Title

Starting small in project choice: A discrete-time setting with a continuum of types

Authors

Hua, Xiameng
Watson, Joel

Publication Date
2022-09-01

DOI
10.1016/j.jet.2022.105490

Starting Small in Project Choice: a Discrete-Time Setting with a Continuum of Types

Xiameng Hua and Joel Watson*

June 2022

Abstract

We add to the literature on long-term relationships with variable stakes and incomplete information by analyzing a discrete-time trust game between a principal and agent, with a continuum of types. In each period the principal selects the level of a project and the agent then decides whether to cooperate or betray; payoffs in the period scale with the level. The agent's benefit of betraying is privately known. The discrete-time framework allows for analysis of renegotiation in terms of an internal consistency condition that compares actual equilibria in the continuation of the game from any period, improving on the prior literature. Our condition assumes the principal has full power to alter the equilibrium selection. Our main result shows that the resulting perfect Bayesian equilibria converge as the period length shrinks to zero, and we provide a closed-form solution. In equilibrium, the relationship starts small and the level gradually rises until it reaches its maximum; cooperation is viable regardless of the type distribution. JEL codes: C72, C73, C78, D82, D86.

*UC San Diego. The authors thank the following colleagues for providing very helpful input: Snehal Banerjee, Renee Bowen, Songzi Du, Matthias Fahn, Jack Fanning, Simone Galperti, David Kreps, Mark Machina, Roberto Serrano, Joel Sobel, participants at 2021 North American Summer Meeting of the Econometrics Society, and seminar participants at Brown University, New York University, and the University of Arizona. The authors are especially grateful to the anonymous referees and the editor. First version: November 2020.

1 Introduction

Long-term relationships in business and greater society often begin with asymmetric information, where the parties are unsure of each others' incentives, and they have choices regarding how to build their relationships. For instance, a manager may not know to what extent a new employee will have the incentive to shirk on his assignments, and the manager can decide what kinds of responsibilities to give this worker over time. Should the manager assign the worker to important projects, where effort would generate substantial profit for the firm but where shirking would translate into great losses?

Conventional wisdom suggests that it is better to start a relationship cautiously with small-stakes projects and then, conditional on good performance, increase the stakes as time goes on. In this way, a manager may be able to induce "bad" types of workers (those who inevitably will shirk at some point) to reveal themselves by shirking when the stakes are low. But if the manager would increase the stakes quickly over time, then a bad type worker would prefer to delay shirking until the stakes are high. Thus, the manager faces a trade-off between the rate at which she increases the worker's responsibilities (conditional on good performance) and when the worker's type will be revealed. Complicating matters, the manager may wish to adjust her plan mid-stream, based on what she learns about the worker.

We explore these dynamics by developing a new game-theoretic model of the interaction between a principal and an agent with private information and a continuum of types. The parties interact in discrete periods. In each period the principal selects the level of a project, and the agent then chooses whether to cooperate or betray. We characterize the model's perfect Bayesian equilibria, and we propose a renegotiation condition, which we call *alteration-proofness*, that narrows the set of equilibrium outcomes. Alteration-proofness is a notion of internal consistency that assumes the principal has full power to coordinate the players on an altered equilibrium.[1] By way of motivation, in line with the large literature on renegotiation in contractual settings, we think it is natural to assume that the players can revisit and change their equilibrium continuation. Also, it is useful to work with models that generate narrow equilibrium predictions.

[1] Internal consistency is the weakest version of Pareto-perfection that underlies definitions of renegotiation-proof equilibrium in the repeated-game literature, specifically those of Rubinstein (1980), Bernheim and Ray (1989), and Farrell and Maskin (1989).

Although there are multiple alteration-proof equilibria, our main result establishes that these equilibria converge as the period length shrinks to zero, meaning that the model has a unique prediction in the limit.[2] The limit outcome is characterized by a differential initial-value problem. We provide an example for which the solution is easily found in closed form, along with analysis of comparative statics.

Our modeling exercise is most closely related to Watson's (1999, 2002) analysis of relationships in continuous time with variable stakes and with two types of players: a "good" type, for whom cooperation would be possible in a setting of complete information, and a single bad type. In Watson's model, an exogenously provided level function gives the stakes of the relationship at every instant of time. The level function is interpreted as jointly determined by the players, and thus the game is not fully noncooperative. These articles show that, by *starting small*, long-term cooperation is always viable between good types of players, regardless of the initial type probabilities. Further, Watson (1999) puts forth renegotiation-proofness conditions that uniquely select a level function and outcome of the game.

We contribute to the literature in three ways. First, because our model is fully noncooperative and in discrete time, the alteration-proofness condition compares actual equilibria in the continuation of the game from any period. This setting provides a better foundation for renegotiation than was possible in Watson (1999). Second, we allow for a continuum of bad types, and we obtain a novel characterization of the alteration-proof perfect Bayesian equilibria, along with comparative statics. Contrary to the result in Watson (1999), we find a multiplicity of alteration-proof equilibria in the discrete-time setting. This leads to our third contribution, which is to devise a method of bounding the set of equilibria and to characterize the bounds as the period length shrinks. Our method incorporates a new mathematical result on the limit of solutions to discrete-time models defined by transition functions (Watson 2021).

The related literature on starting small in relationships includes both seminal theoretical contributions and experimental evidence. Theoretical origins reside in Sobel (1985), Ghosh and Ray (1996), and Watson (1999, 2002). Sobel (1985) focuses on a "sender-receiver" model in which the level is determined by an exogenous random draw in each period; this paper also describes a "loan model" in which one player

[2] The convergence result pertains to nontrivial equilibria, in which the principal sets a positive level in at least one period. For some parameter values, there may also exist a trivial no-trust equilibrium, but we argue that it would be ruled out by a weak form of external consistency.

chooses the level and the other, who could be a friend or enemy, chooses whether to invest or default. The equilibrium of the loan model entails gradual increase in the loan level over time. Our model has the same form of stage game, but the payoffs are different for the player we call the good type of agent, and in our model incentives relate to an infinite horizon. Ghosh and Ray (1996) examines a setting in which players in a community randomly match to form long-term relationships. Players can exit their relationships at any time and then rematch, and newly matched players receive no information about the past behavior of their partners. A fraction of the population is myopic. Players are motivated to weed out myopic types by reducing the level of cooperation in the first period of new relationships, and this serves as a punishment for non-myopic players who might otherwise cheat and rematch without consequences.

On the empirical side, Andreoni, Kuhn, and Samuelson (2019) reports an experiment in which subjects are able to choose the stakes in a two-period prisoners' dilemma, finding that players utilize a starting-small strategy to achieve cooperation. Likewise, Ye et al. (2020) studies a multi-period weakest-link game in the laboratory, where treatments differ in the exogenously set sequence of levels, finding that cooperation is associated with gradualism (starting small and gradual increase of the level). Kartal, Müller, and Tremewan (2019) provides experimental results on an infinite-horizon partnership game, where treatments differ in the set of level options. This paper finds in settings of severe information asymmetry that subjects are able to build trust when they have the option of starting small and gradually raising the stakes of their relationships, and the subjects act accordingly.

Rauch and Watson (2003) develops a model of relationships in which the players have common information but are uncertain of their prospects as a partnership. The article shows theoretically and empirically that it is sometimes optimal to start small.[3] Bowen, Georgiadis, and Lambert (2019) examines starting small in a setting where two heterogeneous agents contribute over time to a joint project and collectively decide its level, finding that, in equilibrium, the effective control over the project scale

[3] Horstmann and Markusen (1996) models the choice by a multinational firm seeking to enter a new (foreign) market between direct investment and contracting with a local sales agent. Information gained from the agency contract is useful in the decision of whether to pursue direct investment. Hence, the agency contract is analogous to starting small in a variable-stakes games (though it may be desirable to extend it indefinitely). Horstmann and Markusen (2018) analyzes a similar model but relaxes the commitment assumption and studies both moral hazard and adverse selection.

relates to the realized types of players. Atakan, Koçkesen, and Kubilay (2020) studies repeated cheap talk and demonstrates that when the conflict of interest between the receiver and the sender is large, starting-small to communicate is the unique equilibrium arrangement.[4]

Also related is the model of Malcomson (2016, 2020), in which a principal and agent with persistent private information have an ongoing relationship governed by a relational contract (the principal makes voluntary payments to reward the agent's effort choice). Malcomson (2016) shows that if agent's type is on a continuum, then there does not exist a fully separating equilibrium, and Malcomson (2020) characterizes the finest partition equilibria. Separation requires a low effort level at the beginning of the relationship, so in this sense some equilibria exhibit a form of starting small. Renegotiation-proofness in the form of external consistency (looking at the frontier of the set of equilibrium payoffs) is also studied in the latter paper.

This paper is organized as follows. In Section 2 we formally describe the model and equilibrium concept, and we analyze the agent's incentive conditions. The renegotiation condition is defined and analyzed in Section 3, where we report our main result on the limit of equilibria. Section 4 presents the case of uniformly distributed bad types. Section 5 provides additional technical notes, including on off-equilibrium-path alterations and how our results extend to the "no-gap case" of agent types assumed away earlier, and discusses additional connections with literature. Section 6 offers concluding comments. The appendices contain details of the analysis and proofs.

2 Model

We examine a model of a relationship between a principal and agent in discrete time and with one-sided incomplete information. In this section we first describe the complete-information version of the game, followed by the incomplete-information version. We then establish notation, review the equilibrium conditions, and provide a partial characterization of "trusting equilibria" along with examples.

[4] Other articles that model some manner of starting small in relationships and/or trust building include Kranton (1996), Blonski and Probst (2001), Rob and Fishman (2005), and Chassang (2010).

2.1 Trust game with complete information

The complete-information version of the model is a repeated game that terminates under some conditions. There are two players, called player 1 (the principal) and player 2 (the agent). The time period is denoted by $k \in \{1, 2, \ldots\}$. We assume the players have a common discount factor $\delta \equiv e^{-r\Delta}$, where $r \in (0, 1)$ is the discount rate and $\Delta > 0$ is the length of each period in real time.

In each period, as long as the game was not terminated earlier, players interact in the stage game shown in Figure 1, where player 1 selects a *trust level* $\alpha \in [0, 1]$ and then player 2 observes α and chooses whether to *betray* or *cooperate*. If player 2 cooperates then both players get the payoff $\alpha\Delta$ in the current period and play continues in the next period. On the contrary, if player 2 betrays then the game ends with terminal payoffs of $-\alpha c$ for player 1 and αx for player 2, where $c > 0$. Players seek to maximize the discounted sum of their period payoffs.

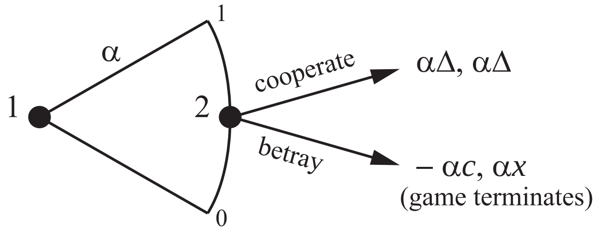


Figure 1: Stage game

It is easy to verify that cooperation can be sustained if and only if $x \leq \Delta/(1 - \delta)$. Under this condition there is a subgame-perfect equilibrium in which, in every period, player 1 chooses $\alpha = 1$ and player 2 cooperates. Player 2's continuation value of playing this way from the start of any period is $\Delta/(1-\delta)$, which exceeds the payoff of betraying. Furthermore, if $0 < x < \Delta/(1 - \delta)$, then there are many other equilibria. In fact, for any sequence $\{\alpha^k\}$ of feasible levels there is an equilibrium in which, on the equilibrium path, this sequence of levels is chosen by player 1 and player 2 always cooperates, so long as the following condition holds for each period k: $\alpha^k x \leq \alpha^k \Delta + \delta v_2^{k+1}$, where $v_2^{k+1} = \sum_{\tau=k+1}^{\infty} \delta^{\tau-k-1} \alpha^\tau \Delta$ is player 2's continuation value from the start of period $k + 1$.[5]

[5] Player 1 can be deterred from deviating by specifying that, following any deviation, the players coordinate on $\alpha = 0$ and betrayal from that point regardless of any further deviations (which is an equilibrium in all future subgames).

To summarize, there are a lot of equilibria featuring trust and cooperation if x is not too large. The best equilibrium for both players is clearly that in which player 1 chooses $\alpha = 1$ in every period. On the contrary, if $x > \Delta/(1-\delta)$ then there is no equilibrium in which cooperation occurs at a positive level in any period.

2.2 Trust game with incomplete information

We are interested in the trust game with incomplete information regarding the payoff parameter x. Specifically, suppose that before the relationship begins, Nature chooses x according to a given probability distribution F that is common knowledge, with support denoted by $X \subset \mathbb{R}$. Player 2 privately observes x, which we therefore refer to as *player 2's type*. Let us label every $x \leq \Delta/(1-\delta)$ a *good type* and every $x > \Delta/(1-\delta)$ a *bad type*. We generally express F as a cumulative probability function, so that $F(x')$ denotes the probability that $x \leq x'$.

In this game, player 1 may be able to establish perpetual cooperation with a good type, but every bad type must eventually betray. The level of the relationship affects both player 2's betrayal gain and the players' flow payoff of cooperation, so by varying the level over time, player 1 may be able to coax the bad types to betray in periods when the level is small. However, there is a trade-off: A bad type of player 2 would be willing to betray in a given period only if this player does not expect that player 1 would choose a much higher level in near future, contingent on player 2 cooperating until then. That is, it may be optimal for a bad type to cooperate for some number of periods and then betray later when α is large. Therefore, player 1 cannot screen out the bad types at a low level and also expect to soon cooperate at a high level with good types. Further, types with higher values of x are essentially less patient than are those with lower values of x, so player 1's choice of levels over time could lead different types of player 2 to betray in different periods.

We assume that the types are bounded and there is a gap between the sets of good types and bad types.[6] We will later examine sequences of games for Δ converging to zero, and we want the type labels to hold for every Δ close to zero. Because $\Delta/(1-\delta)$ decreases and converges to $1/r$ as $\Delta \to 0^+$, we therefore assume that the good types are below $1/r$ and the lowest bad type is strictly above $1/r$. Additional technical assumptions are included in the following assumption.

[6] This is analogous to the "gap" case of the durable-good-monopoly problem (Coase (1972), Gul, Sonnenschein, and Wilson (1986)).

Assumption 1. *Distribution function F is continuous. There are numbers $\hat{\Delta}$, a, and b satisfying $0 < \hat{\Delta} < 1/r < a < b$ such that $F(\hat{\Delta}) = 0$; $F(a) = F(1/r) > 0$; $F(b) = 1$; and restricted to subdomain $[a, b]$, F is twice continuously differentiable with a strictly positive density function f. Finally, $\Delta \leq \min\{\hat{\Delta}, \overline{\Delta}\}$, where*

$$\overline{\Delta} \equiv a \left(\frac{1}{a \min_{x \in [a,b]} f(x)} + 1 \right)^{-1}.$$

Note that the set of bad types is the interval $[a, b]$. We define the derivative of F at endpoint a as its right derivative, and at endpoint b as its left derivative, so that f is well-defined and can have the assumed properties when restricted to $[a, b]$. We define the derivative of f similarly. Because f is strictly positive and continuous on subdomain $[a, b]$, it reaches a minimum that is strictly positive, and so $\overline{\Delta}$ is well-defined. The assumption that the set of good types is bounded away from 0 ensures the existence of a class of simple equilibria but is not needed for existence or used in our characterization theorem. From here, "game" and "trust game" refer to our incomplete-information, discrete-time game with parameters r, Δ, X, and F just now described.

2.3 Strategies and equilibrium conditions

We analyze the game using the weak Perfect Bayesian Equilibrium (PBE) solution concept. In this subsection, we define and provide notation for histories, strategies, and beliefs. We then describe the equilibrium conditions and, noting the plethora of equilibria, motivate the refinement developed in the next section.

For any $k \in \{1, 2, \ldots\}$, a k-period history of level choices is given by $(\alpha^1, \alpha^2, \ldots, \alpha^k)$. This sequence of levels can be interpreted as the public history to the beginning of period $k + 1$ (specifying player 1's information set), where player 2 cooperated in periods $1, 2, \ldots, k$. Likewise, this same sequence $(\alpha^1, \alpha^2, \ldots, \alpha^k)$ represents the public history to player 2's information set in period k, where the public history to the beginning of period k was $(\alpha^1, \alpha^2, \ldots, \alpha^{k-1})$ and then player 1 selected α^k in period k. Note that player 2's personal history includes both $(\alpha^1, \alpha^2, \ldots, \alpha^k)$ and player 2's type x.

Let $H = \cup_{k=0}^{\infty} [0, 1]^k$ be the set of all finite public histories, where $[0, 1]^0$ is taken to be the null history at the beginning of period 1. Let $H_+ = \cup_{k=1}^{\infty} [0, 1]^k$ be the set of non-null public histories. Also, for any k-period public history h and level α, denote by $h' = h\alpha \in H$ the $(k+1)$-period public history realized when h is followed by level α chosen in period $k+1$.

We focus on pure strategies.[7] Player 1's strategy $s_1 : H \to [0,1]$ specifies the level in each period as a function of the public history to this point. Player 2's strategy specifies whether to cooperate or betray in each period, as a function of history to player 2's information sets, including player 2's type. Thus, player 2's strategy is a function $s_2 : H_+ \times X \to \{1, 0\}$, where $s_2(h', x) = 1$ indicates that player 2 cooperates and $s_2(h', x) = 0$ indicates that player 2 betrays.

We describe player 1's beliefs about player 2's type using an assessment function $Q : H \to \mathcal{P}(X)$, where $\mathcal{P}(X)$ denotes the set of probability distributions over X. That is, for any k-period public history $h \in H$, $Q(h)$ is player 1's belief at the beginning of the following period $k+1$.

Given the strategies s_1 and s_2, any public history h, and player 2's type x, let $v_1(h; s_1, s_2, x)$ and $v_2(h; s_1, s_2, x)$ denote the players' continuation values from the period after history h occurs, assuming that x is player 2's actual type and that play will continue according to s_1 and s_2. Because player 1's assessment is $Q(h)$, player 1's expected continuation value is

$$v_1(h; s_1, s_2, Q(h)) \equiv \mathbb{E}_{Q(h)}[v_1(h; s_1, s_2, x)],$$

where $\mathbb{E}_{Q(h)}$ denotes expectation over $x \sim Q(h)$; this assumes that player 1 continues to believe after history h that player 2's strategy is s_2.

We extend player 2's strategy $s_2(h', x)$ to the space of type distributions by taking the expectation, so that for any $h \in H_+$ and any type distribution $\hat{F}$,

$$s_2(h, \hat{F}) \equiv \mathbb{E}_{\hat{F}}[s_2(h, x)].$$

Note that if player 1 chooses level α in the period following public history h, then player 1 expects player 2 to cooperate with probability $s_2(h\alpha, Q(h))$.

We next review the notion of sequential rationality, stated here in terms of single deviations, and the equilibrium definition. The one-deviation principle applies.

[7] Restricting attention to pure strategies for player 2 is without loss in our analysis. Because we have a continuum of types, indifference conditions would occur for only a subset of measure zero. Accounting for randomization by player 1 would complicate the statement of the alteration-proofness conditions, we think with no effect on the main results, as discussed in Section 5.

Definition 1. *Given Q and s_2, player 1's strategy s_1 is called* **sequentially rational** *if for every public history $h \in H$, $s_1(h)$ maximizes*

$$s_2(h\alpha, Q(h))(\alpha\Delta + \delta v_1(h\alpha; s_1, s_2, Q(h\alpha))) + (1 - s_2(h\alpha, Q(h)))(-\alpha c)$$

by choice of $\alpha \in [0, 1]$. Given s_1, player 2's strategy s_2 is called **sequentially rational** *if for every $h \in H_+$ and $x \in X$, $s_2(h, x) = 1$ only if*

$$\alpha\Delta + \delta v_2(h; s_1, s_2, x) \geq \alpha x$$

and $s_2(h, x) = 0$ only if the reverse weak inequality holds.

Definition 2. *A pure-strategy weak* **Perfect Bayesian Equilibrium (PBE)** *is a strategy profile (s_1, s_2) and beliefs Q such that s_1 and s_2 are sequentially rational and Q obeys Bayes' Rule for all histories reached with positive probability given F, s_1, and s_2.*

Note that in periods in which a positive mass of types is supposed to cooperate, player 1 cannot detect a deviation by a type that was meant to betray, and so standard Bayes updating applies. Weak PBE does not constrain belief updating following any "public deviation," where either player 1 deviated or player 2 cooperated in a contingency in which all types were supposed to betray, because the conditional probability formula does not apply in such a contingency. We could impose stronger consistency conditions, such as Watson (2017) defines, but it would be of no consequence because in the game studied here, we can modify any weak PBE to satisfy strong consistency conditions off the equilibrium path where needed.

For any PBE, let $\{\alpha^k\}_{k=1}^K$ denote the sequence of levels chosen by player 1 *on the equilibrium path*, where there is no public deviation. In this expression, K denotes the last period that occurs in equilibrium; K is finite if all types of player 2 betray in bounded time, and $K = \infty$ if for every period k, a positive mass of types cooperate through period k on the equilibrium path. We will show shortly that $K = \infty$ for any PBE, but for now we must allow for the possibility of K finite.[8]

Let us characterize player 2's incentive conditions on the equilibrium path. Type x

[8] Equilibrium strategies must specify behavior after all histories, including ones in which a period $k > K$ is reached following a public deviation. We could describe, for instance, the infinite sequence of levels that would result from player 1 never deviating and player 2 always cooperating; this sequence would be on the equilibrium path through period K and off the equilibrium path thereafter. Such a sequence will not be needed for our analysis.

optimally betrays in some period in the set

$$\beta(x) \equiv \underset{k \in \{1,2,\ldots,K\}}{\arg\max} \sum_{\tau=1}^{k-1} \delta^{\tau-1} \alpha^{\tau} \Delta + \delta^{k-1} x \alpha^{k}.$$

Note that β is defined relative to a given PBE and it constrains player 2 to betray at or before the equilibrium K. In the case of $K = \infty$, $\infty \in \beta(x)$ is allowed and means that type x optimally cooperates forever.[9] For each $k \in \{1, 2, \ldots, K\}$, let $h^k = (\alpha^1, \alpha^2, \ldots, \alpha^k)$ denote the equilibrium-path public history to player 2's information set in period k. Player 2's equilibrium strategy must have the property that, for $x \in X$ and for the lowest k for which $s_2(h^k, x) = 0$, it is the case that $k \in \beta(x)$.

Regarding player 1's incentives, observe that the following specification of beliefs and behavior for off-path continuations achieves a continuation value of zero for both players. After any deviation by player 1, all types of player 2 would immediately betray, ending the game. If player 2 instead cooperates, which constitutes a further public deviation, then player 1's updated belief would assign probability 1 to a bad type. Then in every period thereafter, regardless of the interim history, player 1 is supposed select $\alpha = 0$ and all types of player 2 are supposed to betray. These continuation strategies are sequentially rational.[10] Because player 1 can guarantee a payoff of zero by choosing $\alpha = 0$ forever, player 1's incentive conditions on the equilibrium path amount to having a nonnegative continuation value.

2.4 Trusting PBE

We are particularly interested in PBE in which, on the equilibrium path, the level is strictly positive in at least one period.

Definition 3. *A perfect Bayesian equilibrium in the trust game is called a* **trusting PBE** *if* $\alpha^k > 0$ *for some* $k \in \{1, 2, \ldots, K\}$.

A trusting equilibrium exhibits some degree of cooperation at positive levels of trust, for otherwise player 1 would strictly prefer to set the level to zero in every

[9] The set $\beta(x)$ is nonempty even if $K = \infty$ due to discounting. Also, if $\beta(x)$ contains an infinite number of periods then it must also contain ∞. If K is finite, then it would be feasible for player 2 to cooperate through period K and perhaps betray later, and we would need to check such a deviation to determine whether player 2 best responds.

[10] In fact, they are sequentially rational regardless of player 1's beliefs about player 2's type, but the particular belief specified here will be helpful later for an extension of the model.

period. We first characterize trusting PBE in terms of the relation between the strategy of player 2 and the sequence of levels on the equilibrium path.

Lemma 1. *Every trusting PBE has the following properties: $K = \infty$. There is an integer L and a weakly decreasing sequence $\{x^k\}_{k=0}^{\infty}$ such that (i) for every $x \in X$, if player 2 of type x betrays on the equilibrium path then this betrayal occurs in a period k that satisfies $x \in [x^k, x^{k-1}]$, and (ii) type $x = a$ betrays in period L.*

Proof of Lemma 1. Consider any trusting PBE and let $\{\alpha^k\}_{k=1}^{K}$ be the sequence of levels chosen on the equilibrium path. Let us define $\omega(k, x)$ as the objective function for the definition of β:

$$\omega(k, x) \equiv \sum_{\tau=1}^{k-1} \delta^{\tau-1} \alpha^{\tau} \Delta + \delta^{k-1} x \alpha^{k}.$$

We first show that for any two types x' and x'' such that $x' > x''$, on the equilibrium path type x' betrays in a period weakly earlier than does type x''.

To prove this claim, suppose there exist types x' and x'' and periods k' and k'' such that $x' > x''$, $k' \in \beta(x')$, $k'' \in \beta(x'')$, and yet $k' > k''$, and we will find a contradiction. Imagine that player 2 compares betraying in period k' with betraying in period k'', ignoring other periods. By the definition of β, type x' prefers betraying in period k' whereas type x'' prefers betraying in period k'' only if

$$\omega(k', x') - \omega(k'', x') \geq 0 \ \text{ and } \ \omega(k'', x'') - \omega(k', x'') \geq 0,$$

and the preference is strict if the relevant inequality holds strictly. Using the definition of ω and simplifying terms, we obtain

$$\left(\alpha^{k''} - \delta^{k'-k''} \alpha^{k'}\right) x'' \geq \delta^{-k''} \sum_{\tau=k''}^{k'-1} \delta^{\tau} \alpha^{\tau} \Delta \geq \left(\alpha^{k''} - \delta^{k'-k''} \alpha^{k'}\right) x'.$$

Because the level is strictly positive in at least one period on the equilibrium path, player 2's incentive condition implies that $\alpha^{k''} > 0$, which further implies that the middle term in the above expression is strictly positive. Using the left inequality and $x'' > 0$, we obtain $\alpha^{k''} - \delta^{k'-k''} > 0$. Combining the inequalities and dividing by $\alpha^{k''} - \delta^{k'-k''}$, we get $x'' \geq x'$, contradicting our presumption that $x' > x''$.

Next, we show that $\beta(a)$ is bounded above. Define $\overline{\alpha} = \sup\{\alpha^1, \alpha^2, \ldots\}$. Then for any $\varepsilon > 0$, there exists a period κ such that $\alpha^\kappa \geq \overline{\alpha} - \varepsilon$. If player 2 of type a betrays in period κ, then the game ends and he gets terminal payoff $a\alpha^\kappa$, which weakly exceeds $a(\overline{\alpha} - \varepsilon)$. If $\beta(a)$ were unbounded then $K = \infty$ and $\infty \in \beta(a)$. By cooperating forever, this type's continuation value from period κ is $\sum_{k=\kappa}^{\infty} \delta^{k-1}\alpha^k\Delta$, which is bounded above by $\sum_{k=\kappa}^{\infty} \delta^{k-1}\overline{\alpha}\Delta = \overline{\alpha}\Delta/(1-\delta)$. Because $a > \Delta/(1-\delta)$, we know that $a(\overline{\alpha} - \varepsilon) > \overline{\alpha}\Delta/(1-\delta)$ for sufficiently small values of ε, which contradicts that it is rational for type a to cooperate forever. We conclude that the lowest bad type betrays in some period L on the equilibrium path.

Finally, we show that $K = \infty$. Assume otherwise, meaning that on the equilibrium path all types of player 2 betray at or before period K and some types wait until K to do so. It must be that $\alpha^K > 0$, for otherwise the types that are supposed to betray in period K would strictly prefer to betray in an earlier period where the level is strictly positive (a time which must exist in a trusting equilibrium). But then in period K player 1's continuation value must be strictly negative because he expects player 2 to betray with probability one. This contradicts player 1's rationality because he would strictly gain by selecting $\alpha^k = 0$ for all $k \geq K$. □

Lemma 1 does not pin down the periods of betrayal for the countable number of types of player 2 that may be indifferent between betraying in one period and the next. Because this is a set of measure zero, equilibria that differ in this regard are essentially equivalent.

To summarize the analysis so far, every trusting PBE has an infinite equilibrium path and is partially characterized by its sequence of levels $\{\alpha^k\}_{k=1}^{\infty}$ and its sequence of cutoff types $\{x^k\}_{k=0}^{\infty}$. On the equilibrium path, for any integer k, all types below x^{k-1} cooperate through period $k-1$ and then types in the subinterval (x^k, x^{k-1}), and possibly one or both endpoints, will betray in period k at level α^k. The monotonicity of betrayal dates established by Lemma 1 applies to all types, good types included. All bad types betray in or before period L. Note that the lemma does not indicate whether any good types betray in equilibrium.

As the analysis continues, we will need to keep track of continuation values. Given any trusting PBE and any period k, we let v_1^k denote the expected continuation value for player 1 from the start of period k on the equilibrium path. Likewise, we let $v_2^k(x)$ denote the continuation value of player 2 of type x from the start of period k conditional on player 2 having always cooperated in the past and player 1 not having deviated from the equilibrium level sequence.

We conclude this subsection with an existence result for trusting PBE, which is a corollary of our main existence result in the next section.

Theorem 1. *Under Assumption 1, a trusting PBE exists in the trust game.*

This result extends what was found by previous papers in the literature, in particular Watson (1999, 2002), so it is not surprising. It is worth noting what this means in economic terms. First, an ongoing cooperative relationship between player 1 and good types of player 2 is viable, and value is created regardless of the type distribution. Second, this conclusion relies on the ability of the players to start small in their relationship. That is, if player 1 had only the choice of, say, $\alpha = 0$ or $\alpha = 1$ then there would be no trusting PBE for a sufficiently small mass of good types.

2.5 Intuition and Illustrations

To get a flavor of the relation between the level sequence and player 2's optimal choices, let us examine the trade-off that player 2 faces locally in time. Because in equilibrium type x^k weakly prefers to cooperate through period k, and is in fact the highest type to do so, we have

$$\alpha^k x^k \leq \Delta\alpha^k + \delta v_2^{k+1}(x^k).$$

In some equilibria, type x^k is indifferent between cooperating and betraying in period k, so that the above inequality holds as an equation. In the event that the indifference condition holds until this type actually betrays, an implication is that $v_2^{k+1}(x^k) = \alpha^{k+1}x^k$. Using this expression to substitute for $v_2^{k+1}(x^k)$, we obtain:

$$\alpha^k x^k = \Delta\alpha^k + \delta\alpha^{k+1}x^k. \tag{1}$$

The refinement developed in the next section will be shown to imply that Equation (1) holds in every period k for which $x^k > a$; that is, this indifference condition holds until all bad types have betrayed.

Before proceeding to the equilibrium refinement in the next section, we illustrate the multiplicity of trusting equilibria, which differ in terms of when bad types betray, how the level changes over time, and players 1's payoff. Figures 2–4 depict three equilibria that we constructed for the same specification of parameters: $\Delta = 1$,

$r = 0.1$ (so that $\delta = e^{-r\Delta} = 0.9048$), $a = 11.5083$, and $b = 30$. The distribution F of player 2's type has a mass of 0.3836 of good types and specifies a uniform distribution of bad types. The value of c matters only for player 1's incentives, and the equilibria pictured exist as long as c is not too large. In each of these equilibria, on the equilibrium path all good types cooperate in every period.

For the equilibrium in Figure 2, Equation (1) does not hold in some periods. Every type of player 2 strictly prefers to cooperate in early periods when the level is low, looking forward to betraying in later periods when the level is high. Figures 3 and 4 illustrate equilibria for which Equation (1) holds for all periods. In the equilibrium shown in Figure 3, all bad types betray in period 1 at the beginning of the game, so $L = 1$. In the equilibrium shown in Figure 4, no bad type betrays until the level reaches 1 in period $L = 30$.[11]

It turns out that none of the equilibria pictured satisfy the renegotiation-proofness condition developed in the next section. In the first equilibrium, there are periods in which the level can be increased without affecting player 2's incentives, and this increases player 1's payoff.[12] In the second equilibrium, after observing cooperation in the first period, player 1 would be sure that player 2 is a good type that will never betray. Therefore, in the second period player 1 has the incentive to "jump ahead" to the continuation of the equilibrium from period $L = 30$ where the level is maximal. In the third equilibrium, player 1's payoff decreases as period $L = 30$ approaches and so, in any period before L, player 1 would have the incentive to "stall" as though restarting from the previous period.

3 Alteration Proofness

In this section, we define and analyze a minimal notion of renegotiation that we call *alteration proofness*, where player 1 has the power to dictate an alteration of current

[11] We can show that if $c > b$ then player 1's favorite equilibrium is as pictured in Figure 3, where all bad types of player 2 betray in the first period, whereas if $c < a$ then player 1's favorite equilibrium is as pictured in Figure 4, where all bad types of player 2 wait until period $L - 1$ to betray. These findings match with what Watson (2002) demonstrates in a continuous-time model with a single bad type.

[12] This feature of the first illustration is familiar, for many signaling models have separating equilibria with nonbinding incentive constraints. Consider, for instance, the standard labor-market signaling game and a separating equilibrium in which the high-ability type chooses an education level that is higher than needed for separation. In our model, renegotiation-proofness forces some constraints to bind.

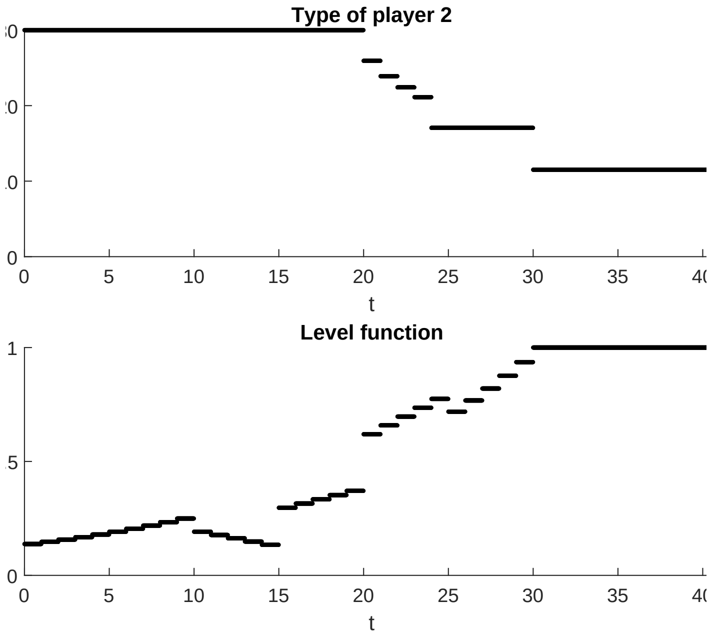


Figure 2: First equilibrium illustration

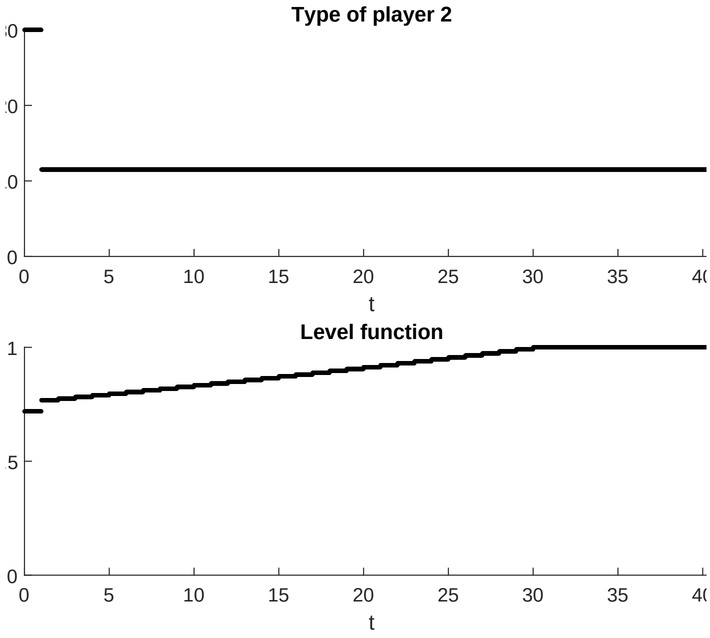


Figure 3: Second equilibrium illustration

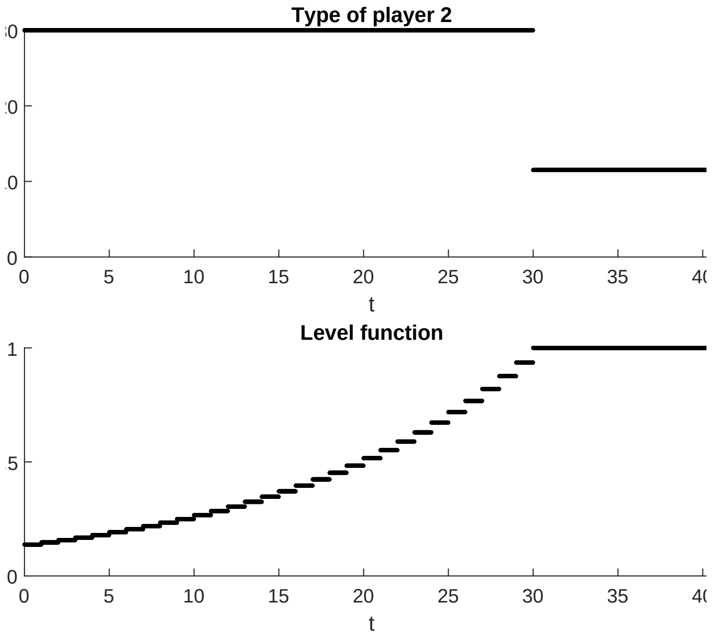


Figure 4: Third equilibrium illustration

equilibrium in the continuation of the game from any period.[13] The concept imposes a form of internal consistency: In a given period k the equilibrium continuation may be altered in any way, so long as in period $k+1$ it returns to a path consistent with the current equilibrium. The new path from period $k+1$ can pick up the current equilibrium as though in any other period $k' \in \{k, k+1, k+2, \ldots, K\}$.

For instance, if $k' = k$ then the players are stalling, essentially postponing the equilibrium path by one period. Any $k' > k+1$ amounts to "jumping ahead" by $k' - k - 1$ periods, and $k' = k+1$ means that the alteration affects only the current period k. Restrictions are inherent in how an equilibrium can be altered in this way. In particular, to pick up on the current equilibrium as though in period k', player 1's belief must be exactly as it would be at the start of period k', so the alteration must specify for the current period k behavior that would lead to such a belief at the end of this period.[14]

[13] That is, we assume that player 1 is the organizational leader; in terms of mechanics, we could imagine that there is pre-play communication at the beginning of each period, players use these messages to coordinate on a continuation path, and only player 1 can speak.

[14] We could allow alterations with $k' < k$, but it would not change the implications of our theory. In this case, Lemma 1 implies $x^{k'} \geq x^{k-1}$. If this inequality is strict then the alteration is not

To focus on what drives our main characterization result and to avoid complicated notation, we shall define alteration-proofness in reference only to continuations of the game *on the equilibrium path.* It is appropriate to also apply alteration-proofness to continuations of the game following public deviations, so that the conditions are imposed both on and off the equilibrium path. In fact, our results extend to this wider application of alteration-proofness, as explained in Section 5 and in the Appendix. The wider imposition of alteration-proofness turns out to not further constrain equilibrium outcomes.

Our alteration-proofness condition is along the lines of the condition developed in Watson (1999) but has two significant advantages. First, Watson (1999) imposes two separate conditions for a stall and a jump, and these are local in nature; our definition here is a single global condition. Second, because Watson (1999) studies a continuous-time model with a jointly selected level, the conditions there are described as limit conditions that go outside the game being analyzed. In the discrete-time framework here, every feasible alteration is an equilibrium in the continuation game.

3.1 The alteration-proofness condition

Consider any trusting PBE, partly characterized by $\{\alpha^k\}_{k=1}^{\infty}$ and $\{x^k\}_{k=0}^{\infty}$, and suppose that period k is reached on the equilibrium path. We imagine that player 1 may dictate that the equilibrium is to be altered in the continuation of the game, in such a way as to have the path of play from period $k+1$ be as though in the original equilibrium from period $k+1+m$, where $m \in \{-1, 0, 1, \ldots, K-k-1\}$ denotes by how many periods the altered equilibrium skips ahead in relation to the original equilibrium from period $k+1$. In the altered equilibrium, play from period k will be described by level and cutoff sequences $\{\tilde{\alpha}^\tau\}_{\tau=k}^{\infty}$ and $\{\tilde{x}^\tau\}_{\tau=k-1}^{\infty}$ where $\tilde{\alpha}^\tau = \alpha^{\tau+m}$ for all $\tau > k$ and $\tilde{x}^\tau = x^{\tau+m}$ for all $\tau \geq k$. Note that the $\tilde{\alpha}^k$ is not nailed down here and so it and m define the alteration.

The level $\tilde{\alpha}^k$ is constrained by the requirement that the original and altered equilibrium continuations fit together in terms of player 2's incentives in period τ. To understand the constraint, observe that the alteration is feasible only if, at the beginning of period $k+1$, player 1's belief is exactly what it would have been at the beginning of period $k+1+m$ in the original equilibrium (on the equilibrium path).

feasible; otherwise, player 1 would strictly prefer the alteration only if an alteration with $k' \geq k$ is strictly preferred.

That is, the continuation game in the altered equilibrium from period $k+1$ must be identical to the continuation game in the original equilibrium from period $k+1+m$. In the latter continuation, player 1's belief about player 2's type is exactly the updated version of F conditional on $x \leq x^{k+m}$ because only these types remain in the game at this point.

Therefore, $\tilde{\alpha}^k$ must be set so that types less than or equal to x^{k+m} prefer to cooperate in period k given the altered level sequence, and types in the interval $(x^{k+m}, x^{k-1}]$ prefer to betray in period k. The first condition is

$$\tilde{\alpha}^k x \leq \Delta\tilde{\alpha}^k + \delta v_2^{k+1+m}(x) \text{ for all } x \in X \text{ such that } x \leq x^{k+m}, \tag{2}$$

and the second is

$$\tilde{\alpha}^k x \geq \Delta\tilde{\alpha}^k + \delta v_2^{k+1+m}(x) \text{ for all } x \in X \text{ such that } x \in (x^{k+m}, x^{k-1}], \tag{3}$$

where v_2^{k+1+m} refers to player 2's continuation value in the original equilibrium.

Note that the constraints can be vacuous depending on how values x^{k-1} and x^{k+m} relate to X. For instance, if no types are scheduled to betray between periods k and $k+m$, which would be the case if $x^{k+m} = x^{k-1}$ or if these values are both between $1/r$ and a, then the second constraint is trivially satisfied.

Definition 4. *Take as given a trusting PBE, with player 2's equilibrium continuation values denoted by* $\{v_2^k(\cdot)\}_{k=1}^{\infty}$. *For any period* k, *integer* $m \in \{-1, 0, 1, \ldots\}$, *and level* $\tilde{\alpha}^k$, *call the triple* $(k, m, \tilde{\alpha}^k)$ *an* **alteration** *of the equilibrium. Call* $(k, m, \tilde{\alpha}^k)$ *a* **feasible alteration** *if Inequalities* (2) *and* (3) *are satisfied.*

Two beneficial alterations were illustrated at the end of Section 2.5. In Figure 3, because all bad types betray in the first period in equilibrium, cooperation in this period would lead player 1 in period 2 to desire the alteration $(2, 28, 1)$; that is, player 1 would jump ahead from period 2 to period 30. In Figure 4, because all bad types betray in period 30 in equilibrium, when period 30 is reached, player 1 would desire an alteration $(30, -1, \alpha^{29})$, effectively going back to period 29 where all types cooperate.

Recall that, for a given trusting PBE and any period k, v_1^k denotes the expected continuation value for player 1 from the start of period k on the equilibrium path, and at the beginning of this period, player 1 believes that player 2's type is weakly below x^{k-1}. Further, after selecting the level α^k that the equilibrium prescribes for period k, player 1 expects player 2 to cooperate with probability $F(x^k)/F(x^{k-1})$ and to betray with complementary probability. We can thus express player 1's expected continuation value recursively as

$$v_1^k = \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right)\left(-c\alpha^k\right) + \frac{F(x^k)}{F(x^{k-1})}\left(\alpha^k \Delta + \delta v_1^{k+1}\right). \tag{4}$$

If at period k player 1 demands that the players coordinate on a feasible alteration $(k, m, \tilde{\alpha}^k)$, then player 1's continuation value would instead be

$$\left(1 - \frac{F(x^{k+m})}{F(x^{k-1})}\right)\left(-c\tilde{\alpha}^k\right) + \frac{F(x^{k+m})}{F(x^{k-1})}\left(\tilde{\alpha}^k \Delta + \delta v_1^{k+1+m}\right).$$

Definition 5. *Call a PBE* **alteration proof** *if it is trusting and no feasible alteration improves player 1's continuation value. That is, for every feasible alteration $(k, m, \tilde{\alpha}^k)$,*

$$v_1^k \geq \left(1 - \frac{F(x^{k+m})}{F(x^{k-1})}\right)\left(-c\tilde{\alpha}^k\right) + \frac{F(x^{k+m})}{F(x^{k-1})}\left(\tilde{\alpha}^k \Delta + \delta v_1^{k+1+m}\right). \tag{5}$$

We refer to these as alteration-proof equilibria.

3.2 Partial characterization

We next partially characterize alteration-proof equilibria. Recall that L denotes the last period in which bad types betray on the equilibrium path, so that $x^{L-1} \geq a \geq x^L$.

Lemma 2. *In every alteration-proof equilibrium, good types never betray and the level is maximal after all bad types have betrayed. That is, $\alpha^k = 1$ for every $k > L$, and it can be assumed that $x^k = a$ for every $k \geq L$.*

Proof. Consider any alteration-proof PBE and let $\eta \equiv \sup\{\alpha^k \mid k > L\}$. We first prove that $\eta = 1$ by assuming otherwise and finding a contradiction. Presuming $\eta < 1$, let $\varepsilon > 0$ be small enough to satisfy $\eta + \varepsilon \leq 1$ and

$$\frac{1}{r} < \frac{\Delta}{1-\delta} + \varepsilon \cdot \frac{\Delta - (1+\delta)(1/r)}{\eta(1-\delta)}. \tag{6}$$

There is such a value of ε because $1/r < \Delta/(1-\delta)$. Let $\tau > L$ be a period at which

$\alpha^\tau > \eta - \varepsilon$, and note that player 1's belief at the beginning of period τ puts positive probability on only good types, which are weakly below $1/r$. By definition of η, we know that $v_1^\tau \leq \eta\Delta/(1-\delta)$.

Then consider alteration $(\tau, -1, \eta+\varepsilon)$. Observe that $(\eta+\varepsilon)x < (\eta+\varepsilon)\Delta + \delta(\eta-\varepsilon)x$ follows from Inequality (6) and $x < 1/r$. Further, the left side is type x's value of betraying immediately in the alteration, whereas the right side is weakly less than the value of waiting until period $\tau+1$ to betray. This implies that all types weakly below $1/r$ strictly prefer to cooperate in period τ given the altered sequence of levels, and so the alteration is feasible. The alteration gives player 1 the continuation value $(\eta+\varepsilon)\Delta + \delta v_1^\tau$, which strictly exceeds v_1^τ, contradicting alteration-proofness.

Having established that $\eta = 1$, we can use a similar argument to show that there is a period $\ell > L$ at which $\alpha^\ell = 1$. If there were no such period, then for any $\varepsilon > 0$ we could find a period $\tau > L$ such that $\alpha^\tau \in (1-\varepsilon, 1)$ where it would have to be the case that $v_1^\tau < \Delta/(1-\delta)$. For sufficiently small ε, the alteration given by $(\tau, -1, 1)$ is feasible and yields player 1 a strictly higher continuation value than v_1^τ. Thus there is a period $\ell > L$ where $\alpha^\ell = 1$. The same logic implies also that $v_1^\ell = \Delta/(1-\delta)$, and so $\alpha^k = 1$ for all $k \geq \ell$.

The penultimate step is to realize that, in the case of $\ell > L+1$, it must also be true that $\alpha^{\ell-1} = 1$ and $v^{\ell-1} = \Delta/(1-\delta)$. This follows from the fact that good types strictly prefer to cooperate in period $\ell - 1$ regardless of the level, given that their continuation value is $\Delta/(1-\delta)$ from period ℓ. If $\alpha^{\ell-1} < 1$ then alteration $(\ell-1, 0, 1)$ is trivially feasible and strictly increases player 1's continuation payoff from period $\ell - 1$. It follows by induction that $\alpha^k = 1$ for all $k > L$. Finally, note that, because $\alpha^L \leq 1$, all good types strictly prefer to cooperate forever rather than betray in period L or any later period. No good type betrays prior to period L in equilibrium, and therefore the good types never betray. □

Lemma 2 establishes that the cutoff sequence $\{x^k\}_{k=0}^\infty$ never falls below $1/r$, and without loss of generality we can assume that $x^0 = b$ and $x^k = \Delta/(1-\delta)$ for every $k \geq L$. That is, the sequence starts at $x^0 = b$ and no bad types betray until the first period k at which $x^k < b$. The last period in which bad types betray is L, where the value of the sequence drops to $\Delta/(1-\delta)$, which is below a, and is then constant.

Lemma 3. *In every alteration-proof equilibrium, $\alpha^k x^k = \Delta\alpha^k + \delta\alpha^{k+1}x^k$ for all $k < L$.*

Recall that this relation between the level sequence and cutoff types was discussed and appears as Equation (1) in the previous section. It means type x^k is indifferent between betraying in period k and betraying in period $k+1$. Rearranging a bit gives an expression for the rate of increase in the level over time, relative to the cutoff type:

$$\frac{\alpha^{k+1}}{\alpha^k} = \frac{x^k - \Delta}{x^k \delta}. \tag{7}$$

The right side strictly exceeds 1, implying that the equilibrium level sequence is strictly increasing. The rate of increase from period to period is itself increasing in the cutoff type and therefore decreasing in k.

Proof of Lemma 3. For convenience in this proof, let us extend v_2^{L+1} to be defined for $x = \Delta/(1-\delta)$ by specifying $v_2^{L+1}(\Delta/(1-\delta)) = \Delta/(1-\delta)$, which would be the continuation value of type $\Delta/(1-\delta)$ in the continuation from period $L+1$ given that the level is 1 thereafter. Of course, there is no type $\Delta/(1-\delta)$ in the model. The extension gives us the starting point for an induction argument.

We begin by proving that, in any alteration-proof equilibrium,

$$v_2^{k+1}(x^k) = \alpha^{k+1} x^k \tag{8}$$

for all $k \leq L$. Note first that this equation holds for $k = L$ because $x^L = \Delta/(1-\delta)$ and $\alpha^{L+1} = 1$. We proceed with an inductive argument.

Suppose that, for a given period $k > 1$, Equation (8) holds. We shall demonstrate that $v_2^k(x^{k-1}) = \alpha^k x^{k-1}$. If $x^{k-1} > x^k$, meaning that type x^{k-1} betrays in period k, we immediately obtain $v_2^k(x^{k-1}) = \alpha^k x^{k-1}$. So let us assume that $x^{k-1} = x^k$, whereby in equilibrium no types betray in period k. Because type x^{k-1} betrays in a future period, it must be that

$$\alpha^k x^{k-1} \leq \Delta\alpha^k + \delta v_2^{k+1}(x^{k-1}) = \Delta\alpha^k + \delta\alpha^{k+1} x^{k-1}. \tag{9}$$

The equality holds because of $x^{k-1} = x^k$ and Equation (8).

Suppose that Inequality (9) is strict. We can find a level $\tilde{\alpha}^k \in (\alpha^k, 1)$ for which, uniquely,

$$\tilde{\alpha}^k x^{k-1} = \Delta\tilde{\alpha} + \delta v_2^{k+1}(x^{k-1}) = \Delta\tilde{\alpha}^k + \delta\alpha^{k+1} x^{k-1}. \tag{10}$$

The existence of this level is implied by the fact that $x^{k-1} > \Delta/(1-\delta) > \Delta$. In fact

$(k, 0, \tilde{\alpha}^k)$ is a feasible alteration. Demonstrating feasibility just requires checking that types below x^{k-1} strictly prefer to cooperate in period k in the alteration, which is straightforward.[15] Because no types betray in period k in the original equilibrium and in the alteration, and because the level in period k is higher in the altered equilibrium, player 1's continuation payoff strictly increases.

Therefore it must be that Inequality (9) holds as an equality. That is, in period k type x^{k-1} is indifferent between betraying and cooperating. This implies that $v_2^k(x^{k-1}) = \alpha^k x^{k-1}$, completing the inductive argument.

Next, using Identity (8) we prove the claim of the lemma. Consider any $k < L$ and let us look at two cases. First, if $x^{k-1} = x^k$ then, by the above argument, weak Inequality (9) binds and we have $\alpha^k x^{k-1} = \Delta\alpha^k + \delta\alpha^{k+1}x^{k-1}$. Replacing x^{k-1} with x^k yields $\alpha^k x^k = \Delta\alpha^k + \delta\alpha^{k+1}x^k$. In the second case, we have $x^{k-1} > x^k$. Then types in the nonempty interval $(x^k, x^{k-1}]$ prefer to betray in period k, whereas types in the nonempty interval $[a, x^k]$ prefer to cooperate. Type x^k must be indifferent between cooperation and betrayal in period k, because player 2's continuation value is continuous in player 2's type for any given level sequence.[16] Hence, $\alpha^k x^k = \Delta\alpha^k + \delta v_2^{k+1}(x^k)$. Using Equation (8) to substitute for $v_2^{k+1}(x^k)$ once again yields $\alpha^k x^k = \Delta\alpha^k + \delta\alpha^{k+1}x^k$. □

Together Lemmas 2 and 3 imply that, in an alteration-proof equilibrium, the level increases gradually until it reaches 1, and then remains at 1 thereafter. The level increases in relation to the rate at which the bad types betray, so that in a given period the cutoff bad type is indifferent between betraying in the current period and betraying in the next period. Good types cooperate forever.

Note that Equation (7) and monontonicity of the sequences $\{\alpha^k\}_{k=1}^{\infty}$ and $\{x^k\}_{k=0}^{\infty}$ are necessary but not sufficient conditions for alteration-proofness. The set of equilibria that satisfy these properties is quite large and varied. For example, Equation (7) holds in the equilibria illustrated in Figures 3 and 4, but these equilibria fail to be alteration-proof.[17]

[15] For $x < x^{k-1}$, $\tilde{\alpha}^k x < \Delta\tilde{\alpha}^k + \delta\alpha^{k+1}x < \Delta\tilde{\alpha} + \delta v_2^{k+1}(x)$ because $(\tilde{\alpha}^k - \delta\alpha^{k+1})x < (\tilde{\alpha}^k - \delta\alpha^{k+1})x^{k-1} = \Delta\tilde{\alpha}^k$.

[16] It is easy to verify that $\omega(k, x)$ is continuous in x for fixed k.

[17] Incidentally, Lemmas 2 and 3 do not indicate exactly the equilibrium level in period L. It is not difficult to show that α^L must be in the interval $(a\delta/(a - \Delta), 1]$. The lower endpoint of this interval would make type a indifferent between betraying at L and waiting to do so at $L+1$, whereas the upper endpoint would make an artificial type $\Delta/(1 - \delta)$ indifferent.

As a step toward our main result, we next derive bounds on equilibrium continuation values for every period $k < L$, by considering alterations in which $m = -1$ or $m = 1$. Our first observation is that in a given period $k < L$ and for $m \in \{-1, 1\}$, the only feasible alteration $(k, m, \tilde{\alpha}^k)$ that player 1 could possibly find attractive is that for which $\tilde{\alpha}^k$ satisfies

$$\tilde{\alpha}^k x^{k+m} = \Delta \tilde{\alpha}^k + \delta \alpha^{k+1+m} x^{k+m}.$$

That is, the alteration is supposed to make type x^{k+m} indifferent, so that any higher types betray in period k and bad types in $[a, x^{k+m}]$ remain into the next period.

We'll use this equation, Identity (4), Equation (7), and Inequality (5). In the case of $m = -1$, we obtain

$$v_1^k \geq \frac{\Delta}{1-\delta} \alpha^{k-1}. \tag{11}$$

In the case of $m = 1$, we get

$$v_1^k \leq \frac{\Delta}{1-\delta} \alpha^k \left(\frac{F(x^k) - F(x^{k-1})}{F(x^{k-1})} \cdot \frac{c}{x^k} + \frac{F(x^k)}{F(x^{k-1})} \right). \tag{12}$$

We thus have upper and lower bounds on player 1's continuation value, which constrain how fast the level increases and bad types betray. For these two conditions to hold, the equilibrium must satisfy

$$\frac{F(x^{k-1}) - F(x^k)}{\Delta F(x^{k-1})} \leq \frac{\alpha^k - \alpha^{k-1}}{(c + \Delta)\alpha^k - \delta c \alpha^{k+1}}. \tag{13}$$

The derivation of Inequalities (11), (12), and (13) is shown in Appendix A.

3.3 Existence and multiplicity

We pointed out earlier that the set of trusting PBE is large. There are even multiple alteration-proof equilibria. The following definition describes a class of equilibria in which v_1^k / α^{k-1} is constant over time:

Definition 6. *Call an alteration-proof equilibrium a* **constant-proportion equilibrium** *(CPE) if there exists a number $\gamma \geq 1$ such that*

$$v_1^k = \gamma \cdot \frac{\Delta}{1-\delta} \alpha^{k-1} \tag{14}$$

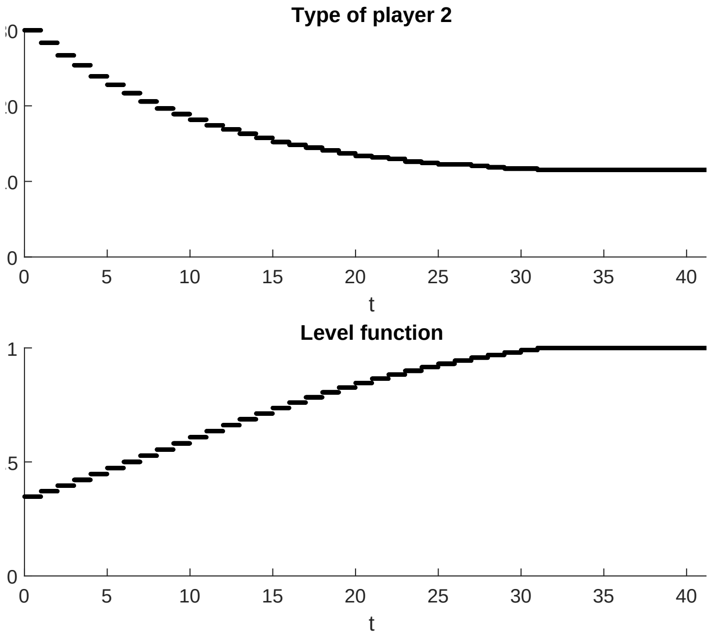


Figure 5: constant-proportion equilibrium with $\gamma = 1$

for all $k \in \{2, 3, \dots, L\}$.

Note that $\gamma = 1$ is the case in which Inequality (11) binds in each period. The next theorem, proved by construction in Appendix B, establishes existence of CPE.

Theorem 2. *Under Assumption 1, there is a number* $\bar{\gamma} > 1$ *such that for all* $\gamma \in [1, \bar{\gamma}]$*, the trust game has a constant-proportion equilibrium with parameter* γ*.*

Let make three remarks on Theorem 2. First, $\bar{\gamma}$ depends on the parameters of the game including Δ. Fixing the other parameters, as Δ decreases toward zero, the value of $\bar{\gamma}$ derived in the proof decreases. Second, a constant-proportion equilibrium with $\gamma = 1$ always exists, regardless of Δ and the other parameters of the game. Substituting $v_1^k/\alpha^{k-1} = v_1^{k+1}/\alpha^k = \Delta/(1-\delta)$ into the equilibrium Identity (4) and rearranging terms yields

$$\frac{\alpha^k - \alpha^{k-1}}{\alpha^k} = -\left(c\frac{1-\delta}{\Delta} + 1\right)\frac{F(x^k) - F(x^{k-1})}{F(x^{k-1})}. \tag{15}$$

This equation, together with indifference Condition (7), characterizes the $\gamma = 1$

constant-proportion equilibrium. Figure 5 illustrates this equilibrium for the same set of parameters used in Section 2.5.

Third, Theorem 2 says nothing about alteration-proof equilibria that are outside the constant-proportion class. We have been able to calculate some other equilibria, but we have been unable to show that they are bounded in some way by constant-proportion equilibria. Thus, although the constant-proportion class provides a useful illustration of alteration-proof equilibria and demonstrates multiplicity, we cannot restrict attention to this class in the next stage of our analysis.

3.4 Characterization theorem

Our main theorem characterizes alteration-proof equilibria as the period length Δ shrinks to zero. Fix a, b, r, and F. We shall consider any sequence of games indexed by positive integer j, where the period length of game j is denoted by $\Delta(j)$, such that $\Delta(j)$ converges to zero as $j \to \infty$. For each game j, we consider an arbitrary alteration-proof equilibrium, described by a level sequence $\{\alpha^k(j)\}_{k=1}^{\infty}$, a sequence of type cutoffs $\{x^k(j)\}_{k=0}^{\infty}$, and a sequence of player 1's continuation values $\{v^k(j)\}_{k=1}^{\infty}$. By Theorem 2, we know that there exists an alteration-proof PBE for each $\Delta(j)$.

To describe what happens as $j \to \infty$, for every j we need to translate the discrete-time equilibrium sequences (levels, type cutoffs, and continuation values) into functions of continuous time. Letting t denote time on the continuum, define $M(t,j) = \min\{k \mid k\Delta(j) \geq t\}$ to be the period in discrete-time game j that contains time t. Then define step functions $\hat{\alpha}(\cdot;j) : [0,\infty] \to [0,1]$, $\hat{x}(\cdot;j) : [0,\infty] \to [0,b]$, and $\hat{v}_1(\cdot;j) : [0,\infty] \to [0,\infty)$ by

$$\hat{\alpha}(t;j) = \alpha^{M(t,j)}(j), \;\; \hat{x}(t;j) = x^{M(t,j)}(j), \;\; \hat{v}_1(t;j) = v_1^{M(t,j)}(j).$$

The theorem establishes that each of these functions converges to a continuous-time limit that is independent of the exact sequence of period lengths and the selection of alteration-proof equilibria. That is, alteration-proofness uniquely pins down the equilibrium when the period length is small. We will denote the limit functions by $\alpha(\cdot)$, $x(\cdot)$, and $v_1(\cdot)$, which alters notation in a way that will hopefully not be confusing. Our theorem also shows that these functions uniquely solve a specific initial-value problem (differential equation) that depends on the parameters.

Theorem 3. *Fix the parameters of a trust game, with the exception of Δ, satisfying Assumption 1. There exist functions $\alpha : [0, \infty) \to [0, 1]$, $x : [0, \infty) \to [0, b]$, and $v_1 : [0, \infty) \to [0, \infty)$, and a positive number T such that the following hold. For any sequence of games given by $\{\Delta(j)\}_{j=1}^{\infty}$, such that $\lim_{j\to\infty} \Delta(j) = 0$, and for any sequence of alteration-proof equilibria given by $\{\hat{\alpha}(\cdot; j), \hat{x}(\cdot; j), \hat{v}_1(\cdot; j)\}_{j=1}^{\infty}$, it is the case that $(\hat{\alpha}(\cdot; j), \hat{x}(\cdot; j), \hat{v}_1(\cdot; j))$ converges uniformly to $(\alpha(\cdot), x(\cdot), v_1(\cdot))$. The limit functions and T are uniquely characterized by:*

(a) Level function $\alpha(\cdot)$ is strictly increasing and differentiable on $(0, T)$, $\alpha(0) > 0$, $\lim\limits_{t\to T^-} \alpha(t) = 1$, and $\alpha(t) = 1$ for every $t \geq T$;

(b) The cutoff-type function $x(\cdot)$ is strictly decreasing and differentiable on $(0, T)$, $x(0) = b$, $\lim\limits_{t\to T^-} x(t) = a$, and $x(t) = 1/r$ for every $t \geq T$; and

(c) On interval $(0, T)$, $\alpha(\cdot)$ and $x(\cdot)$ solve the following system of differential equations:

$$\frac{\alpha'}{\alpha} = -(rc + 1)\frac{f(x)}{F(x)}x', \tag{16}$$

$$\frac{\alpha'}{\alpha} = r - \frac{1}{x} \ . \tag{17}$$

Further, for every $t \geq 0$, player 1's continuation value satisfies $v_1(t) = \alpha(t)/r$.

The existence of multiple and varied alteration-proof equilibria in the discrete-time setting presents a substantial challenge for the limit characterization. Our proof of Theorem 3, provided in Appendix C, uses novel techniques in two steps. First, we use equilibrium identities and alteration-proofness conditions to find bounds on $\{x^k\}_{k=1}^{L}$, working backward from period L and using a calculation that maximizes and minimizes the cutoff type in a given period by making adjustments to the cutoff type in the next period, along with other variables in these periods. We discover a monotone relation on these bounds for successive periods, which allows us to use an inductive argument to construct bounds. Second, we apply the new convergence result proved by Watson (2021) to show that these two sequences of bounds converge uniformly to the same continuous time limit initial-value problem. The corresponding bounds of sequences $\{\alpha^k, v_1^{k+1}\}_{k=1}^{L}$ are constructed similarly. The proof that T is finite depends on the assumption $a > \Delta/(1 - \delta)$, which becomes $a > 1/r$ when $\Delta \to 0$.

Here is a heuristic argument. The implications of alteration-proofness expressed above as Inequalities (11) and (12) provide upper and lower bounds on player 1's continuation values on the equilibrium path. Assuming well-behaved convergence, we have

$$\lim_{j\to\infty}\left[\frac{F(\hat{x}(t;j))-F(\hat{x}(t-\Delta(j);j))}{F(\hat{x}(t-\Delta(j);j))}\cdot\frac{c}{\hat{x}(t;j)}+\frac{F(\hat{x}(t;j))}{F(\hat{x}(t-\Delta(j);j))}\right]=1, \tag{18}$$

and α^k and α^{k-1} converge, so the lower and upper bounds of player 1's continuation value have the same continuous-time limit,

$$\lim_{j\to\infty}\frac{\Delta(j)}{1-e^{-r\Delta(j)}}\hat{\alpha}(t-\Delta(j);j)=\frac{\alpha(t)}{r}.$$

This expression uniquely determines the player 1's continuation value in equilibrium, in relation to the level function. Furthermore, the continuous-time limits of Equation (15) and player 2's indifference Condition (7) lead to Equations (16) and (17), respectively.

The system of equations shown in Theorem 3 can be solved as follows. First, use Equation (17) to substitute for α'/α in Equation (16) to get the following univariate initial-value problem:

$$\frac{dx}{dt}=\frac{(1-rx)F(x)}{x(rc+1)f(x)},\quad x(0)=b. \tag{19}$$

Denote

$$I_x(x)=\int\frac{x(rc+1)f(x)}{(1-rx)F(x)}dx. \tag{20}$$

Then we solve Equation (19) to obtain $x(t)=I_x^{-1}\left(I_x\left(b\right)+t\right)$. Second, to calculate T, we use the terminal condition $a=I_x^{-1}\left(I_x\left(b\right)+T\right)$. Third, we substitute the solution $x(t)$ into Equation (17) to obtain the following univariate initial-value problem:

$$\frac{d\alpha}{\alpha}=\left(r-\frac{1}{x(t)}\right)dt,\quad \alpha(T)=1, \tag{21}$$

which yields $\alpha(t)=\exp\left(I_\alpha(t)-I_\alpha(T)\right)$, where

$$I_\alpha(t)=\int\left(r-\frac{1}{I_x^{-1}\left(I_x\left(b\right)+t\right)}\right)dt.$$

Last, we evaluate the level function at time 0 to obtain $\alpha(0)=\exp\left(I_\alpha(0)-I_\alpha(T)\right)$.

Note that for any distribution F satisfying Assumption 1, the system of differential equations in Theorem 3 can be easily solved numerically. Furthermore, if the distribution function F is in polynomial form on the interval $[a, b]$, so that $F(x) = a_1 x + a_2 x^2 + \ldots + a_n x^n$ for real numbers $a_1, a_2, ..., a_3$, a closed-form analytical solution can be derived from a potentially complicated integration.

4 Uniformly Distributed Bad Types

In this section, we provide an example of the limit of alteration-proof equilibria when the bad types of player 2 are uniformly distributed. Fix $c = 1$ and denote the probability of the good types as $q \equiv F(1/r)$. For any bad type $x \in [a, b]$, we have $F(x) = q + (1-q)(x-a)/(b-a)$ and the density function is $f(x) = (1-q)/(b-a)$. In this special case, we can use the algorithm to solve the equilibrium analytically.

In this example, Initial-Value Problems (19) and (21) become

$$\frac{dx}{dt} = \frac{1-rx}{r+1} \cdot \frac{(b-a)q + (x-a)(1-q)}{(1-q)x}, \quad x(0) = b \tag{22}$$

$$\frac{d\alpha}{\alpha} = \left(r - \frac{1}{x}\right) dt, \quad \alpha(T) = 1. \tag{23}$$

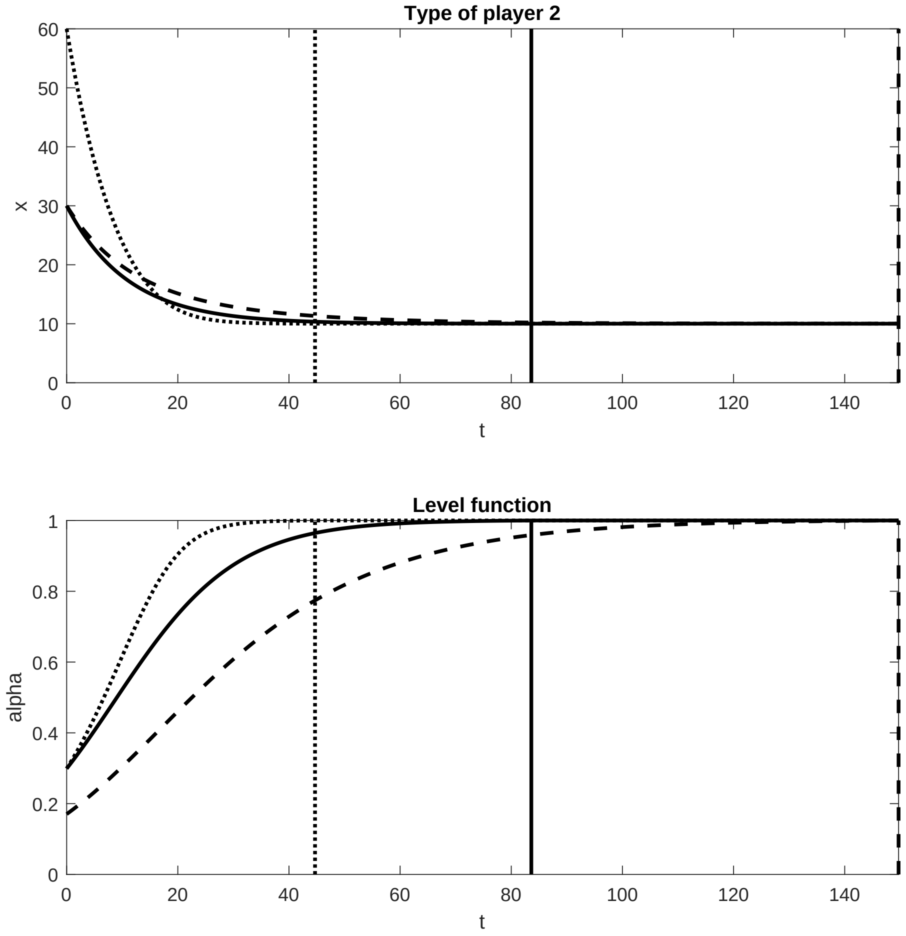
Solving the equations, we have the explicit characterization of the equilibrium:

$$t = \frac{(1-q)(1+r)}{1-q+r(bq-a)} \left(-\frac{1}{r} \ln \frac{1-rx}{1-rb} + \frac{a-bq}{1-q} \ln \frac{(1-q)x + bq - a}{b-a} \right), \tag{24}$$

$$\alpha(t) = \left(\frac{q(b-a)}{(1-q)x(t) + bq - a} \right)^{r+1}, \tag{25}$$

$$T = \frac{(1-q)(1+r)}{1-q+r(bq-a)} \left(-\frac{1}{r} \ln \frac{1-ra}{1-rb} + \frac{a-bq}{1-q} \ln q \right), \quad \alpha(0) = q^{r+1}. \tag{26}$$

Figure 6 graphs the limit level and cutoff-type functions for three cases of parameter values where r, a, and c are fixed. The figure illustrates comparative statics with respect to q and b: Starting with the solid curve, the dashed curve shows the effect of decreasing q while the dotted curve shows the effect of increasing b. Straightforward calculations in Appendix D produce the following comparative statics conclusions. Regarding the last two statements, for any type $\chi \in [a, b]$ we define $\Gamma(\chi)$ to be the time at which the cutoff type is χ; that is, $x(\Gamma(\chi)) = \chi$.



Continuous-time level-curve limit for alteration-proof equilibria with parameters $r = 0.1$, $a = 10.0105$, and $c = 1$. For the comparative statics, we use the following parameters: $q = 0.3337$ and $b = 30$ for solid curves, $q = 0.2002$ and $b = 30$ for dashed curves, and $q = 0.3337$ and $b = 60$ for dotted curves. Vertical lines highlight the time T when the level first hits 1 in each equilibrium.

Figure 6: The limit of alteration-proof equilibria.

Proposition 1. *Consider the case of uniformly distributed bad types and let q denote the probability of the good type. The equilibrium limit has the following properties, where T denotes the time when the level first reaches 1.*

- $\partial T/\partial q < 0$, $\partial T/\partial b < 0$, $\partial \alpha(0)/\partial q > 0$, *and* $\partial \alpha(0)/\partial b = 0$.
- *For a fixed type $\chi \in [a, b]$, the slope of $x(\cdot)$ at time $\Gamma(\chi)$ is decreasing in q and b, and the same is true for the slope of $\ln \alpha(t)$ at time $\Gamma(\chi)$.*

Player 1's equilibrium payoff is increasing in the quality of player 2's type distribution. Lowering the probability of the good type causes player 1 needs to start the relationship at a smaller level and gradualism slows, so it takes longer to build trust. Increasing b, the worst possible type of player 2, has the same implications.

This result may have empirical implications. For example, consider the interaction between a venture capitalist (player 1) and an entrepreneur (player 2). The venture capitalist controls the investment in a project in successive periods, which is like selecting α in our model. The entrepreneur chooses how to allocate the funds, either to productive use (cooperate) or skewed to private benefit (betray). Based on the model, we would expect that the venture capitalist starts small and gradually increases funding before taking the concern public. If the project is in an industry with higher informational barriers, due for instance to sophisticated technologies or geographic distance, then we predict a low initial investment and long period before a public offering. These implications of Proposition 1 are consistent with empirical studies of venture-capital staged financing by Gompers (1995) and Tian (2011).

5 Technical Notes and Extensions

In this section, we elaborate on the foundations of the model, we discuss extensions of the modeling exercise related to both the alteration-proofness concept and the parameters, and we comment on connections with literature.

5.1 Further selection and foundations

In this subsection, we elaborate on the scope and foundation of alteration-proofness. First, recall that our definitions in Section 3 impose alteration-proofness only on the equilibrium path. To the extent that renegotiation is also plausible in off-equilibrium contingencies, we should further require alteration-proofness after histories in which player 1 deviated (the only histories entailing a public deviation in a trusting PBE).

In fact, the equilibria that we construct to prove Theorem 2 satisfy the extended version of alteration-proofness. In a period in which player 1 deviates, every type of player 2 is supposed to betray, ending the game. If player 2 deviates by cooperating, then player 1 subsequently believes with certainty that player 2 is a bad type, and the strategies specify that player 1 set $\alpha = 0$ and player 2 betray in all future periods regardless of the interim history. From every period following player 1's deviation, there are no trusting equilibria in the continuation game due to player 1's posterior belief, and therefore no feasible alteration can give player 1 a positive payoff.

Second, recall that we defined alteration-proofness for trusting equilibria, and this is the class of equilibria to which our convergence result applies. There are also non-trusting equilibria in which player 1 always chooses level 0. Further, using the logic noted in the previous paragraph, we can find a non-trusting equilibrium that satisfies the alteration-proofness conditions. This leads to the question of whether there is an argument along the lines of alteration-proofness that would rule out non-trusting equilibria, so that our convergence implies a unique selection.

Non-trusting equilibria would not survive a reasonable notion of *external consistency* imposed on top of our alteration-proofness concept, because player 1 and every type of player 2 strictly prefer the alteration-proof PBE that we construct for Theorem 2 to every non-trusting equilibrium (which gives a payoff of zero to both players), both at the beginning of the relationship and in later periods. For instance, suppose in addition to the alterations studied in Section 3, the players view as viable a suggestion to switch to another alteration-proof equilibrium that would improve the payoff of every player-type by at least ε for some fixed $\varepsilon > 0$. Then for sufficiently small Δ, none of the alteration-proof equilibria could dominate others in this way (due to Theorem 3) but they all dominate every non-trusting equilibrium.[18]

Continuing with the topic of non-trusting equilibria, it is worth noting that the existence of such equilibria depends on good types having positive betrayal benefits ($x > 0$), so that they have an incentive to betray at positive levels if they expect that the level would be zero in future periods. We conjecture that inclusion of good types with $x \leq 0$ would narrow the set of PBE in interesting ways without even imposing alteration-proofness. This may be a good topic for future work. We have been able to show that, in this setting, there is no non-trusting equilibrium, and all good types cooperate forever in every PBE.[19] This is also true for continuations following a deviation of player 1. Therefore, in any equilibrium construction, we cannot spec-

[18] This comparison is similar to the "Pareto external agreement consistency" condition of Miller and Watson (2013). We utilize the bound ε because we have not determined whether and how alteration-proof equilibria are ranked for fixed Δ.

[19] From a non-trusting strategy profile, if player 1 deviates to a positive level, good types with $x \leq 0$ optimally must cooperate regardless of the anticipated future play. If player 1 deviates in this way and player 2 then cooperates, the probability that player 1 puts on the good types must weakly increase.

ify reversion to a non-trusting path to punish player 1, complicating the argument for existence. As a bridge to further research, in the Supplementary Appendix, we provide a new existence result that does not rely on non-trusting equilibria following deviations.

Finally, it would not be difficult to provide non-cooperative foundations for the alteration-proofness condition, along the lines of Watson (2013) and Miller and Watson (2013). We could model renegotiation at the beginning of each period as a simple dictator game, whereby player 1 has the option to declare an alteration, which the players then coordinate on if feasible. In fact, since, in the construction behind Theorem 4, deviations in the level chosen for the current period are associated with feasible alterations, one could imagine that player 1 triggers an alteration by simply deviating from the level that the equilibrium specifies for the current period.

5.2 No-gap case

Recall that we assumed distribution F is constant in an open interval containing $\Delta/(1-\delta)$; that is, there is a gap at the value where player 2 would be indifferent between betraying and cooperating if the level were constant. Our analysis can be extended to the "no-gap" case, which we can describe by setting $a = 1/r$ and where bad types are those above $\Delta/(1-\delta)$. Without going into details, here is a summary of our findings in this case.

The PBE characterization developed in Section 2 carries over to the no-gap case except that Lemma 1 must be modified to allow $L = \infty$. The intuition is that bad types with x close to $\Delta/(1-\delta)$ are willing to cooperate in periods where α rises slowly, so a strictly increasing sequence $\{\alpha^k\}$ that converges to a number in the interval $(0, 1]$ is associated with a strictly decreasing sequence of cutoff types $\{x^k\}$ that converges to $\Delta/(1-\delta)$. Thus, every bad type betrays at some point, but there are PBE in which, in every period, some bad types have yet to betray.[20]

Likewise, the analysis of alteration-proofness extends to the no-gap case. Lemmas 2 and 3 hold, allowing for $L = \infty$. Calculations underlying our convergence theorem are valid, but in the limit T becomes infinite and Expression (20) becomes an improper integral, with terminal condition $x^L = a = \Delta/(1-\delta)$ at $L = \infty$. The convergence result now identifies a class of solutions that is unique up to a constant

[20] The no gap case here is similar to the no-gap case of the durable-good-monopoly model; see Ausubel and Deneckere (1989).

$\alpha^* \in (0, 1]$. Specifically, in the limit as $\Delta \to 0$, alteration-proof equilibria all share the same path of cutoff types. For every $\alpha^* \in (0, 1]$ there is a sequence of alteration-proof equilibria whose level sequences converge to a function that approaches α^* as $t \to \infty$.

This conclusion is in contrast to alteration-proofness in the gap case, where the level converges to 1 (meaning $\alpha^* = 1$ for every alteration-proof equilibrium) and does so in finite time. But the equilibria are Pareto-ranked in α^*, so an appeal to external consistency would justify selection of the equilibria that entail $\alpha^* = 1$ as a unique limit prediction. Another way to restore our unique prediction is to expand the definition of "feasible alteration" to include player 1 setting $\alpha = 1$ in every period of the continuation, with good types cooperating perpetually and bad types betraying immediately.[21] In an equilibrium with $\alpha^* < 1$, player 1 would strictly prefer such an alteration once only a small fraction of bad types remain. Further, in line with the various notes made here about including good types with $x \leq 0$, we conjecture that in settings with such types, there would be no substantive difference between the analysis of the gap case and the no-gap case.

5.3 More on related literature

Let us expand a bit on our discussion of related literature in the Introduction. As noted, the most closely related paper is Watson (1999). While our modeling exercise shares heuristic logic with Watson's modeling exercise, they have distinct structures and very different analytical approaches.

Watson's model has an exogenously provided level function and therefore is not a noncooperative game, whereas we develop a fully noncooperative model of a principal-agent setting in which the principal selects the level in each period. We define alteration-proofness with respect to actual alternative equilibria in the continuation of the game from any period, and we are able also to study alteration-proofness at contingencies off the equilibrium path. Further, we study a discrete-time setting with a continuum of types, whereas Watson (1999) looked at a continuous-time setting with just two types. The analytical methods developed herein are novel and unique to our setting, with particular challenges owing to discrete time. The analysis of convergence (as well as the existence of multiple alteration-proof equilibria) has no

[21] Such an alteration is not included in the definition of "feasible" given in Section 3 because the continuation from the next period does not coincide with any continuation on the path of the current equilibrium.

counterpart in Watson (1999). Our work in this regard involves a new method of characterizing bounds on the set of equilibria and the first application of a result on the limit of solutions to discrete-time models.

A number of past game theoretic modeling exercises have delivered notable results on the convergence of equilibrium outcomes as the frequency of interaction increases. For instance, Gul, Sonnenschein, and Wilson (1986) substantiate the Coase conjecture on dynamic pricing by a durable-good monopolist (similarly bargaining under incomplete information with one-sided offers by the uninformed-party). The seller in a given period evaluates the rate at which sales increase as the price is lowered (attracting additional lower-valuation buyers) and this interacts with the relative patience of different buyer types. As the period length shrinks, the balance tilts in favor of the seller's interest in lowering the price to increase present-period sales, and in the limit the seller sets a low price from the beginning.

There is a similar trade-off in our model. By raising the level in an equilibrium alteration, the principal can induce more bad types to betray, hastening the time when the principal enjoys cooperation with the good types. One might expect, in line with the Coase conjecture, that as the period length shrinks, the principal would be resigned to start with a high level and suffer the consequences in the event of a bad type of player 2. This is not the case, however, because the relative intertemporal trade-offs for the principal and agent in the trust game are different than for the seller and buyer in the dynamic pricing game. The principal's choice of project level scales the payoffs of both players rather than splitting the the surplus for the seller and buyer, and the agent's action is a choice of whether to divert gains at the principal's expense. Importantly, betrayal by the bad types imposes a loss on the principal, whereas selling to a high-valuation buyer at a low price still generates value to a monopolist. Further, in periods when the level is low, the principal still earns a cooperative flow payoff from the good types, whereas for the durable-good monopolist, pricing high earns no flow benefit from low-valuation buyers. For these reasons, the principal's incentive to increase the level is tempered as the period length shrinks, and the result is gradualism.

Also relevant to the theme of frequent play is the topic of "repeated games with incomplete information" (though technically not repeated games), such as analyzed by Hart (1985) and Shalev (1994). One strand of this literature looks at how incomplete information about stage-game payoffs leads to a "reputation-based refinement" of equilibrium predictions relative to a complete-information benchmark in settings with sufficiently patient players.[22] In the typical setting, a long-run player with private information is, in the limit, infinitely more patient than the other players (a special case being a sequence of short-run players). Any type of long-run player can patiently pretend to be any other type, and the short-run players must either best respond to the mimicked type's strategy in the short run or "learn" that the long-run player is this type, which implies bounds on the long-run player's payoffs. Our setting is quite different because the flow payoff of cooperation is scaled by the period length, whereas the terminal payoff of betrayal is not scaled. The intertemporal incentives are fixed (the discount rate is held constant) while we shrink the period length. This smooths the equilibrium behavior, but neither player becomes infinitely more patient than the other, and so there is no extreme reputation effect.[23]

5.4 Other extensions

Our analysis took place under the assumption that $c > 0$, but it easily extends to the case of $c = 0$, and curiously it still has interesting things to say. With no cost of betrayal, player 1's evaluation of alterations trades off discounting with the probability of cooperation. Player 1 prefers that bad types not betray immediately and so starts small. Alteration-proofness implies a unique equilibrium in discrete time, where player 1 is always indifferent between continuing on the equilibrium path and altering with $m = -1$ and $m = 1$. The limit differential equations can be derived from the expressions shown in Theorem 3 by setting $c = 0$.

The assumption that player 1 can dictate the selection of an equilibrium alteration makes alteration-proofness a tight condition. A renegotiation-proofness condition requiring agreement between the players would be weaker, because there are stall alterations that appeal to player 1 but would not appeal to any type of player 2. Any alteration with $m > 0$ that is desired by player 1 would also be desired by all types of player 2. If we were to assume that player 2 dictates the terms of alterations, then in an alteration-proof equilibrium, the level would start higher and rise at a rate that

[22] Prominent entries include Fudenberg and Levine (1989), Cripps, Schmidt, and Thomas (1996), and Cripps and Thomas (2003). Watson (1993) and Battigalli and Watson (1997) examine implications for rationalizable beliefs. Pei (2021) explores reputations based on randomization that straddles multiple types.

[23] Another line in the literature on incomplete information examines the effect of "behavioral types," but this is farther afield from our project.

holds player 1's continuation value to 0 until all bad types have betrayed.

That the game ends after betrayal is an assumption made for convenience. We think that our results would not be different in a model in which play continues following betrayal. For instance, to extend the trusting equilibria we have studied, we need to deal with histories in which betrayal has occurred in the past. If the betrayal occurred on the equilibrium path, player 1's updated belief would be that player 2 is a bad type and player 1 would choose $\alpha = 0$ thereafter. For a public-deviation betrayal, such as in a period in which all types were supposed to cooperate at the equilibrium level, then we could specify that player 1's posterior belief is concentrated on bad types and player 1 selects $\alpha = 0$ thereafter. Further, if player 1 were to deviate, we could prescribe that all types of player 2 betray in the current period, so player 1's belief is unchanged after this betrayal, and continuation play from the next period is exactly what the players were supposed to do from the current period (a stall of sorts). These provisions would not complicate alteration-proofness.

6 Conclusion

We have added to the literature on relationship building by characterizing alteration-proof equilibria in a discrete-time principal-agent setting with a continuum of bad types. We hope that our closed-form characterization of equilibrium will motivate further analysis of the dynamics of relationships under asymmetric information, in particular in more applied settings where multidimensional realistic ingredients are modeled (such as production technology and monitoring). We think the our alteration-proofness condition may be usefully applied to other settings with incomplete information where a notion of internal consistency is desired (further expanding beyond the standard application of repeated games). A key to its applicability is that posterior beliefs have a threshold form and are monotone over time in equilibrium. In dynamic games with this property, it may be possible to describe an altered equilibrium path in terms of an adjustment in one period and a continuation that essentially jumps ahead or stalls relative to the original equilibrium.

Appendix A Derivation of (11), (12) and (13)

In the original equilibrium, player 1's continuation value from period k satisfies

$$v_1^k = \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right)\left(-c\alpha^k\right) + \frac{F(x^k)}{F(x^{k-1})}\left(\alpha^k \Delta + \delta v_1^{k+1}\right).$$

In the main text, we argue that if player 1 demands an alteration $m \in \{-1.1\}$ in period k, then it suffices to consider alteration (k, m, α^{k+m}).

First consider the case in which $m = -1$. If at period k player 1 demands that the players coordinate on a feasible alteration $(k, -1, \alpha^{k-1})$, then player 1's continuation value would instead be

$$\left(1 - \frac{F(x^{k-1})}{F(x^{k-1})}\right)\left(-c\alpha^{k-1}\right) + \frac{F(x^{k-1})}{F(x^{k-1})}\left(\alpha^{k-1}\Delta + \delta v_1^k\right) = \alpha^{k-1}\Delta + \delta v_1^k.$$

Therefore, alteration-proofness Condition (5) in this case simplifies to

$$v_1^k \geq \alpha^{k-1}\Delta + \delta v_1^k,$$

which yields Inequality (11).

Second consider the case of $m = 1$. If at period k player 1 demands that the players coordinate on a feasible alteration $(k, 1, \alpha^{k+1})$, then player 2 with type $x \in (x^k, x^{k-1})$ betrays at level α^{k+1}, while player 1 and player 2 with type $x \leq x^k$ continue in period k as if they were in period $k+1$ of the original equilibrium. Therefore, player 1's continuation value is

$$v_1^{k+1}\frac{F(x^k)}{F(x^{k-1})} - \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right)c\alpha^{k+1}, \tag{27}$$

and we need this continuation value to be no greater than the continuation value without alteration, v_1^k.

To apply the Condition (5), we need to express x_1^{k+1} in terms v_1^{k+1}. From Equation (4), we can solve for v_1^{k+1} and get

$$v_1^{k+1} = \frac{1}{\delta}\left(v_1^k + \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right)c\alpha^k - \frac{F(x^k)}{F(x^{k-1})}\Delta\alpha^k\right)\frac{F(x^{k-1})}{F(x^k)}.$$

Using this to substitute for v_1^{k+1} in Expression (27) and comparing it with v_1^k, we obtain the following alteration-proofness condition:

$$v_1^k \geq \frac{1}{\delta}\left(v_1^k + \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right) c\alpha^k - \frac{F(x^k)}{F(x^{k-1})}\Delta\alpha^k\right) - \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right) c\alpha^{k+1}.$$

Then we multiply both sides by $-\delta$ and add v_1^k to get

$$-(1-\delta)v_1^k \geq \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right) c\alpha^k - \frac{F(x^k)}{F(x^{k-1})}\Delta\alpha^k - \delta\left(1 - \frac{F(x^k)}{F(x^{k-1})}\right) c\alpha^{k+1},$$

which simplifies to

$$v_1^k \leq \frac{1}{1-\delta}\left[-(\alpha^k - \delta\alpha^{k+1})c + \frac{F(x^k)}{F(x^{k-1})}\left(\alpha^k - \delta\alpha^{k+1}\right)c + \frac{F(x^k)}{F(x^{k-1})}\alpha^k\Delta\right].$$

Notice that Equation (7) can be written $\alpha^k - \delta\alpha^{k+1} = \Delta\alpha^k/x^k$. This substitution yields Inequality (12).

Last, we combine Inequalities (11) and (12). Note that player 1's continuation value satisfying both conditions only if

$$\frac{\Delta}{1-\delta}\alpha^{k-1} \leq \frac{\Delta}{1-\delta}\alpha^k\left(\frac{F(x^k) - F(x^{k-1})}{F(x^{k-1})}\frac{c}{x^k} + \frac{F(x^k)}{F(x^{k-1})}\right).$$

Using player 2's indifference Condition (7) and rearranging terms, we obtain Inequality (13).

Appendix B Proof of Theorem 2

Local and global alteration-proofness

We start with a useful lemma on the sufficiency of local alteration-proofness, which refers to values $m \in \{-1, 0, 1\}$.

Lemma 4. *If a trusting equilibrium is "locally alteration-proof," defined as applying the conditions for $m \in \{-1, 0, 1\}$ only, then it is alteration-proof.*

Proof. For any integer $m \geq 0$, we first write player 1's continuation value v_1^k in the original equilibrium as

$$v_1^k = \sum_{n=0}^{m} \left((c+\Delta)\frac{F(x^{k+n})}{F(x^{k-1})} - c\frac{F(x^{k+n-1})}{F(x^{k-1})} \right) \delta^n \alpha^{k+n} + \delta^{m+1} \frac{F(x^{k+m})}{F(x^{k-1})} v_1^{k+m+1}.$$

Denote player 1's continuation value in a $(\tau, m, \alpha^{\tau+m})$ alteration as $\tilde{v}_1^\tau(m)$:

$$\tilde{v}_1^k(m) = \left((c+\Delta)\frac{F(x^{k+m})}{F(x^{k-1})} - c \right) \alpha^{k+m} + \delta \frac{F(x^{k+m})}{F(x^{k-1})} v_1^{k+m+1}.$$

We will show that if player 1 has no incentive to alter the equilibrium from period k by jumping to period $k+m-1$ and also no incentive to alter from period $k+m-1$ by jumping to period $k+m$, then player 1 also has no incentive to alter from period k by jumping to period $k+m$. We prove this by contradiction: assuming $v_1^k \geq \tilde{v}_1^k(m-1)$ and $v_1^k < \tilde{v}_1^k(m)$, we will derive $v_1^{k+m-1} < \tilde{v}_1^{k+m-1}(1)$ meaning that, in the original equilibrium, a local alteration-proofness condition in period $k+m-1$ is violated.

Alteration-proofness condition $v_1^k \geq \tilde{v}_1^k(m-1)$ can be simplified to

$$(1-\delta^{m-1})\delta \frac{F(x^{k+m-1})}{F(x^{k-1})} v_1^{k+m} \leq \sum_{n=0}^{m-1} \left((c+\Delta)\frac{F(x^{k+n})}{F(x^{k-1})} - c\frac{F(x^{k+n-1})}{F(x^{k-1})} \right) \delta^n \alpha^{k+n} - \left((c+\Delta)\frac{F(x^{k+m-1})}{F(x^{k-1})} - c \right) \alpha^{k+m-1}. \quad (28)$$

Likewise, alteration-proofness condition $v_1^k < \tilde{v}_1^k(m)$ can be simplified to

$$(1-\delta^{m})\delta \frac{F(x^{k+m})}{F(x^{k-1})} v_1^{k+m+1} > \sum_{n=0}^{m} \left((c+\Delta)\frac{F(x^{k+n})}{F(x^{k-1})} - c\frac{F(x^{k+n-1})}{F(x^{k-1})} \right) \delta^n \alpha^{k+n} - \left((c+\Delta)\frac{F(x^{k+m})}{F(x^{k-1})} - c \right) \alpha^{k+m}. \quad (29)$$

Next we rewrite the equilibrium Identity (4) for period $k+m$,

$$v_1^{k+m} = \left((c+\Delta)\frac{F(x^{k+m})}{F(x^{k+m-1})} - c \right) \alpha^{k+m} + \frac{F(x^{k+m})}{F(x^{k+m-1})} \delta v_1^{k+m+1},$$

and then multiply both sides by $(1-\delta^m)F(x^{k+m-1})/F(x^{k-1})$, allowing us to substitute the lower bound of $(1-\delta^m)\delta v_1^{k+m+1} F(x^{k+m})/F(x^{k-1})$ using Inequality (29). After collecting terms, we get a lower bound of v_1^{k+m}:

$$v_1^{k+m}(1-\delta^m)\frac{F(x^{k+m-1})}{F(x^{k-1})} > \left(c - c\frac{F(x^{k+m-1})}{F(x^{k-1})}\right)\alpha^{k+m} + \sum_{n=0}^{m-1}\left((c+\Delta)\frac{F(x^{k+n})}{F(x^{k-1})} - c\frac{F(x^{k+n-1})}{F(x^{k-1})}\right)\delta^n\alpha^{k+n}. \quad (30)$$

Note that the summation term on the right here also appears in Inequality (28). Using Inequality (28) to substitute for the summation term in Inequality (30) yields the following:

$$v_1^{k+m}(1-\delta^m)\frac{F(x^{k+m-1})}{F(x^{k-1})} > \left(c - c\frac{F(x^{k+m-1})}{F(x^{k-1})}\right)\alpha^{k+m} + (\delta-\delta^m)\frac{F(x^{k+m-1})}{F(x^{k-1})}v_1^{k+m} + \left((c+\Delta)\frac{F(x^{k+m-1})}{F(x^{k-1})} - c\right)\alpha^{k+m-1}.$$

Applying Condition (7) to substitute $\alpha^{k+m-1}(x^{k+m-1}-\Delta)/(x^{k+m-1}\delta)$ for α^{k+m}, this inequality simplifies to

$$v_1^{k+m} > \alpha^{k+m-1}\left(\frac{\Delta}{1-\delta} + \frac{F(x^{k-1})-F(x^{k+m-1})}{F(x^{k+m-1})}\cdot\frac{x^{k+m-1}-\Delta/(1-\delta)}{x^{k+m-1}\delta}c\right). \quad (31)$$

On the other hand, by Inequality (12), local alteration-proofness in period $k+m-1$ requires

$$v_1^{k+m-1} \leq \frac{\Delta}{1-\delta}\alpha^{k+m-1}\left(\frac{F(x^{k+m-1})-F(x^{k+m-2})}{F(x^{k+m-2})}\cdot\frac{c}{x^{k+m-1}} + \frac{F(x^{k+m-1})}{F(x^{k+m-2})}\right),$$

which translates into an upper bound on v_1^{k+m} using the period $k+m-1$ equilibrium Identity

$$v_1^{k+m-1} = \left((c+\Delta)\frac{F(x^{k+m-1})}{F(x^{k+m})} - c\right)\alpha^{k+m-1} + \frac{F(x^{k+m-1})}{F(x^{k+m-2})}\delta v_1^{k+m}.$$

Therefore, local alteration-proofness requires

$$v_1^{k+m} \leq \alpha^{k+m-1}\left(\frac{\Delta}{1-\delta} + \frac{F(x^{k+m-2})-F(x^{k+m-1})}{F(x^{k+m-1})}\cdot\frac{x^{k+m-1}-\Delta/(1-\delta)}{\delta x^{k+m-1}}c\right). \quad (32)$$

Recall Inequalities (28) and (29), together with equilibrium Conditions (4) and (7), imply that v_1^{k+m} is bounded below by the right side of Inequality (31). We also showed that the local alteration-proofness Condition (12) and equilibrium Identity (4) imply that v_1^{k+m} is bounded above by the right side of Inequality (32).

However, for all $m \geq 1$, we have $x^{k+m-2} \leq x^{k-1}$, so the upper bound of v_1^{k+m} (the right side of Inequality (32)) is weakly lower than the lower bound of v_1^{k+m} (the right side of Inequality (31)), which is a contradiction. Therefore, we conclude that if player 1 has no incentive to jump from period k to period $k+m-1$, and player 1's local alteration-proof conditions are satisfied, then player 1 must have no incentive to jump from period k to period $k+m$. Applying this argument recursively for $m \in \{1, 2, ...\}$, we have proved that if player 1's alteration-proof conditions for $m \in \{-1, 0, 1\}$ are satisfied, then the global conditions for $m > 1$ are also satisfied. □

PBE construction: equilibrium sequence

Fix any trust game with period length $\Delta < \overline{\Delta}$. Define

$$\tilde{\gamma} \equiv \frac{a - \Delta}{a\delta}, \tag{33}$$

and consider any $\gamma \in [1, \tilde{\gamma}]$. We will construct a constant-proportion (alteration-proof) equilibrium with constant of proportion γ.

We begin by defining a sequence $\{x^k, \alpha^k, v_1^{k+1}\}_{k=0}^{L}$, where L will be determined in the construction. For $k > 0$ the meaning of x^k, α^k, and v_1^k is the same as in the text; α^k is the level chosen in period k on the equilibrium path, x^k is the type cutoff, and v_1^k is player 1's continuation value. The sequence gives these values from period 1 up to the last period L in which bad types betray in equilibrium. The initial cutoff type x^0 will equal b as with every PBE, whereas α^0 will be an artificial value that aids in the equilibrium construction. The sequence is defined by induction, starting with $k = L$ and working backward in time. The period offset for v_1 helps to organize the variables in the recursive step.

The inductive procedure uses Equations (4), (7) and (14), which we restate here:

$$v_1^k = \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right)\left(-c\alpha^k\right) + \frac{F(x^k)}{F(x^{k-1})}\left(\alpha^k\Delta + \delta v_1^{k+1}\right), \tag{34}$$

$$\frac{\alpha^k}{\alpha^{k-1}} = \frac{x^{k-1} - \Delta}{x^{k-1}\delta}, \tag{35}$$

$$v_1^k = \gamma \cdot \frac{\Delta}{1-\delta}\alpha^{k-1}, \tag{36}$$

for $k \in \{1, 2, ..., L\}$. Recall that the first equation is the identity relating player 1's continuation values in adjacent periods, the second equation is the indifference condition for the cutoff type (here stated for periods $k-1$ and k), and the third is the constant-proportion condition that defines the equilibrium we are working to form.

We next describe the induction procedure to construct the sequence. Letting L be an arbitrary positive integer, set $x^L = a$, $\alpha^L = 1$, and $v_1^{L+1} = \Delta/(1-\delta)$. Then, for any k for which $(x^k, \alpha^k, v_1^{k+1})$ has been set and if the procedure has not yet terminated, derive $(x^{k-1}, \alpha^{k-1}, v_1^k)$ as follows. If for the given $(x^k, \alpha^k, v_1^{k+1})$, there is a vector $(x^{k-1}, \alpha^{k-1}, v_1^k)$ that solves the system of Equations (34)-(36), then $(x^{k-1}, \alpha^{k-1}, v_1^k)$ is taken to be this vector; further, if $x^{k-1} < b$ then the procedure continues by lowering the value of k by one unit and restarting the calculations. Otherwise (if there is no solution or if the solution has $x^{k-1} = b$), set $x^{k-1} = b$, set $\alpha^{k-1} = \delta\alpha^k b/(b-\Delta)$, redefine L so that the current value of k is 1, and terminate the procedure.[24]

The construction can be put in terms of the type cutoffs only, where a transition function relates x^k to x^{k-1}, without reference to the other variables α and v_1. Here are the calculations that yield the transition function:

For $k = L$, we plug in the terminal values $\alpha^L = 1$, $x^L = a$, and $v_1^{L+1} = \Delta/(1-\delta)$ into Equations (34)-(36), use Equation (35) to substitute for α^{k-1} in Equation (36), and then use the resulting equation to substitute for v_1^k in Equation (34). This yields the following equation that implicitly identifies x^{L-1}:

$$\Delta\frac{\frac{1-\gamma\delta}{1-\delta} - \frac{\gamma\delta}{x^{L-1}-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\frac{\Delta}{1-\delta}} = \frac{F(x^{L-1}) - F(a)}{F(x^{L-1})}. \tag{37}$$

For $k \in \{1, 2, ..., L-1\}$, we can use Equation (36) to substitute for both v_1^k and v_1^{k+1} in Equation (34), which yields

$$\gamma\alpha^{k-1}\frac{\Delta}{1-\delta} = \left(1 - \frac{F(x^k)}{F(x^{k-1})}\right)\left(-c\alpha^k\right) + \frac{F(x^k)}{F(x^{k-1})}\left(\alpha^k\Delta + \delta\gamma\alpha^k\frac{\Delta}{1-\delta}\right).$$

[24] By definition, type b would be indifferent between betraying in a given period at level α^0 and waiting until the next period to betray at level α^1.

Dividing both sides by α^k, using Equation (35) to substitute for α^{k-1}/α^k, and rearranging terms yields

$$\Delta \frac{1 - \frac{\gamma\delta}{x^{k-1}-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}} = \frac{F(x^{k-1}) - F(x^k)}{F(x^{k-1})}. \tag{38}$$

Thus, the sequence $\{x^k\}_{k=0}^L$ is formed by first letting L be an arbitrary number to be defined later, setting $x^L = a$, and finding x^{L-1} to solve Equation (37). For $k < L$, x^{k-1} is defined inductively by Equation (38). At the point where there is no solution to the transition function or where the solution is exactly b, the procedure terminates, L is set to the number of rounds that occurred, and x^0 is set to b. Once $\{x^k\}_{k=0}^L$ has been determined, the corresponding values of α^k and v_1^{k+1} are easily calculated using Equations (35) and (36) and the initial value $\alpha^L = 1$.

We next show that the procedure to construct $\{x^k\}_{k=0}^L$ is well defined in that the solution to the system of equations, when it exists, is unique. For values of Δ, γ, and x^k for which Equation (38) has an interior unique solution, denote this solution by $x^{k-1} = \mu(x^k; \gamma, \Delta)$.

Lemma 5. *Under Assumption 1, for any $\gamma \in [1, \tilde{\gamma}]$ and $x^k \in [a, b]$, Equation* (38) *has at most one solution. At such a point where $\mu(x^k; \gamma, \Delta) \in (a, b)$, the solution is uniquely defined on a neighborhood of γ and x^k, the conditions of the implicit function theorem hold, and $d\mu(x^k; \gamma, \Delta)/dx^k \geq 0$ and $d\mu(x^k; \gamma, \Delta)/d\gamma \leq 0$. Equations* (37) *has the same properties.*

Proof. We first show that, under Assumption 1, the solution to Equation (38) is unique if it exists. Denote

$$\mathcal{F}(x^{k-1}; x^k, \gamma) \equiv \Delta \frac{1 - \frac{\gamma\delta}{x^{k-1}-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}} - \frac{F(x^{k-1}) - F(x^k)}{F(x^{k-1})},$$

so that $\mathcal{F}(x^{k-1}; x^k, \gamma) = 0$ at a solution point x^{k-1}. Observe that for all $x^k \in (a, b)$ and $\gamma \in [1, \tilde{\gamma}]$,

$$\mathcal{F}(x^k; x^k, \gamma) = \Delta \frac{1 - \frac{\gamma\delta}{x^k-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}} > \Delta \frac{1 - \frac{\tilde{\gamma}\delta}{x^k-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\tilde{\gamma}\frac{\Delta}{1-\delta}} = \Delta \frac{1 - \frac{a-\Delta}{x^k-\Delta}\frac{\Delta/(1-\delta)}{a}}{c + \Delta + (a-\Delta)\frac{\Delta/(1-\delta)}{a}} > 0.$$

The derivative of $\mathcal{F}$ with respect to x^{k-1} exists for every $x^{k-1} \in (a,b)$ and is equal to

$$\frac{\partial \mathcal{F}}{\partial x^{k-1}} = \frac{\Delta \frac{\gamma\delta}{(x^{k-1}-\Delta)^2} \frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}} - \frac{F(x^k) f(x^{k-1})}{(F(x^{k-1}))^2} = \frac{\Delta \frac{\gamma\delta}{(x^{k-1}-\Delta)^2} \frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}} - \frac{f(x^{k-1})}{F(x^{k-1})} \frac{c + \gamma\delta\frac{x^{k-1}}{x^{k-1}-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}},$$

which is strictly negative under the assumption $\Delta \leq \overline{\Delta}$. This implies that at most one value of x^{k-1} solves Equation (38).

Function $\mathcal{F}$ is continuously differentiable on the set of $x^k, x^{k-1} \in (a,b)$ and $\gamma \in (0, \tilde{\gamma})$, and the derivative with respect to x^{k-1} is nonzero. Applying the implicit function theorem to the identity $\mathcal{F} = 0$ yields

$$\frac{dx^{k-1}}{dx^k} = \frac{f(x^k)}{F(x^{k-1})} \cdot \left(\frac{F(x^k) f(x^{k-1})}{(F(x^{k-1}))^2} - \Delta \frac{\frac{\gamma\delta}{(x^{k-1}-\Delta)^2} \frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}} \right)^{-1}$$

and

$$\frac{dx^{k-1}}{d\gamma} = \left(\frac{\Delta}{x^{k-1} - \Delta} + \frac{F(x^{k-1}) - F(x^k)}{F(x^{k-1})} \right) \cdot \\ \left(\frac{\Delta\gamma}{(x^{k-1} - \Delta)^2} - \frac{F(x^k) f(x^{k-1})}{F^2(x^{k-1})} \frac{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}}{\delta\frac{\Delta}{1-\delta}} \right)^{-1}.$$

Because

$$\frac{F(x^k) f(x^{k-1})}{(F(x^{k-1}))^2} > \Delta \frac{\frac{\gamma\delta}{(x^{k-1}-\Delta)^2} \frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}}, \text{ and } \frac{\Delta\gamma}{(x^{k-1} - \Delta)^2} \leq \frac{F(x^k) f(x^{k-1})}{F^2(x^{k-1})} \frac{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}}{\delta\frac{\Delta}{1-\delta}}$$

for $\Delta < \overline{\Delta}$, we have $d\mu(x^k; \gamma, \Delta)/dx^k \geq 0$ and $d\mu(x^k; \gamma, \Delta)/d\gamma \leq 0$.

The calculations for Equation (37) are similar to those above for Equation (38).

□

We conclude the constructive step by showing that L is finite. Because $\gamma \leq \tilde{\gamma} \leq (x^{k-1} - \Delta)/(x^{k-1}\delta)$, the left side of Equation (38) is positive and bounded away from zero. Assumption 1 guarantees that the slope of F is bounded away from zero. This implies that there is a number $\varepsilon > 0$ such that $x^{k-1} - x^k > \varepsilon$ for each k, proving that the inductive step terminates in a finite number of rounds, and this number of rounds is defined to be L so that the resulting sequence runs from $k = 1$ to $k = L$.

PBE construction: strategies

We next specify the strategy profile and verify that it is a PBE. On the equilibrium path through period L, the sequence of levels chosen by player 1 and the cutoff types for player 2 will be $\{\alpha^k, x^k\}_{k=1}^{L}$ as constructed above, and then $\alpha^k = 1$ and $x^k = a$ for every $k > L$. For any history to period k on this path, player 1's strategy prescribes level α^k in period k. Likewise, on this path through the middle of any period k, every type $x \geq x^k$ is supposed to betray and types below x^k cooperate. Player 1's updated belief at the beginning of each period $k+1$ is then given by F conditional on $x < x^k$.

For every history of play in which player 1 had at some point deviated from the prescribed sequence of levels, all types of player 2 are prescribed to immediately betray. For every history of play in which player 1 had at some point deviated from the prescribed sequence of levels and yet player 2 continued to cooperate (a further public deviation), player 1's updated belief assigns probability 1 to the bad type $x = a$ and player 1 is supposed to select $\alpha = 0$.

The specifications just described cover all histories. Beliefs accord to the conditional probability formula on the equilibrium path. It is easy to see that the strategies are sequentially rational. By the construction of $\{\alpha^k, x^k\}_{k=1}^{\infty}$, on the equilibrium path every type of player 2 optimally behaves as prescribed, with bad types betraying at the appointed periods and good types cooperating forever, and player 1 cannot gain by deviating (player 1's continuation value is strictly positive, whereas deviating would lead to a continuation value of zero). Clearly the prescribed behavior is rational off the equilibrium path.

Alteration-proofness of the constructed PBE

We next show that, for γ close to 1, the PBE constructed above is alteration-proof. We do this by proving that it is locally alteration-proof and then applying Lemma 4.

Lemma 6. *For any $\Delta > 0$, there exists $\bar{\gamma} > 1$, such that for all $\gamma \in [1, \bar{\gamma}]$, the constructed sequences $\{\alpha^k, x^k, v_1^{k+1}\}_{k=0}^{L}$ constitute a locally alteration-proof PBE.*

Proof. To verify the local alteration-proof Conditions (11) and (12), we first show the constructed $\{x^k, \alpha^k, v_1^{k+1}\}_{k=0}^{L}$ satisfies

$$\frac{\Delta}{1-\delta}\alpha^{k-1} \leq v_1^k \leq \frac{\Delta}{1-\delta}\alpha^k \left(\frac{F(x^k) - F(x^{k-1})}{F(x^{k-1})} \cdot \frac{c}{x^k} + \frac{F(x^k)}{F(x^{k-1})} \right), \tag{39}$$

for $k \in \{1, 2, ..., L+1\}$.

First, Inequality (39) holds for $k = L+1$, given $v_1^{L+1} = \Delta/(1-\delta)$, $\alpha^L = \alpha^{L+1} = 1$, and $x^L = x^{L+1} = a$. For $k = L$, substituting in $\alpha^L = 1$, $x^L = a$, $v_1^L = \gamma\alpha^{L-1}\Delta/(1-\delta)$, and $\alpha^{L-1} = \delta x^{L-1}/(x^{L-1} - \Delta)$, Inequality (39) becomes

$$\frac{\Delta}{1-\delta}\frac{\delta x^{L-1}}{x^{L-1}-\Delta} \leq \gamma\frac{\Delta}{1-\delta}\frac{\delta x^{L-1}}{x^{L-1}-\Delta} \leq \frac{\Delta}{1-\delta}\left(\frac{F(a)-F(x^{L-1})}{F(x^{L-1})}\cdot\frac{c}{a} + \frac{F(a)}{F(x^{L-1})}\right).$$

The left side of this condition is satisfied with $\gamma \geq 1$, and the right side of this condition can be simplified to

$$\frac{a}{c+a}\left(1-\gamma\frac{\delta x^{L-1}}{x^{L-1}-\Delta}\right) \geq \frac{F(x^{L-1})-F(a)}{F(x^{L-1})}$$

Note that the right side of this inequality is the same as the right side of Equation (37), so a sufficient condition is

$$\frac{a}{c+a}\left(1-\gamma\frac{\delta x^{L-1}}{x^{L-1}-\Delta}\right) \geq \Delta\frac{\frac{1-\gamma\delta}{1-\delta} - \frac{\gamma\delta}{x^{L-1}-\Delta}\frac{\Delta}{1-\delta}}{c+\Delta+\delta\frac{\Delta}{1-\delta}}.$$

For $\gamma \leq \tilde{\gamma}$, this condition can be simplified to

$$\frac{a}{c+a} \geq \frac{\frac{\Delta}{1-\delta}}{c+\frac{\Delta}{1-\delta}},$$

which is satisfied for all $a > \Delta/(1-\delta)$. Hence, we conclude that Inequality (39) holds for $k = L$.

Next, we verify Inequality (39) for $k \in \{2, 3, ..., L-1\}$. The left inequality of this condition is trivially satisfied with $\gamma \geq 1$. Simplifying the right inequality of this condition using Equation (35), we have

$$\frac{x^k}{c+x^k}\left(1-\frac{x^{k-1}\gamma\delta}{x^{k-1}-\Delta}\right) \geq \frac{F(x^{k-1})-F(x^k)}{F(x^{k-1})}.$$

We use the right side of Equation (38) to substitute for the right side of this inequality to get

$$\frac{x^k}{c+x^k}\left(1-\frac{x^{k-1}\gamma\delta}{x^{k-1}-\Delta}\right) \geq \Delta\frac{1-\frac{\gamma\delta}{x^{k-1}-\Delta}\frac{\Delta}{1-\delta}}{c+\Delta+\delta\gamma\frac{\Delta}{1-\delta}},$$

which simplifies to

$$0 \geq (\delta\gamma)^2 - \delta\gamma\left(\delta + \frac{\Delta}{x^{k-1}}\frac{c}{x^k} - \frac{1-\delta}{\Delta}c\right) - c\frac{1-\delta}{\Delta}\left(1 - \frac{\Delta}{x^k}\right)\left(1 - \frac{\Delta}{x^{k-1}}\right). \tag{40}$$

Denote the right side of this inequality as $\mathcal{G}(\gamma, x^{k-1}; x^k, \Delta)$. Recall that x^{k-1} relates to γ and x^k according to $\mu(x^k; \gamma, \Delta)$, so we can write $\mathcal{G}(\gamma, \mu(x^k; \gamma, \Delta); x^k, \Delta)$ to substitute for x^{k-1}. We will show that $\mathcal{G}(\gamma, \mu(x^k; \gamma, \Delta); x^k, \Delta) \leq 0$ for all $x^k \in [a, b]$ and for all $\gamma \in [1, \min\{\tilde{\gamma}, \hat{\gamma}\}]$, where

$$\hat{\gamma} \equiv \frac{1}{2\delta}\left(\left(\delta + \Delta\frac{c}{a^2} - \frac{1-\delta}{\Delta}c\right) + \sqrt{\left(\delta + \Delta\frac{c}{a^2} - \frac{1-\delta}{\Delta}c\right)^2 + 4c\frac{1-\delta}{\Delta}\left(1 - \frac{\Delta}{a}\right)^2}\right).$$

When $\gamma = 1$, $\mathcal{G}(1, \mu(x^k; 1, \Delta); x^k, \Delta)$ is strictly negative. For $\gamma \in [1, \tilde{\gamma}]$, we apply the chain rule to obtain

$$\frac{d\mathcal{G}}{d\gamma} = \delta\left(2\delta\gamma - \delta + \frac{1-\delta}{\Delta}c - \frac{\Delta}{x^{k-1}}\frac{c}{x^k}\right) - \Delta\left(\frac{1-\delta}{\Delta} - \frac{1-\delta}{x^k} - \frac{\delta\gamma}{x^k}\right)\frac{c}{(x^{k-1})^2}\frac{dx^{k-1}}{d\gamma},$$

which is positive when $\Delta < \overline{\Delta}$. Note $\mathcal{G}(\gamma, \mu(x^k; \gamma, \Delta); x^k, \Delta)$ is continuous in γ, so there exists $\hat{\gamma}(x^k)$, such that $\mathcal{G}(\gamma, \mu(x^k; \gamma, \Delta); x^k, \Delta) \leq 0$ for all $\gamma \in [1, \hat{\gamma}(x^k)]$.

In fact, it suffices to set $\hat{\gamma}(x^k) = \hat{\gamma}$, so that this upper bound on γ implies $\mathcal{G}(\gamma, \mu(x^k; \gamma, \Delta); x^k, \Delta) \leq 0$ for all x^k. To see this, let us apply the chain rule to calculate the derivative of $\mathcal{G}$ with respect to x^k:

$$\frac{d\mathcal{G}}{dx^k} = (1-\delta)\frac{c}{(x^k)^2}\left(\frac{1-\delta+\delta\gamma}{1-\delta}\frac{\Delta}{x^{k-1}} - 1\right) + (1-\delta)\frac{c}{(x^{k-1})^2}\left(\frac{1-\delta+\delta\gamma}{1-\delta}\frac{\Delta}{x^k} - 1\right)\frac{dx^{k-1}}{dx^k}.$$

This value is negative when $\Delta < \overline{\Delta}$. The same conclusion holds for x^{k-1} because it enters the expression in a way that is symmetric to x^k. Thus $\mathcal{G}$ is decreasing in both x^k and x^{k-1}, and so plugging in a for these variables yields the highest value of $\mathcal{G}$ for a given γ. Because $\hat{\gamma}$ is the root of $\mathcal{G}(\gamma, a; a, \Delta) = 0$, we conclude that Inequality (40) is satisfied for all $\gamma \in [1, \min\{\hat{\gamma}, \tilde{\gamma}\}]$.

Finally, to complete the proof, we verify Inequality (39) for $k = 1$. We define α^0 as the artificial level such that type $x^0 = b$ is indifferent between betraying in period 0 with level α^0 and in period 1 with level α^1, that is $\alpha^0 \equiv \alpha^1\delta b/(b - \Delta)$. Together with $x^0 = b$ and Equation (34), Inequality (39) evaluated at $k = 1$ becomes

$$\frac{\delta b}{b-\Delta} \leq -c\frac{1-\delta}{\Delta} + F(x^1)\left(c\frac{1-\delta}{\Delta} + 1 - \delta + \delta\gamma\right) \leq F(x^1)\frac{c+x^1}{x^1} - \frac{c}{x^1},$$

which rearranges to

$$\frac{\frac{(1-\delta)b-\Delta}{b-\Delta} + \delta(\gamma-1)}{c\frac{1-\delta}{\Delta} + 1 + \delta(\gamma-1)} \geq 1 - F(x^1) \geq \frac{\delta(\gamma-1)}{c\left(\frac{1-\delta}{\Delta} - \frac{1}{x^1}\right) + \delta(\gamma-1)}. \tag{41}$$

Recall that the sequence is constructed backward in k and terminates at $x^0 = b$ if the remaining mass of bad types is too small for Equation (38) to hold—that is

$$\Delta\frac{1 - \frac{\gamma\delta}{b-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}} \geq 1 - F(x^1). \tag{42}$$

The left inequality of Expression (41) is implied by Inequality (42), as the following is easily verified:

$$\frac{\frac{(1-\delta)b-\Delta}{b-\Delta} + \delta(\gamma-1)}{c\frac{1-\delta}{\Delta} + 1 + \delta(\gamma-1)} \geq \Delta\frac{1 - \frac{\gamma\delta}{b-\Delta}\frac{\Delta}{1-\delta}}{c + \Delta + \delta\gamma\frac{\Delta}{1-\delta}}.$$

The right inequality of Expression (41) is satisfied if

$$\gamma \leq 1 + \frac{1-F(x^1)}{F(x^1)}\frac{c}{\delta}\left(\frac{1-\delta}{\Delta} - \frac{1}{x^1}\right).$$

The right side of this inequality is always greater than 1 because $a \leq x^1 < x^0 = b$ by construction. For $\gamma = 1$, this condition is satisfied as a strict inequality. Because μ is continuous in γ and x^1 results from a finite number of applications of μ, x^1 is continuous in γ except possibly at discrete values where L changes in the sequence construction. Therefore, there is a bound $\check{\gamma}$ such that the right inequality of Expression (41) is satisfied for all $\gamma \in [1, \min\{\check{\gamma}, \tilde{\gamma}\}]$. So we can define

$$\overline{\gamma} \equiv \min\{\tilde{\gamma}, \hat{\gamma}, \check{\gamma}\}.$$

and then for all $\gamma \in [1, \overline{\gamma}]$, the constructed sequence $\{x^k, \alpha^k, v_1^{k+1}\}_{k=0}^{L}$ satisfies the local alteration-proof Condition (39). □

Appendix C Proof of Theorem 3

Two lemmas

We start with a couple of lemmas that will be used at a few points in the main analysis.

Lemma 7. *For any alteration-proof PBE and for all $k \in \{2, ..., L\}$,*

$$\frac{v_1^k}{\alpha^{k-1}} \in \left[\frac{\Delta}{1-\delta}, \frac{\Delta}{1-\delta} \cdot \frac{x^{k-1}-\Delta}{x^{k-1}\delta}\right].$$

Further, $v_1^k/\alpha^{k-1} - \Delta/(1-\delta)$ is on the order of Δ.

Proof. From Inequality (11), we have $v_1^k/\alpha^{k-1} \geq \Delta/(1-\delta)$. From Inequality (12) and Condition (35), we have

$$\frac{v_1^k}{\alpha^{k-1}} \leq \frac{\Delta}{1-\delta} \cdot \frac{x^{k-1}-\Delta}{x^{k-1}\delta}\left(1 - \frac{F(x^{k-1})-F(x^k)}{F(x^{k-1})} \cdot \frac{c+x^k}{x^k}\right) \leq \frac{\Delta}{1-\delta} \cdot \frac{x^{k-1}-\Delta}{x^{k-1}\delta},$$

which proves the uniform upper bound of v_1^k/α^{k-1}. On the second claim, note that the difference between the upper and lower bounds on v_1^k/α^{k-1} can be rewritten by rearranging terms as follows:

$$\frac{\Delta}{1-\delta}\frac{x^{k-1}-\Delta}{x^{k-1}\delta} - \frac{\Delta}{1-\delta} = \frac{\Delta}{\delta} \cdot \frac{\Delta}{1-\delta} \cdot \left(\frac{1-\delta}{\Delta} - \frac{1}{x^{k-1}}\right).$$

The term in parentheses is bounded away from zero for $x^{k-1} \in [a, b]$ and $\Delta/(1-\delta)$ converges to $1/r$ as Δ approaches zero, and so the entire expression is on the order of Δ. □

Lemma 8. *For any alteration-proof PBE, $x^{k+1} - x^k$ and $\alpha^{k+1} - \alpha^k$ are both on the order of Δ.*

Proof. From player 2's indifference Condition (7), rearranging terms gives us

$$x^k = \frac{\Delta\alpha^k}{\alpha^k - \delta\alpha^{k+1}} = \frac{\Delta}{1 - \delta\alpha^{k+1}/\alpha^k}.$$

Because x^k is in $[a,b]$ for all $\Delta > 0$, so is the right most term, which simplifies to

$$\frac{\Delta}{\delta}\left(\frac{1-\delta}{\Delta}-\frac{1}{a}\right) \leq \frac{\alpha^{k+1}-\alpha^k}{\alpha^k} \leq \frac{\Delta}{\delta}\left(\frac{1-\delta}{\Delta}-\frac{1}{b}\right). \tag{43}$$

Note the left and right most terms are on the order of Δ, and when we have α^k strictly positive, then $\alpha^{k+1}-\alpha^k$ must be on the order of Δ.

Second, rewrite Inequality (13) to get

$$0 \geq \frac{F(x^k)-F(x^{k-1})}{F(x^{k-1})} \geq \frac{c(\alpha^k - \delta\alpha^{k+1}) + \Delta\alpha^{k-1}}{(c+\Delta)\alpha^k - \delta c \alpha^{k+1}} - 1 = \frac{\alpha^{k-1}-\alpha^k}{\alpha^k(c/x^k+1)}, \tag{44}$$

where the left-most term is 0 and right-most term is on the order of Δ, so $F(x^k) - F(x^{k-1})$ is also on the order of Δ. By Assumption 1 (in particular that F' is bounded away from 0) and because $F(x^{k-1})$ is bounded away from zero owing to the probability of the good type, we conclude that $x^k - x^{k-1}$ is on the order of Δ. □

Change of variables and adjacent periods

We next move on to the main analysis for the proof of Theorem 3. Our calculations are simplified by introducing a new variable $w^k \equiv v_1^k/\alpha^{k-1}$, for all $k \in \{2, ..., L+1\}$, which reduces the number of dimensions by alleviating v_1^k and α^{k-1} as separate variables. Note that, with the new notation, any alteration-proof PBE sequence $\{x^k, \alpha^k, v_1^{k+1}\}_{k=1}^L$ has a corresponding sequence $\{x^k, w^{k+1}\}_{k=1}^L$.

In fact, the PBE and alteration-proofness Conditions (4), (7), (11) and (12) can be expressed in terms of only the sequence $\{x^k, w^{k+1}\}_{k=1}^L$ as follows:

$$w^k = \frac{x^{k-1}-\Delta}{\delta x^{k-1}}\left(-c + \frac{F(x^k)}{F(x^{k-1})}(\Delta + c + \delta w^{k+1})\right) \tag{45}$$

$$\frac{\Delta}{1-\delta} \leq w^k \leq \frac{\Delta}{1-\delta}\frac{x^{k-1}-\Delta}{\delta x^{k-1}}\left(\frac{F(x^k)-F(x^{k-1})}{F(x^{k-1})}\cdot\frac{c}{x^k} + \frac{F(x^k)}{F(x^{k-1})}\right). \tag{46}$$

Equation (45) is derived by starting from Equation (4), substituting for v_1^k and v_1^{k+1} using $w^k = v_1^k/\alpha^{k-1}$ and $w^{k+1} = v_1^{k+1}/\alpha^k$, dividing by α^k, and using Equation (7) to substitute for α^{k-1}/α^k. The inequalities in Expression (46) are Inequalities (11) and (12) after substituting in w^k.

We define function $g\colon [a,b]\times[a,b]\times[0,\infty) \to [0,\infty)$ to give w^k as a function of

x^k, x^{k-1}, and w^{k+1} according to Equation (45). That is,

$$w^k = g(x^k, x^{k-1}, w^{k+1}) \equiv \frac{x^{k-1} - \Delta}{\delta x^{k-1}} \left(-c + \frac{F(x^k)}{F(x^{k-1})} (\Delta + c + \delta w^{k+1}) \right).$$

By next studying the properties of g, we will be able to construct bounds on equilibrium sequences. In particular, we need to look at how the values of w^{k+1} and x^{k-1} relate over the two adjacent periods k and $k+1$, while fixing w^k, w^{k+1}, and x^{k+1}, and allowing x^k to vary.

Lemma 9. *Fix numbers w^k, $x^{k+1} \in [a, b]$ and w^{k+2}. Under Assumption 1, suppose that the following equations hold for some values $\hat{x}^k, \hat{x}^{k-1} \in (a, b)$ and $\hat{w}^{k+1} \in \mathbb{R}$:*

$$w^k = g(x^k, x^{k-1}, g(x^{k+1}, x^k, w^{k+2})) \text{ and } w^{k+1} = g(x^{k+1}, x^k, w^{k+2}). \tag{47}$$

These equations implicitly define x^{k-1} and w^{k+1} as functions of x^k on a neighborhood of $(\hat{x}^k, \hat{x}^{k-1}, \hat{w}^{k+1})$. The implicit function theorem applies and $dw^{k+1}/dx^k < 0$ and $dx^{k-1}/dx^k > 0$.

Proof. The first part of the lemma follows directly from equilibrium Identity (45) and Assumption 1. To show that w^{k+1} decreases in x^k, we differentiate

$$g(x^{k+1}, x^k, w^{k+2}) = \frac{x^k - \Delta}{\delta x^k} \left(-c + \frac{F(x^{k+1})}{F(x^k)} (\Delta + c + \delta w^{k+2}) \right)$$

with respect to x^k to get

$$\frac{dw^{k+1}}{dx^k} = -c \frac{\Delta}{\delta (x^k)^2} + \frac{F(x^{k+1})}{F(x^k)} \left(\frac{\Delta}{x^k} - \frac{f(x^k)}{F(x^k)} (x^k - \Delta) \right) \frac{\Delta + c + \delta w^{k+2}}{\delta x^k}.$$

Under Assumption 1 and $\Delta \leq \overline{\Delta}$, it is easy to verify that $dw^{k+1}/dx^k < 0$.

Next, we show that x^{k-1} increases in x^k. Denote by g_j the partial derivative of function g with respect to its jth argument. Applying the implicit function theorem to condition $w^k = g(x^k, x^{k-1}, g(x^{k+1}, x^k, w^{k+2}))$ and rearranging terms, we get

$$\frac{dx^{k-1}}{dx^k} = -\frac{g_1(x^k, x^{k-1}, w^{k+1}) + g_3(x^k, x^{k-1}, w^{k+1}) g_2(x^{k+1}, x^k, w^{k+2})}{g_2(x^k, x^{k-1}, w^{k+1})}, \tag{48}$$

where

$$
\begin{aligned}
g_1(x^k, x^{k-1}, w^{k+1}) &= \frac{x^{k-1}-\Delta}{\delta x^{k-1}} \frac{f(x^k)}{F(x^{k-1})}(\Delta + c + \delta w^{k+1}), \\
g_2(x^k, x^{k-1}, w^{k+1}) &= \frac{\Delta}{\delta (x^{k-1})^2} \left(-c + \frac{F(x^k)}{F(x^{k-1})}(\Delta + c + \delta w^{k+1}) \right) \\
&\qquad - \frac{x^{k-1}-\Delta}{\delta x^{k-1}} \frac{F(x^k) f(x^{k-1})}{F^2(x^{k-1})}(\Delta + c + \delta w^{k+1}), \\
g_3(x^k, x^{k-1}, w^{k+1}) &= \frac{x^{k-1}-\Delta}{x^{k-1}} \frac{F(x^k)}{F(x^{k-1})}, \text{ and} \\
g_2(x^{k+1}, x^k, w^{k+2}) &= \frac{\Delta}{\delta (x^k)^2} \left(-c + \frac{F(x^{k+1})}{F(x^k)}(\Delta + c + \delta w^{k+2}) \right) \\
&\qquad - \frac{x^k-\Delta}{\delta x^k} \frac{F(x^{k+1}) f(x^k)}{F^2(x^k)}(\Delta + c + \delta w^{k+2}).
\end{aligned}
$$

Note that Identity (45) can be rearranged to form

$$c + w^k \frac{\delta x^{k-1}}{x^{k-1}-\Delta} = \frac{F(x^k)}{F(x^{k-1})}(\Delta + c + \delta w^{k+1}),$$

which we use to simplify g_2 and get

$$g_2(x^k, x^{k-1}, w^{k+1}) = \frac{\Delta}{x^{k-1}} \frac{w^k}{x^{k-1}-\Delta} - \frac{f(x^{k-1})}{F(x^{k-1})} \left(c \frac{x^{k-1}-\Delta}{\delta x^{k-1}} + w^k \right)$$

and

$$g_2(x^{k+1}, x^k, w^{k+2}) = \frac{\Delta}{x^k} \frac{w^{k+1}}{x^k-\Delta} - \frac{f(x^k)}{F(x^k)} \left(c \frac{x^k-\Delta}{\delta x^k} + w^{k+1} \right).$$

Finally, we substitute the partial derivatives into Equation (48) and collect terms to obtain

$$\frac{dx^{k-1}}{dx^k} = -\frac{\frac{x^{k-1}-\Delta}{\delta x^{k-1}} \frac{\Delta}{x^k} \frac{1}{F(x^{k-1})} \left(f(x^k)(x^k + c) + \frac{\delta w^{k+1}}{x^k-\Delta} F(x^k) \right)}{\frac{\Delta}{x^{k-1}} \frac{w^k}{x^{k-1}-\Delta} - \frac{f(x^{k-1})}{F(x^{k-1})} \left(c \frac{x^{k-1}-\Delta}{\delta x^{k-1}} + w^k \right)}.$$

The numerator is always positive and the denominator is negative when $\Delta \leq \overline{\Delta}$, so we have found that x^{k-1} increases in x^k. □

Lemma 10. *Under Assumption 1 and assuming that the values of function g's first and second arguments satisfy $x^k \in (a, b)$ and $x^{k-1} \in (a, b)$, g is strictly increasing in* *its first and third arguments and strictly decreasing in its second argument.*

Proof. Clearly g is increasing in its third argument, w^{k+1}, given that $x^{k-1} > a > \Delta$. In the proof of Lemma 9 we calculated g_1 and g_2. The first of these is strictly positive and the second, under Assumption 1, is strictly negative. □

Construction of bounds

We will construct two sequences, $\{\overline{x}^\ell\}_{\ell=0}^\infty$ and $\{\underline{x}^\ell\}_{\ell=0}^\infty$, that bound the type cutoffs of all alteration-proof equilibria. These sequences will be indexed in *reverse time* by integer ℓ, which counts the number of periods before period $L-1$ in any alteration-proof equilibrium.[25] The construction incorporates bounds on the w^k values. By construction, for any given alteration-proof equilibrium and its associated sequence $\{x^k, w^{k+1}\}_{k=1}^L$, it will be the case that $x^k \in [\underline{x}^{L-1-k}, \overline{x}^{L-1-k}]$ for each $k \in \{1, 2, ..., L-1\}$. There are two parts of the construction. The first constructs $\{\overline{x}^\ell\}_{\ell=0}^\infty$ and the second in similar fashion constructs $\{\underline{x}^\ell\}_{\ell=0}^\infty$.

Each of the two bounding sequences will be derived starting from an arbitrary alteration-proof equilibrium sequence $\{x^\tau, w^{\tau+1}\}_{\tau=1}^L$ under Assumption 1. We will adjust the sequence in recursive steps and then reverse the time index. So as not to overly complicate the presentation, we continue to call the adjusted sequence $\{x^\tau, w^{\tau+1}\}_{\tau=1}^L$ after every modification (that is, we redefine the sequence as the result of each adjustment), rather than create new notation to track the adjustments. It will turn out that the resulting bounds do not depend on the equilibrium that we started with.

The adjustments needed to construct $\{\overline{x}^\ell\}_{\ell=0}^\infty$ will utilize the following two operations, indexed by a given period k:

Operation 1: For a given sequence $\{x^\tau, w^{\tau+1}\}_{\tau=1}^L$ and integer $k \leq L$ satisfying $w^k \leq g(x^k, x^{k-1}, w^{k+1})$, hold fixed all values in the sequence except x^{k-1}. Raise the value of x^{k-1} to the point x' at which $w^k = g(x^k, x', w^{k+1})$, and then redefine $x^{k-1} \equiv x'$. If no such $x' \in [a, b]$ exists, then stop the procedure and set $x^{k-1} \equiv b$. Note that Lemma 10 ensures that x' is uniquely determined and weakly exceeds the starting value x^{k-1}.

Operation 2: For a given sequence $\{x^\tau, w^{\tau+1}\}_{\tau=1}^L$ and integer $k \leq L-1$ satisfying $w^\tau = g(x^\tau, x^{\tau-1}, w^{\tau+1})$ for every $\tau \in \{k, k+1, \ldots, L\}$, hold fixed all values in the

[25] From Lemma 2 we know that every alteration-proof equilibrium has $x^L = a$, so the bounding sequences will be for the periods $L-1$ and earlier.

sequence except x^{k-1}, x^k, and w^{k+1}. Raise the value of x^k to x' and simultaneously raise x^{k-1} to x'' and lower w^{k+1} to w', such that the system of Equations (47) is maintained, to the point at which $w' = \Delta/(1-\delta)$. Then redefine $x^k \equiv x'$, $x^{k-1} \equiv x''$, and $w^{k+1} \equiv w'$. If no such point exists (because x'' goes above b), then stop the procedure and set $x^{k-1} \equiv b$ and set x^k and w^{k+1} to the values that satisfy (47). Note that Lemma 9 ensures (x', x'', w') is uniquely determined, with x' and x'' weakly exceeding their starting values of x^k and x^{k-1}.

Here are the steps to construct $\{\overline{x}^\ell\}_{\ell=0}^\infty$. Take any alteration-proof equilibrium and let $\{x^\tau, w^{\tau+1}\}_{\tau=1}^L$ be its sequence of levels and w values. From Equation (45), we know that $w^\tau = g(x^\tau, x^{\tau-1}, w^{\tau+1})$ for every $\tau \leq L$. From Lemmas 2 and 7, and because $(x-\Delta)/\delta x$ is increasing in $x > \Delta$, we know that $x^L = a$, $w^{L+1} \leq (\Delta/(1-\delta)) \cdot ((a-\Delta)/\delta a)$, and $w^\tau \geq \Delta/(1-\delta)$ for every τ. Our adjustments will eventually push w^τ down to exactly this lower bound, for each $\tau \leq L$.

The first step in the construction is to reset w^{L+1} to equal $(\Delta/(1-\delta)) \cdot ((a - \Delta)/\delta a)$, which causes $g(x^L, x^{L-1}, w^{L+1})$ to rise. Then perform Operation 1 for $k = L$, which restores $w^L = g(x^L, x^{L-1}, w^{L+1})$ but in raising x^{L-1} causes $g(x^{L-1}, x^{L-2}, w^L)$ to increase. Perform Operation 1 for $k = L-1$, restoring $w^{L-1} = g(x^{L-1}, x^{L-2}, w^L)$.

The second step is to apply Operations 1 and 2 recursively as follows, starting with $k' = L-1$. For any integer $k' \leq L-1$ satisfying $w^\tau = g(x^\tau, x^{\tau-1}, w^{\tau+1})$ for every $\tau \in \{k', k'+1, \ldots, L\}$, perform Operation 2 for $k = k'$, which results in $w^{k'+1} = \Delta/(1-\delta)$ and $w^{k'-1} < g(x^{k'-1}, x^{k'-2}, w^{k'})$, and then perform Operation 1 for $k = k'-1$, which restores $w^{k'-1} = g(x^{k'-1}, x^{k'-2}, w^{k'})$. Decrease k by one and repeat this function until either operation triggers the process to stop. Let N denote the period at which the procedure stops, where $x^M = b$.

Note that Operations 1 and 2 adjust the cutoff-type sequence only by raising values of x^τ, so we have constructed upper bounds on the type cutoffs in any alteration-proof equilibrium. The procedure also results in $w^\tau = g(x^\tau, x^{\tau-1}, w^{\tau+1})$ and $w^\tau = \Delta/(1-\delta)$ for each $\tau \in \{N, N+1, \ldots, L\}$. Letting $\ell = L-1-\tau$, so we count backward in time, we thus have found a bounding sequence $\{\overline{x}^\ell\}_{\ell=0}^\infty$ defined recursively by $\overline{x}^0 = x^{L-1}$,

$$\frac{\Delta}{1-\delta} = \frac{\overline{x}^{\ell+1} - \Delta}{\delta \overline{x}^{\ell+1}} \left(-c + \frac{F(\overline{x}^\ell)}{F(\overline{x}^{\ell+1})} \left(\Delta + c + \delta \frac{\Delta}{1-\delta} \right) \right) \tag{49}$$

for each $\ell \in \{0, 1, \ldots, L-N-2\}$, and $\overline{x}^\ell = b$ for $\ell \geq L-N-1$. Note that Equation (49) comes from plugging $w^\tau = w^{\tau+1} = \Delta/(1-\delta)$ into $w^\tau = g(x^\tau, x^{\tau-1}, w^{\tau+1})$ and replacing

τ with ℓ and replacing $\tau - 1$ with $\ell + 1$ to reverse the index direction. This equation gives $\overline{x}^{\ell+1}$ implicitly as a function of $\overline{x}^{\ell}$.

We construct $\{\underline{x}^{\ell}\}_{\ell=0}^{\infty}$ using the same steps but working in the opposite direction for the adjustments. Note that from Lemmas 2 and 7, we know that $x^L = a$ and $\Delta/(1-\delta) \leq w^{\tau} < (\Delta/(1-\delta)) \cdot ((x^{\tau-1} - \Delta)/\delta x^{\tau-1})$ for every τ. Let us define Operations 1R and 2R just as we did Operations 1 and 2 except with adjustments in the opposite direction. That is, Operation 1R begins with a sequence satisfying $w^k \geq g(x^k, x^{k-1}, w^{k+1})$ and lowers x^{k-1} to the point x' at which $w^k = g(x^k, x', w^{k+1})$. Operation 2R lowers x^k to x' and simultaneously lowers x^{k-1} to x'' and raises w^{k+1} to w', such that the system of Equations (47) is maintained, to the point at which $w' = (\Delta/(1-\delta)) \cdot ((x^{k-1} - \Delta)/\delta x^{k-1})$. It is not difficult to verify that both operations are well defined, yielding cutoff-type values satisfying $x^{k-1} \geq x^k$.

Starting with an arbitrary alteration-proof equilibrium sequence $\{x^{\tau}, w^{\tau+1}\}_{\tau=1}^{L}$, we first reset w^{L+1} down to equal $(\Delta/(1-\delta))$, which causes $g(x^L, x^{L-1}, w^{L+1})$ to fall. Then we perform Operation 1R for $k = L$ to restore $w^L = g(x^L, x^{L-1}, w^{L+1})$, and again for $k = L-1$ to restore $w^{L-1} = g(x^{L-1}, x^{L-2}, w^L)$. We proceed to the recursive step, applying Operations 1R and 2R, starting with $k' = L - 1$ and ending when the boundary b is reached for $x^{\tau-1}$, at a period denoted by N. All operations adjust the cutoff-type sequence only by lowering values of x^{τ}, so we have constructed lower bounds on the type cutoffs in any alteration-proof equilibrium. The procedure also results in $w^{\tau} = g(x^{\tau}, x^{\tau-1}, w^{\tau+1})$ and $w^{\tau} = (\Delta/(1-\delta)) \cdot ((x^{\tau-1} - \Delta)/\delta x^{\tau-1})$ for each $\tau \in \{N, N+1, \ldots, L\}$. Letting $\ell = L - 1 - \tau$, we thus have found a bounding sequence $\{\underline{x}^{\ell}\}_{\ell=0}^{\infty}$ defined recursively by $\underline{x}^0 = x^{L-1}$,

$$\frac{\Delta}{1-\delta} = -c + \frac{F(\underline{x}^{\ell})}{F(\underline{x}^{\ell+1})} \left(\Delta + c + \delta \frac{\Delta}{1-\delta} \cdot \frac{\underline{x}^{\ell} - \Delta}{\delta \overline{x}^{\ell}} \right) \tag{50}$$

for each $\ell \in \{0, 1, \ldots, L-N-2\}$, and $\overline{x}^{\ell} = b$ for $\ell \geq L-N-1$. Equation (50) comes from plugging $w^{\tau} = (\Delta/(1-\delta)) \cdot ((x^{\tau-1} - \Delta)/\delta x^{\tau-1})$ and $w^{\tau+1} = \Delta/(1-\delta) \cdot ((x^{\tau} - \Delta)/\delta x^{\tau})$ into $w^{\tau} = g(x^{\tau}, x^{\tau-1}, w^{\tau+1})$ and replacing τ with ℓ and replacing $\tau - 1$ with $\ell + 1$. This equation gives $\underline{x}^{\ell+1}$ as a function of $\underline{x}^{\ell}$.

Convergence of bounding sequences

In summary, we have constructed sequences $\{\overline{x}^\ell\}_{\ell=0}^\infty$ and $\{\underline{x}^\ell\}_{\ell=0}^\infty$ that bound all alteration-proof equilibrium type-cutoff sequences. The final step of the proof is to apply a convergence result of Watson (2021) to show that as $\Delta \to 0^+$, the upper and lower bounds converge uniformly to the same continuous-time function defined by a differential equation. To be precise, let us make explicit the dependence of the bounding sequences on Δ by writing $\{\overline{x}^\ell(\Delta)\}_{\ell=0}^\infty$ and $\{\underline{x}^\ell(\Delta)\}_{\ell=0}^\infty$, and define step functions $\hat{\overline{x}}\colon [0,\infty) \times (0,\infty) \to [a,b]$ and $\hat{\underline{x}}\colon [0,\infty) \times (0,\infty) \to [a,b]$ by $\hat{\overline{x}}(t;\Delta) = \overline{x}^{[t/\Delta]}$ and $\hat{\underline{x}}(t;\Delta) = \underline{x}^{[t/\Delta]}$, where $[t/\Delta]$ denotes the largest integer that is weakly below t/Δ. As shown below, $\hat{\overline{x}}(\cdot;\Delta)$ and $\hat{\underline{x}}(\cdot;\Delta)$ converge, which implies the convergence of equilibrium type cutoffs stated in Theorem 3.

Lemma 11. *As $\Delta \to 0$, step functions $\hat{\overline{x}}(\cdot;\Delta)$ and $\hat{\underline{x}}(\cdot;\Delta)$ uniformly converge to the same function $z : [0,\infty) \to [0,b]$ that solves this initial-value problem:*

$$\frac{dz}{dt} = \frac{F(z)}{f(z)} \frac{rz-1}{z(1+cr)}, \quad z(0) = a. \tag{51}$$

Proof. We simplify Equations (49) and (50) by rearranging terms to obtain, respectively,

$$\frac{\Delta}{1-\delta} \frac{\delta \overline{x}^{\ell+1}}{\overline{x}^{\ell+1} - \Delta} + c = \frac{F(\overline{x}^\ell)}{F(\overline{x}^{\ell+1})} \left(c + \frac{\Delta}{1-\delta} \right), \tag{52}$$

and

$$F(\underline{x}^{\ell+1}) = F(\underline{x}^\ell) \frac{\Delta + c + \frac{\Delta}{1-\delta} \cdot \frac{\underline{x}^\ell - \Delta}{\underline{x}^\ell}}{\frac{\Delta}{1-\delta} + c}. \tag{53}$$

Define the transition function $\overline{\sigma}\colon [a,b] \times (0,\overline{\Delta}) \to [a,b]$ so that, for every $x^\ell \in [a,b]$ and $\Delta \in (0,\overline{\Delta})$, $\overline{\sigma}(x^\ell,\Delta)$ is the value of $x^{\ell+1}$ that solves (52). Likewise, define $\underline{\sigma}\colon [a,b] \times (0,\overline{\Delta})$ so that $\underline{x}^{\ell+1} = \underline{\sigma}(\underline{x}^\ell,\Delta)$ solves (53).

We next extend the domain of functions $\overline{\sigma}$ and $\underline{\sigma}$ to $\mathbb{R} \times \mathbb{R}$. Regarding Δ, Expressions (52) and (53) are already well-defined for $\Delta \in \mathbb{R} \setminus \{0\}$ and the limits as $\Delta \to 0$ exist because $\lim_{\Delta \to 0} \Delta/(1-\delta) = 1/r$. So for $\Delta = 0$ we simply replace $\Delta/(1-\delta)$ with $1/r$ in (52) and (53), which extends $\overline{\sigma}$ and $\underline{\sigma}$ to $\Delta \in \mathbb{R}$. The extension to $x^\ell \in \mathbb{R}$ can be done arbitrarily.

Under Assumption 1, the extended functions $\overline{\sigma}$ and $\underline{\sigma}$ are twice continuously differentiable on $(a,b) \times \mathbb{R}$. Clearly, the initial states $\overline{x}^0(\Delta)$ and $\underline{x}^0(\Delta)$ converge to a.

Note as well that $\overline{\sigma}(x,0) = x$ and $\underline{\sigma}(x,0) = x$ for all $x \in [a,b]$. Finally, the implicit function theorem applies to calculate $d\overline{\sigma}/d\Delta$ and $d\underline{\sigma}/d\Delta$, and these derivatives are bounded on a neighborhood of $\Delta = 0$ and for all $x^\ell \in [a,b]$. The properties just stated allow us to apply Theorem 2 of Watson (2021), which establishes that $\hat{\overline{x}}(\cdot;\Delta)$ and $\hat{\underline{x}}(\cdot;\Delta)$ uniformly converge to, respectively, functions $\overline{z} : [0,\infty) \to [a,b]$ and $\underline{z} : [0,\infty) \to [a,b]$ that solve initial-value problems given by

$$\frac{d\overline{z}}{dt} = \frac{d\overline{\sigma}}{d\Delta}(\overline{z},0), \quad \frac{d\underline{x}}{dt} = \frac{d\underline{\sigma}}{d\Delta}(\underline{z},0), \qquad \text{and } \overline{z}(0) = \underline{z}(0) = a.$$

To complete the proof, we evaluate $d\overline{\sigma}(\overline{z},0)/d\Delta$ and $d\underline{\sigma}(\underline{z},0)/d\Delta$.

To derive $d\overline{\sigma}/d\Delta$, we apply implicit function theorem by differentiating both sides of Equation (52) with respect to Δ, which yields

$$\frac{1 - \frac{\Delta}{1-\delta} r\delta}{1-\delta} \frac{\delta \overline{x}^{\ell+1}}{\overline{x}^{\ell+1} - \Delta} + \frac{\Delta}{1-\delta} \frac{(-r\delta\overline{x}^{\ell+1} + \delta\frac{d\overline{\sigma}}{d\Delta})(\overline{x}^{\ell+1} - \Delta) - \delta\overline{x}^{\ell+1}(\frac{d\overline{\sigma}}{d\Delta} - 1)}{(\overline{x}^{\ell+1} - \Delta)^2}$$
$$= -\frac{F(\overline{x}^\ell) f(\overline{x}^{\ell+1})}{F^2(\overline{x}^{\ell+1})} \left(c + \frac{\Delta}{1-\delta}\right) \frac{d\overline{\sigma}}{d\Delta} + \frac{F(\overline{x}^\ell)}{F(\overline{x}^{\ell+1})} \frac{1 - \frac{\Delta}{1-\delta} r\delta}{1-\delta}.$$

We solve for $d\overline{\sigma}/d\Delta$ and evaluate it at $\Delta = 0$, replacing $\Delta/(1-\delta)$ with $1/r$ as required by the extension, and setting $\overline{x}^\ell = \overline{z}$ and $\overline{x}^{\ell+1} = \overline{\sigma}(\overline{z},0) = \overline{z}$. This yields

$$\frac{d\overline{z}}{dt} = \frac{d\overline{\sigma}}{d\Delta}(\overline{z},0) = \frac{F(\overline{z})}{f(\overline{z})} \frac{r\overline{z} - 1}{\overline{z}\,(cr+1)}.$$

Similarly, we differentiate both sides of Equation (53) with respect to Δ,

$$\frac{d\underline{\sigma}}{d\Delta} = \frac{F(\underline{x}^\ell)}{f(\underline{x}^{\ell+1})} \frac{(\frac{\Delta}{1-\delta} + c)\left(1 + \frac{1-\frac{\Delta}{1-\delta}r\delta}{1-\delta} \frac{\underline{x}^{\ell+1} - \Delta}{\underline{x}^{\ell+1}} - \frac{\Delta}{1-\delta}\frac{1}{\underline{x}^{\ell+1}}\right) - \left(\Delta + c + \frac{\Delta}{1-\delta} \cdot \frac{\underline{x}^\ell - \Delta}{\underline{x}^\ell}\right) \frac{1-\frac{\Delta}{1-\delta}r\delta}{1-\delta}}{(\frac{\Delta}{1-\delta} + c)^2},$$

and evaluate it at $\Delta = 0$, $\underline{x}^\ell = \underline{z}$, and $\underline{x}^{\ell+1} = \underline{\sigma}(\underline{z},0) = \underline{z}$. This yields

$$\frac{d\underline{z}}{dt} = \frac{d\underline{\sigma}}{d\Delta}(\underline{z},0) = \frac{F(\underline{z})}{f(\underline{z})} \frac{r\underline{z} - 1}{\underline{z}(1 + cr)}.$$

Clearly functions $\overline{z}$ and $\underline{z}$ are the same, identical to the function z described in the statement of the lemma. □

Note that the initial-value problem described in Lemma 11 is identical to that described for x in Theorem 3,

$$\frac{dx}{dt} = -\frac{F(x)}{f(x)}\frac{rx-1}{x(1+cr)},$$

except with the direction of time reversed, so we have proved the result with respect to the cutoff-type sequences. As shown in the text, the value T is derived by integrating the differential equation and solving for the time at which $x = b$.

To show that the equilibrium level sequences and continuation values for player 1 are also characterized as the theorem states, it is enough to observe that we can trivially rewrite the transition functions that define sequences $\{\overline{x}^\ell\}_{\ell=0}^\infty$ and $\{\underline{x}^\ell\}_{\ell=0}^\infty$ as vector-valued functions that include the level and player 1's continuation value. In any alteration-proof equilibrium, the transitions of the level and player 1's continuation value obey

$$\alpha^{k+1} = \frac{x^k - \Delta}{\delta x^k}\alpha^k \quad \text{and} \quad v_1^{k+1} = w^{k+1}\alpha^k,$$

$\alpha^{L+1} = 1$, and $v_1^{L+1} = 1/r$. Corresponding to the lower-bound sequence $\{\underline{x}^\ell\}_{\ell=0}^\infty$ is a lower-bound sequence for α and an upper-bound sequence for w; likewise, the upper-bound sequence $\{\overline{x}^\ell\}_{\ell=0}^\infty$ corresponds to an upper-bound sequence for α and a lower-bound sequence for w. The convergence theorem of Watson (2021) applies to vector sequences. Thus, the characterization of the limit of level sequences and player 1's continuation values then follows from the characterization of the limit of type-cutoff sequences derived above.

Appendix D Proof of Proposition 1

We first derive comparative statics of T. From Equation (26), we use the fact $\ln(z) \leq z - 1$ and get

$$\begin{aligned}
\frac{dT}{dq} &= \frac{(b-a)(1+r)}{(1-q+r(bq-a))^2}\left(\ln\frac{1-ra}{q(1-rb)} - \frac{1-q+r(bq-a)}{q}\frac{bq-a}{b-a}\right) \\
&\leq \frac{(b-a)(1+r)}{(1-q+r(bq-a))^2}\left(\frac{1-ra}{q(1-rb)} - 1 - \frac{1-q+r(bq-a)}{q}\frac{bq-a}{b-a}\right) \\
&= \frac{(b-a)(1+r)}{1-q+r(bq-a)}\left(\frac{1}{1-rb} - \frac{bq-a}{b-a}\right)\frac{1}{q} = \frac{b(1+r)}{q(1-rb)} < 0,
\end{aligned}$$

and

$$\begin{aligned}\frac{dT}{db} &= -\frac{(1-q)(1+r)}{1-q+r(bq-a)}\left(\frac{rq\left(-\frac{1}{r}\ln\frac{1-ra}{1-rb}+\frac{a-bq}{1-q}\ln q\right)}{1-q+r(bq-a)}+\left(\frac{1}{1-rb}+\frac{q}{1-q}\ln q\right)\right)\\ &= \frac{q(1-q)(1+r)}{(1-q+r(bq-a))^2}\left(\ln\frac{1-ra}{q(1-rb)}-\frac{1-ra}{q(1-rb)}+1\right)\leq 0.\end{aligned}$$

Second, from Equation (26), comparative statics of $\alpha(0)$ is

$$\frac{d\alpha(0)}{dq}=(r+1)q^r>0,\quad \frac{d\alpha(0)}{db}=0.$$

Third, for the comparative statics of the slope of x for fixed $\chi\in[a,b]$ at time $\Gamma(\chi)$, we use Equation (22)

$$\frac{d\Gamma}{d\chi}=\frac{(1+r)\chi}{1-r\chi}\cdot\frac{1-q}{(b-a)q+(\chi-a)(1-q)}\equiv g(b,q;\chi),$$

and take partial derivatives of g, we get

$$\begin{aligned}\frac{dg(b,q;\chi)}{dq} &= \frac{(1+r)\chi}{r\chi-1}\cdot\frac{b-a}{((b-a)q+(\chi-a)(1-q))^2}>0,\\ \frac{dg(b,q;\chi)}{db} &= \frac{(1+r)\chi}{r\chi-1}\cdot\frac{q(1-q)}{((b-a)q+(\chi-a)(1-q))^2}>0.\end{aligned}$$

Last, we consider the comparative statics of the slope of $\ln\alpha$ for fixed $\chi\in[a,b]$ at time $\Gamma(\chi)$. Similarly, with Equation (25), we have

$$\frac{d\ln\alpha}{d\chi}=-\frac{(1-q)(1+r)}{(b-a)q+(\chi-a)(1-q)}\equiv h(b,q;\chi).$$

Therefore,

$$\begin{aligned}\frac{dh(b,q;\chi)}{dq} &= \frac{(1+r)(b-a)}{((b-a)q+(\chi-a)(1-q))^2}>0,\\ \frac{dh(b,q;\chi)}{db} &= \frac{q(1-q)(1+r)}{((b-a)q+(\chi-a)(1-q))^2}>0.\end{aligned}$$

References

Andreoni, James, Michael A. Kuhn, and Larry Samuelson. 2019. “Building Rational Cooperation on Their Own: Learning to Start Small.” *Journal of Public Economic Theory* 21 (5): 812–825.

Atakan, Alp, Levent Koçkesen, and Elif Kubilay. 2020. “Starting Small to Communicate.” *Games and Economic Behavior* 121:265–296.

Ausubel, Lawrence M., and Raymond J. Deneckere. 1989. “Reputation in Bargaining and Durable Goods Monopoly.” *Econometrica* 57 (3): 511–531.

Battigalli, Pierpaolo, and Joel Watson. 1997. “On "Reputation" Refinements with Heterogeneous Beliefs.” *Econometrica* 65 (2): 369–374.

Bernheim, B Douglas, and Debraj Ray. 1989. “Collective dynamic consistency in repeated games.” *Games and Economic Behavior* 1 (4): 295–326.

Blonski, Matthias, and Daniel André Probst. 2001. “The Emergence of Trust.”

Bowen, T. Renee, George Georgiadis, and Nicolas S. Lambert. 2019. “Collective Choice in Dynamic Public Good Provision.” *American Economic Journal: Microeconomics* 11, no. 1 (February): 243–98.

Chassang, Sylvain. 2010. “Building Routines: Learning, Cooperation, and the Dynamics of Incomplete Relational Contracts.” *American Economic Review* 100, no. 1 (March): 448–65.

Coase, R. H. 1972. “Durability and Monopoly.” *The Journal of Law and Economics* 15 (1): 143–149.

Cripps, Martin W, and Jonathan P Thomas. 2003. “Some asymptotic results in discounted repeated games of one-sided incomplete information.” *Mathematics of Operations Research* 28 (3): 433–462.

Cripps, Martin W., Klaus M. Schmidt, and Jonathan P. Thomas. 1996. “Reputation in Perturbed Repeated Games.” *Journal of Economic Theory* 69 (2): 387–410.

Farrell, Joseph, and Eric Maskin. 1989. “Renegotiation in repeated games.” *Games and economic behavior* 1 (4): 327–360.

Fudenberg, Drew, and David K. Levine. 1989. “Reputation and Equilibrium Selection in Games with a Patient Player.” *Econometrica* 57 (4): 759–778.

Ghosh, Parikshit, and Debraj Ray. 1996. “Cooperation in Community Interaction Without Information Flows.” *The Review of Economic Studies* 63, no. 3 (July): 491–519.

Gompers, Paul A. 1995. “Optimal Investment, Monitoring, and the Staging of Venture Capital.” *The Journal of Finance* 50 (5): 1461–1489.

Gul, Faruk, Hugo Sonnenschein, and Robert Wilson. 1986. “Foundations of Dynamic Monopoly and the Coase Conjecture.” *Journal of Economic Theory* 39 (1): 155–190.

Hart, Sergiu. 1985. “Nonzero-Sum Two-Person Repeated Games with Incomplete Information.” *Mathematics of Operations Research* 10 (1): 117–153.

Horstmann, Ignatius, and James Markusen. 1996. “Exploring New Markets: Direct Investment, Contractual Relations and the Multinational Enterprise.” *International Economic Review* 37 (February): 1–19.

———. 2018. “Learning to Sell in New Markets: A Preliminary Analysis of Market Entry by a Multinational Firm.” *Review of International Economics* 26 (5): 1040–1052.

Kartal, Melis, Wieland Müller, and James Tremewan. 2019. “Building Trust: The Costs and Benefits of Gradualism.” *Available at SSRN 3324993.*

Kranton, Rachel E. 1996. “The Formation of Cooperative Relationships.” *The Journal of Law, Economics, and Organization* 12, no. 1 (April): 214–233.

Malcomson, James M. 2016. “Relational Incentive Contracts With Persistent Private Information.” *Econometrica* 84 (1): 317–346.

———. 2020. “Grouping Agents with Persistent Types.” *unpublished draft.*

Miller, David A., and Joel Watson. 2013. “A Theory of Disagreement in Repeated Games With Bargaining.” *Econometrica* 81 (6): 2303–2350.

Pei, Harry. 2021. “Trust and betrayals: Reputational payoffs and behaviors without commitment.” *arXiv preprint arXiv:2006.08071.*

Rauch, James, and Joel Watson. 2003. “Starting Small in an Unfamiliar Environment.” *International Journal of Industrial Organization* 21 (7): 1021–1042.

Rob, Rafael, and Arthur Fishman. 2005. “Is Bigger Better? Customer Base Expansion through Word-of-Mouth Reputation.” *Journal of Political Economy* 113 (5): 1146–1162.

Rubinstein, Ariel. 1980. “Strong perfect equilibrium in supergames.” *International Journal of Game Theory* 9 (1): 1–12.

Shalev, Jonathan. 1994. “Nonzero-sum two-person repeated games with incomplete information and known-own payoffs.” *Games and Economic Behavior* 7 (2): 246–259.

Sobel, Joel. 1985. “A Theory of Credibility.” *The Review of Economic Studies* 52, no. 4 (October): 557–573.

Tian, Xuan. 2011. “The causes and consequences of venture capital stage financing.” *Journal of Financial Economics* 101 (1): 132–159.

Watson, Joel. 1993. “A "Reputation" Refinement without Equilibrium.” *Econometrica* 61 (1): 199–205.

———. 1999. “Starting Small and Renegotiation.” *Journal of Economic Theory* 85 (1): 52–90.

———. 2002. “Starting Small and Commitment.” *Games and Economic Behavior* 38 (1): 176–199.

———. 2013. “Contract and game theory: Basic concepts for settings with finite horizons.” *Games* 4:457–496.

———. 2017. “A General, Practicable Definition of Perfect Bayesian Equilibrium.” *unpublished draft.*

———. 2021. “Convergence of Discrete-Time Models with Small Period Lengths.” *unpublished draft.*

Ye, Maoliang, Jie Zheng, Plamen Nikolov, and Sam Asher. 2020. "One Step at a Time: Does Gradualism Build Coordination?" *Management Science* 66 (1): 113–129.

Supplementary Appendix

In this appendix, we provide an additional existence result: an alteration-proof equilibrium that exhibits gradualism, trust, and cooperation by good types in every continuation (including after a deviation by player 1).

Theorem 4. *Under Assumption 1, the trust game has a constant-proportion equilibrium with parameter $\gamma = 1$ that specifies a trusting equilibrium in the continuation game following any history. In fact, after a deviation by player 1, within two periods equilibrium play coincides with a continuation on the original equilibrium path.*

Incidentally, in this equilibrium, when considering an alteration the players anticipate no further alterations in the future.

Proof of Theorem 4:

In reference to the constant-proportion equilibrium definition, consider the case of $\gamma = 1$. We use the same on equilibrium path sequence as in Appendix B, but different off-equilibrium-path specifications.

In the equilibrium we now construct, for every history of play to the beginning of any period, player 1's updated belief about player 2's type will be given by the posterior of F conditioned on $x \leq \overline{x}$ for some number $\overline{x} \geq \Delta/(1-\delta)$. In other words, every continuation game from the start of any period will be defined by an upper-truncated type space. For a given number $\overline{x}$, let us call this continuation game the *$\overline{x}$-truncation continuation game.* Thus, we can fully describe player 1's equilibrium strategy by stating the level player 1 is prescribed to choose in the first period of the $\overline{x}$-truncation game, for every $\overline{x} \geq \Delta/(1-\delta)$. Likewise, player 2's equilibrium strategy will be fully described by stating the set of types that betray in the first period of the $\overline{x}$-truncation game after player 1's choice α in this period, for every $\overline{x} \geq \Delta/(1-\delta)$ and for every $\alpha \in [0, 1]$.

Before describing the strategies, let us make a few notes. Recall that indifference Condition (35) means that type x^{k-1} is indifferent between betraying at level α^{k-1} in one period and waiting to betray at level α^k in the next period. Rearranging this equation yields

$$\alpha^{k-1} = \alpha^k \cdot \frac{x^{k-1}\delta}{x^{k-1} - \Delta}.$$

Now think about the level $\breve{\alpha}^k$ such that type x^k of player 2 would be indifferent between betraying at level $\breve{\alpha}^k$ in one period and waiting to betray at level α^k in the next period. This level is given by

$$\breve{\alpha}^k = \alpha^k \cdot \frac{x^k \delta}{x^k - \Delta},$$

and clearly $\breve{\alpha}^k \in (\alpha^{k-1}, \alpha^k)$. Observe that $\breve{\alpha}^k$ is increasing in k.

Here is the specification of strategies. Consider any $\overline{x}$-truncation continuation game and let ℓ be such that $\overline{x} \in (x^\ell, x^{\ell-1}]$. If $\overline{x} < x^{\ell-1}$ then player 1 is prescribed to choose $\alpha = \alpha^\ell$ in the current period. If $\overline{x} = x^{\ell-1}$ then player 1's specified behavior depends on whether player 1 deviated in the previous period. If player 1 did not deviate in the previous period then player 1 is supposed to choose $\alpha = \alpha^\ell$. If player 1 deviated in the previous period then player 1 is supposed to randomize between $\alpha = \alpha^\ell$ and $\alpha = \alpha^{\ell-1}$, with the probabilities described below.

For whatever level α' that is actually chosen by player 1, player 2's prescribed behavior is determined as follows. If $\alpha' \leq \alpha^{\ell-1}$ then all types above $\Delta\alpha^{\ell-1}/(\alpha^{\ell-1} - \delta\alpha^\ell)$ cooperate and types below betray. If $\alpha' > \alpha^{\ell-1}$ then find the integer ℓ' such that $\alpha' \in [\breve{\alpha}^{\ell'-1}, \breve{\alpha}^{\ell'})$ and $\alpha' \geq \alpha^{\ell-1}$. Player 2's action is then specified as follows:

If $\alpha' \in [\breve{\alpha}^{\ell'-1}, \alpha^{\ell'-1})$ then all types strictly greater than $x^{\ell'-1}$ betray and all types weakly below $x^{\ell'-1}$ cooperate. This is rational because player 1 in the following period will randomize between $\alpha^{\ell'}$ and $\alpha^{\ell'-1}$ with exactly the probabilities that make type $x^{\ell'-1}$ indifferent between betraying at level α' in the current period and waiting to betray in the next period. Note that in this case the continuation game from the next period is a truncation with cutoff $\overline{x}' \equiv x^{\ell'-1}$.

Let us calculate the probability p that player 1 must put on level $\alpha^{\ell'-1}$ in the next period to make type $x^{\ell'-1}$ indifferent. That $\alpha' \in [\breve{\alpha}^{\ell'-1}, \alpha^{\ell'-1})$ ensures that such a probability exists because, given the definition of $\breve{\alpha}^{\ell'-1}$, type $x^{\ell'-1}$ would strictly prefer to betray immediately if player 1 would choose $\alpha^{\ell'-1}$ in the next period, and would strictly prefer to wait if player 1 would choose $\alpha^{\ell'}$ in the next period. Type $x^{\ell'-1}$'s indifference condition is:

$$x^{\ell'-1}\alpha' = \Delta\alpha' + \delta x^{\ell'-1}(p\alpha^{\ell'-1} + (1-p)\alpha^{\ell'}),$$

which yields

$$p = \frac{\frac{x^{\ell'-1} - \Delta}{\delta x^{\ell'-1}}\alpha' - \alpha^{\ell'}}{\alpha^{\ell'-1} - \alpha^{\ell'}}. \tag{54}$$

If $\alpha' \in [\alpha^{\ell'-1}, \breve{\alpha}^{\ell'})$ then all types strictly greater than $\overline{x}' \equiv \alpha'\Delta/(\alpha' - \delta\alpha^{\ell'})$ betray and all types weakly below $\overline{x}'$ cooperate. This is rational because player 1 in the following period will choose $\alpha^{\ell'}$ for sure, making type $\overline{x}'$ indifferent between betraying at level α' in the current period and waiting to betray at level $\alpha^{\ell'}$ in the next period. Note that in this case the continuation game from the next period is a truncation with cutoff $\overline{x}' \in (x^{\ell}, x^{\ell-1}]$.

Denote $\hat{v}_1(\overline{x}; \alpha')$ as player 1's continuation value in $\overline{x}-$truncation game, when player 1 chooses α', assuming players follow prescribed strategies after player 1 deviates. Therefore, player 1's continuation value is

$$\begin{aligned}\hat{v}_1(\overline{x}; \alpha') = &\left(1 - \frac{F(\overline{x}')}{F(\overline{x})}\right)(-c\alpha') + \frac{F(\overline{x}')}{F(\overline{x})}\Delta\alpha' \\ &+ \delta\frac{F(\overline{x}')}{F(\overline{x})}\left(\left(1 - \frac{F(x^{\ell'})}{F(\overline{x}')}\right)(-c\alpha^{\ell'}) + \frac{F(x^{\ell'})}{F(\overline{x}')}\left(\Delta\alpha^{\ell'} + \delta v_1^{\ell'+1}\right)\right), \quad (55)\end{aligned}$$

for $\alpha' \in [\alpha^{\ell'-1}, \breve{\alpha}^{\ell'})$, and

$$\hat{v}_1(\overline{x}; \alpha') = \left(1 - \frac{F(x^{\ell'-1})}{F(\overline{x})}\right)(-c\alpha') + \frac{F(x^{\ell'-1})}{F(\overline{x})}\left(\Delta\alpha' + \delta\frac{\Delta}{1-\delta}\alpha^{\ell'-1}\right), \quad (56)$$

for $\alpha' \in [\breve{\alpha}^{\ell'-1}, \alpha^{\ell'-1})$.

Lemma 12. *In an $\overline{x}$-truncation continuation game and given the strategy, player 1's continuation value from any deviation α' will be weakly lower than some alteration.*

Proof. We define ℓ such that $\overline{x} \in (x^{\ell}, x^{\ell-1}]$. Suppose player 1 deviates to $\alpha' \geq \alpha^{\ell-1}$, we first find the $\ell' \geq \ell$ such that $\alpha' \in [\breve{\alpha}^{\ell'-1}, \breve{\alpha}^{\ell'})$. Next according to the strategy, we discuss the following two cases: $\alpha' \geq \alpha^{\ell'-1}$ and $\alpha' < \alpha^{\ell'-1}$.

In the case of $\alpha' \in [\alpha^{\ell'-1}, \breve{\alpha}^{\ell'})$, all types strictly greater than $\overline{x}' = \alpha'\Delta/(\alpha' - \delta\alpha^{\ell'})$ betray and all types weakly below $\overline{x}'$ cooperate. By choosing α' in current period and $\alpha^{\ell'}$ in the following period, player 1's continuation value becomes Equation (55). We substitute $v_1^{\ell'+1} = \alpha^{\ell'}\Delta/(1-\delta)$, and $\alpha'/\alpha^{\ell'} = \delta\overline{x}'/(\overline{x}' - \Delta)$ into Equation (55) rearrange terms and get

$$\hat{v}_1(\overline{x}; \alpha') = \delta\alpha^{\ell'}\left(c + \frac{\Delta}{1-\delta}\right)\frac{F(x^{\ell'})}{F(\overline{x})} + \left(\left(-c + (c+\Delta)\frac{F(\overline{x}')}{F(\overline{x})}\right)\frac{\delta\overline{x}'}{\overline{x}' - \Delta} - c\delta\frac{F(\overline{x}')}{F(\overline{x})}\right)\alpha^{\ell'}.$$

To find player 1's best deviation in this case, we differentiate $\hat{v}_1(\overline{x}; \alpha')$ with respect to $\overline{x}'$ and get

$$\alpha^{\ell'}\frac{\delta\Delta}{(\overline{x}'-\Delta)^2}\left(c\frac{F(\overline{x})-F(\overline{x}')}{F(\overline{x})}+\frac{f(\overline{x}')}{F(\overline{x})}(\overline{x}'+c)(\overline{x}'-\Delta)-\frac{F(\overline{x}')}{F(\overline{x})}\Delta\right),$$

which is positive for $\Delta \leq \overline{\Delta}$. Hence, we conclude that $\hat{v}_1(\overline{x};\alpha')$ increases in $\overline{x}'$. Further, because of $\overline{x}' = \alpha'\Delta/(\alpha'-\delta\alpha^{\ell'})$, this implies that $\hat{v}_1(\overline{x};\alpha')$ decreases in α' and the optimal deviation for $\alpha' \in [\alpha^{\ell'-1}, \breve{\alpha}^{\ell'})$ is $\alpha' = \alpha^{\ell'-1}$.

In the case of $\alpha' \in [\breve{\alpha}^{\ell'-1}, \alpha^{\ell'-1})$, all types weakly below $x^{\ell'-1}$ cooperate, and player 1 randomizes in the following period by putting probability $1-p$ on $\alpha^{\ell'}$ and probability p on $\alpha^{\ell'-1}$, with p given by Equation (54). One can verify that $\alpha' \in [\breve{\alpha}^{\ell'-1}, \alpha^{\ell'-1})$ implies $p \in (0,1]$. Player 1 's continuation value from the α' deviation, Equation (56), simplifies to

$$\hat{v}_1(\overline{x};\alpha') = \alpha'\left(-c+\frac{F(x^{\ell'-1})}{F(\overline{x})}(c+\Delta)\right)+\delta\frac{F(x^{\ell'-1})}{F(\overline{x})}\frac{\Delta}{1-\delta}\alpha^{\ell'-1}.$$

Note that $\hat{v}_1(\overline{x};\alpha')$ is linear in α', allowing us to conclude that the best way to deviate within the interval $[\breve{\alpha}^{\ell'-1}, \alpha^{\ell'-1}]$ is to set α' equal to one of the boundaries. Since the interval is open at the upper boundary, we are using the fact that player 1's continuation value is continuous there. To see this, let us look at the lower boundary and the interval below it. At $\alpha' = \breve{\alpha}^{\ell'-1}$, the continuation value is

$$\hat{v}_1(\overline{x};\alpha^{\ell'-1}) = \alpha^{\ell'-1}\frac{x^{\ell'-1}\delta}{x^{\ell'-1}-\Delta}\left(-c+\frac{F(x^{\ell'-1})}{F(\overline{x})}\left(c+\Delta+\frac{\Delta}{1-\delta}\frac{x^{\ell'-1}-\Delta}{x^{\ell'-1}}\right)\right),$$

which is the same as

$$\lim_{\alpha'\to\breve{\alpha}^{\ell'-1}}\hat{v}_1(\overline{x};\alpha' \in [\alpha^{\ell'-2}, \breve{\alpha}^{\ell'-1})).$$

Recall in the case of $\alpha' \in [\alpha^{\ell'-2}, \breve{\alpha}^{\ell'-1})$, we have proved that $\hat{v}_1(\overline{x};\alpha')$ monotonically decreases in α', so we conclude that comparing to $\alpha' = \breve{\alpha}^{\ell'-1}$, player 1 is able to obtain a higher payoff by choosing $\alpha' = \alpha^{\ell'-2}$. Therefore, combing the two cases, we have that player 1's optimal deviation in the interval $[\alpha^{\ell'-1}, \alpha^{\ell'-2}]$ is one of the endpoints $\{\alpha^{\ell'-1}$ and $\alpha^{\ell'-2}\}$. Finally, because deviations with $\alpha' = \alpha^{\ell'-1}$ and $\alpha^{\ell'-2}$ are equivalent to alterations with $(\ell, \ell'-\ell-1, \alpha^{\ell'-1})$ and $(\ell, \ell'-\ell-2, \alpha^{\ell'-2})$ respectively, we conclude that player 1 has no incentive to deviate in an alteration-proof PBE. □

Lemma 12 implies that if player 1 has no incentive to alter the game, then she also has no incentive to deviate. It remains to show that the PBE with the specified strategy is alteration-proof. However, as the on equilibrium path outcome for the this equilibrium is a special case of the alteration-proof PBE in Appendix B, we apply Lemma 4 and 6 with $\gamma = 1$ to conclude that the prescribed strategy constitutes an alteration-proof PBE.