from happening at all. It is not a childcare option.

While our survey results do not reproduce the "fine is a price" effect of Gneezy and Rustichini's original, anecdotal evidence like this suggests the effect may still exist in the field. While our results suggest most people would not intend to behave this way, and that it is unlikely they do so because the simple existence of the fine displaces social controls, a many studies approach is likely needed to fully explain the behaviour.

Appendix

Table 1: Daycare Study – Sensitivity of Response & Individual Characteristics

Δ *Late Time*	G&R Fine		Alt Fine		Combined Fines	
	Coef. (SE)	**t (P>|t|)**	**Coef. (SE)**	**t (P>|t|)**	**Coef. (SE)**	**t (P>|t)**
G&R Cond	N/A		N/A		1.241 (0.459)	2.71*** (P<0.007)
Age	-0.059 (0.038)	-1.53 (p<0.127)	-0.0127 (p<0.029)	-0.43 (p<0.668)	-0.040 (0.0245)	-1.65* (p<0.10)
Female	0.852 (0.745)	1.14 (p<0.254)	0.283 (0.621)	0.46 (p<0.649)	0.614 (0.482)	1.27 (p<0.204)
Minority	0.909 (0.874)	1.04 (p<0.299)	0.482 (0.681)	0.71 (p<0.480)	0.754 (0.544)	1.39 (p<0.166)
Education	-0.024 (0.262)	0.090 (p<0.927)	0.256 (0.173)	1.48 (p<0.139)	0.115 (0.155)	0.74 (p<0.458)
Employment	0.053 (0.316)	0.17 (p<0.866)	-0.129 (0.183)	-0.71 (p<0.481)	-0.028 (0.178)	0.16 (p<0.877)
Medium Y	-1.109 (0.874)	-1.27 (p<0.205)	0.477 (0.599)	0.80 (p<0.426)	-0.344 (0.533)	-0.65 (p<0.519)
High Y	-1.745 (1.093)	-1.60 (p<0.111)	0.652 (1.152)	0.57 (p<0.572)	-0.638 (0.774)	-0.82 (p<0.410)
Parent	-1.027 (0.965)	-1.06 (p<0.288)	-0.166 (0.914)	0.18 (p<0.856)	-0.654 (0.665)	-0.98 (p<0.320)
Daycare User	-0.396 (1.114)	0.36 (p<0.723)	-0.787 (1.043)	-0.75 (p<0.451)	-0.474 (0.754)	-0.63 (p<0.529)
Pick-up Parent	0.899 (1.134)	0.79 (p<0.428)	0.949 (0.946)	1.00 (p<0.316)	0.948 (0.736)	1.29 (p<0.198)
Pro-Social	-0.320 (0.187)	-1.71* (p<0.088)	-0.193 (0.151)	-1.28 (p<0.203)	-0.251 (0.120)	-2.09** (p<0.037)
Constant	11.04 (2.26)	4.88*** (P<0.000)	5.42 (1.94)	2.79*** (P<0.006)	7.71 (1.451)	5.31*** (p<0.0000)
N	400		403		803	
F-Stat	$F_{(11,388)}$=1.77*		$F_{(11,391)}$=1.26		$F_{(12,790)}$=2.40***	
Prob>F	0.0570		0.24856		0.0047	

* = Significant at 10% or lower; ** = Significant at 5% or less; ***=Significant at 1% or less.

Table 2: Tax Study – Sensitivity of Response & Individual Characteristics

Δ *Pr Report*	G&R Fine		Alt Fine		Combined Fines	
	Coef. (SE)	**t (P>|t|)**	**Coef. (SE)**	**t (P>|t|)**	**Coef. (SE)**	**t (P>|t)**
G&R Cond	N/A		N/A		-0.386 (0.127)	3.03*** (P<0.003)
Age	-0.011 (0.008)	-1.39 (p<0.166)	0.010 (0.011)	0.93 (p<0.356)	-0.002 (0.007)	-0.23 (p<0.815)
Female	0.219 (0.171)	1.29 (p<0.199)	0.221 (0.200)	1.10 (p<0.272)	0.202 (0.132)	1.53 (p<0.126)
Minority	-0.035 (0.183)	-0.19 (p<0.848)	-0.218 (0.209)	-1.04 (p<0.299)	-0.097 (0.141)	-0.69 (p<0.492)
Education	-0.008 (0.053)	-0.16 (p<0.875)	0.065 (0.062)	1.05 (p<0.297)	0.019 (0.042)	0.48 (p<0.631)
Employment	-0.017 (0.049)	-0.35 (p<0.730)	0.021 (0.074)	0.29 (p<0.774)	-0.004 (0.044)	0.11 (p<0.914)
Medium Y	-0.216 (0.184)	-1.18 (p<0.240)	-0.162 (0.206)	-0.79 (p<0.433)	-0.192 (0.139)	-1.38 (p<0.169)
High Y	-0.224 (0.298)	-0.75 (p<0.453)	-0.088 (0.315)	-0.28 (p<0.780)	-0.142 (0.214)	-0.67 (p<0.506)
Self-Report	0.202 (0.171)	1.18 (p<0.239)	0.115 (0.202)	0.57 (p<0.572)	0.194 (0.130)	1.49 (p<0.137)
Under-reported	0.537 (0.331)	1.63 (p<0.105)	0.291 (0.278)	1.05 (p<0.297)	0.372 (0.210)	1.77* (p<0.077)
Audited	0.091 (0.305)	0.30 (p<0.764)	0.091 (0.255)	0.35 (p<0.723)	0.118 (0.197)	0.60 (p<0.550)
Fined	-0.405 (0.344)	-1.18 (p<0.240)	-0.705 (0.331)	-2.13** (p<0.034)	-0.611 (0.239)	-2.55** (p<0.011)
Pro-Social	-0.068 (0.184)	-0.37 (p<0.714)	-0.541 (0.206)	-2.62*** (p<0.009)	-0.291 (0.138)	-2.12** (p<0.035)
Constant	1.21 (0.392)	3.11*** (P<0.002)	0.978 (0.420)	2.33** (P<0.021)	1.32 (0.284)	4.65*** (p<0.000)
N	216		215		431	
F-Stat	$F_{(12,203)}$=1.09		$F_{(12,202)}$=2.89***		$F_{(13,417)}$=3.04***	
Prob>F	0.371		0.0010		0.0003	

* = Significant at 10% or lower; ** = Significant at 5% or less; ***=Significant at 1% or less.

Table 3: Daycare Study – Probability of "Fine is a Price" Behaviour

$P(y_i = 1)$ Late Time	G&R Fine		Alt Fine		Combined Fines	
	Coef. (P>|z|)	$\frac{dP(y_i = 1)}{dx}$	Coef. (P>|z|)	$\frac{dP(y_i = 1)}{dx}$	Coef. (P>|z|)	$\frac{dP(y_i = 1)}{dx}$
G&R Cond	N/A		N/A		0.069 (p<.599)	0.010
Age	-0.019 (p<.192)	-0.002	-0.023 (p<.107)	-0.003	-0.019 (p<.058)	-0.003*
Female	0.079 (p<.667)	0.012	0.012 (p<.955)	-0.001	0.035 (p<.803)	0.005
Minority	0.008 (p<.966)	0.001	-0.059 (p<.793)	-0.007	-0.014 (p<.923)	-.002
Education	0.123 (p<.029)	0.018**	0.161 (p<.020)	0.021**	0.136 (p<.002)	0.020***
Employment	-0.046 (p<.594)	-0.006	-0.706 (p<.003)	-0.095***	-0.160 (p<.057)	-0.023*
Medium Y	-0.054 (p<.783)	-0.007	-0.067 (p<.758)	-0.009	-0.085 (p<.561)	-0.012
High Y	-0.842 (p<.088)	-0.124*	-0.496 (p<.229)	-0.067	-0.677 (p<.025)	-0.010**
Parent	0.477 (p<.079)	0.070*	-0.175 (p<.583)	-0.023	0.204 (p<.300)	0.029
Daycare User	-0.510 (p<.184)	-0.075	0.425 (p<.271)	0.057	-0.052 (p<.839)	-0.007
Pick-up Parent	0.789 (p<.030)	0.116**	-0.028 (p<.931)	-0.004	0.370 (p<.108)	0.053
Pro-Social	-0.111 (p<.001)	-0.017***	-0.135 (p<.000)	-0.018***	-0.120 (p<.000)	-0.017***
Constant	-0.428 (p<.457)		0.026 (p<.964)		-0.304 (p<.462)	
N	400		403		803	
χ^2-Stat	$\chi^2_{(11)}$=35.47***		$\chi^2_{(11)}$=46.97***		$\chi^2_{(12)}$=73.56***	
Prob>χ^2	0.0002		0.0000		0.0000	

* = Significant at 10% or lower; ** = Significant at 5% or less; ***=Significant at 1% or less.

Table 4: Tax Study – Probability of "Fine is a Price" Behaviour

$P(y_i = 1)$ Pr report Y	G&R Fine		Alt Fine		Combined Fines	
	Coef. (P>|z|)	$\frac{dP(y_i = 1)}{dx}$	Coef. (P>|z|)	$\frac{dP(y_i = 1)}{dx}$	Coef. (P>|z|)	$\frac{dP(y_i = 1)}{dx}$
G&R Cond	N/A		N/A		0.041 (p<.802)	0.007
Age	-0.020 (p<.122)	-0.004*	-0.005 (p<.733)	-0.001	-0.014 (p<.123)	-0.002
Female	0.028 (p<.909)	0.005	-0.422 (p<.103)	-0.065	-0.180 (p<.296)	-0.032
Minority	0.099 (p<.678)	0.017	-0.192 (p<.533)	-0.030	-0.050 (p<.781)	-.009
Education	-0.026 (p<.736)	-0.005	0.186 (p<.023)	0.029**	0.051 (p<.358)	0.009
Employment	0.101 (p<.219)	0.018	-0.137 (p<.039)	-0.021**	0.035 (p<.565)	0.006
Medium Y	-0.076 (p<.762)	-0.014	-0.061 (p<.817)	-0.009	-0.074 (p<.687)	-0.013
High Y	-0.045 (p<.899)	-0.008	0.294 (p<.437)	0.045	0.076 (p<.777)	0.013
Self-Reported	0.217 (p<.358)	0.039	-0.636 (p<.021)	-0.098**	-0.179 (p<.304)	-0.032
Under-Reported Y	-0.036 (p<.920)	-0.006	0.805 (p<.022)	0.125**	0.365 (p<.108)	0.065
Audited	0.548 (p<.162)	0..099	0.536 (p<.131)	0.083	0.523 (p<.049)	0.092**
Fined	-.001 (p<.998)	-.000	0.089 (p<.861)	0.014	0.091 (p<.789)	0.016
Pro-Social	0.083 (p<.725)	-0.015	-0.186 (p<.452)	-0.029	0.053 (p<.753)	0.009
Constant	-0.861* (p<.100)		-1.225** (p<.038)		-0.933** (p<.018)	
N	216		215		431	
χ^2-Stat	$\chi^2_{(12)}$=9.51		$\chi^2_{(12)}$=37.60***		$\chi^2_{(13)}$=22.37**	
Prob>χ^2	0.6589		0.0002		0.0499	

* = Significant at 10% or lower; ** = Significant at 5% or less; ***=Significant at 1% or less.