Conformity and out of equilibrium beliefs

Cartwright, Edward

Postprint / Postprint
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
www.peerproject.eu

Empfohlene Zitierung / Suggested Citation:
Cartwright, E. (2009). Conformity and out of equilibrium beliefs. *Journal of Economic Behavior & Organization*, *70*(1-2), 164-185. https://doi.org/10.1016/j.jebo.2008.11.003

JOURNAL OF Economic Behavior & Organization

Title: Conformity and out of equilibrium beliefs

Author: Edward Cartwright

PII: S0167-2681(08)00222-9
DOI: doi:10.1016/j.jebo.2008.11.003
Reference: JEBO 2282

To appear in: *Journal of Economic Behavior & Organization*

Received date: 31-5-2006
Revised date: 18-11-2008
Accepted date: 19-11-2008

Please cite this article as: Cartwright, E., Conformity and out of equilibrium beliefs, *Journal of Economic Behavior and Organization* (2008), doi:10.1016/j.jebo.2008.11.003

Conformity and out of equilibrium beliefs

Edward Cartwright
Department of Economics,
Keynes College,
University of Kent,
Canterbury,
Kent. CT2 7NP. UK.
E.J.Cartwright@kent.ac.uk

December 2007

Abstract

We analyze a model of conformity with contrasting inferences. Given a form of 'strong inferences', any non-conforming agent is believed to have 'extreme preferences' and can expect to receive low esteem. With a weaker form of inferences, a non-conforming agent could be inferred to have 'average preferences' and can expect a smaller fall in esteem. We find that the type of inferences need not influence whether a conformist equilibrium exists. It will, however, impact on the size of the set of conformist equilibria and thus weakening inferences acts as an equilibrium selection device.

1 Introduction

A social norm is a prescription of how a person should behave. Examples, include 'wear a suit at work', 'do not live off other people' or 'send a Christmas card to someone who sends one to you' (Elster 1989). Typically conformity to a norm involves sacrifice such as wearing a suit when jeans and a T-shirt would be preferred, so why do people adhere to norms? One reason is that actions send signals to others of a 'persons type' (Bernheim 1994). If non-conformity is seen as a signal of someone with 'extreme preferences', then conforming may be better than non-conforming because conformity, while immediately costly, leads to better treatment by others (Kreps 1997). There can exist, therefore, a 'conformist' Bayesian Nash equilibrium in which people conform to the norm and anyone who does not conform is inferred as having some 'extreme preferences'. The problem for anyone wishing to model this type of conformity is that the definition of Bayesian Nash equilibrium does not tie down 'out of equilibrium beliefs' (Banks and Sobel 1987, Cho and Kreps 1987). Basically, because no person should not conform, Nash equilibrium allows that anything could be inferred about a person who does not conform. Clearly, however, out of equilibrium beliefs are a crucial aspect of the equilibrium. Our goal in this paper, using the model of conformity developed by Bernheim, is to explore how out of equilibrium beliefs impact conformity.

To explain the issues consider, informally, the norm of 'how much to tip at a restaurant'. Suppose that the size of tip is seen as a signal of generosity. The 'ideal type of person' is someone who would like to tip 10%. People who would want to tip less are considered greedy, and people who would want to tip more are too generous. Suppose that there exists a norm to tip 15% and people adhere to this norm. If someone tips 10%, then what should others infer about this person? Even if we restrict attention to so-called 'reasonable beliefs' (Banks and Sobel 1987, Cho and Kreps 1987), there are plenty of possibilities. One possibility, let's call it *strong inferences*, is to say that anyone who tips less than 15% must be very greedy; the fact that he has not adhered to the norm is a signal that he has the 'most extreme preferences' (Bernheim). Another possibility, *weak inferences*, is to say that if someone tips 10%, then they may be very greedy, but equally they might just be the type of person who likes to tip 10%.

Intuitively, it should make a difference whether people have strong or weak inferences. If people have strong inferences, then there are strong incentives to conform because the costs of deviation are high. If people have weak inferences, then the motivations to conform are much less. Indeed a person may deviate from the norm precisely to signal that they have the 'ideal type'. To question whether it does make a difference whether inferences are weak or strong we first need to formally capture the notion of weak and strong inferences. We use the D1 Criterion, as used by Bernheim, to capture strong inferences. Informally, if inferences satisfy the D1 Criterion then any deviation from the norm is inferred to have been done by the person with the *most* incentive to deviate. To capture weak inferences, we introduce the IWD1 Criterion, which is closely related to divinity (as defined by Banks and Sobel). If inferences satisfy the IWD1 Criterion then *any* person with an incentive to deviate is inferred to be equally likely to deviate from the norm. Using the model of conformity introduced by Bernheim, and contrasting the outcome when inferences satisfy the D1 Criterion to that when inferences satisfy the IWD1 Criterion, we show whether it does make a difference if inferences are strong or weak. Our results can be summarized as follows:

Equilibrium existence: Bernheim demonstrated that *there always exists a conformist equilibrium (in his model) if inferences satisfy the D1 Criterion.* In this paper *we provide a necessary and sufficient condition, called the worse than condition, for the existence of a conformist equilibrium that can be supported by inferences satisfying IWD1.* We demonstrate, with an example, that the worse than condition need not hold, so there need not exist a conformist equilibrium when inferences satisfy IWD1. This demonstrates that using the D1 Criterion, as Bernheim did, is not innocuous. We prefer, however, to focus on the more positive conclusion that in many cases a conformist equilibrium can be sustained by weaker inferences than those of the D1 Criterion. The worse than condition will, for example, be satisfied if the desire to be inferred as a 'good type' is sufficiently high or many people have types near the ideal. Bernheim's existence result was a seminal contribution to the literature in showing how 'harsh enough penalties' for non-conformity can be produced endogenously rather than simply assumed. Our results demonstrate that weaker inferences can still produce 'harsh' penalties to non-conformity.

Equilibrium selection: One possible shortcoming of the results of Bernheim is the multiplicity of conformist equilibria. In particular, *there may exist multiple conformist equilibria supported by inferences satisfying the D1 Criterion, all based on a different norm.* For example, there could be a conformist equilibrium where the norm is to 'tip 5%', one where the norm is to 'tip 20%', and so on. If there are many equilibria, each with a different norm, then one may question how we can think of conformity arising if no-one knows what the norm is, and everybody knows that no one knows what the norm is, and so on? Bernheim suggests, as seems reasonable, that this indeterminacy may be resolved by a focal point, possibly determined by history or a policy maker. We demonstrate, however, that weaker inferences act as an equilibrium selection device and reduce the set of actions that could potentially become norms. Indeed, *there may be at most one conformist equilibrium, and therefore at most one norm, that can be supported by any inferences satisfying IWD1.*[1] This norm will correspond to the preference of what we call the median type. The median type is characterized by a symmetry in which the 'costs' of being seen as 'above' or 'below' this type are the same. For example, if the median type is 'to tip 15%', then it is the same to be seen as 'someone who likes to tip more than 15%' as to be seen as 'someone who likes to tip less than 15%'.

Equilibrium 'efficiency': There will always exist a conformist equilibrium supported by inferences satisfying the D1 Criterion where the norm is the action preferred by the 'ideal type' (Bernheim). Intuitively one might expect that the equilibrium selection of IWD1 selects such an equilibrium. In general, however, the median type differs from the ideal type, so *there need not exist a conformist equilibrium supported by inferences satisfying IWD1 where the norm is the action preferred by the 'ideal type'.* For example, if 'tip 10%' is preferred by the ideal type but to 'tip 15%' is preferred by the median

[1] Related results are due to Azar (2004). Azar considers a model of tipping where slight deviations from a tipping norm result in only mild consequences. This could be equated to weak inferences. Azar finds that a tipping norm can only be sustained on specific tip values that will depend on the preferences of agents. There is therefore not the multiplicity of equilibria that one finds in Bernheim.

type, then there may exist a conformist equilibrium with a norm to ‘tip 15%’ but not one to ‘tip 10%’. The median type will differ from the ideal type if it is, say, better to be inferred as more generous rather than more greedy than the ideal. In this case, a norm of 10% could not be an equilibrium because some would want to tip more in order to signal that they are more generous than the ideal type.

We proceed as follows: Section 2 outlines the model, Section 3 discusses out of equilibrium beliefs, Section 4 treats equilibrium existence, Section 5 equilibrium selection, Section 6 q-uniform inferences and Section 7 concludes with an Appendix containing remaining proofs and full derivations of the examples used.

2 Model of conformity

We use a model of conformity introduced by Bernheim. The model is characterized by a separability between an agent’s intrinsic utility, determined by his own action, and esteem, determined by the type others infer him to have. It is also characterized by incomplete information about type. An agent’s action, thus, serves as a signal to others of his type. [For complete details of the model see Bernheim.]

There are a continuum of agents. Each agent chooses a publicly observable *action* x from the set $X = [0, 2]$ and has a type from the set of *agent types* $T = [0, 2]$. The type of an agent indicates his *intrinsic bliss point.* Specifically, there exists an *intrinsic utility* function $g : [0, 2] \to \mathbb{R}$ and an agent of type t receives intrinsic utility $g(x - t)$ from playing action x where

Assumption 1: Function g is twice continuously differentiable, strictly concave, symmetric and achieves a maximum at 0.

Thus, an agent of type t maximizes intrinsic utility by choosing action $x = t$, and the further his action from type, then the lower his intrinsic utility.

The distribution of types within the population is described by a cumulative density function F defined on set T and a corresponding probability density function f.

Assumption 2: The *support*$[f] = T$ and f is continuous.

The type of an agent is private information. Agents receive esteem according to the type that they are inferred to be. Specifically, there exists an *esteem function* $h : [0, 2] \to \mathbb{R}$ where $h(b)$ is the esteem of an agent who is inferred to be of type b.

Assumption 3: Function h is twice continuously differentiable, strictly concave, symmetric ($h(1+z) = h(1-z)$) and achieves a maximum at $b = 1$.

Type 1 is the *ideal type* in the sense that someone inferred to be of type 1 receives the maximum esteem. The further is inferred type from 1, then the less is esteem; so someone inferred to be of type 0 or 2 receives the least esteem.

The tension that will exist in the model between conforming and not conforming should now be clear. An agent faces the trade off between choosing an action that gives high intrinsic utility but may result in low esteem versus sacrificing intrinsic utility to earn esteem. The key feature is that each type of agent has her own preferred action that maximizes intrinsic utility, but there exists a unique type that agents wish to be inferred as.

Action will be used as a signal of type. Specifically, there exists an inference function $\phi(b, x)$ where, informally, $\phi(b, x)$ denotes the probability that an agent who chooses x is inferred to be of type b. More formally, $\phi(\cdot, x)$ is a probability density function defined on set X. Let

$$T^{\phi}(x) := \{b \in T : \phi(b, x) > 0\}$$

be the set of types that are inferred as having potentially chosen x. Let

$$H^{\phi}(x) := \int_0^2 \phi(b, x) h(b) db \qquad (1)$$

denote the esteem of an agent who chooses x. Note that esteem is a weighted average based on the esteem function and inferences.

The payoff of an agent is a weighted sum of intrinsic utility and esteem. Specifically, the payoff of an agent of type t from playing action x given inference function ϕ is

$$U(x, t, \phi) := g(x - t) + \lambda H^{\phi}(x)$$

where λ is an index of how important is esteem for the agent.

2.1 Signalling equilibria

All agents of the same type are assumed to choose the same action.[2] An *action function* μ maps T into X where $\mu(t)$ denotes the action chosen by agents of type t. The *pair* (μ, ϕ) consisting of action function μ and inference function ϕ are sufficient to determine the payoffs of all agents. Pair (μ, ϕ) is a *signalling equilibrium* if actions are optimal given inferences and inferences can be deduced from the action function using Bayes' Rule. Following Bernheim we focus on a specific type of pure strategy signalling:

A *signalling equilibrium* (μ, ϕ) is characterized by a tuple (x_p, t_l, t_h, μ_s) consisting of real numbers x_p, t_l and t_h, where $0 \leq t_l \leq x_p \leq t_h \leq 2$, and a continuous, strictly increasing function $\mu_s : [0, 1] \to X$ where[3]

$$\mu(t) = \begin{cases} \mu_s(t) \text{ if } t < t_l \\ x_p \text{ if } t \in [t_l, t_h] \\ 2 - \mu_s(2 - t) \text{ if } t > t_h \end{cases} . \qquad (2)$$

[2] Essentially this is an assumption that agents use pure strategies. Allowing mixed strategies significantly complicates the analysis and also makes the interpretation of an equilibrium much more difficult. It need not, however, be an innocuous assumption. In particular, when we look at equilibrium existence in Section 4 it should be born in mind that we are looking for the existence of a pure strategy equilibrium.

[3] Note that it is more convenient for us to use the function μ_s mapping types to actions. Bernheim uses function ϕ_s mapping actions to types. Thus, $\mu_s(t) = \phi_s^{-1}(t)$ and $\mu_s^{-1}(x) = \phi_s(x)$.

The inference function ϕ will satisfy

$$\phi(b,x)=\begin{cases}1 \text{ if } x=\mu(b) \text{ and } b<t_l \text{ or } b>t_h\\ 0 \text{ if } x=\mu(b')\neq x_p \text{ for some } b'\neq b\\ f(b)\left[F(t_h)-F(t_l)\right]^{-1} \text{ if } x=x_p \text{ and } b\in[t_l,t_h]\\ 0 \text{ if } x=x_p \text{ and } b\neq[t_l,t_h]\end{cases}. \tag{3}$$

Bernheim demonstrates that any (pure strategy) signalling equilibrium with inferences satisfying the D1 Criterion (to be explained below) can be characterized by such a tuple (x_p,t_l,t_h,μ_s). Thus, an agent with type $t<t_l$ chooses action $\mu(t)$ and is (correctly) inferred to be of type t, so receives esteem $h(t)$. An agent with type $t>t_h$ chooses action $\mu(t)$ and is (correctly) inferred to be of type t, so receives esteem $h(t)$.[4] Agents with types $t\in[t_l,t_h]$ choose action x_p. Consequently, they receive esteem

$$H^{\phi}(x_p):=\int_{t_l}^{t_h}\frac{h(b)f(b)}{F(t_h)-F(t_l)}db. \tag{4}$$

Agents with types $t\notin[t_l,t_h]$ fully separate while those with types $t\in[t_l,t_h]$ constitute a *central pool* who choose unique action x_p. Action x_p can be interpreted as the *norm*, and agents with types $t\in[t_l,t_h]$ *conform to the norm.* Note that x_p need not equal 1, but a type 1 agent will conform to the norm, so $t_l\leq 1\leq t_h$ (see Theorem 3 of Bernheim). An inference function satisfying (3) is consistent with Bayes Rule. It remains to check that actions are optimal. For this we require that

$$g(\mu(t)-t)+\lambda H^{\phi}(\mu(t))\geq g(x-t)+\lambda H^{\phi}(x) \tag{5}$$

for all $t\in T$ and $x\in X$.[5]

Throughout the following we shall use examples to illustrate the analysis. All examples will be based on the *spherical case* of $g(z)=-z^2$ and $h(b)=(1-b)^2$. Even though we have not quite finished the description of a signalling equilibrium (we do this in the next Section), it seems worthwhile to provide our first example of such an equilibrium in order to illustrate what one may look like. The example will also prove useful in discussing the further issues that will arise as we proceed.

In **Example 1** we set $\lambda=1.25$ and $f(t)=0.5$ for all $t\in[0,2]$. There exists (see the Appendix for more details on all of the examples) a signalling equilibrium (μ,ϕ) with inferences satisfying the D1 Criterion where $x_p=1, t_l\approx 0.076$ and $t_h\approx 1.924$. The norm, therefore, is 1, and any agent with type $t\in[0.076,1.924]$ conforms to the norm. Given (3) this means that $\phi(b,1)=(t_h-t_l)^{-1}$ for all $b\in[t_l,t_h]$ and is 0 otherwise. Using (4) we can then calculate that an agent who conforms to the norm receives esteem $H^{\phi}(1)=-0.2845$. An agent with type $t\in[0,0.076)$ does not conform to the norm but, as we can see in Figure 1 below, does choose an action closer to the norm than his type. An agent of type

[4] Note that the action function is symmetric in the sense that $\mu(2-t)=\mu(t)$ if $t<t_l$ and $2-t_l>t_h$. This is the case even if the distribution over types f is asymmetric (see p. 852 of Bernheim).

[5] If (μ,ϕ) is a signalling equilibrium, then μ can be recovered from ϕ, and thus, where it will cause no confusion, we shall use ϕ to characterize the equilibrium, as in the notation $H^{\phi}(t)$.

$t = 0.05$, for example, chooses action $\mu(0.05) \approx 0.4$. The type 0.05 agent will, however, be correctly inferred as a type 0.05 agent and thus receive esteem $-(1-0.05)^2$.

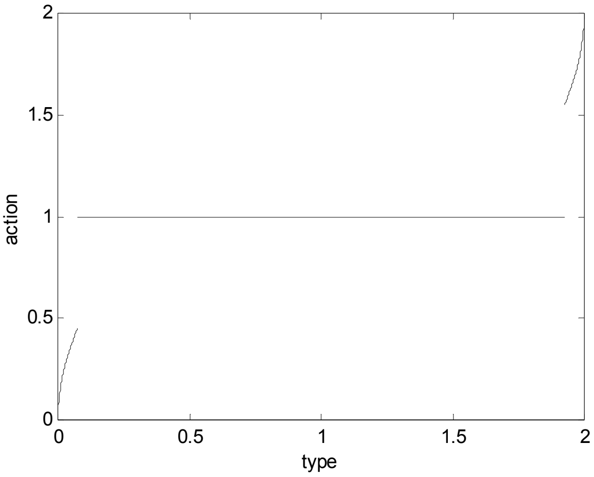

Figure 1: The action function in Example 1 when $x_p = 1$.

Three special cases of a conformist equilibrium are (1) a *fully separating equilibrium* where $t_l = t_h = 1$ implying that there is no central pool, (2) a *fully conformist equilibrium* where $t_l = 0$ and $t_h = 2$, implying that all agents conform to the norm and conformity is no signal of type, (3) a *partially conformist equilibrium* where either $t_l \in (0,1)$ and/or $t_h \in (1,2)$. Example 1 is an example of a partially conformist equilibrium. This lies somewhere between the two extremes of full separation and full conformity where there is a central pool but not all types of agent conform to the norm.

3 Out of equilibrium beliefs and the IWD1

Given a signalling equilibrium (μ, ϕ) characterized by tuple (x_p, t_l, t_h, μ_s), let

$$x_l := \lim_{t \uparrow t_l} \mu(t) \text{ and } x_h := \lim_{t \downarrow t_h} \mu(t)$$

be the actions chosen by those 'at the edge of the central pool' (where we set $x_l = 0$ or $x_h = 2$ if appropriate). Let

$$XE^{\mu} = \left\{ x \in X : \mu^{-1}(x) = \varnothing \right\}$$

be the set of actions that should not be chosen if agents behave according to action function μ. From Bernheim we know that $XE^{\mu} = [x_l, x_h] - \{x_p\}$.[6] If (μ, ϕ) is a fully separating equilibrium, then $XE^{\mu} = \varnothing$. Otherwise, $x_l < x_h$ and the set XE^{μ} is non-empty (Theorem 6 of Bernheim). There are therefore actions that should not be played in equilibrium. In Example 1, for instance, we find that

[6] That actions $x \in [0, x_l)$ and $x \in (x_h, 2]$ are chosen follows because $\mu(0) = 0$ (if $t_l > 0$), $\mu(2) = 2$ (if $t_h < 2$), and the action function is a continuous function of type for $t \in [0, t_l)$ and $t \in (t_h, 2]$.

$x_l \approx 0.4512$ and $x_h \approx 1.5488$, so $XE^{\mu} = [0.4512, 1) \cup (1, 1.5488]$. No agent, for example, should choose $x = 0.9$.

As is standard (see Fudenberg and Tirole 1991), the definition of a signalling equilibrium ties down (see in particular (3) and (4)), inferences about actions that are chosen with positive probability. It does not, however, impose restrictions on inferences about actions 'off the equilibrium path', namely $x \in XE^{\mu}$. Basically, if an agent 'deviates' and chooses an action $x \in XE^{\mu}$, then Bayes Rule is no guide as to what type of agent it should be inferred has deviated.

Clearly inferences about actions off the equilibrium path are a crucial aspect of equilibrium as they determine the incentives or lack of incentive for an agent not to conform to the norm. It determines, for instance, in Example 1 whether a type 0.9 agent would want to conform to the norm and choose $x_p = 1$ or 'deviate' and choose his internal bliss point of $x = 0.9$. Thus, one looks to impose criterion on the inference function ϕ to obtain 'reasonable' beliefs (Banks and Sobel 1997, Cho and Kreps 1997). Bernheim uses the *D1 criterion.* We shall formally define the Criterion below but can state here

Fact 1: If (μ^*, ϕ^*) is a signalling equilibrium characterized by tuple (x_p, t_l, t_h, μ_s) and inferences satisfy the D1 Criterion, then $T^{\phi^*}(x) = \{t_l\}$ for all $x_l \leq x < x_p$ and $T^{\phi^*}(x) = \{t_h\}$ for all $x_h \geq x > x_p$.

Thus, any agent who deviates from the central pool is assumed to have either of the types 'at the edge of the central pool'. This provides a strong incentive for an agent to conform. In Example 1, for instance any agent who chooses $x = 0.9$ is inferred to have type $t_l = 0.0761$, meaning that they receive esteem $H^{\phi^*}(0.9) = -(1 - 0.9)^2 \approx -0.854$. Given that the esteem from conforming is $H^{\phi^*}(1) = -0.2845$, there is no incentive to deviate.

The D1 Criterion imposes a 'strong' criterion on out of equilibrium beliefs. Intuitively, for instance, it seems reasonable that a type $t = 0.9$ agent would potentially deviate to $x = 0.9$ as this is his intrinsic bliss point. This motivates asking *whether a signalling equilibrium (μ^*, ϕ^*) supported by inferences satisfying the D1 Criterion remains an equilibrium if 'weaker' conditions are assumed of the inference function.* To answer this question we clearly must define 'weaker' conditions. We do this below to give inference function ϕ'. Given this we then fix μ^* and ask whether the pair (μ^*, ϕ') is also a conformist equilibrium.

Before proceeding one point is worth noting. For each action $x \in X$, there is at most one signalling equilibrium with inferences satisfying the D1 Criterion where the norm is $x_p = x$ (Theorem 4 of Bernheim). Thus, fix a norm x_p and let (μ^*, ϕ^*) denote a signalling equilibrium with inferences satisfying the D1 Criterion. It is a simple matter to show that any signalling equilibrium with inferences satisfying (3) and where the norm is x_p must also have action function μ^*, whether inferences satisfy the D1 Criterion or not. Fact 2 (below) will formalize this for the inference functions that we shall use, so we do not spend time formally proving this more general claim here.[7] What this does mean, however, is that we

[7] However here is the intuition: In constructing the unique equilibrium action function μ^* when inferences satisfy the D1 Criterion, the out of equilibrium beliefs are irrelevant (see pages 849-857 of Bernheim or the derivation of μ^* for Example 1 in the current paper). The only point at which out or equilibrium beliefs prove important is in checking whether μ^* actually is consistent with equilibrium or not (see page 857 of Bernheim). Different out of equilibrium beliefs cannot therefore lead to different equilibrium actions, but they can change whether μ^* is consistent with an equilibrium.

can fix the action function μ^* without any need to worry whether a weakening of inferences should lead to a change in equilibrium actions.

3.1 Incentive to deviate

Take as given a signalling equilibrium (μ^*, ϕ^*) where inferences ϕ^* satisfy the D1 Criterion. If we consider some 'weaker' inferences ϕ' then we know that pair (μ^*, ϕ') can only be a signalling equilibrium if inference functions ϕ^* and ϕ' 'agree' for actions $x \notin XE^{\mu^*}$. Thus, we shall impose that $\phi^*(b,x) = \phi'(b,x)$ for all b and $x \notin XE^{\mu^*}$. This implies that $H^{\phi^*}(x) = H^{\phi'}(x)$ for all $x \notin XE^{\mu^*}$.[8] It also implies that equilibrium payoffs $U(\mu^*(t), t, \phi^*)$ can be fixed independently of ϕ'. That is $U(\mu^*(t), t, \phi^*) = U(\mu^*(t), t, \phi')$. What may change when we consider inferences ϕ' are inferences about actions $x \in XE^{\mu^*}$ that are off the equilibrium path. Thus, $H^{\phi^*}(x)$ and $H^{\phi'}(x)$ may differ as may $U(x,t,\phi^*)$ and $U(x,t,\phi')$ for $x \in XE^{\mu^*}$. This can change the 'incentives' to conform.

Given pair (μ^*, ϕ') if an agent of type t were to choose some action $x \in XE^{\mu^*}$, then he would receive payoff $g(x-t) + \lambda H^{\phi'}(x)$. We know that $g(x-t) + \lambda H^{\phi^*}(x)$ is less than $U(\mu^*(t), t, \phi^*)$, so he would not want to deviate given inference function ϕ^*. But what about $g(x-t) + \lambda H^{\phi'}(x)$? If $g(x-t) + \lambda H^{\phi'}(x) \geq U(\mu^*(t), t, \phi^*)$ then he would want to deviate to x and not conform. If $g(x-t) + \lambda H^{\phi'}(x) \leq U(\mu^*(t), t, \phi^*)$ then he would not want to deviate. The first possibility is not consistent with equilibrium while the second is. Crucial, therefore, is the value of[9]

$$\varepsilon^{\phi^*}(t,x) := \frac{U(\mu^*(t), t, \phi^*) - g(x-t)}{\lambda}. \tag{6}$$

If $H^{\phi'}(x) < \varepsilon^{\phi^*}(t,x)$, then an agent of type t would not gain from choosing x rather than $\mu^*(t)$. If $H^{\phi'}(x) > \varepsilon^{\phi^*}(t,x)$, then an agent of type t would do better choosing x than $\mu^*(t)$. Given this we shall interpret $\varepsilon^{\phi^*}(t,x)$ as the *incentive to conform* and say that a type t agent has *more incentive to deviate to* x than a type t' agent if $\varepsilon^{\phi^*}(t,x) < \varepsilon^{\phi^*}(t',x)$.[10] Clearly, if (μ^*, ϕ') is a signalling equilibrium, then $H^{\phi'}(x) \leq \varepsilon^{\phi^*}(t,x)$ for all t and $x \in XE^{\mu^*}$. This is equivalent to condition (5) and will provide a simple check of whether (μ^*, ϕ') is a signalling equilibrium or not. Note that $\varepsilon^{\phi^*}(t,x)$ is independent of ϕ' and can thus be fixed given ϕ^*.

Figures 2a and b plot $\varepsilon^{\phi^*}(t,x)$ for values of $x = 0.9$ and 1.4 respectively. In interpretation we can see from Figure 2a that $\varepsilon^{\phi^*}(0.5, 0.9) \approx -0.35$, implying that a type $t = 0.5$ agent would prefer $x = 0.9$ to $\mu^*(0.5) = 1$ if the esteem from playing $x = 0.9$ exceeded -0.35. As we have already seen, if inferences satisfy the D1 Criterion, then $H^{\phi^*}(0.9) = -0.854$, so a type 0.5 agent does not want to deviate. We need to question whether 'weaker' conditions on inferences still guarantee that $H^{\phi'}(0.9) < -0.35$.

[8] This explains why we shall not be interested in fully separating equilibria where $XE^{\mu^*} = \varnothing$ and, thus, the D1 Criterion is never used.

[9] Bernheim uses the notation $I(x,t)$ where $I(x,t) = \lambda\varepsilon^{\phi}(t,x)$.

[10] Note, however, that, fixing a value of $H^{\phi}(x)$ an agent will typically either do better to choose x or do better not to choose x, so it is not immediately clear that a higher $\varepsilon^{\phi}(t,x)$ would equate with less likelihood of actually choosing x.

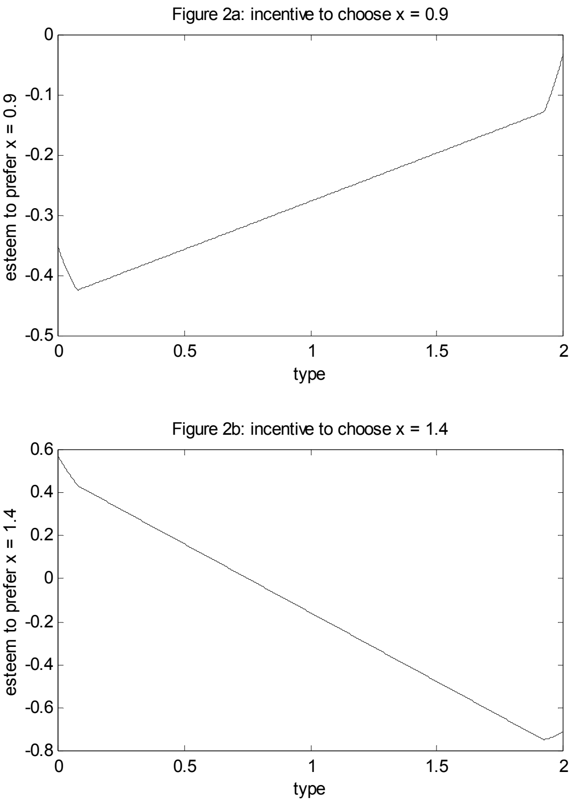

Figure 2: The value of $\varepsilon^{\phi^*}(t, 0.9)$ and $\varepsilon^{\phi^*}(t, 1.4)$ in Example 1 when $x_p = 1$.

We can see from Figure 2 that when $x = 0.9$ agents with types near t_l have the most incentive to deviate while the converse holds when $x = 1.4$. [Also note the change in the scale of the y-axis.] This is a general property as shown by Bernheim (see the Proof of Theorem 3):

Lemma 1 (Bernheim): For any $x \in (x_l, x_p)$ the value $\varepsilon^{\phi^*}(t, x)$ is strictly decreasing in t for $t < t_l$ and strictly increasing in t for $t > t_l$.[11] For any $x \in (x_p, x_h)$ the value $\varepsilon^{\phi^*}(t, x)$ is strictly decreasing in t for $t < t_h$ and strictly increasing in t for $t > t_h$.

The type of agent who has most incentive to deviate proves important, so let

$$\underline{\varepsilon}^{\phi^*}(x) = \left\{ t \in T : \varepsilon^{\phi^*}(t, x) = \min_{b \in T} \varepsilon^{\phi^*}(b, x) \right\}.$$

3.2 Weak D1 Criterion

Before defining weaker conditions on the inference function we can now formally define the D1 Criterion (Cho and Kreps).

[11] More formally, if $t' < t'' \leq t_l$ then $\varepsilon(t', x) > \varepsilon(t'', x)$, and if $t_l \leq t' < t''$, then $\varepsilon(t'', x) > \varepsilon(t', x)$.

D1 Criterion: Inference function ϕ^* satisfies the D1 Criterion if $\phi^*(b,x) = 0$ for each $x \in XE^{\mu^*}$ and every $b \notin \underline{\varepsilon}^{\phi^*}(x)$.

Fact 1 is immediate from Lemma 1. To contrast the D1 Criterion we shall follow the approach of Cho and Kreps (1987) and Banks and Sobel (1987) by modelling inferences ϕ' as resulting from an iterative process of reasoning. More specifically, we shall consider a sequence of inference functions $\phi_0, \phi_1, ..., \phi_y, ...$ where informally $\phi' = \lim_{y\to\infty} \phi_y$.

One condition we shall impose in constructing the inference functions ϕ_y is that those agents with relatively more incentive to deviate to action x are, at least, equally likely to be inferred as having chosen x.

Incentive Condition: Inference function ϕ satisfies the incentive condition if for each $x \in XE^{\mu^*}$ and any $t, t' \in [t_l, t_h]$ or $t, t' \in [0, t_l]$ or $t, t' \in [t_h, 2]$, if $\varepsilon^{\phi^*}(t,x) \leq \varepsilon^{\phi^*}(t',x)$, then

$$\frac{\phi_y(t,x)}{\phi_y(t',x)} \geq \frac{f(t)}{f(t')}. \tag{7}$$

Clearly, an inference function satisfying the D1 Criterion satisfies the incentive condition. We shall discuss other possibilities as we proceed.

To construct the inference functions ϕ_y we need one definition. Given some real number A let

$$T(x,A) := \{t \in T : \varepsilon^{\phi^*}(t,x) \leq A\} \cup \underline{\varepsilon}^{\phi^*}(x). \tag{8}$$

If $t \in T(x,A)$, then an agent of type t would want to deviate to action x if the esteem $H^{\phi}(x)$ from choosing x is greater than or equal to A. If no agent would wish to deviate, then we have $T(x,A) = \underline{\varepsilon}^{\phi^*}(x)$. The necessity for this condition will soon become clear.

The D1 Criterion essentially fixes A at $h(t_l)$ or $h(t_h)$. Given that (μ^*, ϕ^*) is a signalling equilibrium we know, for instance, that $T(x, h(t_l)) = \underline{\varepsilon}^{\phi^*}(x)$ for all $x \in (x_l, x_p)$. To derive weaker inferences we need to set A at some level greater than $H^{\phi^*}(x)$. In equilibrium, those who conform and receive esteem $H^{\phi^*}(x_p)$ have the highest esteem.[12] An intuitive starting point, therefore, in constructing weaker inferences is to set $H^{\phi^*}(x_p)$ as an upper bound on the esteem an agent could expect if he were to choose action $x \in XE^{\mu^*}$.[13] If $t \notin T(x, H^{\phi^*}(x_p))$, then a type t agent would not want to deviate to action x if he expected to receive esteem $H^{\phi^*}(x_p)$. This suggests a weaker condition on inferences.

[12] To demonstrate this we need to show that $H^{\phi^*}(x_p) > h(t_l), h(t_h)$. Given that $t_l < x_l < x_p$ a type t_l agent would do strictly better to choose action x_l than x_p if $h(t_l) \geq H^{\phi^*}(x_p)$. This contradicts that (μ^*, ϕ^*) is a signalling equilibrium. A similar argument shows that $h(t_h) < H^{\phi^*}(x_p)$.

[13] A weaker starting point would be to set A equal to the maximum esteem of $h(1)$. It turns out that doing so would not significantly alter our results. In particular, as can be seen from reading the proofs, the results of Section 5 (equilibrium selection) and the results of Section 4 (equilibrium existence) concerning the necessary conditions for equilibrium would be unaffected. The sufficient conditions for equilibrium existence would change because any type of agent would likely want to deviate to any x for an esteem of $h(1)$. It, therefore, becomes harder to tie down inferences. This, however, only seems to motivate why $h(1)$ is unreasonably high as an intuitive starting point in constructing inferences.

Weak D1 Criterion:[14] Inference function ϕ_0 satisfies the weak D1 criterion if it satisfies the incentive condition and $\phi_0(b,x)=0$ for every $x \in XE^{\mu^*}$ and every $b \notin T(x, H^{\phi^*}(x_p))$.

Inference function ϕ^*, satisfying the D1 Criterion, also satisfies the weak D1 Criterion. There are, however, many other inference functions that do not satisfy the D1 Criterion, but do satisfy the weak D1 Criterion. In Example 1 for instance we have that $H^{\phi^*}(1) = -0.2845$, so, as can be seen from Figure 2a, $T(0.9, H^{\phi^*}(1)) = [0, 0.95]$. This implies that inference function ϕ_0 in which all agents with types $t \in [0, 0.95]$ are inferred as equally likely to have chosen $x = 0.9$ satisfies the weak D1 Criterion. This is what we shall define below as uniform inferences and implies that $\phi_0(b, 0.9) = 1.053$ for all $b \in [0, 0.95]$ and is 0 otherwise. [See Table 1 below.] With this inference function the esteem from choosing $x = 0.9$ can be calculated, using (4), as $H^{\phi_0}(0.9) = -0.3507$. This can be contrasted with the esteem of $H^{\phi^*}(0.9) = -0.854$ that would result if inferences satisfy the D1 Criterion. The 'weaker inferences' of the weak D1 Criterion result in a higher esteem from deviating and, in this case, are not consistent with equilibrium. This can be seen from Figure 2a where $-0.3507 > \varepsilon^{\phi^*}(t, 0.9)$ for many types, implying that there are agents with an incentive to deviate to 0.9. Consequently there are inferences that satisfy the weak D1 Criterion but are not consistent with a signalling equilibrium (in Example 1 with norm $x_p = 1$).

The Weak D1 Criterion can, however, be thought of as the starting point (rather than the end point) in forming inferences. In particular, there is an inconsistency between the esteem $H^{\phi^*}(x_p)$ that agents 'expect' to receive and actual esteem $H^{\phi_0}(x)$. If inferences are given by ϕ_0, then one could reason that the most esteem an agent can expect from choosing action x is $H^{\phi_0}(x)$. For instance in Example 1 the most esteem an agent could expect from choosing $x = 0.9$ is $H^{\phi_0}(0.9) = -0.3507$. This suggests that it may be more appropriate to say that a type t agent is only deemed likely to choose x if he had an incentive to do so given esteem of $H^{\phi_0}(x)$ (rather than $H^{\phi^*}(x_p)$). This leads to

1-*step weak D1 Criterion (1-WD1):* Inference function ϕ_1 satisfies the 1-step weak D1 criterion if it satisfies the incentive condition and $\phi_1(b,x) = 0$ for every $x \in XE^{\mu^*}$ and $b \notin T(x, H^{\phi_0}(x))$ and ϕ_0 satisfies the weak D1 criterion.

From Figure 2a we can see that when $x = 0.9$ and $H^{\phi_0}(0.9) = -0.3507$, we get that $T(0.9, H^{\phi_0}(0.9)) = [0, 0.5373]$. We can thus revise the set of agent types who could be inferred as having deviated to x. Given the 1-WD1 we can also revise the esteem an agent can expect to receive from choosing x to $H^{\phi_1}(x)$. Iterating this argument produces more refined beliefs where the more iterations are performed, the more introspection is required of agents.

y-step weak D1 Criterion (y-WD1): Inference function ϕ_y satisfies y-WD1 if it satisfies the incentive condition and $\phi_y(b,x) = 0$ for every $x \in XE^{\mu^*}$ and $b \notin T(x, H^{\phi_{y-1}}(x))$ and ϕ_{y-1} satisfies $y-1$-WD1.

[14] This condition can be seen as in the spirit of equilibrium domination (Cho and Kreps). We adopt the term weak D1 criterion reflecting the discussion of Cho and Kreps about the D1 criterion.

Iterated weak D1 Criterion (IWD1): Inference function ϕ' satisfies IWD1 if $\phi'(b,x) = 0$ for every $x \in XE^{\mu^*}$ and $b \notin T(x, H^{\phi_\infty}(x))$ where $H^\infty(x) := \lim_{y\to\infty}\{H^{\phi_y}(x)\}$.[15]

In the y-step IWD1, starting with beliefs satisfying the weak D1 criterion, y iterations are performed sequentially, eliminating types of agents deemed as potentially choosing each action. Taking y to infinity we obtain beliefs satisfying IWD1.

At this stage we have not specified how ϕ_y will be derived from ϕ_{y-1}. If inferences satisfy the D1 Criterion, then we set $\phi_0(b,x) = 0$ for all $b \notin \underline{\varepsilon}^{\phi^*}(x)$ and $x \in XE^{\mu^*}$. Given that (μ^*, ϕ^*) is an equilibrium, it is trivial that inferences ϕ_y satisfying the D1 Criterion also satisfy y-WD1. Consequently the D1 Criterion is a special case of IWD1. A second special case, already discussed informally above, represents the opposite extreme where all agents who had an incentive to deviate are deemed equally likely to have chosen x.

Uniform y-step weak D1 Criterion (uniform y-WD1): Inference function ϕ_y satisfies uniform y-WD1 if

$$\phi_y(b,x) = \begin{cases} 0 \text{ for all } b \notin T(x, H^{\phi_{y-1}}(x)) \\ f(b)\left[\int_{T(x,H^{\phi_{y-1}}(x))} f(b)db\right]^{-1} \text{ otherwise} \end{cases} \tag{9}$$

and ϕ_{y-1} satisfies uniform $y-1$-WD1.[16]

Thus, all agents with a positive incentive to deviate are seen as equally likely to deviate. Such inferences are essentially equivalent to divine beliefs as introduced by Banks and Sobel. The D1 Criterion and uniform IWD1 can be seen as opposite ends of the spectrum of inference functions consistent with IWD1 (and the incentive condition). In Section 6 we discuss a more general inference function that covers the entire spectrum. In Example 1 with uniform inferences we get

Table 1: Deriving uniform inferences for Example 1.

	$x = 0.9$		$x = 1.4$	
y	$T(x, H^{\phi_{y-1}}x))$	$H^{\phi_y}(x)$	$T(x, H^{\phi_{y-1}}(t^*))$	$H^{\phi_y}(x)$
$-$		-0.2845		-0.2845
0	$[0, 0.95]$	-0.3507	$[1.2, 2]$	-0.4136
1	$[0, 0.54]$	-0.5590	$[1.4017, 2]$	-0.5210
2	$\{t_l\}$	-0.854	$[1.5698, 2]$	-0.6315
3			$[1.7429, 2]$	-0.7649
4			$\{t_h\}$	-0.854

Thus, pair (μ^*, ϕ_0) is not a signalling equilibrium because agents of type $[0, 0.95]$ would want to deviate to action $x = 0.9$ and agents of type $[1.2, 2]$ would want to deviate to action $x = 1.4$. The more iterations we perform, however, the more types are eliminated as potentially having chosen both $x = 0.9$ and 1.4.

[15] If the $\lim_{y\to\infty}\{H^{\phi_y}(x)\}$ does not exist (see Example 3 in Section 5.1 below), set $H^\infty(x) = \min_y\{H^{\phi_y}(x)\}$.

[16] We use the natural extension of a uniform weak D1 Criterion to derive ϕ_0.

Inference function ϕ_2 is such that no agent would want to deviate to $x = 0.9$ because the esteem from doing so is only $H^{\phi_1}(0.9) = -0.5590$. In this case we set esteem at $h(\underline{\varepsilon}^{\phi^*}(x))$. The inference function must satisfy $\phi_2(b, 0.9) > 0$ for some b, so once all types are eliminated (as was the case above), we have to make an assumption. Setting $H^{\phi'}(x) = h(\underline{\varepsilon}^{\phi^*}(x))$ seems natural but means that inferences in this case are equivalent to those of the D1 Criterion (for this particular x).

More generally if (μ, ϕ') is a signalling equilibrium and inferences satisfy y-WD1 then $T^{\phi'}_{y+1}(x)$ is $\{t_l\}$ or $\{t_h\}$. Otherwise, by construction, an agent of type $t \in T_{y+1}(x)$ would do strictly better to choose x than $\mu(t)$. Thus, a signalling equilibrium can be supported by inferences satisfying IWD1 only if inferences 'collapse' to those obtained with the D1 criterion. This need not be the case as we shall show below. It does, however, motivate an important observation:

Fact 2: If (μ, ϕ') is a signalling equilibrium and inference function ϕ' satisfies y-WD1 then there exists a signalling equilibrium (μ, ϕ^*) where inference function ϕ^* satisfies the D1 Criterion.

Informally, the set of signalling equilibria with inferences satisfying the D1 Criterion nests the set of signalling equilibria with inferences satisfying the y-WD1. In looking, therefore, for signalling equilibria with an inference function that satisfies y-WD1 we can restrict our attention to action functions μ^* that result from signalling equilibria with an inference function that satisfies the D1 Criterion. We can conclude, for example, from Table 1 that there does not exist a signalling equilibrium in Example 1 when $x_p = 1$ if inferences satisfy uniform 2-WD1.

4 Equilibrium Existence

Our first result provides necessary and sufficient conditions for the existence of a signalling equilibrium (μ^*, ϕ') where inferences ϕ' satisfy IWD1. The following result is due to Bernheim:

Fact 3 (Bernheim Theorems 2 and 5)**:** Given Assumptions 1-3 there does exist a signalling equilibrium (μ^*, ϕ^*) where inference function ϕ^* satisfies the D1 Criterion.

Given Facts 2 and 3 our task is to look at each signalling equilibria (μ^*, ϕ^*) with an inference function satisfying the D1 Criterion and question whether the existence of equilibrium (μ^*, ϕ^*) implies the existence of equilibrium (μ^*, ϕ') where inference function ϕ' satisfies IWD1. We begin with an example demonstrating that a signalling equilibrium need not exist if inferences satisfy uniform IWD1.

In **Example 2** we set $\lambda = 2.5$ and $f(t) = 0.9$ for $t \in [0, 0.5]$ and $t \in [1.5, 2]$ and $f(t) = 0.1$ for $t \in [0.5, 1.5]$. The distribution of agent types is therefore skewed towards agents with 'more extreme types'. To illustrate that there does not exist a signalling equilibrium with inferences satisfying the uniform IWD1 we consider norm $x_p = 1$. When $x_p = 1$ there is a unique signalling equilibrium (μ^*, ϕ^*) supported by inferences satisfying the D1 Criterion. The pair (μ^*, ϕ') is not, however, a signalling equilibrium if inferences satisfy uniform IWD1.

To see why, we first note that (μ^*, ϕ^*) is a fully conformist equilibrium, so $\mu(t) = 1$ for all t. Next, consider the motivation for a type $t = 0.5$ agent to choose $x = 0.5$. One can easily calculate, as we do in the Appendix, that $\varepsilon^{\phi^*}(0.5, 0.5) = -0.633 < -0.6$. In other words, a type $t = 0.5$ agent would prefer $x = 0.5$ to $x = 1$ if the esteem from choosing 0.5 exceeded -0.6. We know, however, (by Lemma 1 and the fact that $t_l = 0$) that if inference function ϕ_y satisfies the uniform y-WD1, then $T^{\phi_y} = [0, t_y]$ for some t_y. If $t_y \geq 0.5$, then we can also show that $H^{\phi_y}(x) \geq -0.6$. This gives the circularity that we require: a type 0.5 agent would want to deviate to 0.5 if he gets esteem of more than -0.6 from doing so, but if a type 0.5 agent is inferred as potentially deviating to 0.5, then the esteem from deviating will be more than -0.6. Consequently, if inferences satisfy uniform IWD1, a type $t = 0.5$ agent would prefer $x = 0.5$ to conformity and the pair (μ^*, ϕ') cannot be a signalling equilibrium. The table below illustrates this.

Table 2: Deriving inferences for Example 2.

y	$T^{\phi_y}(0.5)$	$H^{\phi_y}(0.5)$
		-0.5334
0	$[0, 0.75]$	-0.560
1	$[0, 0.683]$	-0.567
$\vdots$	$\vdots$	$\vdots$
∞	$[0, 0.66]$	-0.569

The above treats the case of a norm $x_p = 1$. To demonstrate formally that there does not exist a signalling equilibrium where inferences satisfy the D1 Criterion for Example 2 we also need to consider possible signalling equilibria with norms $x_p \neq 1$. The conclusion, however, is the same, and because the exercise is somewhat tedious we relegate the details to the Appendix.

The reason that there does not exist an equilibrium (μ^*, ϕ') with inferences satisfying uniform IWD1 in Example 2 is that a type $t = 0.5$ agent would want to deviate to $x = 0.5$ even if inferences were such that any agent who deviates to $x = 0.5$ is inferred to be of type 0.5 or less. This motivates our first result and the following notation. Let

$$H(t_1, t_2) := \frac{1}{F(t_2) - F(t_1)} \int_{t_1}^{t_2} f(b)h(b)db.$$

Note that if inferences ϕ' are uniform and $T^{\phi'}(x) = [t_1, t_2]$, then $H^{\phi'}(x) = H(t_1, t_2)$. Define

$$E_-(x) := H(t_l, x) \text{ and } E_+(x) := H(x, t_h)$$

for all $x \in (x_l, x_h)$ as the esteem from playing action x if an agent who chooses x is inferred to be equally likely to have a type $t \in [t_l, x]$ or $t \in [x, t_h]$ respectively. Figures 3a and 3b plot $E_-(x)$ and $E_+(x)$ for Examples 1 and 2.

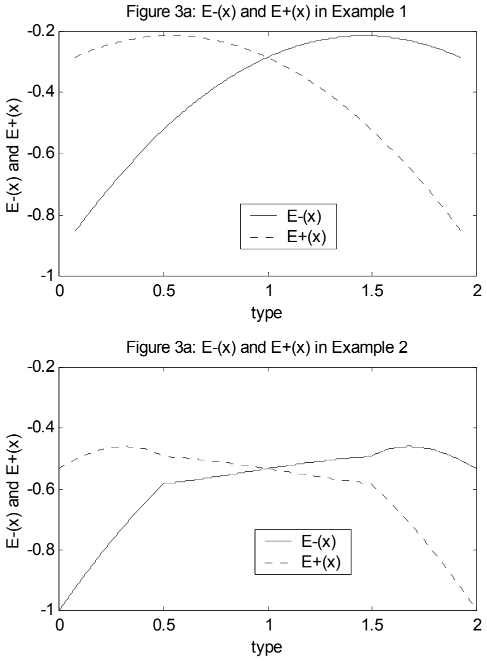

Figure 3: Values of $E_{-}(x)$ and $E_{+}(x)$ for example 1 and 2.

We now define an important condition:

Worse-than Condition: We say that the *worse than condition* is satisfied if $E_{-}(x) < \varepsilon^{\phi^*}(x, x)$ for all $x \in (x_l, x_p)$ and $E_{+}(x) < \varepsilon^{\phi^*}(x, x)$ for all $x \in (x_p, x_h)$.[17]

The worse than condition requires that an agent of type $t < x_p$ would not want to play action $x = t$ rather than norm x_p if any agent who plays x is inferred to have a type between t_l and x. Thus, a type t agent would only want to play action $x = t$ if in doing so he is inferred as potentially having a type $b > t$. In Example 2 we saw that the worse than condition was not satisfied for $x = 0.5$. This is confirmed in Figure 4b and was given as the reason an equilibrium did not exist. Figure 4a shows that the worse than condition is satisfied in Example 1 where we have evidence to suggest that an equilibrium does exist.

[17] In the case of a fully separating equilibrium we think of the worse than condition as being satisfied 'by default' as there exists no such x.

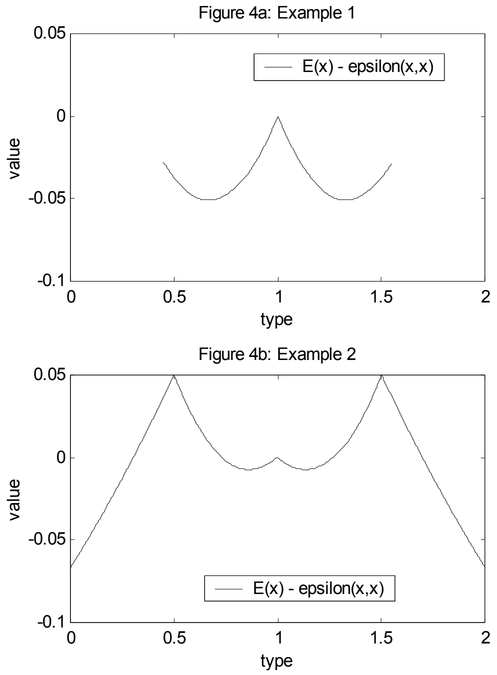

Figure 4: The value of $E_-(x) - \varepsilon^{\phi^*}(x,x)$ for $x \in (x_l, x_p)$ and of $E_+(x) - \varepsilon^{\phi^*}(x,x)$ for all $x \in (x_p, x_h)$ in Examples 1 and 2.

The following result demonstrates that the worse than condition is a necessary and sufficient condition for equilibrium existence.

Theorem 1: Let pair (μ^*, ϕ^*) be a signalling equilibrium where inference function ϕ^* satisfies the D1 Criterion. If the worse than condition is satisfied and ϕ' satisfies IWD1, then (μ^*, ϕ') is a signalling equilibrium. If the worse than condition is not satisfied, then there exists inference function ϕ' that satisfies IWD1 such that (μ^*, ϕ') is not a signalling equilibrium.

Proof of Theorem 1: Pick a $x \in (x_l, x_p)$ with a symmetric argument treating $x \in (x_p, x_h)$. We begin by showing that if the worse than condition is satisfied, no agent would wish to play x rather than x_p. By Lemma 1, $T^{\phi_y}(x) = [t_y^-, t_y^+]$ for some $t_y^- \leq t_l \leq t_y^+$. We conjecture (*) that $\lim_{y \to \infty} t_y^- = \lim_{y \to \infty} t_y^+ = t_l$. If so, $T^{\phi'}(x) = \{t_l\} = T^{\phi^*}(x)$, and if playing x is not individually rational for an agent given inferences ϕ^*, then it cannot be given inferences ϕ'.

To prove the conjecture (*) we begin by noting that, given Assumption 4 and Lemma 1, the maximal esteem for choosing x that is consistent with inferences satisfying the Weak D1 Criterion would put $\phi_0(b,x) = kf(x)$ for some constant k and all $b \in T^{\phi_0} = [t_l, t_0^*]$ for some $t_0^* \leq t_0^+$. Thus, if inferences ϕ_0 satisfy the weak D1 Criterion we have $H^{\phi_0}(x) \leq E_-(t_0^*)$. The worse than condition implies that $E_-(t_0^*) < \varepsilon^{\phi^*}(t_0^*, t_0^*)$. Given that g achieves a maximum at 0, we also know that $\varepsilon^{\phi^*}(t_0^*, t_0^*) \leq \varepsilon^{\phi^*}(x, t_0^*)$. Thus, $H^{\phi_0}(x) < \varepsilon^{\phi^*}(x, t_0^*)$, implying that $t_0^* \notin T(x, H^{\phi_0}(x))$ and $t_0^* \notin T^{\phi_1}(x)$ if inferences satisfy the 1-WD1. Thus, we must have $t_1^+ < t_0^*$. By the choice of t_0^* we also know that $H^{\phi_1}(x) < E_-(t_0^*)$. Repeating the above argument we see that t_y^+ strictly decreases to limit t_l. That is $\lim_{y\to\infty} t_y^+ = t_l$. For t_y^+ sufficiently close to t_l we know by the worse than condition that $t_y^+ \notin T(x, H^{\phi_y}(x))$, so $T^{\phi_{y+1}}(x) = \{t_l\}$ by Lemma 1. This implies that $\lim_{y\to\infty} t_y^- = t_l$ as desired.

To demonstrate the 'only if' element of the result, suppose that there exists some $x < x_p$ where $E_-(x) \geq \varepsilon^{\phi^*}(x, x)$. We know that $x \in T(x, H^{\phi^*}(x))$, so, we can set $T^{\phi_0} = [t_l, x]$, implying that $H^{\phi_0}(x) = E_-(x)$. Inferences ϕ_0 satisfy the weak D1 Criterion. Given that $E_-(x) \geq \varepsilon^{\phi^*}(x, x)$, we know that $x \in T(x, H^{\phi_0}(x))$, so we can set $T^{\phi_1} = [t_l, x]$ and $H^{\phi_1}(x) = E_-(x)$. Iterating this argument implies that there exists an inference function ϕ_y that satisfies y-WD1 and where $x \in T(x, H^{\phi_y}(x))$ for all y. There exist therefore inferences ϕ' that satisfy IWD1 such that (μ^*, ϕ') is not a signalling equilibrium.■

Example 2 demonstrated that the worse than condition need not be satisfied and thus there need not exist a signalling equilibrium with inferences satisfying IWD1. From Figure 4a we see that for Example 1 the worse than condition is satisfied, so there does exist a signalling equilibrium if inferences satisfy IWD1. In looking for when the worse than condition will hold more generally, we make two observations.

First, the worse than condition will be satisfied if λ is sufficiently large. To see why note that for λ sufficiently large there will only exist fully conformist equilibria.[18] Using $t_l = 0$ and $t_h = 2$ the worse than condition can be rewritten

$$g(0) - g(x_p - x) < \lambda \left[H^{\phi^*}(x_p) - H(0, x) \right] \tag{10}$$

for all $x \in (0, x_p)$ and

$$g(0) - g(x_p - x) < \lambda \left[H^{\phi^*}(x_p) - H(x, 2) \right] \tag{11}$$

for all $x \in (x_p, 2)$. It can be shown (see Lemma 2 below) that there always exist some $x_p \in X$ such that $H^{\phi^*}(x_p) > H(0, x)$ for all $x \in (0, x_p)$ and $H^{\phi^*}(x_p) > H(x, 2)$ for all $x \in (x_p, 2)$. Given that, in this case, the right hand side of both (10) and (11) will be positive, the condition will hold for sufficiently large λ. In Example 2, for instance, the worse than condition is satisfied if $\lambda > 5$ (see the Appendix).

Second, we note that whether or not the worse than condition holds will depend on the distribution of agent types. Recall that the worse than condition requires an agent of type $t < x_p$ not to want to choose action $x = t$ rather than norm x_p if any agent who chooses x is inferred to have a type between t_l and x. This is more likely to hold if the esteem from conforming to the norm is relatively high while the esteem of being inferred to have a type between t_l and x is relatively low. The esteem from conforming will be relatively high if proportionally many agents who conform have types near to the ideal type. This suggests that the worse than condition is likely to hold when proportionally many agents have types near to the ideal type. The following result supports this by demonstrating that the worse than condition is satisfied in the spherical model if $\lambda \geq 1.5$ and f is unimodal around 1 in the sense that f is symmetric ($f(t) = f(2 - t)$) and $f(t) \leq f(t')$ for all $t < t' \leq 1$. Note that f was not unimodal in Example 2 where the 'mass of agents' with types at the two extremes lowered the esteem from conforming and meant the worse than condition was not satisfied.

[18] A type t agent would conform to norm x_p if $g(0) + \lambda \max\{h(t_l), h(t_h)\} < g(t - x_p) + \lambda H(t_l, t_h)$. Given that $h(t_l), h(t_h) < H(t_l, t_h)$, this is satisfied for sufficiently high λ.

Corollary 1: Suppose that $g(z) = -z^2, h(t) = -(1-t)^2, \lambda \geq 1.5$ and f is unimodal. Then, there exists action function μ^* such that (μ^*, ϕ') is a signalling equilibrium for any inference function ϕ' that satisfies IWD1.

The proof is in the Appendix. One important point to note about Theorem 1 and Corollary 1 is that they only guarantee existence of a signalling equilibrium with inferences satisfying IWD1. There may not exist a signalling equilibrium with inferences satisfying the y-step IWD1 for some finite y, so, we do require 'introspection' on the part of agents to be able to 'sequentially eliminate deviating types'. In reality, however, it may require only few iterations to eliminate types as we see in Table 1.

5 Equilibrium Selection

If inferences satisfy the D1 Criterion, then for sufficiently high λ, there can exist a signalling equilibrium (μ^*, ϕ^*) with any norm.[19]

Fact 4: For any $x \in X$ there exists λ_x such that if $\lambda > \lambda_x$ there exists a signalling equilibrium (μ^*, ϕ^*) where inference function ϕ^* satisfies the D1 Criterion and $x_p = x$.

In this Section we shall demonstrate that if inferences satisfy IWD1, then equilibria need only exist for specific norms. The 'weaker inferences' of IWD1 thus act as an equilibrium selection device. To explain why we require a second lemma.

Lemma 2: (i) There exists a unique type $t^m \in [t_l, t_h]$ such that $E_-(t^m) = E_+(t^m)$. (ii) $E_-(x)$ is an increasing function of x for all $x \in (t_l, t^m]$ and $E_+(x)$ is a decreasing function of x for all $x \in [t^m, t_h)$. (iii) $E_-(x) \geq H(t_l, t_h)$ for all $x \in (t^m, t_h)$ and $E_+(x) \geq H(t_l, t_h)$ for all $x \in (t_l, t^m)$.

Proof: Given Assumptions 2 and 3 (in particular that f is continuous) it is immediate that $E_-(x)$ is a continuous function of x and there exists real number x_- such that $E_-(x)$ is increasing in x for $x < x_-$ and decreasing in x for $x > x_-$. Clearly $E_-(t_l) = h(t_l)$ and $E_-(t_h) = H(t_l, t_h)$. Similarly $E_+(x)$ is a continuous function of x that is increasing in x for $x < x_+$ and decreasing in x for $x > x_+$ for some real number x_+. Here $E_+(t_l) = H(t_l, t_h)$ and $E_+(t_h) = h(t_h)$. Finally, if $E_-(x) = E_+(x)$, then because

$$H(t_l, t_h) = \frac{E_-(x)\left[F(x) - F(t_l)\right] + E_+(x)\left[F(t_h) - F(x)\right]}{F(t_h) - F(t_l)}$$

we know that $E_-(x) = H(t_l, t_h)$. The three statements of the Lemma now follow.■

Figures 3 (and 5 to follow) illustrate by plotting $E_-(x)$ and $E_+(x)$ for Examples 1, 2 (and 3). The unique type t^m where $E_-(t^m) = E_+(t^m)$ will prove important in the following and be referred to as the *median type*. The median type is characterized by a symmetry in which there is the same esteem to being inferred as of a type 'above t^m' as of a type 'below t^m'. The following result (a corollary of

[19] See Footnote 18 for an explanation why.

Theorem 2 to come later) illustrates the potential for weaker inferences to select equilibria and also illustrates the importance of the median type.

Corollary 2: If (μ^*, ϕ') is a fully conformist equilibrium and inferences satisfy the uniform y-WD1 or uniform IWD1, then $x_p = t^m$.

Below we shall consider the consequences of non-uniform inferences and partially separating equilibrium. Before doing so, however, to illustrate Corollary 2, we look at what happens if f is not symmetric. If f is symmetric and there is a fully conformist equilibrium with inferences satisfying uniform IWD1, then the norm must be the action preferred by the ideal type. This seems intuitive given symmetry and that 1 is the ideal type. If f is not symmetric, then t^m is unlikely to equal 1, and there may exist no signalling equilibrium supported by inferences satisfying IWD1 where the norm is 1. Example 3 will hopefully show why this is also intuitive.

5.1 Selection with an asymmetric distribution

In **Example 3** we set $F(0) = 0.1$ and $F(t) = 0.1 + 0.45t$ for all $t \in (0, 2]$.[20] There is an obvious asymmetry in the distribution where many agents have type $t = 0$. In the Appendix we show that there exists a unique $t^m \simeq 1.17$ where $E_-(t^m) = E_+(t^m)$. Setting $\lambda = 2.5$ there does exist a fully conformist equilibrium (μ^*, ϕ^*) supported by inferences satisfying the D1 Criterion if either $x_p = 1$ or $x_p = t^m$. Applying Corollary 2, however, there exists a fully conformist equilibrium (μ^*, ϕ') supported by inferences satisfying uniform IWD1 if and only if $x_p = t^m$.

We can illustrate the 'if' part of Corollary 2 by setting $x_p = t^m$ and asking whether any agent would want to deviate to $x = 1$. Intuitively a type $t = 1$ agent may want to deviate to signal that he does have the ideal type. The following table shows, however, that if inferences satisfy the uniform 3-WD1, then no agent would want to deviate.

Table 3: Deriving uniform inferences in Example 3 with $x_p = t^m$.

y	$T^{\phi_y}(1)$	$H^{\phi_y}(1)$
		-0.4
0	$[0, 1.085]$	-0.425
1	$[0, 0.9]$	-0.495
2	$[0, 0.309]$	-0.785
2	$\{0\}$	-1

Deviating to $x = 1$ when the norm is $x = t^m$ is a signal that the agent has some type in the interval $[0, t_y^+]$ for some t_y^+. In this example that is an 'undesirable' signal to send as it suggests the agent is of type $b = 0$ with high probability. This is why no agent would wish to deviate to $x < t^m$. Figure 5b confirms that there is indeed a signalling equilibrium if inferences satisfy IWD1 and $x_p = t^m$. To

[20] Clearly Assumption 2 is not satisfied. This, however, is not important for the example as we could 'make f continuous' without consequence for our analysis.

illustrate the ‘only if’ part of Corollary 2, we set $x_p = 1$ and show that some agent would want to deviate to $x = t^m$.

Table 4: Deriving uniform inferences in Example 3 with $x_p = 1$.

y	$T^{\phi_y}(t^m)$	$H^{\phi_y}(t^m)$
		-0.4
0	$[1.085, 2]$	-0.364
1	$[0.821, 2]$	-0.284
$\vdots$	$\vdots$	$\vdots$
9	$[0, 2]$	-0.4

The cycle of Table 4 repeats implying that some types of agent would always want to deviate to $x = t^m$. Deviating to $x = t^m$ when the norm is $x_p = 1$ is a signal that the agent has some type in the interval $[t_y^-, 2]$ for some t_y^-. This is a ‘desirable’ signal to send, in this example, as it can be inferred the agent cannot have type $b = 0$. This is illustrated in Figure 5c where we see that $E_+(x) > \varepsilon^{\phi^*}(x, x)$ for $x = t^m$, so the worse than condition is not satisfied.

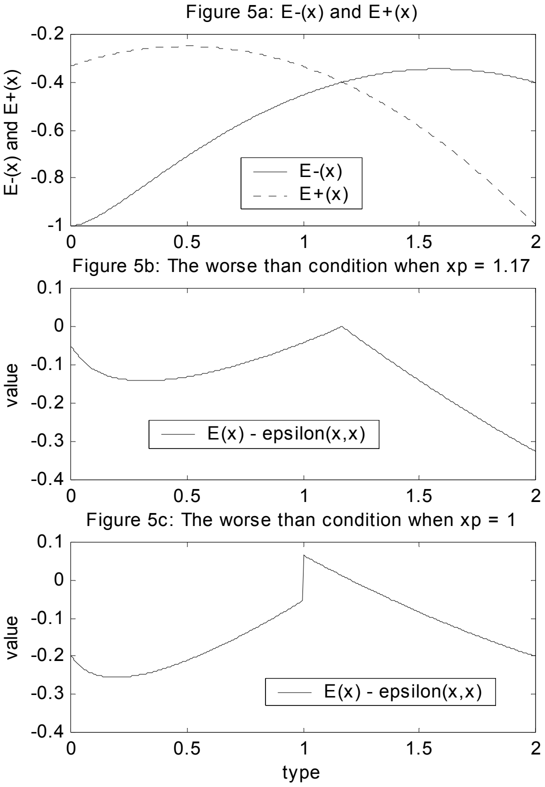

Figure 5: The values of $E_-(x)$ and $E_+(x)$ and of $E_-(x) - \varepsilon^{\phi^*}(x, x)$ for $x \in (x_l, x_p)$ and of $E_+(x) - \varepsilon^{\phi^*}(x, x)$ for all $x \in (x_p, x_h)$ in Example 3.

In Example 2 it is better to be inferred as having a type above 1 than below 1, so the norm cannot exist on action $x_p = 1$ if inferences satisfy uniform IWD1. It seems reasonable given the 'mass of agents' with type 0 that the norm should be above 1. More generally, the norm must exist on an action where $E_-(x)$ and $E_+(x)$ coincide because only at this point will there be no incentives to deviate either above or below x_p. A weakening of inferences to IWD1 thus suggests that the action preferred by the median type, and not that of the ideal type, may be an appropriate norm.

5.2 Equilibrium selection a partially conformist equilibrium

Section 5.1 and Corollary 2 focused on fully conformist equilibrium. We now briefly consider equilibrium selection in the case of a partially conformist equilibrium. Uniform IWD1 does not act as such a powerful equilibrium selection device in this case. To illustrate we can return to Example 1 but set $x_p = 1.05$. Now we find that $t_l = 0.177$ and $t_h = 2$. Figure 6 plots the action function μ^*.

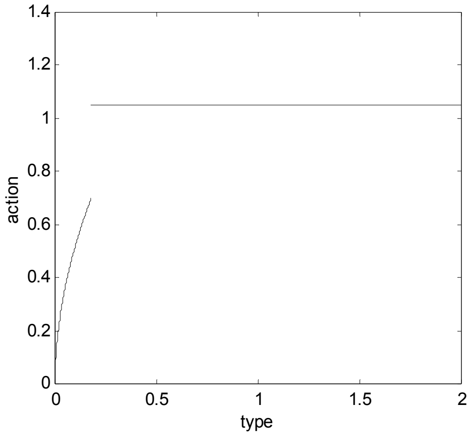

Figure 6: The action function in Example 1 when $x_p = 1.05$.

Given that $E_-(x_p) \neq E_+(x_p)$ we may expect that a signalling equilibrium need not exist if inferences satisfy uniform IWD1. There does, however, exist such an equilibrium. The following table illustrates why by showing that no agent would have an incentive to deviate to $x = 1.04$ if inferences satisfy uniform IWD1.

Table 5: Deriving uniform inferences in Example 1 with $x_p = 1.05$.

y	$T^{\phi_y}(1.04)$	$H^{\phi_y}(1.04)$
		-0.285
0	$[0.154, 1.045]$	-0.226
1	$[0.078, 2]$	-0.309
2	$\{t_l\}$	-0.677

The important point to recognize is that with a partially conformist equilibrium (μ^*, ϕ^*) where ϕ^* satisfies the D1 Criterion, the esteem from deviating to $x = 1.04$ is $h(0.177) = -0.677$, not $h(0) = -1$

(as would be the case with a fully conformist equilibrium). That inferences satisfy uniform IWD1 thus provides contrasting effects on the esteem to agents who deviate because they may be inferred as having a type closer to 1 than t_l but may also be inferred as having a type closer to 0 than t_l. This implies that uniform IWD1 is not so clearly 'weaker' than the D1 Criterion. That no agent would want to deviate to $x = 1.04$ confirms this. Figure 7 further illustrates by contrasting the worse than condition with a revised worse than condition. The revision uses $H(0,x) - \varepsilon^{\phi^*}(x,x)$ rather than the $H(t_l, x_l) - \varepsilon^{\phi^*}(x,x)$ of the worse than condition to recognize that agents could potentially be inferred to have types less than t_l.

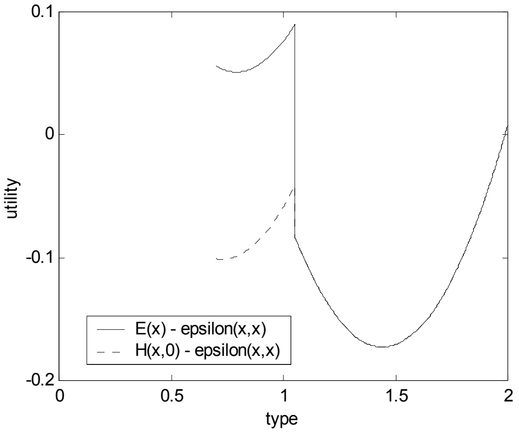

Figure 7: The values of $E_-(x) - \varepsilon^{\phi^*}(x,x)$ or $H(0,x) - \varepsilon^{\phi^*}(x,x)$ for $x \in (x_l, x_p)$ and of $E_+(x) - \varepsilon^{\phi^*}(x,x)$ for $x \in (x_p, x_h)$ in Example 1 when $x_p = 1.05$.

The discussion above highlights that in the case of a partially conformist equilibrium, inferences satisfying uniform IWD1 are not necessarily much weaker than the D1 Criterion. We do, however, still obtain an equilibrium selection result if the requirement of uniform inferences is relaxed.

Theorem 2: If (μ^*, ϕ^*) is a signalling equilibrium where $E_-(x_p) \neq E_+(x_p)$, then there exist inferences ϕ' satisfying IWD1 such that (μ^*, ϕ') is not a signalling equilibrium.

Proof: Suppose that $x_p > t^m$ with a symmetric argument treating $x_p < t^m$. Choose some $x \in (x_l, x_p)$ such that $x \geq t^m$. Clearly, $t^m \in T(x, H^{\phi^*}(x_p))$. Thus, by Lemma 1, inferences ϕ_0 such that $T^{\phi_0} = [t_l, t_0^+]$ for some $t_0^+ \in [t^m, t_h)$ satisfy IWD1. By Lemma 2 we have that $E_-(t_0^+) \geq H(t_l, t_h)$. Given that $H^{\phi^*}(x_p) = H(t_l, t_h)$ we have that $t^m \in T(x, H^{\phi_0}(x))$. Repeating this argument we get $t^m \in T(x, H^{\phi_y}(x))$. There exist therefore inferences satisfying y-WD1 such that a type t^m agent would wish to deviate to x rather than conform.■

The **proof of Corollary 2** is immediate by noting that the inferences given in the proof of Theorem 2 correspond to uniform inferences if $t_l = 0$ (as would be the case with a fully conformist equilibrium).

6 Non-uniform inferences

We have seen that if inferences satisfy uniform-IWD1, then there need not exist an equilibrium or can exist at most one action that can be the norm. We also know (Facts 3 and 4) that if inferences satisfy the D1 Criterion, then there exists a signalling equilibrium and any action could be the norm. Between the extremes of uniform IWD1 and the D1 Criterion are criteria that satisfy IWD1 and result in more or less signalling equilibrium and more or less actions that could be norms.

To illustrate we consider a simple ‘intermediate’ criterion of q uniform y-WD1. Recall that set $T(x, A)$ contains the types of agents who would wish to deviate to x if the esteem from choosing x is A. We refine the uniform y-WD1 by selecting the proportion q who have the most incentive to deviate. Formally, define

$$\beta(q, x, A) := \min\left\{\beta : F\left[T(x, \beta)\right] \geq qF\left[T(x, A)\right]\right\}$$

where $F\left[T\right]$ denotes the proportion of agents with types in set T.

q uniform y-WD1: Inference function ϕ_y satisfies q-uniform y-WD1 if

$$\phi_y(b, x) = \begin{cases} 0 \text{ for all } b \notin T(x, \beta(q, x, H^{\phi_{y-1}}(x)) \\ f(b)\left[\int_{T(x,\beta(q,x,H^{\phi_{y-1}}(x))} f(b)db\right]^{-1} \text{ otherwise} \end{cases}$$

and ϕ_{y-1} satisfies q uniform $y-1$-WD1.

If $q = 1$, then inferences satisfy uniform y-WD1, and any agent with a positive incentive to deviate is inferred as equally likely to have deviated. If $q = 0$, then inferences satisfy the D1 Criterion, and only the type of agent with the most incentive to deviate is inferred as likely to deviate. Between these extremes the proportion q who had the most incentive to deviate are inferred as likely to deviate. This can have consequences for both equilibrium existence and equilibrium selection. We can illustrate by revisiting Examples 2 and 3.

Recall that in Example 2 when $q = 1$ there did not exist a signalling equilibrium. Table 6 illustrates that a type $t = 0.5$ agent would not have an incentive to deviate to $x = 0.5$ when $q = 0.7$ (even though he does when $q = 1$). For example, if inferences satisfy the 0.7-uniform weak D1 Criterion then only agents with types $[0, 0.333]$ and not those with types $[0, 0.75]$ are inferred as likely to deviate. This lowers the esteem to choosing x from -0.56 to -0.704. The following can be compared to Table 2.

Table 6: Deriving q-uniform inferences in Example 2 with $x_p = 1$ and $q = 0.7$.

y	$T(x, H^{\phi_{y-1}}(0.5))$	$T^{\phi_y}(0.5)$	$H^{\phi_y}(0.5)$
			-0.5334
0	$[0, 0.75]$	$[0, 0.333]$	-0.704
1	$[0, 0.216]$	$[0, 0.151]$	-0.856
2	$\{0\}$	$\{0\}$	-1

Figure 8 confirms that when $q = 0.7$ there does exist a signalling equilibrium. The plot revises the worse than condition to reflect $q \neq 1$ and plots, for instance, $E_{-}(qx) - \varepsilon^{\phi^*}(x,x)$ for $x \leq 0.5$.

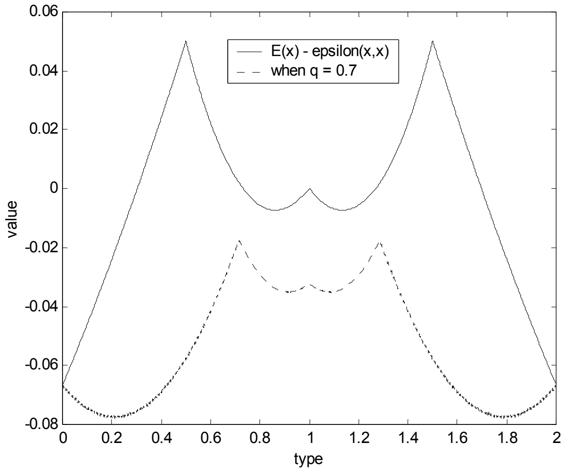

Figure 8: The worse than condition in Example 2 when $q = 1$ and $q = 0.7$.

Moving on, in Example 3 we know that when $q = 1$ there is a unique signalling equilibrium with norm $x_p = t^m$. If, however, $q = 0.8$ and inferences satisfy q-uniform IWD1, then there exists a signalling equilibrium with $x_p = 1$. Table 7 demonstrates why no agent would wish to deviate to $x = t^m$ when $x_p = 1$ and $q = 0.8$ and can be compared to Table 4.

Table 7: Deriving q-uniform inferences in Example 3 with $x_p = 1$ and $q = 0.8$.

y	$T(x, H^{\phi_{y-1}}(t^m))$	$T^{\phi_y}(t^m)$	$H^{\phi_y}(t^m)$
			-0.4
0	$[1.085, 2]$	$[1.268, 2]$	-0.447
1	$[1.4277, 2]$	$[1.542, 2]$	-0.612
2	$\{2\}$	$\{2\}$	-1

Figure 9 plots the revised worse than condition to demonstrates that when $q = 0.8$ there does exist a signalling equilibrium.

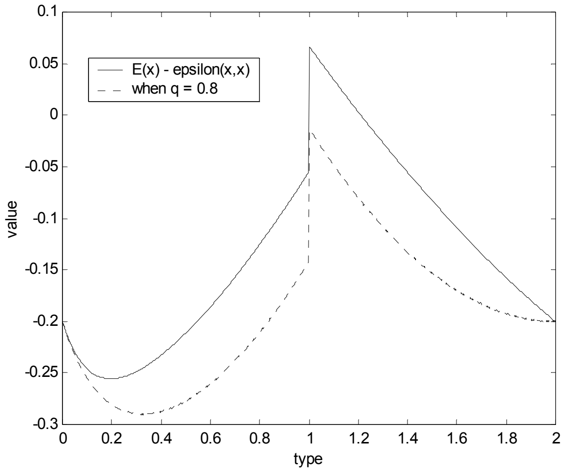

Figure 9: The worse than condition in Example 3 with $x_p = 1$ when $q = 1$ and $q = 0.7$.

More generally, it is clear that if there exists a signalling equilibrium (μ^*, ϕ^*) with inferences satisfying the D1 Criterion, then there exists a signalling equilibrium (μ^*, ϕ') with inferences satisfying q-uniform y-WD1 for all $q \leq \overline{q}$ for some $\overline{q} > 0$. The particular value of $\overline{q}$ will depend on the parameters of the model and nature of the signalling equilibrium. If, therefore, $q < 1$ and inferences satisfy q-uniform y-WD1, then more actions can be norms. To summarize consider the thought experiment of reducing q from 1 to 0. For $q = 1$ we may find that there exists no signalling equilibrium. For some q there will exist a signalling equilibrium with norm $x = t^m$. Reducing q further there will exist many signalling equilibria where any action in some interval $[x_-^q, x_+^q]$ can be a norm where $t^m \in [x_-^q, x_+^q]$. Decreasing q will increase the size of the interval until potentially any action $x \in [0, 2]$ can be the norm for $q = 0$.

7 Conclusions

Using a model of conformity introduced by Bernheim we have contrasted two systems of inferences. If inferences satisfy the D1 Criterion then any non-conformist is perceived to have the 'most extreme preferences'. If inferences satisfy a weaker IWD1 Criterion, then a non-conformist is perceived to be of a type that could have potentially gained from deviating. We have provided necessary and sufficient conditions for the existence of a signalling equilibrium that can be supported by any inferences satisfying IWD1. One consequence of this was to demonstrate that a weakening of inferences acts as an equilibrium selection device by reducing the set of signalling equilibria.

One implication of the equilibrium selection result we obtain is to reconsider the dynamics of conformity proposed by Bernheim who posits that a norm will remain relatively stable to changes in preferences, but a large change in preferences can result in a discontinuous jump to some new norm. This is possible because a particular action can be the norm for a wide range of type distributions when considering signalling equilibria that are supported by inferences satisfying the D1 Criterion. Our results demonstrate that once weaker inferences are assumed, there may exist a unique signalling equilibrium that changes as preferences change. This would suggest a more gradual evolution of the norm as preferences change. Many norms, such as 'how much to tip' are, at least in principle, flexible to changes in preferences, so this is not an unreasonable prediction.

An interesting question is what types of inferences, and what types of punishment mechanisms, do exist to sustain conformity. This requires looking more closely at human behavior in specific economic contexts. A particularly interesting issue is whether different 'belief structures' tend to emerge in different contexts. For example, contrast two social norms. One norm may be to set artificially high wages; the result is unemployment (Akerlof 1980). A second norm may be not to claim from the welfare state; the result is reduced unemployment (Lindbeck et al. 1999). Clearly one of these norms could be seen as good and the other bad. Is the equilibrium selection of 'weak beliefs' more likely to select 'good norms'? Also, are 'weak beliefs' more likely to exist around 'bad norms' as compared to 'good norms' on the basis that deviating from a bad norm is 'more understandable'? We leave these as issues for future work.

8 Appendix

8.1 Details of the Examples

In **Example 1**, As discussed by Bernheim, the solution for μ_s can be found by solving the linear dynamic system

$$\begin{bmatrix} \frac{dt}{dv} \\ \frac{dx}{dv} \end{bmatrix} = \begin{bmatrix} x - t \\ \lambda(1-t) \end{bmatrix}$$

where v is a 'dummy' variable. Working through one obtains the differential equation $x'' + x' + \lambda x = \lambda$. This has roots $-\frac{1}{2} \pm \frac{1}{2}(1-4\lambda)^{\frac{1}{2}}$. Setting $\lambda = 1.25$ and using initial conditions $x(0) = t(0) = 0$ gives particular solution

$$\begin{aligned} x(v) &= 1 + e^{-\frac{1}{2}v}\left(\frac{2\lambda - 1}{2}\sin v - \cos v\right) \\ t(v) &= 1 + e^{-\frac{1}{2}v}\left(\frac{1}{2\lambda}\left(\frac{2\lambda-1}{2} - 2\right)\sin v - \cos v\right). \end{aligned}$$

Tracing out x and t as functions of v provides action $x(v)$ as a function of type $t(v)$ and thus gives μ_s. Type t_l and t_h are found by finding the type of agent who is indifferent between choosing action $\mu_s(t_l)$ and being inferred as type t_l versus choosing action 1 and receiving esteem $H^\phi(1)$. When $x_p = 1$ the symmetry of the problem implies that $t_h = 2 - t_l$, so, from (4),

$$H^\phi(1) = \frac{-1}{2(1-t_l)} \int_{t_l}^{2-t_l} (1-b)^2. \tag{12}$$

Given this we can find t_l by setting

$$-(t_l - \mu_s(t_l))^2 - \lambda(1-t_l)^2 = -(1-t_l)^2 + \lambda H^{\phi}(1). \tag{13}$$

Using numerical methods to find an approximate value for t_l we obtain $t_l = 0.0761$. From (12) this implies that $H^{\phi^*}(1) = -0.2845$. We now have x_p, t_l, t_h and μ_s and thus have characterized the equilibrium. We do, of course, need to check that this is indeed an equilibrium. Section 3 sets out the additional requirements that need be imposed on inferences in order to satisfy the D1 Criterion (see, in particular, Fact 1), but, given x_p, t_l, t_h and μ_s and the requirement that inferences satisfy the D1 Criterion, we can derive the action function μ^* and ϕ^* using (2), (3) and Fact 1. It is then a simple matter to check (5) and verify that there does indeed exist a signalling equilibrium with inferences satisfying the D1 Criterion that can be characterized as above. It should also be clear that this is the unique signalling equilibrium with norm $x_p = 1$.

In order to derive the numbers given in Table 1 of Section 3.2 we first need to calculate $\varepsilon^{\phi^*}(t, 0.9)$ and $\varepsilon^{\phi^*}(t, 1.4)$ using (6). For example,

$$\varepsilon^{\phi^*}(t, 0.9) = \frac{-(t-\mu^*(t))^2 - \lambda H^{\phi^*}(\mu^*(t) + (0.9-t)^2}{\lambda}.$$

Figure 2 plots $\varepsilon^{\phi^*}(t, 0.9)$ and $\varepsilon^{\phi^*}(t, 1.4)$. Given a value for A one can then calculate $T(0.9, A)$ and $T(1.4, A)$. If $T^{\phi^y}(0.9) = [0, t_y]$ for some t_y and inferences are uniform then, using (9), we derive

$$H^{\phi_y}(0.9) = \frac{-1}{t_y}\int_0^{t_y}(1-b)^2 = -1 + t_y - \frac{t_y^2}{3}. \tag{14}$$

Similarly, if $T^{\phi^y}(1.4) = [t_y, 2]$ for some t_y and inferences are uniform, then

$$H^{\phi_y}(1.4) = 1 - t_y - \frac{(2-t_y)^2}{3}. \tag{15}$$

Using this and the values of $\varepsilon^{\phi^*}(t, 0.9)$ and $\varepsilon^{\phi^*}(t, 1.4)$, the numbers in Table 1 are easily obtained.

In Section 5.2 we set $x_p = 1.05$ in which case μ_s remains the same, but the values of t_l and t_h will change. Suppose that $t_h = 2$. Then

$$H^{\phi}(x_p) = \frac{-1}{2-t_l}\int_{t_l}^{2}(1-b)^2. \tag{16}$$

Using relation (13) we then obtain $t_l = 0.1777$ and $H^{\phi}(x_p) = -0.2847$. As when treating $x_p = 1$ we can now use Fact 1 to derive an action function and inference function and check (5) to verify that there does indeed exist a signalling equilibrium with inferences satisfying the D1 Criterion characterized as above. To derive the numbers in Table 5 we first calculate $\varepsilon^{\phi^*}(t, 1.04)$ using (6). The values of $\varepsilon^{\phi^*}(t, 1.04)$ are plotted in Figure 10.

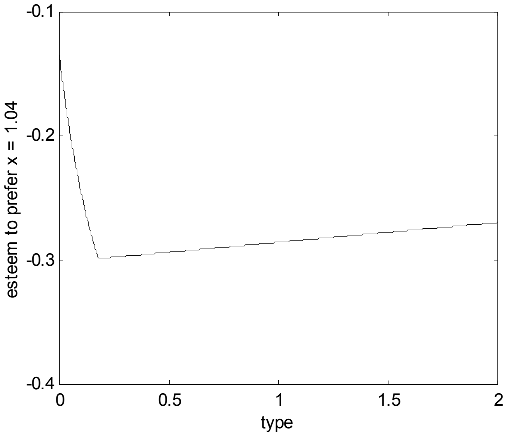

Figure 10: The value of $\varepsilon^{\phi^*}(t, 1.04)$ in Example 1 when $x_p = 1.05$.

If $T^{\phi_y} = [t_d, t_y]$ for some t_d, t_y, then

$$H^{\phi_y}(1.04) = \frac{-1}{t_y - t_d} \int_{t_d}^{t_y} (1-b)^2.$$

Given this we can iteratively derive the figures in Table 5.

Throughout the remainder let $H := \int_0^2 f(b)h(b)db$. In **Examples 2, 3 and 4** where there exists a fully conformist equilibrium, we know that $t_l = 0$ and $t_h = 2$, so

$$U^{\phi^*}(x_p, t, \phi^*) = -(x_p - t)^2 + \lambda H.$$

Using equation (6) this implies that $\varepsilon^{\phi^*}(x, t) \leq A$ if and only if

$$\frac{1}{\lambda}\left[U^{\phi^*}(x_p, t, \phi^*) + (x-t)^2\right] \leq A$$

or

$$t(x_p - x) \leq \frac{1}{2}\left[\lambda(A - H) + x_p^2 - x^2\right]. \tag{17}$$

To check that there exists a signalling equilibrium supported by the D1 Criterion, we only need check that a type 0 or type 2 agent has no incentive to choose $x = 0$ or 2 rather than not conform.[21] In equilibrium a type 0 agent receives payoff $U^{\phi^*}(x_p, 0, \phi^*) = -x_p^2 + \lambda H$, and a type 2 agent $U^{\phi^*}(x_p, 2, \phi^*) = -(2-x_p)^2 + \lambda H$. If a type 0 agent chooses $x = 0$ or a type 2 agent chooses $x = 2$ (and inferences satisfy the D1 Criterion), he receives payoff $U(0, 0, \phi^*) = U(2, 2, \phi^*) = -\lambda$. Thus, we need to check that

$$\max\left\{x_p^2, (2-x_p)^2\right\} \leq \lambda(1 + H). \tag{18}$$

[21] Lemma 1 tells us that type 0 and 2 agents have the most incentive to deviate to any action x. A type 0 or 2 agent has most incetive to deviate to $x = 0$ or $x = 2$ (if inferences satisfy the D1 Criterion).

In **Example 2**

$$H = \int_0^2 f(b)h(b)db = -2\left[0.9\int_0^{0.5}(1-b)^2 db + 0.1\int_{0.5}^1 (1-b)^2 db\right] = -\frac{16}{30}.$$

Putting $x_p = 1$ and $\lambda = 2.5$ into condition (18) demonstrates there does exist a fully conformist equilibrium (μ^*, ϕ^*) if inferences satisfy the D1 Criterion. To verify the figures given in the text we first note that $U^{\phi^*}(0.5) = -(1-0.5)^2 + \lambda H^{\phi^*}(1) = -\frac{19}{12}$, so $\varepsilon^{\phi^*}(0.5, 0.5) = \frac{1}{\lambda}U^{\phi^*}(0.5) = -\frac{19}{30} < -\frac{3}{5}$. Second, if $T^{\phi_y}(0.5) = [0, t_y]$ for some $t_y \geq 0.5$ then

$$H^{\phi'}(0.5) \geq -\frac{20}{9}\int_0^{0.5}\frac{9}{10}(1-b)^2 db = -\frac{7}{12}.$$

To derive the numbers in Table 2 we first use (17) to derive that $\varepsilon^{\phi^*}(x,t) \leq A$ if and only if

$$2t(1-x) \leq \lambda\left(A + \frac{16}{30}\right) + 1 - x^2. \tag{19}$$

If $T^{\phi^y}(0.5) = [0, t_y]$ for some $t_y \in (0.5, 1.5)$ and inferences are uniform, then, using (9),

$$\begin{aligned} H^{\phi_y}(0.5) &= \frac{-1}{0.45 + 0.1(t_y - 2)}\left[0.9\int_0^{0.5}(1-b)^2 + 0.1\int_{0.5}^{t_y}(1-b)^2\right] \\ &= \frac{-1}{0.45 + 0.1(t_y - 2)}\left[\frac{56}{240} + \frac{1}{10}\left(t_y - t_y^2 + \frac{t_y^3}{3}\right)\right]. \end{aligned} \tag{20}$$

Iterative use of (19) and (20) gives the numbers in Table 2. [To derive the numbers in Table 6 of Section 6 we iteratively use (19) and (14) to find $H^{\phi_y}(0.5)$ and $T(x, H^{\phi_{y-1}}(0.5)) = [0, t_y^1]$ with an intermediate step to find $T^{\phi_y}(0.5) = [0, t_y]$ where $F(t_y) = 0.7F(t_y^1)$.]

Thus far we have considered a signalling equilibrium with $x_p = 1$. To verify there are no signalling equilibrium in Example 2 with inferences satisfying uniform IWD1, we do need to consider other potential norms. We find that there does exist a signalling equilibria where inferences satisfy the D1 Criterion for any norm $x_p \in (0.73, 1.27)$. For any signalling equilibria we can repeat the exercise above to show that a type $t = 0.5$ (or $t = 1.5$) agent would wish to deviate from the norm. Figure 11 illustrates that for $x_p > 1$ a type $t = 0.5$ agent would wish to deviate to $x = 0.5$. Symmetry implies that a type $t = 1.5$ agent would wish to deviate for all $x_p < 1$.

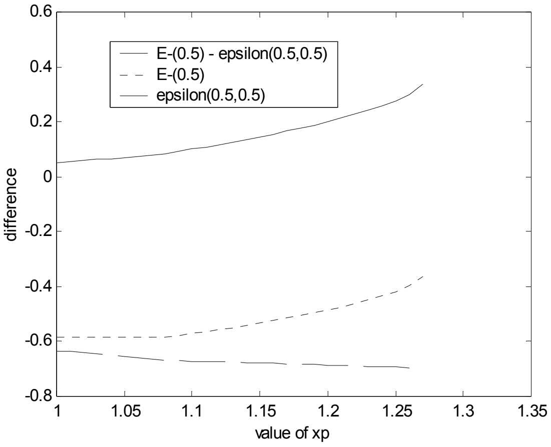

Figure 11: The value of $E_{-}(0.5) - \varepsilon^{\phi^*}(0.5, 0.5)$ against x_p in Example 2.

Finally, To illustrate that the worse than condition is satisfied when $\lambda > 5$, we can derive $\varepsilon^{\phi^*}(x, x) - E_{-}(x)$ for all $x \in (0, 0.5]$ to give

$$\varepsilon^{\phi^*}(t, 0.5) - E_{-}(x) > \frac{-(1-t)^2 + (0.5-t)^2}{5} - \frac{16}{30} + 1 - t + \frac{t^2}{3} = \frac{19}{60} - \frac{4}{5}t + \frac{t^2}{3} \geq 0$$

as desired. When $x \in (0.5, 1)$ the calculations are considerably more tedious because $E_{-}(x)$ is given by (20), so we omit the details.

In **Example 3** we see that $H = -\frac{1}{10} - \frac{9}{10}\frac{1}{3} = -0.4$. Using condition (18) we can check that there does exist a fully conformist equilibrium for $x_p = 1$ and $x_p = 1.17$ if $\lambda = 2.5$. To find t^m we know from Lemma 2 that $E_{+}(t^m) = E_{-}(t^m) = H = -0.4$. Using that

$$E_{+}(2-t) = -\frac{1}{t}\int_0^t (1-b)^2 db = -1 + t - \frac{t^2}{3} \tag{21}$$

for all $t < 2$ and setting $E_{+}(2-t) = -0.4$ gives quadratic $t^2 - 3t + 1.8 = 0$. Solving this and picking the root t^* less than 2 gives $t^m = 2 - t^* = \frac{1}{2} + \frac{3}{2\sqrt{5}} \simeq 1.17$. To derive the numbers in table 3 we first set $x_p = t^m$ and $x = 1$, and using (17) get that $\varepsilon^{\phi^*}(1, t) \leq A$ if and only if

$$t \leq \frac{2.5A + t^{m^2}}{2(t^m - 1)}. \tag{22}$$

Next note that if $T^{\phi_y} = [0, t_y]$ then

$$H^{\phi_y}(t_y) = \frac{-0.1 - 0.45\int_0^{t_y}(1-b)^2 db}{0.1 + 0.45t_y} = \frac{-0.1 - 0.45\left(t_y - t_y^2 + \frac{t_y^3}{3}\right)}{0.1 + 0.45t_y}. \tag{23}$$

Iterative use of (22) and (23) yields $T^{\phi_y}(1)$ and $H^{\phi_y}(1)$ and the numbers in Table 3. To derive the numbers in Table 4 we set $x_p = 1$ and $x = t^m$ to find $\varepsilon^{\phi^*}(1,t) \leq A$ if and only if

$$t \geq \frac{2.5A + 2 - t^{m^2}}{2(1-t^m)}. \tag{24}$$

If $T^{\phi_y} = [t_y, 2]$ then, from (21), we can see

$$H^{\phi_y}(t) = 1 - t_y - \frac{(2-t_y)^2}{3}. \tag{25}$$

Iterative use of (24) and (25) yields $T^{\phi_y}(1)$ and $H^{\phi_y}(1)$ and the numbers in Table 4. To derive the numbers in Table 7 of Section 6 we iteratively use (24) and (25) to find $H^{\phi_y}(t^m)$ and $T(x, H^{\phi_{y-1}}(t^m)) = [t_y^1, 2]$ with an intermediate step to find $T^{\phi_y}(t^m) = [t_y, 2]$ where $1 - F(t_y) = 0.8\left[1 - F(t_y^1)\right]$.

8.2 Proof of Corollary 1

Set $x_p = 1$. Using $H \geq -1/3$ we know by (18) that there exists a fully conformist equilibrium with inferences satisfying the D1 Criterion. Now, fix an $x < 1$ (using a symmetric argument for $x > 1$). Given Theorem 1 we need to show that,

$$g(0) - g(x-1) < \lambda\left[H - E_-(x)\right]. \tag{26}$$

The value of $g(0) - g(x-1) = (1-x)^2$ is a given. The value of $H - E_-(x)$ will depend on f. We claim (*) that $H - E_-(x)$ attains its minimum value when[22]

$$f(t) = \begin{cases} \frac{1}{2(1-x)} \text{ for all } t \in [x,1] \\ 0 \text{ for all } t \in [0,x). \end{cases}$$

Taking the claim as given, we have

$$\begin{aligned} H - E_-(x) &= -\frac{1}{1-x}\int_x^1 (1-b)^2 db + (1-x)^2 \\ &= \frac{-1}{1-x}\left[\frac{1}{3} - x + x^2 - \frac{x^3}{3}\right] + (1-x)^2. \end{aligned}$$

Substituting into (26) we require that

$$(1-x)^2 < \frac{-\lambda}{1-x}\left[\frac{1}{3} - x + x^2 - \frac{x^3}{3}\right] + \lambda(1-x)^2,$$

which simplifies to

$$\frac{2}{3}\lambda - 1 - x(1-x)(2\lambda - 3) - x^3\left(\frac{2}{3}\lambda - 1\right) > 0. \tag{27}$$

[22] The *support*$[f]$ does not equal T, but this is basically irrelevant. Making the support equal to T will only increase $H - E_-(x)$.

Given that $\max x(1-x) = \frac{1}{4}$, equation (27) holds if $\lambda > \frac{3}{2}$.

It remains to prove the claim (*). The intuition is simpler than the formal argument. Essentially we wish to minimize H and maximize $E_{-}(x)$ within the limits that f be unimodal around 1. Given that $h(t)$ is increasing in t for $t < 1$, the value of H will be minimized by setting $f(t) = y$ for some y and all $t \in [x, 1]$. The value of $E_{-}(x)$ is maximized, relative to this, (and the restriction that f be unimodal) by setting $f(t) = y$ for all $t \in [q, x)$ and some q. Given that $F(1) = 0.5$ we know that $(1-q)y = 0.5$, so set

$$f_q(t) = \begin{cases} \frac{1}{2(1-q)} \text{ for } t \in [q, 1] \\ 0 \text{ for } t \in [0, q) \end{cases} .$$

We now need to show that $H - E_{-}(x)$ is minimized when $q = x$. This follows due to the concavity of the h function. Given that $h(x) > h(t) + (x-t)h(1)$ for all $t < x$, putting more weight on types $t > x$ reduces $H - E_{-}(x)$. We can, however, provide a more formal argument. First, note that

$$\begin{aligned} H - E_{-}(x) &= \frac{1}{1-q}\int_q^1 h(b)db - \frac{1}{x-q}\int_q^x h(b)db \\ &= \frac{1}{1-q}\int_x^1 h(b)db - \frac{1-x}{(x-q)(1-q)}\int_q^x h(b)db \\ &= \frac{-1}{3(1-q)} + \frac{1}{x-q}\left(x - x^2 + \frac{x^3}{3}\right) - \frac{1-x}{(x-q)(1-q)}\left(q - q^2 + \frac{q^3}{3}\right) \qquad (28) \\ &= \frac{1}{1-q}\left[x - x^2 + \frac{x^3}{3} + \frac{1-x}{1-q}\left(1 - (q+x) + \frac{(q+x)^2}{3}\right) - \frac{1}{3}\right]. \qquad (29) \end{aligned}$$

Next, note that $x - x^2 + \frac{x^3}{3} \geq q - q^2 + \frac{q^3}{3} \geq 0$ and $1 \geq x \geq q \geq 0$. By inspection, to minimize $H - E_{-}(x)$, we need that $q = x$.■

References

Akerlof, G. A., 1980. A theory of social custom of which unemployment may be one consequence. Quarterly Journal of Economics 94, 749-75.

Azar, O.H., 2004. What sustains social norms and how they evolve? The case of tipping. Journal of Economic Behavior and Organization 54, 49-64.

Banks, J., Sobel, J., 1987. Equilibrium selection in signalling games. Econometrica 55, 647-661.

Bernheim, B. D., 1994. A theory of conformity. Journal of Political Economy 102, 841-877.

Cho, I., Kreps, D., 1987. Signaling games and stable equilibria. The Quarterly Journal of Economics 52, 179-221.

Elster, J., 1989. Social norms and economic theory. Journal of Economic Perspectives 3, 99-117.

Fudenberg, D., Tirole, J., 1991. Game Theory. Cambridge, MA: MIT Press.

Kreps, D., 1997. Intrinsic motivation and extrinsic incentives. The American Economic Review 87, 359-364.

Lindbeck, A., Nyberg, S., Weibull, J., 1999. Social norms and economic incentives in the welfare state. The Quarterly Journal of Economics 114, 1-35.