Web Appendix to accompany "*Signaling Quality via Demand Lockout*"

Date: 08/09/2024

This appendix contains four sections: (T1) Omitted proofs from propositions presented in the paper; (T2) The proofs of the three lemmas presented in the paper; (T3) A discussion of model extensions with the mathematical details; (T4) Analysis to validate R-level predictions; (T5) A discussion of alternative explanations of empirical results, and (T6) Omitted details of estimation of propensity score. There is some reproduction of the material from the main paper to make this appendix self-contained.

T1: Detailed Proofs of the Main Propositions

Proof of Proposition 1:

In the main text we presented the proof for the existence of the separating equilibrium with the high type playing lockout (R) and the low type playing not lockout (N). Here, we analyze the existence of the other equilibria and consider two versions of the intuitive criterion to restrict receivers' beliefs. Following Simester (1995) we assume that the receivers' belief reverts to the prior if they cannot eliminate either type given the most favorable belief. We also consider the standard assumption that beliefs in the second stage of the intuitive criterion reverts to the lowest type for which deviating is profitable, given the most favorable belief.

Here, we will analyze the other possible equilibria in this game. Suppose there exists a separating equilibrium with $\sigma_F(R|h) = 0$ and $\sigma_F(R|l) = 1$, which induces consumer beliefs: $\mu_C(R) = 0$ and $\mu_C(N) = 1$. The on-path payoffs for the low type is given by: $\Pi_l|R = p\lambda^1\left(\int_{\underline{k}}^{E[u_l]} f^1(k)dk\ \left(1 + (1-\alpha)\ \omega\ \int_{E[u_l]}^{1} f^1(k)dk\right)\right)$. The payoffs for the low types when deviating from the equilibrium and mimicking the high type is given by: $\Pi_l|N = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1 + (1-\alpha)\ \omega\ \int_{E[u_h]}^{1} f^i(k)dk\right)\right)$. Because the incentive constraint $\Pi_l|R > \Pi_l|N$ never holds, we can rule out this equilibrium.

Next, suppose there exists a pooling equilibrium with $\sigma_F(R|h) = 0$ and $\sigma_F(R|l) = 0$, which induces consumer beliefs: $\mu_C(N) = \mu_0$ and $\mu_C(R) = 0$.

The payoffs for the two types (on path) are given by: $\Pi_h|N = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1 + \alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$ and $\Pi_l|N = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1 + (1-\alpha)\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$

The payoffs for the two types (off path) when deviating are given by: $\Pi_h|R = p\lambda^1\left(\int_{\underline{k}}^{E[u_l]} f^1(k)dk\left(1+\alpha\ \omega\ \int_{E[u_l]}^{1} f^1(k)dk\right)\right)$ and $\Pi_h|R = p\lambda^1\left(\int_{\underline{k}}^{E[u_l]} f^1(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_l]}^{1} f^1(k)dk\right)\right)$.

This is a PBE because $\Pi_h|N > \Pi_h|R$ and $\Pi_l|N > \Pi_h|R$. To check if the off-path equilibrium belief $\mu_C(R) = 0$ survives the intuitive criterion, we need to first check which type has incentive to deviate under the most favorable consumer belief ($\mu_C(R) = 1$). The high type has incentive to deviate if $\lambda^1\left(\int_{\underline{k}}^{E[u_h]} f^1(k)dk\left(1+\alpha\ \omega\ \int_{E[u_h]}^{1} f^1(k)dk\right)\right) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$. The low type has incentive to deviate if $\lambda^1\left(\int_{\underline{k}}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_h]}^{1} f^1(k)dk\right)\right) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$.

The high type prefers to deviate if: $\frac{\int_{\underline{k}}^{E[u_{\mu_0}]} f^2(k)dk\left(1+\alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^2(k)dk\right)}{\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_\mu]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)}$

The low type has incentive to deviate if: $\frac{\int_{\underline{k}}^{E[u_{\mu_0}]} f^2(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^2(k)dk\right)}{\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\ \omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_\mu]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)}$. The intuitive criterion (Cho and Krops, 1987) potentially eliminates equilibria where only one type has incentive to deviate. Both types have incentive to deviate for a sufficiently large λ.

Next, we need to check in the region where both types have incentive to deviate under the most favorable belief which type also has incentive to deviate under the least favorable belief. The high type has incentive to deviate under the least favorable belief if $\lambda^1\left(\int_{\underline{k}}^{E[u_l]} f^1(k)dk\left(1+\alpha\ \omega\ \int_{E[u_l]}^{1} f^1(k)dk\right)\right) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$. The low type has incentive to deviate under the least favorable belief whenever $\lambda^1\left(\int_{\underline{k}}^{E[u_l]} f^1(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_l]}^{1} f^1(k)dk\right)\right) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$, which both never hold. If, instead of assuming beliefs are the least unfavorable (as is the general assumption of the Intuitive Criterion), we assume that beliefs revert to the prior whenever both types have incentive to deviate in the first stage (Simester 1995), the high type has incentive to deviate if $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\right.\right.$

$\alpha\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\Big)\Big)$. The low type has incentive to deviate if $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+(1-\right.\right.$ $\alpha)\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^1(k)dk\Big)\Big) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$.
Neither type has incentive to deviate under the least favorable or the prior belief, because
$\lambda^1\left(\int_{\underline{k}}^{E[u_l]} f^1(k)dk\left(1+\alpha\ \omega\ \int_{E[u_l]}^{1} f^1(k)dk\right)\right) < \lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\right.\right.$
$\left.\left.\alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^1(k)dk\Big)\Big) < \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)\ \forall\ \alpha\right.$

Therefore, the equilibrium survives the intuitive criterion whenever both types or neither type has incentive to deviate under the most favorable beliefs. If only the low type has incentive to deviate, the receiver needs to assign a belief of $\mu_C(R) = 0$ and we have shown above that no type has incentive to deviate under this low belief. Finally, if only the high type has incentive to deviate under the most favorable belief, the receiver needs to assign a belief of $\mu_C(R) = 1$ to the observed action. Thus, the intuitive criterion eliminates equilibria in which only the high type has incentive to deviate (under the most favorable belief) and the PBE thus survives the Intuitive Criterion whenever: $\frac{\lambda^1}{(1-\lambda^1)} \leq \frac{\int_{\underline{k}}^{E[u_{\mu_0}]} f^2(k)dk\left(1+\alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^2(k)dk\right)}{\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_\mu]} f^1(k)dk\right)\right)}$ or
$\frac{\int_{\underline{k}}^{E[u_{\mu_0}]} f^2(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^2(k)dk\right)}{\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\ \omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_\mu]} f^1(k)dk\right)\right)} \leq \frac{\lambda^1}{(1-\lambda^1)}$

Suppose there exists a pooling equilibrium with $\sigma_F(R|h) = 1$ and $\sigma_F(R|l) = 1$, which induces consumer beliefs: $\mu_C(R) = \mu_0$ and $\mu_C(N) = 0$.

The payoffs for the two types (on path, is given by: $\Pi_h|R = p\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\right.\right.$
$\left.\left.\alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right)$ and $\Pi_l|R = p\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right)$

The payoffs for the two types when deviating are given by: $\Pi_h|N =$
$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+\alpha\ \omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$ and $\Pi_l|N =$
$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+(1-\alpha)\ \omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$

For the equilibrium to hold, we require: (IC 1) $\Pi_h|R > \Pi_h|N$ and (IC 2) $\Pi_l|R > \Pi_l|N$. The constraint on the high type implies that he prefers to not deviate whenever:

$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\ \omega\int_{[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_{\mu_0}]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)}$$

The constraint on the low type implies that he prefers to not deviate whenever:

$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+(1-\alpha)\,\omega\int_{[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_{\mu_0}]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)}$$

The equilibrium holds for $\max\left\{\begin{array}{c}\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\,\omega\int_{[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_{\mu_0}]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)}\\ ,\\ \frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+(1-\alpha)\,\omega\int_{[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_{\mu_0}]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)}\end{array}\right\} < \frac{\lambda^1}{(1-\lambda^1)}$

Next, we check if the beliefs assigned to the off-path action survive the intuitive criterion. First, we check which type has incentive to deviate under the most favorable beliefs (i.e, $\mu(N)=1$. The high type has incentive to deviate if:

$\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\alpha\,\omega\int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) < \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_h]}^{1} f^i(k)dk\right)\right)$ and the low type has incentive to deviate if: $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) < \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_h]}^{1} f^i(k)dk\right)\right)$.

Both types always have incentive to deviate under the most favorable beliefs. Next, we need to check which type has incentive to deviate under the least favorable belief. This implies $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\alpha\,\omega\int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) < \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$ and $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) < \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$. Note that we showed above that this is only a PBE for $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\alpha\,\omega\int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$ and $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) > \sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$. Thus, neither type has an incentive to deviate under the least favorable beliefs and the Intuitive Criterion does not eliminate either type.

If we alternatively assume that beliefs revert to the prior whenever both types have incentive to deviate, the high type has incentive to deviate if $\lambda^1\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk\left(1+\right.\right.$

$\alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^1(k)dk\Big)\Big) < \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk \left(1+\alpha\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$. The low type has incentive to deviate if $\lambda^1 \left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk \left(1+(1-\alpha)\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^1(k)dk\right)\right) < \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk \left(1+(1-\alpha)\ \omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$. Both types always have an incentive to deviate and the pooling equilibrium survives the intuitive criterion.

Finally, we can now summarize the equilibria and conclude that the focal separating equilibrium exists and is unique if:

$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\ \omega\ \int_{E[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+\alpha\ \omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)} \leq \frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\ \omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)},$$

$$\frac{\int_{\underline{k}}^{E[u_{\mu_0}]} f^2(k)dk\left(1+\alpha\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^2(k)dk\right)}{\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\left(1+\alpha\omega\left(1-\left(2\int_{\underline{k}}^{E[u_{\mu_0}]} f^1(k)dk+\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\right)\right)\right)} \leq \frac{\lambda^1}{(1-\lambda^1)} \leq \frac{\int_{\underline{k}}^{E[u_{\mu_0}]} f^2(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_{\mu_0}]}^{1} f^2(k)dk\right)}{\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\ \omega\left(1-\left(2\int_{\underline{k}}^{Eu_{\mu_0}} f^1(k)dk+\int_{E[u_{\mu_0}]}^{E[u_h]} f^1(k)dk\right)\right)\right)}$$ and

$$\frac{\lambda^1}{(1-\lambda^1)} \leq \frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\ \omega \int_{[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk(1+\alpha\omega\ (1-\left(2\int_{\underline{k}}^{E[u_l]} f^1(k)dk+\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk\right)}$$ or

$$\frac{\lambda^1}{(1-\lambda^1)} \leq \frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+(1-\alpha)\ \omega \int_{[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk(1+(1-\alpha)\omega\ (1-\left(2\int_{\underline{k}}^{E[u_l]} f^1(k)dk+\int_{E[u_l]}^{E[u_{\mu_0}]} f^1(k)dk\right)}.$$

It is generally difficult to further characterize the density functions and parameter region in which the equilibrium is unique, but it is possible to find parameters that satisfy the inequalities above.

When $\mu_0 \to 0$, the separating equilibrium is unique in the range in which it is a PBE. $\mu_0 \to 0$ implies $E\left[u_{\mu_0}\right] \to E[u_l]$, and the separating equilibrium is the unique PBE whenever

$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\ \omega\ \int_{E[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+\alpha\ \omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)} \leq \frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+(1-\alpha)\ \omega\ \int_{E[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)}$$

Proof of Proposition T1:

We are looking at a potential separating equilibrium in which the high-quality product spends an amount a on advertising and the low-quality product does not advertise. Recall that prices are exogenously set at $p > 0$ for both products.

Suppose there exists a separating equilibrium with $\sigma_F(A|h) = 1$ and $\sigma_F(A|l) = 0$, which induces consumer beliefs: $\mu_C(A) = 1$ and $\mu_C(NA) = 0$. As before, for notational ease, let $[u_h] = \alpha\,\bar{q} + (1-\alpha)\underline{q} - p$, $E[u_l] = (1-\alpha)\bar{q} + \alpha\underline{q} - p$, and $E\left[u_{\mu_0}\right] = \mu_0 E[u_h] + (1-\mu_0)E[u_l]$

The payoffs for the two types (on path) is given by $\Pi_h|A = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1 + \alpha\,\omega\int_{E[u_h]}^{1} f^i(k)dk\right)\right) - a$ and $\Pi_l|NA = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1 + (1-\alpha)\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$

The payoffs for the two types when mimicking the other type are given by $\Pi_h|NA = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1 + \alpha\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$ and $\Pi_l|A = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1 + (1-\alpha)\,\omega\int_{E[u_h]}^{1} f^i(k)dk\right)\right) - a$

For this equilibrium to hold, we require the following two incentive compatibility to hold:

$$\sum_{i\in\{1,2\}}\lambda^i\int_{E[u_l]}^{E[u_h]} f^i(k)dk\left(1 + (1-\alpha)\omega\left(\int_{E[u_h]}^{1} f^i(k)dk - \int_{\underline{k}}^{Eu_l} f^i(k)dk\right)\right) < \frac{a}{p} < \sum_{i\in\{1,2\}}\lambda^i\int_{E[u_l]}^{E[u_h]} f^i(k)dk\left(1 + \alpha\omega\left(\int_{E[u_h]}^{1} f^i(k)dk - \int_{\underline{k}}^{Eu_l} f^i(k)dk\right)\right)$$

By FOSD, $\int_{E[u_h]}^{1} f^2(k)dk - \int_{\underline{k}}^{Eu_l} f^2(k)dk < \int_{E[u_h]}^{1} f^1(k)dk - \int_{\underline{k}}^{Eu_l} f^1(k)dk$. Thus, $\int_{E[u_h]}^{1} f^2(k)dk - \int_{\underline{k}}^{Eu_l} f^2(k)dk > 0\ is$ a sufficient condition to ensure that there exists an a such that the above inequalities hold.

Next, we consider the other possible pooling and separating equilibria.

Suppose there exists a separating equilibrium with $\sigma_F(A|h) = 0$ and $\sigma_F(A|l\,) = 1$, which induces consumer beliefs: $\mu_C(A) = 0$ and $\mu_C(NA) = 1$. The payoffs for the low type is given by: $\Pi_l|A = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1 + (1-\alpha)\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right) - a$ The payoffs for the low types when mimicking the high type is given by: $\Pi_l|NA = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1 + \alpha\,\omega\int_{E[u_h]}^{1} f^i(k)dk\right)\right)$.

Because the incentive constraint $\Pi_l|A > \Pi_l|NA$ never holds, we can rule out this equilibrium.

Next, suppose there exists a pooling equilibrium with $\sigma_F(A|h) = 0$ and $\sigma_F(A|l) = 0$, which induces consumer beliefs: $\mu_C(NA) = \mu_0$ and $\mu_C(A) = 0$.

The payoffs for the two types (on path) are given by: $\Pi_h|NA =$ $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)$ and $\Pi_l|NA =$ $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)$

The payoffs for the two types off path are given by: $\Pi_h|A = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]}f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_l]}^{1}f^i(k)dk\right)\right)-a$ and $\Pi_l|A = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]}f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_l]}^{1}f^i(k)dk\right)\right)-a.$

This is a PBE because $\Pi_h|NA > \Pi_h|A$ and $\Pi_l|NA > \Pi_h|A$. To check if the off-path equilibrium belief $\mu_C(A)=0$ survives the intuitive criterion, we need to first check which type has incentive to deviate under the most favorable consumer belief ($\mu_C(A)=1$). The high type has incentive to deviate if

$\sum_{i\in\{1,2\}}\lambda^i\left(\int_{E[u_{\mu_0}]}^{E[u_h]}f^i(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1}f^1(k)dk-\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\right)\right)\right)>\frac{a}{p}$. The low type has incentive to deviate if $\sum_{i\in\{1,2\}}\lambda^i\left(\int_{E[u_{\mu_0}]}^{E[u_h]}f^i(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_h]}^{1}f^1(k)dk-\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\right)\right)\right)>\frac{a}{p}$.

The intuitive criterion potentially eliminates equilibria where only one type has incentive to deviate. Both types have incentive to deviate for a sufficiently small a.

Next, for the region where both types have incentive to deviate under the highest favorable belief, we check which type has incentive to deviate under the least favorable belief. The high type has incentive to deviate under the least favorable belief if: $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]}f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_l]}^{1}f^i(k)dk\right)\right)-a >$ $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)$. The low type has incentive to deviate if $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]}f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_l]}^{1}f^i(k)dk\right)\right)-a >$ $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)$. Both types never have incentive to deviate under the least favorable belief. If we alternatively assume that beliefs revert to the prior whenever both types have incentive to deviate under the most favorable beliefs, the high type has incentive to deviate if $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)-a > p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)$. The low type has incentive to deviate if $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)-a > p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{\mu_0}]}^{1}f^i(k)dk\right)\right)$ Both types never have incentive to deviate when assuming their beliefs revert to the prior.

Therefore, the equilibrium survives the intuitive criterion whenever both types or neither type has incentive to deviate under the most favorable beliefs. If only the low type has incentive to deviate, the receiver needs to assign a belief of $\mu_C(A) = 0$ and we have shown above that no type has incentive to deviate under this low belief. Finally, if only the high type has incentive to deviate under the most favorable belief, the receiver needs to assign a belief of $\mu_C(A) = 1$ to the observed action. Thus, the intuitive criterion eliminates equilibria in which only the high type has incentive to deviate (under the most favorable belief):
$\sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk \left(1 + (1-\alpha)\,\omega \int_{E[u_h]}^{1} f^i(k)dk\right)\right) - \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk \left(1 + (1-\alpha)\,\omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right) < \frac{a}{p} < \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk \left(1 + \alpha\,\omega \int_{E[u_h]}^{1} f^i(k)dk\right)\right) - \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk \left(1 + \alpha\,\omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$

The PBE survives the intuitive criterion whenever $\sum_{i\in\{1,2\}} \lambda^i \left(\int_{E[u_{\mu_0}]}^{E[u_h]} f^i(k)dk \left(1 + \alpha\omega \left(\int_{E[u_h]}^{1} f^i(k)dk - \int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\right)\right)\right) < \frac{a}{p}$ or $\frac{a}{p} < \sum_{i\in\{1,2\}} \lambda^i \left(\int_{E[u_{\mu_0}]}^{E[u_h]} f^i(k)dk \left(1 + (1-\alpha)\omega \left(\int_{E[u_h]}^{1} f^i(k)dk - \int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\right)\right)\right)$.

Suppose there is a pooling equilibrium with $\sigma_F(A|h) = 1$ and $\sigma_F(A|l) = 1$, which induces consumer beliefs: $\mu_C(A) = \mu_0$ and $\mu_C(NA) = 0$.

The payoffs for the two types (on path, is given by: $\Pi_h|A = \mathrm{p} \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk \left(1 + \alpha\,\omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right) - a$ and $\Pi_l|A = \mathrm{p} \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk \left(1 + (1-\alpha)\omega \int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right) - a$.

The payoffs for the two types when deviating are given by: $\Pi_h|NA = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk \left(1 + \alpha\,\omega \int_{E[u_l]}^{1} f^i(k)dk\right)\right)$ and $\Pi_l|NA = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk \left(1 + (1-\alpha)\,\omega \int_{E[u_l]}^{1} f^i(k)dk\right)\right)$

For the equilibrium to hold, we require: (IC 1) $\Pi_h|R > \Pi_h|N$ and (IC 2) $\Pi_l|R > \Pi_l|N$. The constraint on the high type implies that he prefers to not deviate whenever:

$$\frac{a}{p} < \sum_{i\in\{1,2\}} \int_{E[u_l]}^{E[u_{\mu_0}]} f^i(k)dk \left(1 + \alpha\omega \left(\int_{E[u_{\mu_0}]}^{1} f^i(k)dk - \int_{\underline{k}}^{E[u_l]} f^i(k)dk\right)\right)$$

The constraint on the low type implies that he prefers to not deviate whenever:

$$\frac{a}{p} < \sum_{i\in\{1,2\}} \int_{E[u_l]}^{E[u_{\mu_0}]} f^i(k)dk \left(1 + (1-\alpha)\omega \left(\int_{E[u_{\mu_0}]}^{1} f^i(k)dk - \int_{\underline{k}}^{E[u_l]} f^i(k)dk\right)\right)$$

The equilibrium holds for

$$\min\left\{\begin{array}{l} \sum_{i\in\{1,2\}}\int_{E[u_l]}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\omega\ \left(\int_{E[u_{\mu_0}]}^{1} f^i(k)dk-\int_{\underline{k}}^{E[u_l]} f^i(k)dk\right)\right), \\ \sum_{i\in\{1,2\}}\int_{E[u_l]}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\omega\ \left(\int_{E[u_{\mu_0}]}^{1} f^i(k)dk-\int_{\underline{k}}^{E[u_l]} f^i(k)dk\right)\right) \end{array}\right\}>\frac{a}{p}$$

Next, we check if the beliefs survive the intuitive criterion. First, we check which type has incentive to deviate under the most favorable beliefs. The high type has incentive to deviate if:

$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a<$
$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1+\alpha\ \omega\int_{E[u_h]}^{1} f^i(k)dk\right)\right)$ and the low type has incentive to deviate if: $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a<$
$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk\left(1+(1-\alpha)\ \omega\int_{E[u_h]}^{1} f^i(k)dk\right)\right)$.

Both types always have incentive to deviate under the most favorable beliefs. Next, we need to check which type has incentive to deviate under the least favorable belief. This implies $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a<\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+\alpha\ \omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$ and the low type has incentive to deviate if:
$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a<$
$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$

Note that we showed above that this is only a PBE for $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a>\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+\alpha\ \omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$ and $p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a>$
$p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$. Thus, neither type has an incentive to deviate under the least favorable beliefs and the Intuitive Criterion does not eliminate either type.

If we alternatively assume that beliefs revert to the prior whenever both types have incentive to deviate, the high type has incentive to deviate if $\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a<\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+\alpha\ \omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$. The low type has incentive to deviate if $\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)-a<p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{\mu_0}]} f^i(k)dk\left(1+(1-\alpha)\omega\int_{E[u_{\mu_0}]}^{1} f^i(k)dk\right)\right)$. Both types always have an incentive to deviate and the pooling equilibrium survives the intuitive criterion.

Finally, we can now summarize the equilibria and conclude that the focal separating equilibrium exists and is unique if:

$\sum_{i\in\{1,2\}}\lambda^i\int_{E[u_l]}^{E[u_h]}f^i(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_h]}^{1}f^i(k)dk-\int_{\underline{k}}^{Eu_l}f^i(k)dk\right)\right)<\frac{a}{p}<$
$\sum_{i\in\{1,2\}}\lambda^i\int_{E[u_l]}^{E[u_h]}f^i(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1}f^i(k)dk-\int_{\underline{k}}^{Eu_l}f^i(k)dk\right)\right)$,

$\sum_{i\in\{1,2\}}\lambda^i\left(\int_{E[u_{\mu_0}]}^{E[u_h]}f^i(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_h]}^{1}f^i(k)dk-\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\right)\right)\right)<\frac{a}{p}<$
$\sum_{i\in\{1,2\}}\lambda^i\left(\int_{E[u_{\mu_0}]}^{E[u_h]}f^i(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1}f^i(k)dk-\int_{\underline{k}}^{E[u_{\mu_0}]}f^i(k)dk\right)\right)\right)$, and

$$\sum_{i\in\{1,2\}}\int_{E[u_l]}^{E[u_{\mu_0}]}f^i(k)dk\left(1+\alpha\omega\left(\int_{E[u_{\mu_0}]}^{1}f^i(k)dk-\int_{\underline{k}}^{E[u_l]}f^i(k)dk\right)\right)<\frac{a}{p},\ or\ \sum_{i\in\{1,2\}}\int_{E[u_l]}^{E[u_{\mu_0}]}f^i(k)dk\left(1\right.$$
$$\left.+(1-\alpha)\omega\left(\int_{E[u_{\mu_0}]}^{1}f^i(k)dk-\int_{\underline{k}}^{E[u_l]}f^i(k)dk\right)\right)<\frac{a}{p}.$$

When $\mu_0\to 0$, the separating equilibrium is unique in the range in which it exists. $\mu\_0\to 0$ implies $E[u_{\mu_0}]{-}{>}\,E[u_l]$, and the separating equilibrium holds and is unique whenever
$\sum_{i\in\{1,2\}}\lambda^i\int_{E[u_l]}^{E[u_h]}f^i(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_h]}^{1}f^i(k)dk-\int_{\underline{k}}^{Eu_l}f^i(k)dk\right)\right)<\frac{a}{p}<$
$\sum_{i\in\{1,2\}}\lambda^i\int_{E[u_l]}^{E[u_h]}f^i(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1}f^i(k)dk-\int_{\underline{k}}^{Eu_l}f^i(k)dk\right)\right)$ and $0<\frac{a}{p}$

Proof of Proposition 3

To show the first part, consider the payoffs from both actions for high and low types:

$\Pi_h|R=\gamma\,\Pi_h|\left(R,r(q_j)=q_j\right)+(1-\gamma)\,\Pi_h|\left(R,r(q_j)\neq q_j\right)=\gamma\,\lambda^1\left(\int_{\underline{k}}^{E[u_{h|+}]}f^1(k)dk\,\left(1+\right.\right.$
$\left.\left.\alpha\,\omega\,\int_{E[u_{h|+}]}^{1}f^1(k)dk\right)\right)+(1-\gamma)\lambda^1\left(\int_{\underline{k}}^{E[u_{h|-}]}f^1(k)dk\,\left(1+\alpha\,\omega\,\int_{E[u_{h|-}]}^{1}f^1(k)dk\right)\right)$

$\Pi_l|N=\gamma\,\Pi_h\left(R,r(q_j)=q_j\right)+(1-\gamma)\,\Pi_h|\left(R,r(q_j)\neq q_j\right)=$
$\gamma\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_{l|-}]}f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_{l|-}]}^{1}f^i(k)dk\right)\right)+(1-$
$\gamma)\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l|+]}f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_l|+]}^{1}f^i(k)dk\right)\right)$

The payoffs for the two types when mimicking the other type are given by:

$$\Pi_h|N = \gamma \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_l|+]} f^i(k)dk \left(1 + \alpha\, \omega \int_{E[u_l|+]}^{1} f^i(k)dk\right)\right) + (1 - \gamma) \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_l|-]} f^i(k)dk \left(1 + \alpha\, \omega \int_{E[u_l|-]}^{1} f^i(k)dk\right)\right)$$

$$\Pi_l|R = \gamma\, \lambda^1 \left(\int_{\underline{k}}^{E[u_{h|-}]} f^1(k)dk \left(1 + (1-\alpha)\omega \int_{E[u_{h|-}]}^{1} f^1(k)dk\right)\right) + (1 - \gamma)\lambda^1 \left(\int_{\underline{k}}^{E[u_{h|+}]} f^1(k)dk \left(1 + (1-\alpha)\omega \int_{E[u_{h|+}]}^{1} f^1(k)dk\right)\right)$$

Combining, the IC conditions imply:

$$\frac{\gamma \int_{\underline{k}}^{E[u_l|+]} f^2(k)dk\left(1+\alpha\,\omega \int_{E[u_l|+]}^{1} f^2(k)dk\right)+(1-\gamma)\int_{\underline{k}}^{E[u_l|-]} f^2(k)dk\left(1+\alpha\,\omega \int_{E[u_l|-]}^{1} f^2(k)dk\right)}{\gamma \int_{E[u_l|+]}^{E[u_h|+]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_h|+]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l|+]} f^1(k)dk\right)\right)+(1-\gamma)\int_{E[u_l|-]}^{E[u_h|-]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_h|-]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l|-]} f^1(k)dk\right)\right)} <$$

$$\frac{\lambda^1}{(1-\lambda^1)} <$$

$$\frac{\gamma \int_{\underline{k}}^{E[u_l|-]} f^2(k)dk\left(1+(1-\alpha)\,\omega \int_{E[u_l|-]}^{1} f^2(k)dk\right)+(1-\gamma)\int_{\underline{k}}^{E[u_l|+]} f^2(k)dk\left(1+(1-\alpha)\,\omega \int_{E[u_l|+]}^{1} f^2(k)dk\right)}{\gamma \int_{E[u_l|-]}^{E[u_h|-]} f^1(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_h|-]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l|-]} f^1(k)dk\right)\right)+(1-\gamma)\int_{E[u_l|+]}^{E[u_h|+]} f^1(k)dk\left(1+(1-\alpha)\omega\left(\int_{E[u_h|+]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l|+]} f^1(k)dk\right)\right)}.$$

This is equivalent to the IC in proposition 1 for $\gamma = \frac{1}{2}$ and the PBE exists for a sufficiently low γ.

To show the first part of the statement, we consider the payoff for the high type playing lockout (R) in a separating equilibrium and compare it to the payoff under the equilibrium without the signaling action being available. As $\gamma \to 1$, this simplifies to $\lambda^1 \left(\int_{\underline{k}}^{1} f^1(k)dk \right) > \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{1} f^i(k)dk\right)$, which never holds.

Comparing profits in the pooling equilibrium and the separating equilibrium:

$$\Pi_h\big|\big(R, \mu_C(R) = 1\big) - \Pi_h\big|\big(N, \mu_C(N) = \mu_0\big) = \gamma\, \lambda^1 \left(\int_{\underline{k}}^{E[u_{h|+}]} f^1(k)dk \left(1 + \alpha\, \omega \int_{E[u_{h|+}]}^{1} f^1(k)dk\right)\right) - (1-\gamma)\lambda^1 \left(\int_{\underline{k}}^{E[u_{h|-}]} f^1(k)dk \left(1 + \alpha\, \omega \int_{E[u_{h|-}]}^{1} f^1(k)dk\right)\right) - \sum_{i\in\{1,2\}} \left(\gamma\lambda^i \left(\int_{\underline{k}}^{E[u_\mu|+]} f^i(k)dk \left(1 + \alpha\, \omega \int_{E[u_\mu|+]}^{1} f^i(k)dk\right)\right) + (1-\gamma)\lambda^1 \left(\int_{\underline{k}}^{E[u_\mu|-]} f^i(k)dk \left(1 + \alpha\, \omega \int_{E[u_\mu|-]}^{1} f^i(k)dk\right)\right)\right) = \gamma \left(\lambda^1 \int_{E[u_\mu|+]}^{E[u_{h|+}]} f^1(k)dk \left(1 + \alpha\, \omega \left(\int_{E[u_{h|+}]}^{1} f^1(k)dk - \int_{\underline{k}}^{E[u_\mu|+]} f^1(k)dk\right)\right) - \lambda^2 \left(\int_{\underline{k}}^{E[u_\mu|+]} f^2(k)dk \left(1 + \alpha\, \omega \int_{E[u_\mu|+]}^{1} f^2(k)dk\right)\right)\right) + (1-\gamma) \left(\lambda^1 \int_{E[u_\mu|-]}^{E[u_{h|-}]} f^1(k)dk \left(1 + \alpha\, \omega \left(\int_{E[u_{h|-}]}^{1} f^1(k)dk - \int_{\underline{k}}^{E[u_\mu|-]} f^1(k)dk\right)\right) - \lambda^2 \left(\int_{\underline{k}}^{E[u_\mu|-]} f^2(k)dk \left(1 + \alpha\, \omega \int_{E[u_\mu|-]}^{1} f^2(k)dk\right)\right)\right)$$

As $\gamma \to \frac{1}{2}$, we have already shown in proposition 1 that $\Pi_h|R > \Pi_h|N$ if μ_0 is sufficiently low. As $\gamma \to 1$, the above expression reduces to $-\lambda^2$. $\Pi_h|R = \Pi_h|N$ can have multiple solutions, depending on the PDF of the opportunity costs. By continuity of the density, whenever the separating equilibrium is a PBE, there exists a $\bar{\gamma}$ such that $\Pi_h|R < \Pi_h|N$ for $\gamma > \bar{\gamma}$ and a $\underline{\gamma}$ such that $\Pi_h|R > \Pi_h|N$ for $\gamma < \underline{\gamma}$.

T2: Proofs of the Lemmas

Lemma 1 (Role of Uncertainty in the main result)

We now consider the role of quality uncertainty for the firm. Clearly, there cannot be a separating equilibrium when the high-type and the low-type firms offer the high and low quality products with equal probability ($\alpha = \frac{1}{2}$). Intuitively, one might expect that the separating equilibrium is most feasible when both types are maximally vertically differentiated and the high-type always offers high-quality products and the low-type always offers low-quality products, as the incentive for the high type to convince consumers that they are high-type is maximized. However, because expected quality also affects WOM, there is no separating equilibrium when the quality difference between the two types is at the maximum ($\alpha \to 1$).

Lemma 1: There exists an $\alpha^* < 1$, such that no separating equilibrium survives if firm uncertainty is too low ($\alpha > \alpha^*$). When μ_0 is sufficiently high, high-type profits under full certainty are lower than high-type profits in a separating equilibrium under uncertainty ($\Pi_h|(R, \alpha = \alpha^*) > \Pi_h(N, \alpha \to 1)$.

Proof: Evaluated at the lower bound of $\alpha \to \frac{1}{2}$, $E[u_l] \to \frac{\underline{q}+\bar{q}}{2} - p$. And $E[u_h] \to \frac{\underline{q}+\bar{q}}{2} - p$ A necessary condition for the equilibrium is $\int_{\frac{\underline{q}+\bar{q}}{2}-p}^{1} f^2(k)dk < \int_{\frac{\underline{q}+\bar{q}}{2}-p}^{1} f^1(k)dk - \int_{\underline{k}}^{\frac{\underline{q}+\bar{q}}{2}-p} f^1(k)dk$, which simplifies to:

$2\int_{\underline{k}}^{\frac{\underline{q}+\bar{q}}{2}-p} f^1(k)dk < \int_{\underline{k}}^{\frac{\underline{q}+\bar{q}}{2}-p} f^2(k)dk.$

As $\alpha \to 1$, $\int_{E[u_l]}^{1} f^2(k)dk \to 1$, $\int_{E[u_h]}^{1} f^1(k)dk \to 0$, and $\int_{\underline{k}}^{E[u_l]} f^1(k)dk \to 0$, and the necessary equilibrium condition implies $1 < 0$, which never holds. Noting that $\frac{\partial E[u_l]}{\partial \alpha} < 0$ and $\frac{\partial E[u_h]}{\partial \alpha} > 0$, by continuity of the probability density: $\frac{\partial}{\partial \alpha}\int_{E[u_l]}^{1} f^2(k)dk > 0$, $\frac{\partial}{\partial \alpha}\int_{E[u_h]}^{1} f^1(k)dk < 0$, and $\frac{\partial}{\partial \alpha}\int_{\underline{k}}^{E[u_l]} f^1(k)dk > 0$, which implies that there exists, at most, one $\alpha^* = \{\alpha | \int_{\alpha\underline{q}+(1-\alpha)\bar{q}-p}^{1} f^2(k)dk = \int_{\alpha\bar{q}+(1-\alpha)\underline{q}-p}^{1} f^1(k)dk - \int_{\underline{k}}^{\alpha\underline{q}+(1-\alpha)\bar{q}-p} f^1(k)dk\}$.

Thus, for $2\int_{\underline{k}}^{\frac{\underline{q}+\bar{q}}{2}-p} f^1(k)dk < \int_{\underline{k}}^{\frac{\underline{q}+\bar{q}}{2}-p} f^2(k)dk$ and $\alpha < \alpha^*$, there exists a range $\lambda \in (\underline{\lambda}, \bar{\lambda})$, such that the separating equilibrium is a PBE.

Comparing the profits between the pooling equilibrium evaluated at $\alpha \to 1$ and the separating equilibrium at some $\frac{1}{2} < \hat{\alpha} < \alpha^*$.

$(\Pi_h|N, \alpha = 1) = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{\mu_0\bar{q}+(1-\mu_0)\underline{q}-p} f^i(k)dk \left(1 + \omega \int_{\mu_0\bar{q}+(1-\mu_0)\underline{q}-p}^{1} f^i(k)dk\right)\right).$

$(\Pi_h|R, \alpha = \hat{\alpha}) = \lambda^1 p \left(\int_{\underline{k}}^{\hat{\alpha}\bar{q}+(1-\hat{\alpha})\underline{q}-p} f^1(k)dk \ \left(1 + \hat{\alpha} \ \omega \ \int_{\hat{\alpha}\bar{q}+(1-\hat{\alpha})\underline{q}-p}^{1} f^1(k)dk\right)\right)$

$(\Pi_h|R, \alpha = \alpha^*)$ is independent of μ_0 and $(\Pi_h|N, \alpha = 1)$ is increasing in μ_0. Evaluated at $\mu_0 = 0$, $(\Pi_h|N, \alpha = 1) = 0$. Evaluated at $\mu_0 = 1$, $(\Pi_h|N, \alpha = 1) = p$. We have shown in Proposition 1 that $0 < (\Pi_h|R, \alpha = \alpha^*) < \lambda^1$. Thus, by continuity of the distribution of k, there exists one $\mu_0^* = \left\{ \mu_0 | \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{\mu_0\bar{q}+(1-\mu_0)\underline{q}-p} f^i(k)dk \left(1 + \alpha \ \omega \int_{\mu_0\bar{q}+(1-\mu_0)\underline{q}-p}^{1} f^i(k)dk\right)\right) = \right.$
$\left. \lambda^1 p \left(\int_{\underline{k}}^{\hat{\alpha}\bar{q}+(1-\hat{\alpha})\underline{q}-p} f^1(k)dk \ \left(1 + \alpha \ \omega \ \int_{\hat{\alpha}\bar{q}+(1-\hat{\alpha})\underline{q}-p}^{1} f^1(k)dk\right)\right)\right\}$ and $\mu < \mu^* \Leftrightarrow$
$(\Pi_h|N, \alpha = 1) < (\Pi_h|R, \alpha = \hat{\alpha})$.

Lemma 2 (Role of uncertainty in the advertising equilibrium)

This part considers the role of uncertainty in the advertising equilibrium.

***Lemma 2**:* There exists no separating equilibrium with positive profits for the high quality type if there is no firm uncertainty ($\alpha \to 1$).

Proof: At any possible separating equilibrium, it follows directly from the IC that the highest possible advertising cost is such that $\Pi_h|A \to 0$. Next, we consider the profit under the lowest possible advertising cost for $\alpha \to 1$ and $\alpha = \hat{\alpha} < 1$

The lowest possible advertising cost in equilibrium is given by the IC
$p \sum_{i\in\{1,2\}} \lambda^i \int_{E[u_l]}^{E[u_h]} f^i(k)dk \left(1 + (1-\alpha)\omega \left(\int_{E[u_h]}^{1} f^i(k)dk - \int_{\underline{k}}^{E[u_l]} f^i(k)dk \right)\right) < a$. As $\alpha \to 1$:
$p \sum_{i\in\{1,2\}} \lambda^i \int_{E[u_l]}^{E[u_h]} f^i(k)dk \left(1 + (1-\alpha)\omega \left(\int_{E[u_h]}^{1} f^i(k)dk - \int_{\underline{k}}^{E[u_l]} f^i(k)dk \right)\right) \to \ p$. Evaluating profit at $\alpha \to \ 1$ and $a \to p$, we get $\Pi_h|A = p - p = 0$. Profit for the low type is zero because $E[u_l] \to \underline{q} - p = \underline{k}$.

$\Pi_l|NA = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk \left(1 + (1-\alpha) \ \omega \int_{E[u_l]}^{1} f^i(k)dk\right)\right) = 0$

For any equilibrium with $\hat{\alpha} < 1$, the IC is given by $p \sum_{i\in\{1,2\}} \lambda^i \int_{\hat{\alpha}\underline{q}+(1-\hat{a})\bar{q}-p}^{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p} f^i(k)dk \left(1 + \right.$
$\left.(1 - \ \hat{\alpha})\omega \left(\int_{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p}^{1} f^i(k)dk - \int_{\underline{k}}^{\hat{\alpha}\underline{q}+(1-\hat{a})\bar{q}-p} f^i(k)dk \right)\right) < a.$

Evaluating profit at $\hat{\alpha}$: $\Pi_h|A = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p} f^i(k)dk\left(1+\right.\right.$
$\left.\left.\hat{\alpha}\,\omega\int_{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p}^{1} f^i(k)dk\right)\right) - p\sum_{i\in\{1,2\}}\lambda^i\int_{E[u_l]}^{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p} f^i(k)dk\left(1+(1-\right.$
$\left.\hat{\alpha})\omega\left(\int_{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p}^{1} f^i(k)dk - \int_{\underline{k}}^{\hat{\alpha}\underline{q}+(1-\hat{a})\bar{q}-p} f^i(k)dk\right)\right) =$
$p\sum_{i\in\{1,2\}}\omega\int_{\underline{k}}^{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p} f^i(k)dk\left((2\hat{\alpha}-1)\left(1-\int_{\underline{k}}^{\hat{\alpha}\bar{q}+(1-\hat{a})\underline{q}-p} f^i(k)dk\right)\right) +$
$\int_{\underline{k}}^{\hat{\alpha}\underline{q}+(1-\hat{a})\bar{q}-p} f^i(k)dk\left(1+(1-\hat{\alpha})\omega\left(1-\int_{\underline{k}}^{\hat{\alpha}\underline{q}+(1-\hat{a})\bar{q}-p} f^i(k)dk\right)\right) > 0$

We have shown that, whenever a separating equilibrium exists, profits are higher for $\alpha < 1$.

The intuition from Lemma 2 is similar to the case with the lockout signaling equilibrium. If the high-type always offers a high-quality product and advertises, no consumer would wait for WOM. As before, this would allow the low-type to mimic the high-type and receive the same profit, which would contradict the equilibrium. Thus, as the expected high-type product quality approaches the upper bound, the lowest cost of advertising prices that allow for a separating equilibrium increases towards the point where the high-type makes no profits.

Proof of Lemma 3*:*

Consumer beliefs for high expert reviews are given by: $E\left[q|\left(S,\gamma,r\left(q_j\right)=\bar{q}\right)\right] = b_h\bar{q} + (1-b_h)\underline{q}$. Expanding, we get , $E\left[q|\left(S,\gamma,r\left(q_j\right)=\bar{q}\right)\right] = \frac{\gamma q(S)}{\gamma q(S)+(1-\gamma)(1-q(S))}\bar{q} + \left(1-\frac{\gamma q(S)}{\gamma q(S)+(1-\gamma)(1-q(S))}\right)\underline{q}$. Plugging in $\gamma = 1$, we get $E\left[q|\left(S,\gamma,r\left(q_j\right)=\bar{q}\right)\right] = \bar{q}$.

Similarly, beliefs for low expert reviews are given by: $E\left[q|\left(S,\gamma,r\left(q_j\right)=\underline{q}\right)\right] = b_l\bar{q} + (1-b_l)\underline{q}$, Expanding, we get $E\left[q|\left(S,\gamma,r\left(q_j\right)=\underline{q}\right)\right] = \frac{(1-\gamma)q(S)}{(1-\gamma)q(S)+\gamma(1-q(S))}\bar{q} + \left(1-\frac{(1-\gamma)q(S)}{(1-\gamma)q(S)+\gamma(1-q(S))}\right)\underline{q}$. Plugging in in $\gamma = 1$, we get $E\left[q|\left(S,\gamma,r\left(q_j\right)=\underline{q}\right)\right] = \underline{q}$.

Recall that we have assumed $F^i(\bar{q}-p) = 1$ and that $F^i(k)$ is strictly increasing. We can find that $E\left[q|\left(S,\gamma,r\left(q_j\right)=\bar{q}\right)\right] > 0 \Leftrightarrow \frac{q(S(1-q(S))(q_h-q_l)}{(1-q(S)+\gamma(2q(S)-1)^2} > 0$, which always holds. For all $\gamma < 1$ $\Leftrightarrow E\left[q|\left(S,\gamma,r\left(q_j\right)=\bar{q}\right)\right] < \bar{q}$. By continuity of the CDF, for $< 1 \Leftrightarrow F^i\left(E\left[q|\left(S,\gamma,r\left(q_j\right)=\bar{q}\right)\right]-p\right) < 1$., which completes the proof that there exists a mass of consumers that do not consume conditional on positive review.

T3: Model Extensions

In this section we first derrive the equilibrium with advertising. We then briefly consider the robustness of the demand lockout equilibrium under assumptions of endogenous cost and partial word-of-mouth spillover. To simplify exposition, we assume that consumers exogenously are assigned to be potential early or late consumers and α denotes the proportion of consumers that potentially consumers in the first period. The remaining $1-\alpha$ consumers enter the market in the second period and make a consumption decision.

In this section, we compare advertising to the demand lockout. But before we do so, we consider advertising in isolation as a possible way to signal to the market.

As before, nature decides the firm type in the initial period. Upon learning of its quality, the firm can opt to spend money on uninformative advertising at an exogenously given cost indexed as a. For simplicity, we assume a discrete action space defined as $s = \{A, NA\}$, where *A* denotes spending an amount of a on advertising, and *NA* indicates no advertising spending. Influentials observe the advertising decision, make (rational) inferences, and consume accordingly in the first period. Finally, in the second period, followers (potentially) get informed about the actual quality through WOM, and they either consume or do not consume.

First, consider an equilibrium where only the high-type firm advertises. For this to be a separating equilibrium, the low type should not have an incentive to mimic the high type and vice-versa. Assuming the equilibrium induces the correct posterior beliefs, the high-type firm's profit is given by

$$\Pi_h|A = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk \left(1 + \alpha\, \omega \int_{E[u_h]}^{1} f^i(k)dk\right)\right) - a\,, \quad (3)$$

where a is the advertising cost, and p is the price. The profit for the low-type firm is

$$\Pi_l|NA = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk \left(1 + (1-\alpha)\, \omega \int_{E[u_l]}^{1} f^i(k)dk\right)\right). \quad (4)$$

If the high type mimics the low type and does not advertise, profits are $\Pi_h|NA = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk \left(1 + \alpha\, \omega \int_{E[u_l]}^{1} f^i(k)dk\right)\right)$. The low type's profits from acting as a high type by advertising are $\Pi_l|A = p \sum_{i\in\{1,2\}} \lambda^i \left(\int_{\underline{k}}^{E[u_h]} f^i(k)dk \left(1 + (1-\alpha)\, \omega \int_{E[u_h]}^{1} f^i(k)dk\right)\right) - a$.

Proposition T1. *In a separating equilibrium, only the high-type firm advertises. In the first period, consumers with relatively low opportunity costs consume both under no advertising and advertising ($k \leq E[q_l]$). Consumers with relatively high opportunity costs ($E[q_l] < k \leq E[q_h]$ consume only when the product is accompanied by costly advertising. In the second period, a fraction of consumers receive WOM and consume if the product quality is high.*

Advertising as a signal has been studied in the literature in many different contexts (e.g., Nelson 1974). In our setting, because consumers in the second period consume the product only upon receiving positive WOM, the value of the high opportunity cost segment differs across the low- and high types. Even if the low-type firm can make the high opportunity cost segment (mistakenly) believe that it is of high quality, the true quality gets revealed upon consumption, and ultimately, positive WOM is less likely to be generated for such a firm. Similar to the demand lockout signal, the high type benefit when

their quality is revealed to the high opportunity cost segment because doing so induces positive WOM. For this equilibrium to hold, advertising needs to be sufficiently expensive, so the low type is better off not advertising. The increase in demand for the low quality via the signal needs to be less than the cost of advertising. At the same time, the advertising price cannot be so high that a high-type firm also cannot benefit from advertising. Similar to the lockout mechanism, information uncertainty for the firm plays an important role in the existence of this equilibrium. (See Lemma 2 and other details in the Web Appendix)

Endogenous cost: In this section, we model the cost a choice variable for the firm and assume that consumers observe it. Instead of nature having full control over the quality of an offering, a firm can endogenously set the cost, which affects the quality. To make things tractable, assume that quality is still discrete, and the probability of the product being high-quality increases as the cost increases. Let the continuous function $\delta(c_j)$ be the probability of product j being high quality, conditional on cost c_j. Further, let $\frac{\partial \delta(c_j)}{\partial c_j} > 0, \frac{d^2\delta(c_j)}{dc_j^2} \leq 0, \delta(0) = \underline{\delta}$, and $\delta(p) = \overline{\delta}$, where $\underline{\delta}$ is the lower limit and $\overline{\delta}$ is the upper limit of the probability of being the high type. We assume $\underline{\delta} = 0$ and $\overline{\delta} = 1$ to simplify the analysis.

In the separating equilibrium, profits for the high type are given by: $\Pi_h^{\mathrm{r}} = \lambda^1 \times (p - c_j)$ and profit for the low-quality product from not using the signal is given by: $\Pi_l^n = (1 - \lambda^1)(p - c_j)$.

Expected profit, conditional on signaling, is given by:

$$E[\pi|c_j] = \delta(c_j)\, \pi_h^r(c_j) + \left(1 - \delta(c_j)\right)\, \pi_l^n(c_j) \tag{T4}$$

Consumers form their beliefs about quality conditional on the observed cost:

$$E[q|c_j] = \delta(c_j) q_h + \left(1 - \delta(c_j)\right) q_l \tag{T5}$$

Proposition 5: *For a sufficiently concave cost-to- quality mapping* $\left(\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0} > \frac{1-\lambda^1}{(2\lambda^1-1)p}\right)$*, the equilibrium cost is positive* $(c^* > 0)$*. The signaling lockout equilibrium exists for a sufficiently high difference between segments' reservation price* $(k^1 > \overline{k^1})$*. For a low difference in reservation prices* $(k^1 \leq \overline{k^1})$*, the firm sets the cost such that all consumers demand the product in the pooling equilibrium without signaling* $(E[q|\delta(c_j)] = p + k^1)$*.*

Proof: The profit function when the firm is using the lockout signal is given by: $E[\pi|c_j] = \delta(c_j)\, \left(\lambda^1 \times (p - c_j)\right) + \left(1 - \delta(c_j)\right)\left((1 - \lambda^1)(p - c_j)\right)$ and the first-order condition is given by: $\frac{dE[\pi|c_j]}{dc_j} = \frac{d\delta(c_j)}{dc_j} * \left(\lambda^1 \times (p - c_j)\right) - \delta(c_j)\, \lambda^1 - \frac{d\delta(c_j)}{dc_j} * \left((1 - \lambda^1)(p - c_j)\right) + \left(1 - \delta(c_j)\right)(\lambda^1 - 1) =$

$$\frac{d\delta(c_j)}{dc_j}\Big((2\lambda^1-1)(p-c_j)\Big)+\delta(c_j)\,(1-2\lambda^1)+\lambda^1-1=0$$

The second derivative, with respect to c_j is given by: $\frac{d^2E[\pi|c_j]}{dc_j^2}=(1-2\lambda^1)\left(2\frac{d\delta(c_j)}{dc_j}-\frac{d^2\delta(c_j)}{dc_j^2}(p-c_j)\right)\leq 0$, because $\frac{\partial\delta(c_j)}{\partial c_j}>0$, $\frac{d^2\delta(c_j)}{dc_j^2}\leq 0$ and $\lambda^1>\frac{1}{2}$. To establish an interior solution, we now consider the FOC at the upper and lower bound of the cost.

First, consider the upper bound, where $c_j=p$. It is easy to verify that $E[\pi|c_j]=\delta(c_j)\,\left(\lambda^1\times(p-c_j)\right)+\left(1-\delta(c_j)\right)\left((1-\lambda^1)(p-c_j)\right)=0\;for\;c_j=p$. Further, set cost to $c_j=p-\epsilon$ to get $E[\pi|c_j]=\delta(c_j)\,(\lambda^1\times\epsilon)+\left(1-\delta(c_j)\right)\left((1-\lambda^1)(\epsilon)\right)>0\;\forall\;q_l\geq p$, which we have assumed to be true. Thus, the equilibrium cost cannot be at the upper bound.

Now, consider the lower bound, where $c_j=0$. Because $\delta(0)=0$, the first-order condition reduces to: $\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0}*\left((2\lambda^1-1)p\right)+\lambda^1-1=0$

In order for an interior solution to exist, the first derivative needs to be increasing at the lower bound: $\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0}\times\left((2\lambda^1-1)p\right)+\lambda^1-1\geq 0\Leftrightarrow\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0}>\frac{1-\lambda^1}{(2\lambda^1-1)p}$.

Recall that $\delta(c_j)$: $[0,p]\rightarrow[0,1]$ and $\frac{d^2\delta(c_j)}{dc_j^2}\leq 0$. By the mean value theorem, we can conclude that $\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0}\geq\frac{1}{p}$. Thus, a sufficient statistic for an interior solution is given by: $\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0}\geq\frac{1}{p}\geq\frac{1-\lambda^1}{(2\lambda^1-1)p}\Leftrightarrow\lambda^1>\frac{2}{3}$

We have established that the solution for cost for the lockout equilibrium is given implicitly by $\frac{d\delta(c_j)}{dc_j}\Big((2\lambda^1-1)(p-c_j)\Big)+\delta(c_j)\,(1-2\lambda^1)+\lambda^1-1=0$, whenever $\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0}>\frac{1-\lambda^1}{(2\lambda^1-1)p}$ or $\lambda^1>\frac{2}{3}$ or $c_j=0$ else.

Next, we need to consider the two possible pooling equilibria where the firm serves both segments without signaling or serves the quality insensitive segment only. Note that the case where the firm sets cost at zero and only serves the quality insensitive segment is equivalent to the signaling equilibrium with cost at zero. This equilibrium only occurs if $\frac{d\delta(c_j)}{dc_j}\Big|_{c_j=0}<\frac{1-\lambda^1}{(2\lambda^1-1)p}$. Next, we consider the pooling equilibrium where the firm invests sufficiently much in quality to increase expected quality to the level where both segments purchase the product absent any signaling.

Consider c_j such that $E[q|\delta(c_j)]=p+k^1$. Profit now is given by: $\pi=p-c_j$, s.t. $E[q|\delta(c_j)]<p+k^1$.

Note that $E\left[q|\delta(c_j)\right]$ is given by: $E\left[q|\delta(c_j)\right] = \delta(c_j)q_h + \left(1-\delta(c_j)\right)q_l = p + k^1 \Leftrightarrow \delta^*(c_j) = \frac{p+k^1-q_l}{q_h-q_l}$. Taking the derivative, we find that

$\frac{d\delta^*(c_j)}{dk^1} = \frac{1}{q_h-q_l} > 0$. Because $\delta(c_j)$ is an increasing function, we find that $\frac{d\pi}{dk^1} < 0$

Plugging in, we find that $\pi = p - c_j = 0\ for\ k^1 = q_h - p$ and $\pi = p - c_j = p$ for $k^1 = q_l - p$.

Because $p > \delta(c_j)\ \left(\lambda^1 \times (p-c_j)\right) + \left(1-\delta(c_j)\right)\left((1-\lambda^1)(p-c_j)\right) > 0$ and $\frac{\mathrm{d}\pi}{\mathrm{dk}^1} < 0$ we can conclude that there exists a $\overline{k^1}$ where the separating equilibrium dominates the pooling equilibrium whenever $k^1 > \overline{k^1}$. For $k^1 \leq \overline{k^1}$, the pooling equilibrium dominates and the firm sets c such that $E\left[q|\delta(c_j)\right] = p + k^1$.

The main result from the proposition is that the lockout remains an equilibrium strategy, even when the firm has control over the cost and the quality. The equilibrium can break under two conditions. First, if the cost-to-quality mapping is not sufficiently concave, it is prohibitively expensive to invest in quality. The firm is then better off serving only the quality insensitive segment with a low-quality product. Secondly, if the difference between the two segments' reservation price is not sufficiently high, the firm is better off investing sufficiently high such that expected utility from consumption is at least as high as the reservation price of both segments.

Spillover of Word of Mouth: We have assumed that there is no communication between the segments. We will now consider the case where there is a limited spillover of word-of-mouth across the segments. Consider that absent consumption, a proportion of consumers ω^i within each segment receives word-of-mouth from the other segment irrespective of its own prior consumption. If the high-quality product excludes segment 2 and restricts demand to segment 1, the profit is given by $\Pi_h^{\mathrm{r}} = \lambda^1 \times p$. Profit for the low-quality product from not using the signal is given by: $\Pi_l^n = (1-\lambda^1)p$

If the low-quality type mimics the high-quality type (via demand lockout), its profit is given by $\Pi_l^{\mathrm{r}} = \alpha^1 \times\ \lambda^1 \times p$, because consumers in period 2 will receive word of mouth, revealing the low-quality and not consume. If the high-quality product mimics the low-quality type, its profits are $\Pi_h^{\mathrm{n}} = (1-\lambda^1)p + \lambda^1\omega^1 p$. It is straightforward to show that the separating equilibrium described in proposition 1 holds, as long as the proportion of consumers receiving spillover information is sufficiently low $\omega^1 \leq 2 - \frac{1}{\lambda^1}$.

An alternative assumption is that consumers do not observe word-of-mouth but can see consumption in the other segment. Rational consumers will make an inference that word-of-mouth revealed a quality above the reservation quality. This assumption ensures that if there is consumption in the second period in segment i, all consumers infer that the quality of the product is such that $q_j \geq p - k^i$. Recall that we have assumed that only consumers in the quality insensitive segment have demand for the low-quality product ($p + k^1 > q_l > p$) and consumers in both segments demand the high type product $(q_h > p + k^1)$. Combining the two assumptions ($q_h \geq p + k^1 > q_l > p$), it is clear that consumption by the quality insensitive

segment does not reveal any useful information. However, consumption in the second period by the quality sensitive segment reveals the product to be of high quality to all consumers. Still, the information is not actionable because the second segment would have either consumed the product in the first period or be locked out in equilibrium. If the segment consumed the product in the first period, word-of-mouth within the segment reveals the quality. If the segment is locked out, there is no use to knowing the quality because of the lockout.

Imperfect Lockout: We now consider the case in which the lockout is imperfect and some consumers are able to consume the product despite being in the locked out segment. To make the problem tractable, we assume that there is a random proportion of consumers (denoted by $b \in (0,1)$) in segment 2 that is not affected by the lockout and always has the opportunity to consume the product. The proportion b is common knowledge and is independent from the opportunity cost[1].

The on-path profits are given by

$$\Pi_h|R = p\lambda^1\left(\int_{\underline{k}}^{E[u_h]} f^1(k)dk\left(1+\alpha\,\omega\int_{E[u_h]}^{1} f^1(k)dk\right)\right)+p\lambda^2\left(b\int_{\underline{k}}^{E[u_h]} f^2(k)dk\left(1+\alpha\,b\,\omega\int_{E[u_h]}^{1} f^2(k)dk\right)\right)$$

$$\Pi_l|N = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$$

The off-path profits if each type deviates from the equilibrium path is given by:

$$\Pi_l|R = \lambda^1 p\left(\int_{\underline{k}}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\,\omega\int_{E[u_h]}^{1} f^1(k)dk\right)\right)+p\lambda^2\left(b\int_{\underline{k}}^{E[u_h]} f^2(k)dk\left(1+(1-\alpha)\,b\omega\int_{E[u_h]}^{1} f^2(k)dk\right)\right)$$

$$\Pi_h|N = p\sum_{i\in\{1,2\}}\lambda^i\left(\int_{\underline{k}}^{E[u_l]} f^i(k)dk\left(1+\alpha\,\omega\int_{E[u_l]}^{1} f^i(k)dk\right)\right)$$

We can summaries the incentive compatibility constraints as:

[1] Alternatively, it could be that consumers with the lowest opportunity cost have a higher probability to evade the lockout (e.g., have a fake ID to watch R-rated movies). However, to keep the problem tractable, we leave that problem to future researchers.

$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\,\omega\,\int_{E[u_l]}^{1} f^2(k)dk\right)-\left(b\int_{\underline{k}}^{E[u_h]} f^2(k)dk\left(1+\alpha\, b\omega\,\int_{E[u_h]}^{1} f^2(k)dk\right)\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)} <$$
$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+(1-\alpha)\,\omega\,\int_{E[u_l]}^{1} f^2(k)dk\right)-\left(b\int_{\underline{k}}^{E[u_h]} f^2(k)dk\left(1+(1-\alpha)\, b\omega\,\int_{E[u_h]}^{1} f^2(k)dk\right)\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+(1-\alpha)\,\omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)}.$$

We can make several interesting observations regarding the equilibrium. First, it always breaks down for a sufficiently high b. Secondly, the equilibrium region "shifts", as now the cost of locking out demand is reduced. Thus, there exists a parameter space in which the separating equilibrium is a PBE for some $b > 0$ but not for $b = 0$. Thirdly, whenever there is slack in the IC of the separating equilibrium, an increase in b increases high type profits and holds low type profits

Proof:

To prove the first statement, observe that $\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+(1-\alpha)\,\omega\,\int_{E[u_l]}^{1} f^2(k)dk\right) -$ $\left(b\int_{\underline{k}}^{E[u_h]} f^2(k)dk\,\left(1+(1-\alpha)\, b\omega\,\int_{E[u_h]}^{1} f^2(k)dk\right)\right)$ evaluated at $b = 1$ is always negative, which means the IC on the low type can never hold. For the second statement, one can easily construct a set of parameters in which the IC on the high type only holds whenever $b > 0$:

$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\,\omega\,\int_{E[u_l]}^{1} f^2(k)dk\right)-\left(b\int_{\underline{k}}^{E[u_h]} f^2(k)dk\left(1+\alpha\, b\omega\,\int_{E[u_h]}^{1} f^2(k)dk\right)\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)} < \frac{\lambda^1}{(1-\lambda^1)} <$$
$$\frac{\int_{\underline{k}}^{E[u_l]} f^2(k)dk\left(1+\alpha\,\omega\,\int_{E[u_l]}^{1} f^2(k)dk\right)}{\int_{E[u_l]}^{E[u_h]} f^1(k)dk\left(1+\alpha\omega\left(\int_{E[u_h]}^{1} f^1(k)dk-\int_{\underline{k}}^{E[u_l]} f^1(k)dk\right)\right)}$$. For the third statement, note that $\frac{\partial}{\partial}\Pi_h|R > 0$ and $\frac{\partial}{\partial}\Pi_l|N = 0$

T4: Additional Empirical Analyses

The reasonable estimation of the R-level is critical for our identification strategy and thus deserves additional scrutiny and robustness checks. We will proceed to check the ratings from the different models for internal consistency and use expert reviews (from *Common Sense Media*) to validate the prediction's external validity.

As a very first check, we display the distributions of the estimated values for R-levels for movies rated R and PG-13 using the five predictive models. The distributions have intuitive appeal. The mass of movies rated PG-13 has estimated values around 0.3-0.4 for all models. Similarly, the mass for movies rated R is significantly above 0.5.

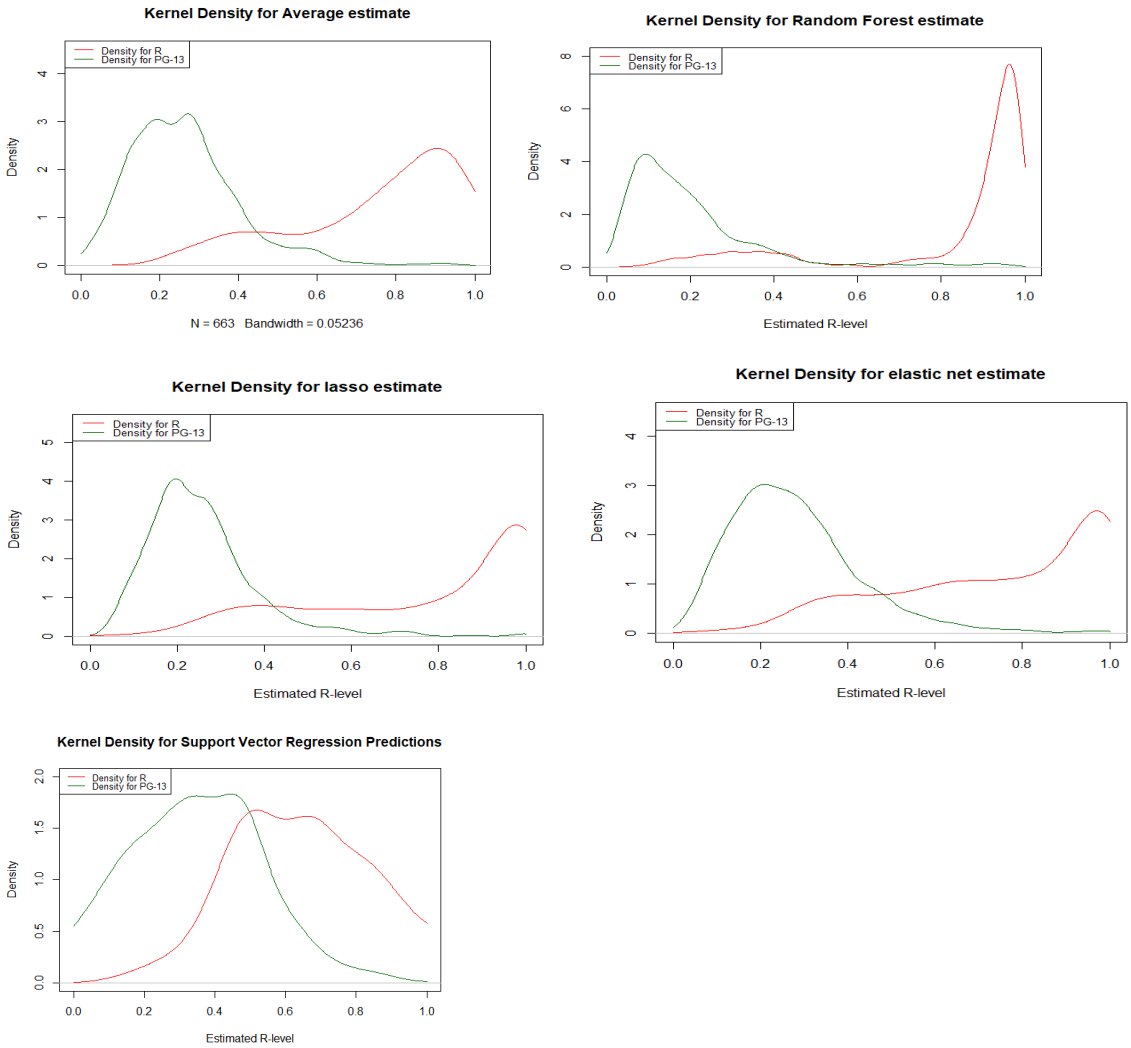


Figure T4: Density plots of estimated values from different models for PG-13 and R-rated movies

Next, we test the internal validity of the recovered estimates. The first robustness check compares the estimates from the different models. The intuition for this test is that, if we have recovered values close to the true underlying distribution of the R-levels, we would expect

different predictive models to give similar estimates. A naïve approach would be to simply compare predictions from the different models and test the correlation between the estimates. This could be misleading because movies with R-ratings will necessarily have higher ratings in all models. To circumvent this issue, we condition the estimates on the MPAA rating and make comparisons only within R and within PG-13 movies. We test the correlations within each MPAA category and find (see table T1) that the correlations are large and significant. These robustness checks do not conclusively prove that the recovered distribution is a good proxy for the underlying distribution. However, given that the models mostly agree, we can rule out that the estimates are due to the idiosyncrasy of one model.

PG-13/PG	average	lasso	elastic net	random forest	SVR
average	1	0.976146	0.969198	0.940351	0.848854
lasso	0.976146	1	0.977808	0.89681	0.764761
elastic net	0.969198	0.977808	1	0.853935	0.807084
random forest	0.940351	0.89681	0.853935	1	0.69567
SVR	0.848854	0.764761	0.807084	0.69567	1

R-rated	average	lasso	elastic net	random forest	SVR
average	1	0.97741	0.975776	0.94757	0.834182
lasso	0.97741	1	0.982702	0.909999	0.746028
elastic net	0.975776	0.982702	1	0.877007	0.800368
random forest	0.94757	0.909999	0.877007	1	0.686323
SVR	0.834182	0.746028	0.800368	0.686323	1

Table T1: Correlations between prediction models and weighted avg predictions

Our next test involves the use of expert age recommendations within MPAA rating categories to validate the estimates within the R and PG-13 group. In figure T5, we plot the relationship between the expert's age recommendation and the estimated value from the ensemble method, with a local polynomial regression line fitted to it for movies rated R and PG-13, and the full sample, respectively. Reassuringly we find a positive relationship within each group.

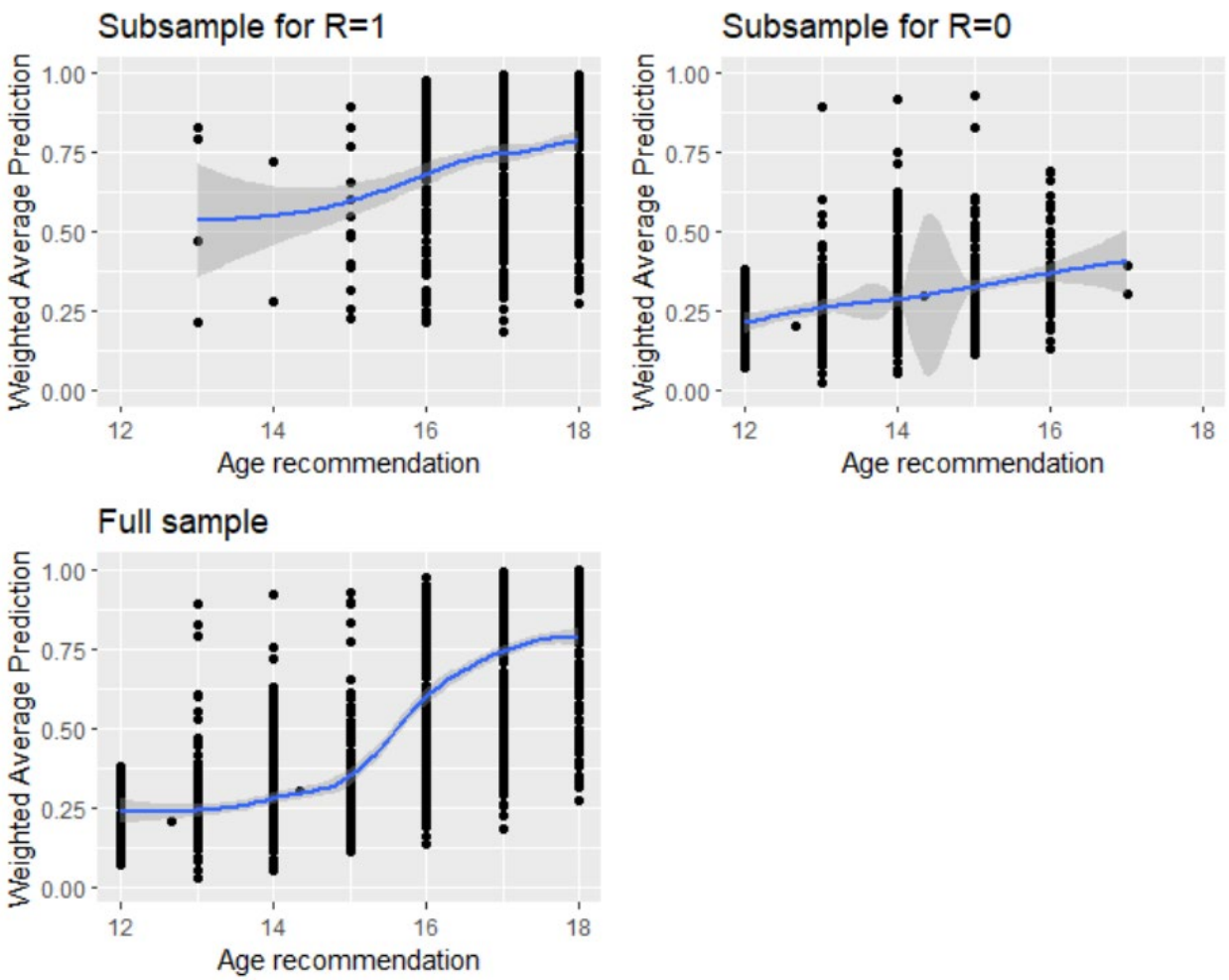


Figure T5: Comparing estimated values from weighted average prediction with age recommendations

Note: Using data from *Common Sense Media* for subsample of rated R movies, PG-13 movies, and full sample

To formally test the estimated values, we estimate the following ordinary linear squares model for each of the predictions (Table T2):

$$\widehat{P_i} = \beta_0 + \beta_1 \times R_i + \beta_2 \times age\ recommendation_i + \beta_3\ R_i \times age\ recommendation_i + e_i$$

Where $\widehat{P_i}$ is the predicted estimate of the R-level, $R_i \in \{0,1\}$ is an indicator variable for R rated movies, and $age\ recommendation \in (5,18)$ is the age recommendation given by Common Sense Media. The coefficients of interest are β_2 and β_3, which captures the relationship between the Common-Sense Media age recommendations and the predicted values. For movies rated PG-1, a positive β_2 indicates that movies with a higher age recommendation also have a higher predicted value. Reassuringly, the coefficients in all models are significantly ($p<0.01$) positive, between 0.017 and 0.041. Similarly, for movies rated R, a positive $(\beta_2 + \beta_3)$ indicates that movies with a higher age recommendation also have a higher predicted value. The coefficients for $(\beta_2 + \beta_3)$ range from 0.05 to 0.063.

	Dependent variable:				
	Weighted Average	**Lasso**	**Elastic Net**	**Random Forrest**	**Support Vector Regression**
R	0.295***	0.296***	0.231***	0.483***	0.114***
	(0.022)	(0.024)	(0.024)	(0.027)	(0.027)
Age recommendation	0.026***	0.021***	0.028***	0.017***	0.041***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
R × Age recommendation	0.036***	0.047***	0.044***	0.027***	0.024**
	(0.008)	(0.009)	(0.008)	(0.009)	(0.010)
Constant	0.301***	0.288***	0.322***	0.224***	0.398***
	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
Observations	1,356	1,356	1,356	1,356	1,356
R^2	0.720	0.673	0.652	0.716	0.500
Adjusted R^2	0.719	0.672	0.651	0.715	0.499
Residual Std. Error	0.159	0.181	0.176	0.196	0.202
F Statistic (df = 3; 1352)	1,158.484***	925.719***	842.794***	1,135.569***	450.353***

Note: CSrating is mean centered. Age recommendations are only available for 1,356 of the observations.

*p**p***p<0.01

Table T2: Validation of predictions

Next, we consider a different approach to matching based on the propensity score. In the main text, we matched each R rated movie to 3 PG rated movies. Alternatively, we consider matching each R-rated movie to either 1,5, or 10 PG movies with appropriate weights. Alternatively, we also use the inverse propensity weight and re-estimate the sample. We re-estimate the three main empirical models, given in equations (17), (18), and (19) using the different samples.

We also present analysis done separately for each genre below. We chose all genre-tags that are associated with over 10% of the movies in our sample. Each movie can have multiple genre tags, so we observe 9 genres that have a sufficiently high number of observations. We use of the inverse propensity weighting to construct each sample weights and present the analysis below.

Below, we also use a different approach to matching, based on the expert ratings provided by Common Sense Media. We use this approach as a robustness check for the main estimation. The estimation of the propensity score using the subtitle data is highly accurate, yet it is possible that there remain systematic biases in the estimated propensity. This could potentially

bias the estimates in the main models (i.e. equations 17, 18, and 19) if the potential bias is correlated with the error terms in those models. While the alternative estimation might have the same problem, the potential bias is likely different from the bias stemming from the machine learning process, because the data is generated by an expert panel that watched the movie. One potential issues are that experts might give better movies more lenient ratings. The number of movies that received a rating by Common Sense media is a significantly smaller than the set of movies for which we observed subtitles, so there could be selection in terms of which movies receive ratings (e.g., only movies with a larger potential audience will be rated). Despite these potential issues and the lower statistical power due to the significantly lower number of observations, the results are largely similar to the ones observed in the full sample, using the subtitle data. We present the results of this alternative estimation below in tables T9, T10, and T11.

Rated R propensity score

	Dependent variable:			
	R -rating			
	IPS	W-1	W-5	W-10
	(1)	(2)	(3)	(4)
Ad spending before release	-0.002	-0.014	-0.014	-0.000
	(0.009)	(0.018)	(0.014)	(0.018)
Budget				
	-0.014***	-0.011*	-0.013***	-0.013***
	(0.003)	(0.006)	(0.005)	(0.004)
Critics' count	0.018	-0.051*	-0.026	-0.019
	(0.012)	(0.028)	(0.021)	(0.017)
Critics' standard deviation	0.140***	0.277***	0.273***	0.250***
	(0.025)	(0.058)	(0.044)	(0.040)
Major studio	-0.585***	-0.046	-0.174	-0.160
	(0.165)	(0.391)	(0.299)	(0.245)
Foreign	0.291	0.487	0.624*	0.659
	(0.197)	(0.439)	(0.332)	(0.302)
Critics' mean	0.023***	0.015	0.023**	0.022**
	(0.006)	(0.013)	(0.010)	(0.009)
Constant	-1.939***	-0.055	-1.930	-1.148
	(0.660)	(1.803)	(1.280)	(1.120)
Year Fixed Effects	Yes	Yes	Yes	Yes
Genre Fixed Effects	Yes	Yes	Yes	Yes
Observations	705	564	642	705

Log Likelihood	-615.826	-137.377	-275.441	-1,234.721
Akaike Inf. Crit.	1,299.651	342.755	618.882	2,515.441

Note: *p **p ***p<0.01

Table T3: Reproducing Table 4 using alternative weighting

The table above reproduces Table 4, using four alternative matching and weighting mechanisms. (1) directly uses the propensity score and gives each movie a weight equal to the propensity of it's movie rating. In (2), (3), and (4), we consider alternative number of matches (i.e., 1,5, and 10). Each R-rated movie is matched to the closest n PG-13 movies, with respect to the propensity score. The weight for each matched PG-13 movie is given by $\frac{1}{\#matches}$.

Revenue Week 1 Model Propensity Score matched data

	Dependent variable:			
	Revenue			
	IPS	W-1	W-5	W-10
	(1)	(2)	(3)	(4)
R	-3.173***	-7.566***	-6.404***	-5.914***
	(1.012)	(2.332)	(1.522)	(1.370)
logBudget	0.561***	0.557**	0.468***	0.445***
	(0.116)	(0.253)	(0.161)	(0.139)
logAd spending before release	0.186***	-0.045	0.100	0.124
	(0.067)	(0.123)	(0.089)	(0.080)
Critics' count	0.030*	0.055	0.063**	0.057***
	(0.016)	(0.040)	(0.025)	(0.022)
Critics' standard deviation	-0.083**	-0.284***	-0.258***	-0.235***
	(0.039)	(0.103)	(0.065)	(0.058)
Critics'_mean	-0.036***	-0.087***	-0.076***	-0.069***
	(0.009)	(0.021)	(0.013)	(0.012)
Foreign	-0.235	0.748	1.015***	0.830***
	(0.309)	(0.568)	(0.372)	(0.320)
Major studio	0.525**	2.778***	2.232***	2.079***
	(0.232)	(0.469)	(0.310)	(0.276)
R × logBudget	0.325**	0.279	0.413**	0.447***

	(0.137)	(0.267)	(0.179)	(0.159)
R× logAd spending before release	0.055	0.297**	0.151	0.125
	(0.080)	(0.131)	(0.099)	(0.091)
R × Critics' count	-0.029	-0.048	-0.057**	-0.050**
	(0.020)	(0.042)	(0.027)	(0.025)
R × Critics' standard deviation	0.124***	0.311***	0.282***	0.256***
	(0.045)	(0.106)	(0.069)	(0.062)
R× Critics' mean	0.012	0.058***	0.046***	0.038***
	(0.011)	(0.022)	(0.015)	(0.014)
R × Foreign	-0.653*	-1.439**	-1.705***	-1.520***
	(0.359)	(0.595)	(0.412)	(0.366)
R × Major studio	0.041	-2.130***	-1.646***	-1.512***
	(0.302)	(0.506)	(0.363)	(0.335)
Constant	16.753***	21.519***	20.210***	19.724***
	(0.909)	(2.294)	(1.468)	(1.307)
Year Fixed Effects	Yes	Yes	Yes	Yes
Genre Fixed Effects	Yes	Yes	Yes	Yes
Observations	701	560	638	664
R^2	0.534	0.587	0.631	0.652
Adjusted R^2	0.505	0.555	0.606	0.629
Residual Std. Error	2.006 (df = 659)	1.585 (df = 518)	1.564 (df = 596)	1.574 (df = 622)
F Statistic	18.409*** (df = 41; 659)	17.972*** (df = 41; 518)	24.907*** (df = 41; 596)	28.415*** (df = 41; 622)

Note: Budget data missing for 77 movies & studio info is unavailable for 41 movies *p**p***p<0.01

Table T4: Reproducing Table 5 using alternative weighting

Note: The table above reproduces Table 5, using four alternative matching and weighting mechanisms. (1) directly uses the propensity score and gives each movie a weight equal to the propensity of it's movie rating. In (2), (3), and (4), we consider alternative number of matches (i.e., 1,5, and 10). Each R-rated movie is matched to the closest n PG-13 movies, with respect to the propensity score. The weight for each matched PG-13 movie is given by $\frac{1}{\#matches}$.

Week 2 Revenue

	Dependent variable:			
	Revenue (Week 2)			
Lag Revenue ×	(1)	(2)	(3)	(4)
Intercept	0.333***	0.537***	0.493***	0.418***
	(0.085)	(0.165)	(0.107)	(0.097)
Weekly advertising spending	0.055***	0.033***	0.008*	-0.001
	(0.004)	(0.005)	(0.004)	(0.004)
R	-0.239*	-0.471**	-0.491***	-0.449***
	(0.122)	(0.184)	(0.131)	(0.122)
Budget	-0.0005***	-0.0003*	-0.0002	0.00000
	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Consumer review mean	0.042*	-0.015	0.010	0.032
	(0.025)	(0.050)	(0.032)	(0.028)
Quality shock $\bar{q}$	0.094*	0.328***	0.235***	0.206***
	(0.053)	(0.097)	(0.064)	(0.057)
R × Consumer review mean	0.079**	0.149***	0.147***	0.133***
	(0.037)	(0.056)	(0.040)	(0.037)
R × Quality shock $\bar{q}$	-0.151**	-0.353***	-0.289***	-0.273***
	(0.075)	(0.106)	(0.078)	(0.072)
Year Fixed Effects	Yes	Yes	Yes	Yes
Genre Fixed Effects	Yes	Yes	Yes	Yes
Observations	2,806	2,235	2,555	2,657
R^2	0.789	0.794	0.833	0.849
Adjusted R^2	0.786	0.791	0.831	0.847
Residual Std. Error	6.300 (df = 2772)	4.444 (df = 2201)	4.431 (df = 2521)	4.468 (df = 2623)
F Statistic	304.005*** (df = 34; 2772)	249.917*** (df = 34; 2201)	370.619*** (df = 34; 2521)	433.847*** (df = 34; 2623)

Note: $\bar{q}_i = PR_i^U - PR_i^C$, where PR_i^U is the percentile rank of the consumer score for movie i, and PR_i^C is the percentile rank of the critics. Budget data is missing for 77 movies in the raw data. # of Reviews in 10,000s *p**p***p<0.01

Table T5: Reproducing Table 6 using alternative weighting

Note: The table above reproduces Table 6, using four alternative matching and weighting mechanisms. (1) directly uses the propensity score and gives each movie a weight equal to the propensity of it's movie rating. In (2), (3), and (4), we consider alternative number of matches (i.e., 1,5,and 10). Each R-rated movie is matched to the closest n PG-13 movies, with respect to the propensity score. The weight for each matched PG-13 movie is given by $\frac{1}{\#matches}$.

Rated R propensity score

	Dependent variable:								
	R -rating								
	Drama	Thriller	Comedy	Crime	Action	Romance	Mystery	Horror	SciFi
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Ad spending before release	-0.013	0.022	0.004	0.020	0.026	-0.035*	0.026	0.088*	0.028
	(0.012)	(0.016)	(0.020)	(0.026)	(0.023)	(0.021)	(0.026)	(0.046)	(0.046)
Budget									
	-0.003	-0.018***	-0.060***	-0.024**	-0.016***	-0.005	-0.013*	-0.022**	-0.022**
	(0.003)	(0.004)	(0.008)	(0.010)	(0.005)	(0.007)	(0.007)	(0.009)	(0.009)
Critics' count	0.022	0.028	0.047*	-0.013	0.035	-0.018	0.009	0.060	0.089*
	(0.018)	(0.018)	(0.026)	(0.030)	(0.027)	(0.029)	(0.029)	(0.037)	(0.050)
Critics' standard deviation	0.074**	0.201***	0.151***	0.290***	0.327***	0.114*	0.138*	0.229***	0.181*
	(0.034)	(0.044)	(0.047)	(0.075)	(0.067)	(0.063)	(0.072)	(0.082)	(0.096)
Major studio	-0.483**	-0.709***	0.114	-0.465	-1.907***	-0.165	-0.483	2.771***	-1.077
	(0.230)	(0.268)	(0.324)	(0.454)	(0.418)	(0.377)	(0.468)	(0.890)	(0.666)
Foreign	0.169	0.267	0.409	0.319	0.739	-0.139	0.960	1.610***	2.898***
	(0.241)	(0.327)	(0.399)	(0.553)	(0.549)	(0.377)	(0.596)	(0.611)	(1.072)
Critics' mean	0.026***	0.022**	0.018	0.097***	-0.015	0.048***	0.024	-0.037*	-0.004
	(0.009)	(0.010)	(0.012)	(0.019)	(0.015)	(0.015)	(0.017)	(0.021)	(0.022)
Constant	-2.111**	-0.901	-2.242	11.002	14.524	-3.839**	-0.939	15.994	-3.824
	(0.846)	(1.169)	(1.391)	(1,515.989)	(743.610)	(1.680)	(1.626)	(2,662.648)	(2.339)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	405	288	234	162	155	145	112	104	74
Log Likelihood	-354.445	-259.391	-166.792	-101.692	-125.953	-134.354	-96.762	-70.631	-47.156
Akaike Inf. Crit.	754.891	564.782	379.584	249.383	297.905	314.708	239.525	187.263	140.313

Note: *p**p***p<0.01

Table T6: Table 4 Reproduction by Genre

The table above reproduces Table 4 separately for multiple genres. The weights for each movie are given by the inverse propensity score and the sample for each genre includes all movies that include that genre tag. Each movie can belong to multiple genres and we select all genres that are attached to more than 10% of our sample in the analysis.

Revenue Week 1 Model Propensity Score matched data

	Dependent variable:								
					Revenue				
	Drama	Thriller	Comedy	Crime	Action	Romance	Mystery	Horror	SciFi
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
R	-5.438***	-2.695*	-1.037	-3.046	0.124	-3.249	-2.901	-3.709	0.754
	(1.721)	(1.448)	(1.832)	(2.405)	(1.727)	(3.564)	(3.095)	(2.315)	(2.728)
logBudget	0.396**	0.227	0.497**	0.748	1.206***	0.726**	0.693*	-0.157	0.083
	(0.162)	(0.201)	(0.203)	(0.718)	(0.216)	(0.312)	(0.388)	(0.354)	(0.378)
logAd spending before release	0.088	0.026	0.276***	0.246	0.093	0.177	0.229	0.430**	0.456
	(0.094)	(0.116)	(0.095)	(0.210)	(0.131)	(0.119)	(0.250)	(0.178)	(0.318)
Critics' count	0.010	0.015	0.061**	0.007	-0.033	0.058	0.006	0.146***	-0.001
	(0.028)	(0.021)	(0.029)	(0.044)	(0.024)	(0.039)	(0.042)	(0.036)	(0.038)
Critics'_standard deviation	-0.083	-0.049	-0.053	-0.095	-0.004	-0.329***	-0.167	-0.295***	0.040
	(0.067)	(0.059)	(0.061)	(0.078)	(0.060)	(0.108)	(0.118)	(0.109)	(0.101)
Major studio	0.800**	0.455	0.457	-0.110	0.423	0.495	0.139	-1.489	0.496
	(0.383)	(0.341)	(0.392)	(0.524)	(0.314)	(0.558)	(0.609)	(1.112)	(0.698)
Foreign	-0.041	-0.714	-0.324	-0.840	0.518	-0.230	-0.524	-0.027	-0.616
	(0.435)	(0.483)	(0.581)	(0.671)	(0.544)	(0.667)	(0.939)	(0.921)	(1.808)
Critics' mean	-0.059***	-0.002	-0.032**	-0.011	0.009	-0.053**	-0.026	-0.030	0.003
	(0.016)	(0.014)	(0.014)	(0.026)	(0.015)	(0.023)	(0.026)	(0.024)	(0.026)
R × logBudget	0.584***	0.894***	0.433*	0.351	-0.328	-0.356	-0.003	0.715*	1.378***
	(0.193)	(0.243)	(0.258)	(0.733)	(0.311)	(0.402)	(0.482)	(0.400)	(0.494)
R × logAd spending before release	0.195*	0.211	-0.070	-0.080	-0.052	0.061	0.295	-0.176	-0.395
	(0.117)	(0.129)	(0.128)	(0.226)	(0.151)	(0.213)	(0.287)	(0.198)	(0.345)
R × Critics' count	-0.004	-0.032	-0.102***	-0.030	0.043	-0.021	-0.043	-0.177***	-0.050

	(0.031)	(0.026)	(0.036)	(0.049)	(0.032)	(0.057)	(0.049)	(0.044)	(0.048)
R × Critics'_standard deviation	0.102	0.047	0.110	0.139	0.005	0.317**	0.175	0.296**	-0.165
	(0.074)	(0.067)	(0.074)	(0.086)	(0.073)	(0.131)	(0.136)	(0.123)	(0.128)
R ×Major studio	-0.317	-0.428	0.348	0.357	-0.539	0.443	0.439	2.076*	-0.454
	(0.468)	(0.435)	(0.536)	(0.658)	(0.466)	(0.794)	(0.763)	(1.207)	(0.922)
R ×Foreign	-0.880*	-0.057	-0.335	-0.447	-1.357**	-1.086	-0.008	-1.281	-0.307
	(0.501)	(0.560)	(0.701)	(0.765)	(0.630)	(0.833)	(1.017)	(1.021)	(1.867)
R ×Critics'_mean	0.027	-0.016	0.007	0.001	0.005	-0.009	0.011	0.037	-0.016
	(0.018)	(0.017)	(0.018)	(0.028)	(0.020)	(0.029)	(0.032)	(0.029)	(0.033)
Constant	18.823***	16.403***	14.350***	15.489***	12.159***	19.814***	18.230***	18.733***	15.696***
	(1.556)	(1.222)	(1.517)	(2.145)	(1.347)	(2.902)	(2.697)	(1.991)	(1.932)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	405	288	234	162	155	145	112	104	74
R^2	0.518	0.450	0.568	0.606	0.572	0.546	0.512	0.652	0.631
Adjusted R^2	0.479	0.386	0.504	0.516	0.468	0.427	0.331	0.509	0.374
Residual Std. Error	2.176 (df = 374)	1.815 (df = 257)	1.972 (df = 203)	1.832 (df = 131)	1.414 (df = 124)	2.306 (df = 114)	1.926 (df = 81)	1.666 (df = 73)	1.633 (df = 43)
F Statistic	13.392*** (df = 30; 374)	7.013*** (df = 30; 257)	8.883*** (df = 30; 203)	6.724*** (df = 30; 131)	5.523*** (df = 30; 124)	4.578*** (df = 30; 114)	2.827*** (df = 30; 81)	4.556*** (df = 30; 73)	2.452*** (df = 30; 43)
Note:									*p**p***p<0.01

Table T7: Reproduced table 5 by genre

The table above reproduces Table 5 separately for multiple genres. The weights for each movie are given by the inverse propensity score and the sample for each genre includes all movies that include that genre tag. Each movie can belong to multiple genres and we select all genres that are attached to more than 10% of our sample in the analysis.

Week 2 Revenue

	Dependent variable:								
	Revenue (Week 2)								
Lag Revenue ×	Drama (1)	Thriller (2)	Comedy (3)	Crime (4)	Action (5)	Romance (6)	Mystery (7)	Horror (8)	SciFi (9)
Intercept	-0.165	0.368***	0.097	0.167	0.084	0.883	0.348**	0.420***	0.762
	(0.165)	(0.103)	(0.138)	(0.204)	(0.223)	(1.131)	(0.141)	(0.161)	(0.671)
Weekly advertising spending	0.061***	0.018***	0.068***	0.062***	0.085***	0.039***	0.001	0.015**	0.208***
	(0.006)	(0.004)	(0.009)	(0.007)	(0.011)	(0.008)	(0.005)	(0.008)	(0.025)
R	0.107	-0.257*	0.078	-0.155	0.234	-0.346	-0.338**	-0.449**	0.134
	(0.199)	(0.139)	(0.207)	(0.237)	(0.339)	(0.307)	(0.143)	(0.182)	(0.954)
Budget	0.0003	-0.001***	-0.00002	-0.001*	-0.001***	0.0002	-0.001***	-0.0001	-0.003***
	(0.0002)	(0.0002)	(0.0004)	(0.0004)	(0.0003)	(0.001)	(0.0003)	(0.0002)	(0.001)
Consumer review mean	0.186***	0.062*	0.136***	0.096	0.126*	0.099	0.051	0.014	-0.022
	(0.049)	(0.034)	(0.035)	(0.064)	(0.068)	(0.074)	(0.038)	(0.050)	(0.206)
Quality shock $\bar{q}$	0.341***	-0.003	0.006	0.043	-0.043	0.333***	-0.073	-0.092	0.403
	(0.071)	(0.065)	(0.065)	(0.095)	(0.175)	(0.109)	(0.085)	(0.086)	(0.371)
R × Consumer review mean	-0.041	0.073*	-0.046	0.046	-0.076	0.120	0.087**	0.132**	-0.038
	(0.060)	(0.043)	(0.066)	(0.075)	(0.104)	(0.096)	(0.043)	(0.059)	(0.294)
R × Quality shock $\bar{q}$	-0.500***	-0.030	0.075	-0.023	-0.007	-0.633***	0.233*	-0.127	-0.707
	(0.093)	(0.091)	(0.119)	(0.121)	(0.232)	(0.167)	(0.121)	(0.107)	(0.429)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,617	1,152	938	645	622	579	442	412	288
R^2	0.806	0.892	0.809	0.925	0.776	0.854	0.960	0.947	0.720
Adjusted R^2	0.803	0.890	0.804	0.922	0.768	0.848	0.958	0.944	0.696
Residual Std. Error	4.933 (df = 1594)	4.195 (df = 1129)	5.025 (df = 915)	3.039 (df = 622)	10.020 (df = 599)	4.124 (df = 556)	2.634 (df = 419)	2.231 (df = 389)	11.344 (df = 265)
F Statistic	288.229*** (df = 23; 1594)	406.113*** (df = 23; 1129)	168.364*** (df = 23; 915)	332.727*** (df = 23; 622)	90.390*** (df = 23; 599)	141.352*** (df = 23; 556)	440.405*** (df = 23; 419)	303.169*** (df = 23; 389)	29.690*** (df = 23; 265)

Note: *p**p***p<0.01

Table T8: Reproduced table 6 by genre

The table above reproduces Table 6 separately for multiple genres. The weights for each movie are given by the inverse propensity score and the sample for each genre includes all movies that include that genre tag. Each movie can belong to multiple genres and we select all genres that are attached to more than 10% of our sample in the analysis.

Rated R propensity score

	Dependent variable: R rating	
	(1)	(2)
Ad spending before release	-0.00004**	-0.0001***
	(0.00002)	(0.00003)
Budget	-0.005	0.013
	(0.005)	(0.008)
Critics' count	0.012	0.005
	(0.024)	(0.039)
Critics' standard deviation	0.215***	0.198**
	(0.060)	(0.080)
Major studio	0.223	0.229
	(0.407)	(0.559)
Foreign	-0.409	-0.059
	(0.378)	(0.574)
Critics' mean	0.043***	0.014
	(0.013)	(0.021)
Constant	-5.032***	3.123
	(1.170)	(5.038)
Year Fixed Effects	Yes	Yes
Genre Fixed Effects	Yes	Yes
Observations	249	249
Log Likelihood	-127.024	-74.261
Akaike Inf. Crit.	270.048	236.522
Note:		*p**p***p<0.01

Table T9: Logistic Regression of R-rating on movie characteristics

Note: This analysis uses a propensity-score matched sample of observations within common support. The propensity score is estimated using subtitle data and age recommendations from *Common Sense Media*.

Revenue Week 1 Model Propensity Score matched data

	Dependent variable: Log Revenue			
	(1)	(2)	(3)	(4)
R	-3.760*	-4.628**	-2.373	-1.301
	(2.137)	(2.137)	(2.487)	(2.774)
logBudget	0.161	0.190	0.379*	0.238
	(0.186)	(0.192)	(0.218)	(0.252)
logAd spending before release	0.340***	0.245**	0.360***	0.343**
	(0.115)	(0.118)	(0.129)	(0.156)
Critics' count	0.073**	0.050	0.056	0.071*
	(0.035)	(0.035)	(0.038)	(0.042)
Critics'_standard deviation	-0.260***	-0.255***	-0.301***	-0.271***
	(0.075)	(0.074)	(0.082)	(0.092)
Critics'_mean	-0.056***	-0.053***	-0.042**	-0.030
	(0.018)	(0.018)	(0.019)	(0.022)
Major Studio		1.322***	1.536***	1.706***
		(0.505)	(0.526)	(0.589)
Foreign		0.487	0.699	0.988*
		(0.397)	(0.460)	(0.526)
R × logBudget	0.609**	0.603**	0.424	0.571*
	(0.241)	(0.246)	(0.274)	(0.329)
R× logAd spending before release	0.154	0.214	0.042	0.020
	(0.137)	(0.142)	(0.158)	(0.190)
R × Critics' count	-0.121***	-0.099**	-0.092**	-0.103**
	(0.042)	(0.042)	(0.044)	(0.050)
R × Critics' standard deviation	0.114	0.120	0.163*	0.128
R× Critics' mean	0.042*	0.040*	0.023	0.017
	(0.021)	(0.021)	(0.023)	(0.026)
R × Major studio		-1.217**	-1.436**	-1.589**
		(0.604)	(0.624)	(0.719)
R × Foreign		-0.967*	-1.257**	-1.829**
		(0.549)	(0.624)	(0.720)

Constant	17.372***	18.257***	16.366***	15.816***
	(1.781)	(1.792)	(2.228)	(2.545)
Year Fixed Effects	No	No	No	No
Genre Fixed Effects	No	No	No	No
Observations	175	175	175	175
R^2	0.607	0.629	0.673	0.711
Adjusted R^2	0.581	0.593	0.605	0.598
Residual Std. Error	1.342 (df = 163)	1.322 (df = 159)	1.303 (df = 144)	1.314 (df = 125)
F Statistic	22.914*** (df = 11; 163)	17.933*** (df = 15; 159)	9.873*** (df = 30; 144)	6.283*** (df = 49; 125)
Note:				*p**p***p<0.01

Table T10: Regression of opening week revenue on movie characteristics

Note: This analysis uses a propensity-score matched sample of observations within common support. The propensity score is estimated using subtitle data and age recommendations from *Common Sense Media*.

Week 2 Revenue Propensity Score matched data

	Dependent variable: Revenue (Week 2-5)			
Lag Revenue ×	(1)	(2)	(3)	(4)
Intercept	0.466***	0.471***	-0.350	-0.225
	(0.054)	(0.114)	(0.309)	(0.318)
Weekly advertising spending		0.00002*	0.00001	0.00001
		(0.00001)	(0.00001)	(0.00001)
R	0.004	0.069	0.776*	0.660
	(0.025)	(0.114)	(0.417)	(0.422)
Budget			-0.00004	-0.0003
			(0.0004)	(0.0005)
Consumer review mean			0.221***	0.173**
			(0.073)	(0.078)
Quality shock $\bar{q}$		-0.161	0.077	0.090
		(0.132)	(0.174)	(0.175)
# of Consumer Reviews				0.147*
				(0.089)
R × Consumer review mean			-0.186*	-0.144
			(0.103)	(0.108)

R × Quality shock $\bar{q}$		-0.059	-0.265	-0.251
		(0.164)	(0.214)	(0.218)
R × # of Consumer Reviews				-0.132
				(0.094)
Year Fixed Effects	Yes	Yes	Yes	Yes
Genre Fixed Effects	Yes	Yes	Yes	Yes
Observations	1,031	1,031	979	979
R^2	0.768	0.778	0.779	0.779
Adjusted R^2	0.764	0.769	0.768	0.769
Residual Std. Error	4.544 (df = 1014)	4.500 (df = 990)	4.598 (df = 935)	4.596 (df = 933)
F Statistic	197.519*** (df = 17; 1014)	84.595*** (df = 41; 990)	74.814*** (df = 44; 935)	71.687*** (df = 46; 933)

Note: $\bar{q}_i = PR_i^U - PR_i^C$, where PR_i^U is the percentile rank of the consumer score for movie i, and PR_i^C is the percentile rank of the critics. Budget data is missing for 77 movies in the raw data. # of Reviews in 10,000s

*p **p ***p<0.01

Table T11: Regression of revenue in weeks 2-5 on movie characteristics

Note: This analysis uses a propensity-score matched sample of observations within common support. The propensity score is estimated using subtitle data and age recommendations from *Common Sense Media*.

T5: Discussion of alternative explanations

Overall, our empirical analysis found the patterns in the data that are consistent with the proposed model. In this section, we discuss some alternative explanations of these results. Any alternative theory needs to be consistent with the following stylized empirical facts:

1. Conditional on "inappropriateness," a movie is more likely to receive an R-rating if it has fewer, more spread out, and higher ratings, or is produced by a foreign studio.
2. In the opening week, revenues are higher for R-rated movies that have less (critic) information available, are not produced by large studios, and movies that have a lower budget.
3. In the following weeks, a quality shock has a smaller impact on rated R movies. The effect of consumer rating on movie revenues is larger for rated R movies

First, we consider if MPAA systematically discriminates against specific movies. Waguespack and Sorenson (2011) find that the MPAA is more likely to give more restrictive classifications to movies produced by smaller and less powerful studios. While this finding is consistent with our proposed signaling explanation, we might be simply providing evidence of this bias emanating from the MPAA. In other words, it might be that the same movies we hypothesize as using the

R-rating as a signal are simply being discriminated against. However, if this were the case, then the results on revenue that we hypothesize and find support for are harder to rationalize. If the bias is systematically against certain types of movies, then based on the observable characteristics, we would expect R-rated movies to perform similar to PG-13 movies at the box office, albeit with a lower intercept because of the excluded segment. If the MPAA uses unobservable movie characteristics, such as high-quality, in their decision, we would not be able to parse this effect from signaling. Particularly, if the MPAA has a bias towards giving movies of higher quality more restrictive age restrictions, we would not be able to falsify this alternative theory. On the other hand, if the MPAA follows a rule where movies of higher quality receive a less restrictive rating[2], we would expect the results from the revenue equation to be opposite to what we observe. Overall, we are comfortable in saying that bias from MPAA ratings generally does not invalidate our results, unless the bias in MPAA ratings is in the same direction as the strategic signaling.

We have largely considered quality on a vertical scale. However, under asymmetric information, consumers need to make an inference about fit as well. R-rated movies are, in general, more realistic, violent, and graphic. A consumer who prefers these attributes will correctly use the fact that R-rated movies are more consistent with her taste. Our econometric approach uses subtitle data to compare movies that are equally likely to receive an R-rating or PG-13 rating. However, consumers in the opening week do not know the actual level of "inappropriateness" and form beliefs based on all available information, including the MPAA rating. It could be hypothesized that the differences in beliefs about horizontal characteristics are driving the results. While this seems intuitive, it is not fully consistent with our findings.

Why is potential horizontal differentiation induced by R-rating and PG-13 unlikely to explain our data? Let us suppose that R-rated movies target a different set of consumers. In the current Model, we assume that the two segments have different vertical preferences (i.e., different opportunity costs to consumption) while abstracting away from horizontal preferences. In the empirical application, we select movies that are similar in terms of "inappropriateness" and exclude extreme movies (i.e., movies with a very high or very low measure of the R-score). However, as pointed out by you, consumers might still make inferences about the underlying appropriateness of R-rated and PG movies. This would be expected because, on average, movies rated R are, by definition, higher on the latent R-score.
However, note that R-rated movies we select for empirical analysis (viz., movies close to the cutoff) are, on average, *less* age-inappropriate than a moviegoer might expect. Similarly, PG-13 movies close to the cutoff are, on average, *more* age-inappropriate than a typical PG-13 movie. To map these features onto a model, we briefly outline a very simple model of horizontal differences.
Imagine two consumer segments wherein consumers in Segment 1 have a horizontal preference for more violent movies[3], while consumers in Segment 2 prefer movies that are less

[2] This seems more intuitive if the MPAA follows some kind of quality vs. rating trade-off rule. For a movie at the margin between PG-13 and R, higher quality might break the indifference because there is a larger benefit, outweighing the cost of "inappropriateness".
[3] We are using "violence" as a shorthand for any inappropriate content.

violent. To represent this, suppose their utility functions are given by $u^1(v)$, and $u^2(v)$, where $\frac{du^1(v)}{dv} > 0$ and $\frac{du^2(v)}{dv} < 0$. Suppose movies are distributed on some distribution of violence $v \sim F(v)$, with a lower bound $F(\underline{v}) = 0$ and an upper bound $F(\bar{v}) = 1$. Movies with inappropriateness above some v^* receive a rating of R, whereas movies below receive a rating of PG-13. Without additional information, the expected level of violence in movies rated R is given by $E[v|R] = \int_{v^*}^{\bar{v}} v\, f(v)\, dv$ and the expected level of violence in movies rated PG is given by $E[v|PG] = \int_{\underline{v}}^{v^*} v\, f(v)\, dv$. Because consumers in Segment 1 prefer more violent movies, their expected utility from a product rated R is higher than for a product rated PG-13 ($u^1(E[v|R]) > u^1(E[v|PG])$. Similarly, consumers in segment 2 prefer movies rated PG because the expected violence is lower ($u^2(E[v|R]) < u^2(E[v|PG])$).

This simple horizontal setup can explain a number of results. First, because consumers in Segment 1 make inferences about violence from the age rating, a movie rated R will receive higher demand (from this segment). Suppose also that, in turn, the critics' (or consumer) rating of quality is their experienced utility, based on the horizontal differentiation. This horizontal setup could then potentially explain differences in perceived quality of movies between movies rated R and PG, depending on the horizontal fit.

Empirically, such a horizontal differentiation argument would also imply that (ceteris paribus) more violent R-rated movies give consumers higher levels of utility ($\frac{dg^1(v)}{dv} > 0$). While we are unable to test this relationship causally, we can use the estimated (latent) v in our empirical application to run the following regression:

$$q_i = \beta_0 + \beta_1 R_i + \beta_2\, v_i + \beta_3 R_i \times v_i + \epsilon_i \qquad \text{(R.1)}$$

In the regression, q_i is the mean of consumer reviews, R_i is the binary indicator for a movie being rated R, and v_i is the predicted latent R-score. Because only consumers in Segment 1 consume R-rated movies, this allows for a direct test of $\frac{dg^1(v)}{dv} > 0$, which would imply a positive coefficient for β_3. It is difficult to interpret β_2 because these reviews potentially come from consumers in both segments. We present the results from this regression in column one of Table T12 and find that β_3 is negative and statistically significant- this goes against the prediction from the above horizontal Model.

We re-estimate the Model with the mean of the critic's review. Here (column 2 in Table T12), we again find that the coefficients of β_2 and β_3 are not significantly positive, which implies that there is no positive correlation between movie's level of inappropriateness and critic's reviews. Thus, none of the main predictions about different quality, the return to alternative signals, or the returns to 3rd party information can be rationalized in this very simple setup without appealing to some additional factors, such as difference in the consumers' sensitivity to quality (or variation in opportunity cost) between the two segments.

A similar alternative to the vertical quality could be that watching R-rated movies is “cool” for some people, because of some peer effects. Because the signal affects the perception of the movie, the rating could be a product attribute that alters the quality of the movie instead of signaling quality. Strategic movie studios would use the R-rating whenever the return to being “cool” exceeds the loss from locking out one segment. During the opening weekend, the return to being a “cool” movie is likely largest when the other attributes are unknown. So far, the

predictions from this alternative theory overlap perfectly with the proposed theory. However, with this alternative theory the R-rating does not convey any information but merely serves as another –always known – attribute. In the weeks after opening, this theory does not predict the effect of the quality shock to differ between rated R and PG-13 movies, as predicted by the information uncertainty reducing signaling theory. Furthermore, as quality gets revealed, "coolness" becomes a less prominent attribute and high-quality movies should see lower returns to being "cool" than low quality movies[4]. While we cannot falsify this alternative theory completely, it cannot account for all the observed empirical facts and does not invalidate the proposed signaling theory.

	Dependent variable:	
	Consumer Review Mean	*Critics Review Mean*
	(1)	(2)
R	0.390***	11.563***
	(0.095)	(3.188)
v	0.081	-3.175
	(0.128)	(4.302)
$R \times v$	-0.317*	-5.689
	(0.168)	(5.656)
Constant	3.092***	55.462***
	(0.032)	(1.081)
Observations	1,325	1,325
R^2	0.043	0.033
Adjusted R^2	0.041	0.030
Residual Std. Error	0.447 (df = 1321)	15.001 (df = 1321)
F Statistic	19.817*** (df = 3; 1321)	14.871*** (df = 3; 1321
Note:		*p<0.1, **p<0.05, p***<0.01

Table T12: Regression of quality proxies on R-rating and level of inappropriateness

[4] Because of diminishing returns to quality, the effect of the "coolness" from the R-rating should diminish for higher quality movies.

Note: R is the binary indicator for a movie being rated R, and v is the predicted latent R-score. The regression estimates the effect of v on the average consumer and critic review, to identify horizontal preferences for higher or lower v within PG-13 or R rated movies. Movies rated R receive higher ratings, but increases in v have no effect in the PG-13 segment and a (marginally significant) negative effect on movies rated R.

T6: Omitted details of estimation of propensity score

To reduce the dimensionality of the text, we apply several transformations to the text corpus. First, we transform all the text to lowercase letters, remove whitespace, remove punctuation, and remove numbers. Next, we remove stop words using a predefined list of 174 words (Lewis et al. 2004). These words include articles ("the," "an"), conjunctions ("and," "or"), short function words ("over," "and," "your"), and other short, uninformative words. These words are important for context, but they convey little meaning in isolation. Similarly, rare words are likely to have little diagnostic value. So, we limit the dataset to words that appear in at least 15% of the documents.[5] The final step involves the stemming of the words (Porter 1980). A standard algorithm replaces words with their roots; for example, the words "argue," "argued," "argues," and "arguing" are all reduced to the stem "argu." Each of these steps lowers the computational burden and facilitates interpretability. The raw data has 543,641 unique terms. After removing whitespace, punctuation, numbers, and stop words, we have 170,671 unique terms. Stemming reduces the number of unique terms to 127,836. Finally, excluding elements that appear in fewer than 15% of the documents leads to 1,715 unique terms.

Although we could, in principle, use any number of n-grams, we use a simple "bag of words" (n=1) to represent the subtitle documents for two reasons. First, the bag of words is simple enough to give the model interpretability and to capture our measure of "inappropriateness" quite well. Second, the dimensionality of the representation increases exponentially in the order of n of the phrases tracked, increasing computational complexity without any significant gains (Gentzkow et al., 2019). However, many words, such as the f-word, are highly informative in the bag of words representation. Given that our proposed models using n=1 are sufficiently accurate, the interpretability gained, and the computational simplicity from not using n-grams of order $n > 1$ outweighs any loss.

Finally, we transform the document from raw frequencies to "term frequency-inverse document frequencies." For a word j in document i, term frequency $tf(D, i, j)$ is the count c_{ij} of occurrences of j in i. The inverse document frequency is defined as:

[5] Rare words typically contain the names of the cast members and directors.

$$idf(D,j) = \log_2 \frac{|D|}{|\{d|t_j \in d\}|}, \quad (11)$$

where, D is the total count of documents, and $|\{d \mid t_j \in d\}$ is the count of documents in which term j appeared. The term frequency-inverse document frequency is given by:

$$tfidf(D,i,j) = tf_{ij} \times idf_j \quad (12)$$

and is calculated for every term in the matrix.

Next, we describe the four models used in more detail, first outlining the estimation strategy for each model. To avoid overfitting, we follow cross-validation, break the data into five random, disjoint subsets, and estimate the model, excluding one subset. After estimating the model five times, each time excluding a different subset, we have an estimate of the probability of a movie being rated R.

First, a natural starting point to predict the R rating is logistic regression. Following Tibshirani (1996), we use a lasso logistic regression, where the coefficient is defined to be the solution to: $\hat{\beta} = \operatorname{argmin}\left\{ l(\beta) + \lambda \sum_{j=1}^{\rho} |\beta_j| \right\}$, (13)

where, $l(\beta)$ is the standard logistic term and $\lambda \sum_{j=1}^{\rho} |\beta_j|$ is the penalty term.

Because λ cannot be optimally defined a priori (Gentzkow et al., 2019), we use 10-fold cross-validation and select the largest value of λ with the mean error of no more than one standard error away from the minimum. This approach leads to a slightly larger λ than the approach of minimizing the mean error and allows for more shrinkage and a simpler model. The cross-validation splits the sample into ten disjoint subsets and then fits the full regularization path each time, excluding each subset in turn. As previously described, we estimate the model five times for each subset and estimate a separate value for λ for each model. In Table 3a, we list the words with the largest coefficient of $\hat{\beta}$; many of these words indicated sexual content, violence, or inappropriate language[6].

Second, we estimate a logistic regression with an elastic net penalty term. This model is similar to the lasso, but the penalty term is more flexible:

$$\hat{\beta} = \operatorname{argmin}\left\{ l(\beta) + \lambda_2 \sum_{j=1}^{\rho} \beta_j^2 + \lambda_1 \sum_{j=1}^{\rho} |\beta_j| \right\}. \quad (14)$$

[6] For example, the word *bed* has a high coefficient. This seems to indicate that visual cues that make movies more likely to receive an R-rating will also be represented in the words spoken.

We again use 10-fold cross-validation to select the largest value of λ that has a mean error of no more than one standard error away from the minimum. We find the words with the highest $\hat{\beta}$ display a considerable overlap with the words from the lasso selection model (see Table 3b). We pick the values for λ_1 and λ_2 so that they sum up to one and give the highest prediction accuracy while not being too similar to the lasso model.

The next model we estimate is a random forest (Breiman 2001). This method is an efficient algorithm for high-dimensional classification problems and does not rely on the functional form assumed in the logistic regression. The principle of random forest is to grow a large number of regression trees from independent subsets of variables. For each tree, randomness is induced when selecting the variables on which to split. We use 1,000 trees for each random forest model. The number of variables randomly sampled at each split is set to be $\frac{J}{3}$, where J is the number of variables in the model. We use the probability of trees that classified an observation as R to obtain a probability estimate from the random forest classification. In Fig. T6, we plot the mean node purity increase by splits on word j, as measured by the decrease in the sum of squares. We find that most words make intuitive sense and capture inappropriateness

The final model we estimate is a support vector regression (SVR) (Joachims 1998). SVR is appealing when working with text data for several reasons. First, its ability to learn is independent of the dimensionality of the feature space, which makes it appealing for high-dimensional data. Second, the SVR is not as aggressive in eliminating covariates with low informational count—particularly compared with the logistic regression. Third, unlike the other models, the SVR does not impose that the estimated probabilities must lie between 0 and 1. We find that 3.3% of the observations have values below zero, and 4.5% are above one. Although truncating the values at zero and one would not change the accuracy, we keep these values to preserve the contained information about the rank order of movies. To estimate the SVR, we use a Radial basis function kernel.

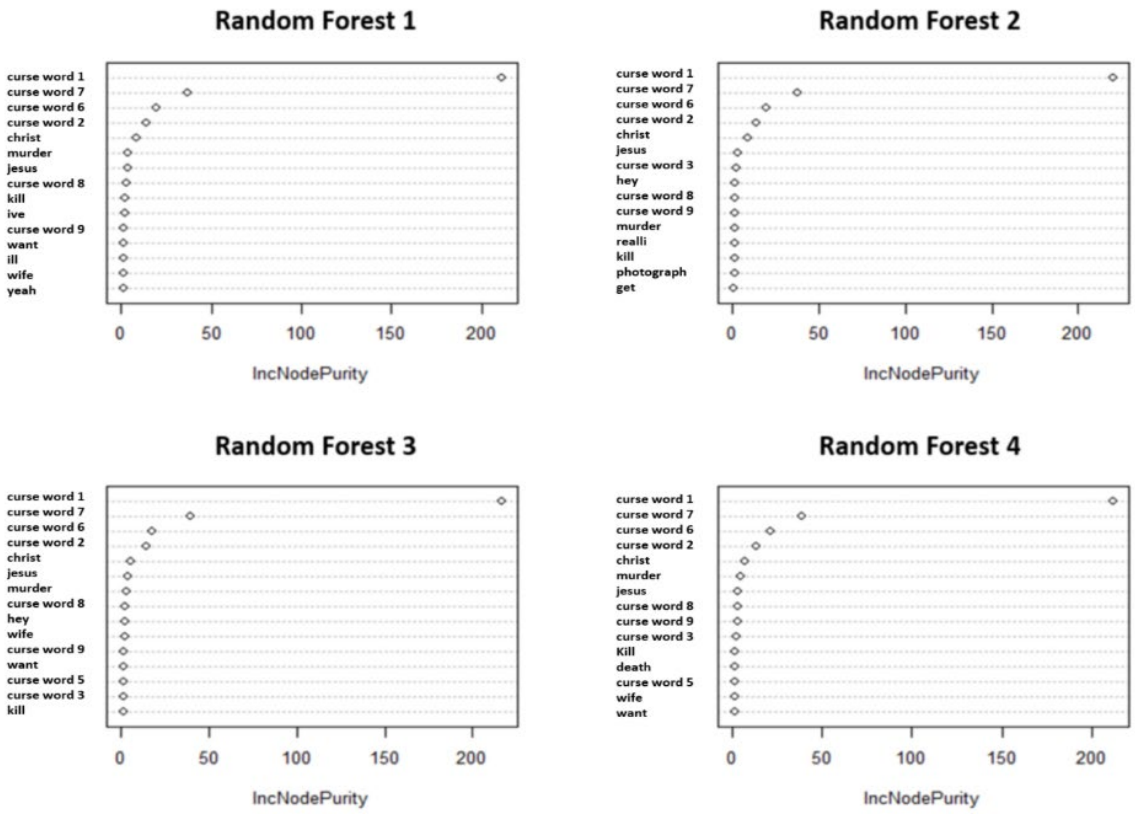
Tables T13A and T13B: Words with the highest coefficient in the Lasso Model (A) and the Elastic Net model (B)

Lasso	Model 1	Model 2	Model 3	Model 4	Model 5
1	want	want	want	want	want
2	curse word 1	curse word 1	sure	curse word 1	curse word 1

3	whi	christ	curse word 1	like	ask
4	christ	curse word 3	christ	whi	christ
5	kill	kill	peopl	christ	kill
6	ill	bed	bodi	kill	peopl
7	god	photograph	coupl	address	god
8	death	death	curse word 3	outsid	cigarett
9	didnt	curse word 4	laid	death	understand
10	murder	none	murder	god	bed
11	photograph	glad	kill	ill	death
12	bed	pain	god	curse word 4	goddamn
13	ani	curse word 2	bed	murder	curse word 3
14	curse word 2	killer	wife	curse word 3	itd
15	close	bleed	death	bed	privat

Elastic Net	Model 1	Model 2	Model 3	Model 4	Model 5
1	want	want	want	want	want
2	whi	christ	sure	like	like
3	curse word 1	curse word 1	like	curse word 1	ask
4	christ	sure	curse word 1	christ	curse word 1
5	kill	curse word 2	christ	curse word 2	christ
6	curse word 2	curse word 3	curse word 3	kill	god
7	ani	kill	peopl	curse word 6	understand
8	ill	bed	bodi	address	kill
9	god	anyway	curse word 2	death	curse word 2
10	bed	pleas	laid	curse word 3	leav
11	death	much	bed	bed	peopl
12	photograph	glad	coupl	outsid	cigarett
13	murder	curse word 6	later	murder	bed
14	curse word 3	close	murder	itd	death
15	curse word 4	outsid	itd	anyway	itd

Figure T6: Words with the highest importance



Note: The importance numbers are generated via random forest models