

## 1 Overview and miscellaneous

I have slightly clarified the discussion around Results 2 and 3 in the four paragraphs following Equation 15 [ine:Tdelta1] on page 28 [check]. I realized that it could be more transparent when re-examining the results under the alternative models.

## 2 Effort in government welfare function

### Editor Point 2:

There is a question in my mind about the definition of welfare and the objective functions of the executive and the legislature. As I understand it, the lobbying effort involves a resource cost (this is not a cash transfer), and this cost in principle should be reflected in the expression for welfare. Viewed from a different perspective: you assume that the executive and the legislature care about import-competing profits, and thus about the lobby in a broad sense, but don't care about the resource cost of effort that the lobby incurs, and this seems hard to justify. I am sorry to be raising this issue at this stage and not in the previous round, but I became aware of it recently.

### 2.1 Dixit, Grossman and Helpman analogy

#### Editor Point 3:

You state that your specification of legislature preferences can be seen as a special case of the Dixit-Grossman-Helpman (DGH) model. It would be helpful if you could substantiate this claim. More specifically, let us consider a simple version of the DGH specification in your setting. Suppose you modify your model only in two ways. First, suppose the legislature preferences are given by  $W+g(e)$ , where  $g(e)$  is an increasing and concave function of contributions (while the lobby preferences remain the same). And second, suppose the lobby offers a contribution schedule  $e(t)$  before the legislature chooses the tariff. The question is: would this model deliver the same results as yours (at least qualitatively)? Note that in this setting you can focus on a simple all-or-nothing contribution schedule, of the kind "if you give me a 5% tariff I give you \$100, otherwise you get nothing." Intuitively the promised contribution will just compensate the legislator for the loss associated with the requested tariff, so the analysis might not be hard. If this DGH version of your model yields similar qualitative results, pointing out this "isomorphism" would help you in several ways. First, it would provide "foundations" for your assumed legislature

preferences, in terms of a model (DGH) that people are familiar with. Second, this would help address my question 2 above: in the DGH model we can think of  $e$  as money, and I think it would be reasonable to stick with your current definition of aggregate welfare (even though in principle one might question this definition of aggregate welfare when utility is not transferrable). And third, you could examine whether your results rely on the presence of diminishing marginal utility from contributions: what would happen if  $g$  is linear (as in the basic Protection for Sale model) rather than strictly concave?

- *I've supplied the suggested results to connect my model to that of Dixit, Grossman and Helpman (1997) in Appendix B.2, along with some text pointing to it at the end of Section 2.2 [page 10]. There results from the DGH-style model are qualitatively different in a minor way, and I discuss this in Appendix B.1 where I present a unitary government/legislature version of the model that most closely matches Maggi and Rodriguez-Clare 2007. To summarize the difference, when the decision-maker is assumed to be unitary as seems necessary in the DGH model, the parameter space over which the optimal value of  $T$  in Section 6 is strictly positive and finite is smaller than in the non-unitary model.*
- *I have also verified that none of the results of the model change if lobbying effort is subtracted from the legislative/executive welfare function (except, of course, the evaluated level of welfare). I have refrained from adding the " $-e$ " to the welfare function out of a desire to be consistent with the existing literature, as I state at the very end of Section 2.2 [page 10]. The only other cost of adding it in is the obvious additional complication to the mathematical expressions.*
  - *It is interesting to note that in the unitary model of Appendix B.1, this change would slightly loosen the legislative constraint by making the future punishment harsher at any given  $\tau^a$ .*

### 3 Interpretation: SOP vs. time inconsistency

#### Referee 1 Point 1:

I believe the interpretation of the model as representing two branches of the government is arbitrary. It is more realistic to interpret the model as representing the preferences of a government with time-inconsistent preferences, such that ex ante the government's objective is to maximize social welfare, but at the time of implementing the agreement it might be influenced by lobby group activities.

I understand that the author has tried to modify the interpretation of the model to address this concern, but in my view the modification was insufficient.

- *I agree that the separation-of-powers interpretation versus a one-branch interpretation along the lines of the one discussed by the editor (see immediately below) are equally valid. I have therefore added a discussion of this one-branch interpretation on pages [14-15].*

#### Editor Point 5:

Regarding Referee 1's idea about time-inconsistent preferences, I guess I am ok with your separation-of-powers interpretation (subject to the caveats I expressed in my first-round letter), but I also think it would be useful to discuss a possible alternative interpretation along the lines of Referee 1's idea. I am thinking of the Maggi and Rodriguez-Clare setting, where there is a unitary government that cares about welfare and contributions, but at the stage of signing the agreement ("ex ante" stage) the lobby has no influence on the government, because the lobby cares only about the short run (due to the fact that specific capital is mobile in the long run). The way I think about this interpretation is slightly different from Referee 1, in that there is no time inconsistency of preferences: the government has the same preferences across time, but ex-ante the lobby is not active. I am not sure whether a Maggi and Rodriguez-Clare specification of this kind would yield similar results as yours. This might be an interesting question to discuss, and if the answer is yes, it would be useful to point out that the model admits also this alternative interpretation.

- *As I have said immediately above, I agree completely. I have added this discussion on pages 14-15 and also used this framing in the introduction of Appendix B. I should point out, as I do in the text, that the model admits this 'one-branch' interpretation only if the one branch is non-unitary. I explore the unitary model in Appendix B.1.*

## 4 Continuation payoffs and changes in the preferences

### Editor Comment 1:

I agree with Referee 2 that there are still problems with the analysis. Correcting these problems is a necessary (though not sufficient) condition for me to move forward with this paper. You will need to convince us beyond the reasonable doubt that the analysis is correct.

- *Please see below.*

### Editor Point 8:

The way you write the key program in (9)-(11) is confusing because you have an “e” floating around, and it is not clear where it should be evaluated. Unless I am missing something, in some places this should be  $\bar{e}(\tau^a)$  and in other places it should be  $e^a$ . It would also be helpful to write  $e^a$  as a function of  $\tau^a$ . Since the choice variable in the program is  $\tau^a$ , you should make clear what is a function of  $\tau^a$  and what is not.

- *I have denoted  $e_a$  as  $e_a(\tau^a)$  throughout the text to make clear the dependence on the trade agreement tariff.*
- *Similarly, I have been careful to use  $\bar{e}(\tau^a)$  throughout (instead of sometimes also using  $\bar{e}$  without displaying the dependence on  $\tau^a$ ). The only exceptions are (1) when  $\bar{e}$  is an argument in  $\gamma$  in the proofs to prevent expressions from becoming too long and (2) when it could be confusing as when discussing explicitly the relationships between tariffs and  $\bar{e}$ .*
- *I have labeled any  $e$  that is not meant to describe a general functional form such as in Assumption 1 as  $e_b$ . This is what you will now see, for instance, in Equations (10) and (11). Note that the value of  $e_b$  that generally of most interest is  $\bar{e}(\tau^a)$ , and that in equilibrium  $e_b = 0$ .*

### Referee 1 Point 3:

It is assumed that the current legislator evaluates future welfare (i.e., the continuation payoffs) based on the expected preferences of the future government, which will be induced by lobbying efforts in the future. Alternatively, it could be assumed that the current legislator evaluates future profits based on its current preferences.

The latter assumption might be more consistent with the premise of the model, which is essentially a decision-making model with time-inconsistent preferences. Moreover, I think the results related to self-enforceability of the agreement will continue to hold if the author adopts the latter assumption.

Some discussion of this point could be illuminating.

- *I believe that the setup in the body of the paper is consistent with the “latter assumption,” and so I’m happy to read this argument that it is more consistent with the premise of the model. I have labeled this difference as an assumption as to whether the decision-maker is unitary or non-unitary. I believe it is what underlies the concerns of Referee 2 below about the analysis. I have therefore added Appendix B.1 to analyze the ‘unitary’ version of the model. All of the results go through unchanged in a qualitative sense with one small exception: the parameter space over which there is a strictly positive and finite optimal punishment length is smaller in the unitary model. This is a direct result of the fact that the lobbying effort at the time of the break decision does not interact with the lobby’s future profits from the point of view of the government.*

### Referee 2 Point 3:

I am confused about one basic aspect of the incentive constraint for the legislature. Consider the RHS of (7) on page 16. If the lobby chooses  $e$  and the legislature decides to select its best response given  $e$ ,  $\tau^R(e)$ , and thereby breaks the agreement, then a T-period punishment is launched in the next period. The notation in (7) (and likewise in (10)) suggests that the lobby continues choosing the same  $e$ , thus generating the same  $\gamma(e)$ , during the punishment phase. I don’t see why this would be the case. Wouldn’t the lobby instead choose the effort level  $e_{tw}$  that is determined in the first-order condition given by (6)? And indeed, if we look at the lobby incentive constraint, as given by the RHS of (8) (and likewise in (11))), we see that that constraint does assume that  $e_{tw}$  is used by the lobby during the punishment phase. I can’t tell exactly what is going on here. There could be an oversight, or I could be misunderstanding the notation. At a minimum, some clarification is needed.

- *I agree with you about the effort levels put forth by the lobby. I think the confusion comes from the fact that I did not clearly explain how the assumption of a non-unitary legislature translates into this incentive constraint. In the non-unitary legislature, if the current period incentive constraint evaluated  $\gamma(\cdot)$  at the future values of lobbying effort, this would be interpreted as the current-period median legislator evaluating the current-period incentive constraint using a mixture of his own weight on import-competing profits with those of the legislators who are median in the future in the trade agreement and trade-war scenarios. The first major step I have taken to clarify the incentive constraint is the analysis in Appendix B.1 of the unitary model (described above), which I would guess is the model you had in mind. I discuss the differences at length in the beginning of that Appendix.*
- *The second major step I have taken to clarify the incentive constraint is the addition of the following paragraph immediately following Equation 12:*

Note that the median legislator, whose identity is determined by the lobby's effort level  $e_b$ , evaluates future payoffs according to her own political economy weight,  $\gamma(e_b)$ . Of course, depending on legislator  $e_b$ 's choice, either legislator  $e_a$  or legislator  $e_{tw}$  will be the decision maker in those future periods. But legislator  $e_b$ , who is the decision maker in the current period, maximizes her own welfare given the predicted behavior of future decision makers.

This consideration also leads to a further concern/question. If my point above is correct and the lobby should be modeled in (7) as choosing  $e_{tw}$  in the punishment phase, and if  $e_{tw} > e$ , then would the legislature ever deviate (even for  $e$  in the non-triggering range as currently defined) in order to trigger a trade war and thereby enjoy the higher  $e_{tw}$  and thus the higher gamma that the trade war elicits? Recall that  $W_{ML}$  is increasing in  $e$  as an independent argument. Is this potential incentive captured?

- *That  $\bar{e}$  must be greater than  $e_{tw}$  is a requirement for a non-trivial trade agreement to be enforceable (see the first full paragraph after Equation 12 and the three paragraphs leading up to Result 1), because the lobby would always prefer  $e_{tw}$  to any lower effort level.*

**Referee 2 Point v:**

Page 20: Related to comment 3 above, I don't follow why in (12) that  $e$  can't change from  $\bar{e}$  as we move into the trade war.

- *Please see response to Referee 2 Point 3.*

## 5 More comparison to MRC

### Referee 1 Point 2:

As part of Result 1, the author states that “The equilibrium trade agreement is never subject to dispute.” But this is true by construction of the equilibrium. In other words, the author finds an equilibrium that is self-enforcing and, thus, no dispute arises. Therefore, I believe that this statement is not sufficiently interesting or insightful to be part of Result 1.

However, I find the discussion immediately following the Result 1 quite interesting. The author shows that at the equilibrium, the applied tariff is equal to the negotiated binding. That is, the lobby group exert sufficient effort to induce the government to apply a tariff as high as the negotiated cap. Moreover, it is interesting that in the equilibrium the self-enforcing constraint for the legislature is not binding.

The former result is reminiscent of Maggi and Rodriguez-Clare’s finding in their commitment model that the optimal form of trade agreement is a tariff cap, rather than a particular tariff rate. The justification is similar in both papers: by allowing the government to set a tariff below the cap, the lobby groups are induced to pay for the increase in the applied tariff, which reduces the incentives of the lobby groups to push for more protection.

I think more insights could be generated by comparing this paper and those of Maggi and Rodriguez-Clare. Therefore, I recommend the author to provide more discussion of how her paper is related to Maggi and Rodriguez-Clare.

### Editor Point 4:

Referee 1 mentions that your results are reminiscent of Maggi and Rodriguez-Clare’s finding that weak bindings are preferable to strong bindings, though for slightly different reasons (there are no issues of self-enforcement in Maggi and Rodriguez-Clare). You do not explicitly compare a weak-binding agreement with a strong-binding agreement (unless I missed it), but if you could show that the former is preferable, this would be an interesting result worthy of emphasis.

- *I’ve removed the statement about the trade agreement not being subject to dispute from Result 1 and added references to the other two features of the equilibrium suggested by Referee 1. I have modified the paragraph that follows Result 1 to reflect this change. I have also added an additional paragraph that discusses the relationship to Maggi and Rodriguez-Clare 2007. To support this discussion, I have added Appendix B.3 with analysis of the strong binding case. I relegated most of this discussion to an appendix because it is quite nuanced and I’d prefer not to distract from the main results.*
- *Section 3.1 also contains a discussion of the relationship between my model and Maggi and*

*Rodriguez-Clare 2007.*

- *I have added a references to both discussions on pages 2-3 of the introduction.*



## 6 Punishment length vs. Dispute length

### Referee 2 Point 1:

My first comment concerns the punishment length,  $T$ . The paper highlights an interesting trade-off between the advantages of a high  $T$  (which helps to deter cheating by the legislature) and a low  $T$  (which helps to deter disruptive lobby effort), where the value for  $T$  in turn interacts with how low the cooperative bound tariff level can be pushed. But an ongoing concern for me is the interpretation of the endogenous determination of punishment length,  $T$ . I don't have a good sense of how to interpret the real-world determination of  $T$ . The author acknowledges that the  $T$ -period punishment never occurs along the equilibrium path but argues that the determination of  $T$  can be related to the choice of design for a dispute settlement system. To me, though, I think that there is some tension involved when using a model where all dispute behavior is off the equilibrium path as a means of interpreting an existing dispute settlement system or evaluating a proposed new system. Disputes actually happen, and dispute settlement systems are designed with that in mind.

As the author notes, dispute activity would occur along the equilibrium path if the model were modified to include shocks. Presumably, such a modification would reinforce the value of punishment phases of limited duration. But I worry that the model with equilibrium-path punishments could bring into play new considerations not currently featured in the author's model, since the frequency and duration of the punishment would then directly affect payoffs. And if shocks were public, then arguably the enrichment of the model to consider escape clause rules (contingent trade policies) becomes more compelling.

### Editor Point 6:

Regarding Referee 2's concerns, I agree on a basic point: it's not clear how much we can learn about the design of a dispute settlement system from a model that has no disputes (nor punishments) on the equilibrium path. My suggestion would be to take a more modest approach in pitching your analysis: rather than speaking to the design of dispute settlement procedures, you are making some more limited points about the optimal severity of punishments. You are in good company, by the way: a number of authors (including myself) have written papers on the optimal severity of punishments in trade agreements, and often using models with no punishments on the equilibrium path.

- *I have removed all claims about optimal dispute resolution mechanisms, referencing instead optimal punishments.*

## 6.1 Compare to Park (2011)

### Referee 2 Point 2:

It is interesting in this regard to compare the author's model with that of Park (2011), who also describes trade-offs that lead to the determination of a finite value for  $T$ . In his model, however, the information structure is such that punishment phases are triggered after a government observes an extreme private signal, corresponding to a private shock, and then selects a public tariff that signals the beginning of a punishment phase. (The critical value for  $T$  balances the benefit of selecting a higher public tariff against the cost of triggering a punishment phase.) In his model, punishment phases occur along the equilibrium path, and alternative dispute systems might be pursued with the goal of reducing the frequency or severity of such punishments, if possible, as Park discusses. These kinds of considerations are hard to contemplate in the author's set up, though, since the disputes are off the equilibrium path.

Park (2011) comments on the literature:

- Bagwell and Staiger (2005) and more recently Bagwell (2008) analyse the issue of implementing trade agreements when each government is privately informed about its own domestic political pressure for protection. Their analysis differs from this paper because it focuses on identifying the structure of trade agreements that can induce the truthful revelation of private political pressure.
- Earlier models developed with respect to this issue, such as Dixit (1987), Bagwell and Staiger (1990), and Riezman (1991), suggest that the WTO may serve the role of helping countries coordinate on more efficient equilibria among the multiple equilibria that typically arise in a repeated game set-up. To model a more explicit role of the WTO, Kovenoch and Thursby (1993) assume that the DSP of the WTO has an informational superiority over trading countries in distinguishing between true violations and mistaken perceptions, which in turn enhances a reputation mechanism that supports cooperation. In a multilateral trading environment, Maggi (1999) shows that the WTO may facilitate cooperation-enhancing third-country sanctions by disseminating information about deviations. While these models introduce more specific roles for the WTO to play in coordinating a cooperative equilibrium, the literature has not resolved the question of why the WTO is necessary for coordination because these previous studies offer no theory of why countries could not coordinate a cooperative equilibrium in a non-WTO environment.

This paper represents the emergence of the WTO as a change in the observation structure of a repeated game. The presence of the WTO changes the nature of punishment-triggering signals from private into public. In the absence of the WTO, the private nature of signals of potential violations limits the flexibility of punishment phases that countries can employ because these phases must provide countries with the incentive for truthful revelation of

private signals in triggering punishments. The WTO can publicize its opinions on violations, which relaxes such a constraint in designing an optimal punishment scheme, enabling a better cooperative equilibrium even in the absence of any informational superiority of the WTO. This result contrasts with the analysis of Ludema (2001) who emphasizes that the DSP of the WTO may require trade agreements to be renegotiation-proof by promoting communication among countries prior to starting punishments. According to his analysis, such communication negatively affects cooperation by forcing countries to rely on weaker punishments.

- concealed trade barriers introduced similar to Riezman (1991); I assume that each country cannot directly observe the other country's local market price of its export. For example, a mixture of a consumption tax and a production subsidy can replicate the effect of a tariff
- Maggi and Staiger (2008) analyse the possible role that the DSP of the WTO plays in completing an incomplete contract and characterize the optimal choice of contractual incompleteness and the DSP design. In a related study, Maggi and Staiger (2009) characterize optimal remedies for breaches of trade agreements in the presence of uncertain political pressure for protection, for which the DSP may generate noisy signals. Beshkar (2008) analyses how the rulings of the DSP can affect renegotiation of trade agreements in the context of designing a direct revelation bargaining mechanism. However, they do not introduce imperfect private signals of potential deviations into their models, and so such signals play no role in their analyses of the DSP of the WTO.
- Hungerford (1991) develops a model in which the WTO plays a negative role in enforcing trade agreements because the model assumes that the DSP of the WTO involves uninformative and costly investigation.
- Forward cite search turns up Bajona and Ederington (2012 working paper), Anesi and Facchini (2016 working paper from Midwest)

## 6.2 Change title

I need to take out dispute settlement

- Should I also take out separation of powers?

## 7 Shorten section 8

### Editor Point 7:

Regarding the new section 8 on alternative punishments, I appreciate the additional work you have done, but given that the paper is quite long, I would suggest reducing this section to a couple of paragraphs that summarize the main points and provide the basic intuition. You can place the analysis in an online appendix, or keep it in the working paper version, if you like.

- *I've shortened the section and opted to keep the full analysis in the working paper only.*

## 8 Citation for optimal punishment length of infinity

### Editor Point 9:

I am not sure I understand the way you cite Klimenko-Ramey-Watson (KRW) in section 6. The point that in a standard model (without lobbying and without punishments on the equilibrium path) more severe punishments are always better is a completely standard result that was made in many papers before KRW.

## 9 Referee 2 Minor Comments

i). Pages 8 and 9: Is  $\pi_x$  in (2) the same as  $\pi$  in (1)?

- *Yes, they are the same. I have added an "X" subscript to the lobby's profits throughout the paper to clarify.*

ii). Page 12: The notation  $W_E$  and  $W^*_E$  is used in (3) before being formally defined.  $W_E$  is subsequently defined in (4).

- *The exposition would be more confusing if I reversed the order of Equations (3) and (4), so I added a verbal description of  $W_E$  and  $W^*_E$  in the paragraph before Equation (3) with a footnote referencing Equation (4).*

iii). Page 16: The author sometimes refers to "continuation values" as corresponding to current payoffs plus discounted future payoffs, but I think the more conventional use of the term refers only to the latter (i.e., to the  $V$  terms in the incentive constraint).

- *Thank you for pointing this out. I have clarified the terminology wherever it is not clear that I'm referring to the future discounted stream of payoffs from period  $t+1$  [pages 16, 19]*

iv). Page 19: The author says that, in the absence of a trade agreement, the lobby has no incentive to be active. I don't know what this means - wouldn't the lobby still choose  $e_{tw}$  in that case?

- *Yes, this is absolutely correct. Thank you for catching this opportunity for clarification. I have changed the sentence (now on page [20]) to read “...the lobby has no incentive to exert effort to break the trade agreement...”*

vi). Page 21: I was a little confused by the statements of Lemmas 1 and 2. Are we assuming here that parameters are in a region such that an equilibrium exists under which  $\bar{e}$  exceeds  $e_{tw}$ ? Existence is treated on the next page, but I wasn't sure what was being assumed on this page and in these lemmas.

- *I see now that the discussion of Lemma 1 was cause for confusion. These lemmas hold for any value of  $\bar{e}(\tau^a) > e_a$ , that is, any value of  $e_b$  that would lead to a tariff higher than the trade agreement tariff. I have simplified the paragraph following Lemma 1 so as not to raise the extraneous point about  $\bar{e}(\tau^a)$  needing to be greater than  $e_{tw}$  in equilibrium. I have also moved discussion of this on the previous page from a footnote into the main text.*

vii). Pages 34-35: I was confused by the discussion of the credibility of the alternative punishment scheme. I probably missed something, but here is where I got confused. In a punishment period, if the lobby deviated with a very high  $e$ , and thereby generated a very high gamma value, would the legislature be able to commit not to respond with a higher tariff? I.e., is the legislature's supposed lack of response to lobbying credible?

- *Note that this discussion has been removed from the text per the editor's instruction (Editor Point 7, Section 7 above). It is not automatic that the legislature can commit to not respond with a high tariff to such a deviation of the lobby. Equation (18) on page 36 of the previous draft is the required condition.*