Test design under **unobservable** falsification

Eduardo Perez-Richet (Sciences Po) Vasiliki Skreta (UCL, UT Austin)

The University of Texas at Austin
Department of Economics

test and faslification

tests seek to uncover some **state**: e.g. student's ability; drugs potency/side effects; car's pollution; bank's systemic risk

decisions based on test results often by (several) third parties ('the market'), non-coordinated, non-contractible

manipulations/ falsification /cheating, sadly, common

- standardized tests: teachers – testers – recruiters
- drugs: pharmaceuticals –FDA – (consumers)
- emissions: car manufacturers – regulator (EPA) – (consumers)
- asset rating: asset issuers – rating agencies – investors
- stress test: banks – Fed – (investors)

On January 11 2017: *"VW agreed to pay a criminal fine of $4.3bn for selling around 500,000 cars fitted with so-called "defeat devices" that are designed to reduce emissions of nitrogen oxide (NOx) under test conditions."*

On January 12 2017: US Environmental Protection Agency (EPA) accused Fiat Chrysler Automobile of using illegal software in conjunction with the engines which, allowed thousand of vehicles to exceed legal limits of toxic emissions

our goal:test design in the presence of cheating

Setup: Test+Falsification

baseline setup

- agent: endowed with 1 or continuum of items
- Receiver(s): choose 'pass' or 'fail'
- agent wants each item to be passed (payoff 1-0)
  - state $s \in S \subseteq [-\underline{s}, \overline{s}]$, with $-\underline{s} < 0 < \overline{s}$, and $\{-\underline{s}, \overline{s}\} \subseteq S$
  - $S = \{-\underline{s}, \overline{s}\}$ as the binary state case
  - items i.i.d. prior π
  - prior mean $\mu_0 \equiv E_\pi(s) < 0$
- Receiver(s) preferences (identical for all receivers)
  - fail $\rightarrow$ 0
  - pass s $\rightarrow$ s
- Receiver pass item i iff $E(s) > 0$

Time

1. **Test:** A test τ is exogenously given and publicly observable.
2. **Falsification:** The agent chooses a *falsification strategy* ϕ (interim same)
3. **State:** The state s is realized according to π
4. **Testing and results:** The falsification strategy generates a falsified state of the world s' according to ϕ, and the test generates a public signal x about the falsified state of the world according to $\tau_{s'}$
5. **Receiver decision:** The receiver forms beliefs and chooses to approve or reject.

Tests. A test is a Blackwell experiment: a measurable space of signals X, and a Markov kernel τ from S to $\Delta(X)$

- π and τ define joint probability measure $X \times S$: $\tau\pi$
- in the absence of falsification
  - conditional on observing x, receiver forms a belief about S: $\tau\pi_x$
  - conditional on s distribution of signals depends on τ

falsification technology

The agent can falsify the state of the world that is fed to the test.

- **falsification** ϕ which is a Markov kernel from S to ΔS
- if T is a Borel subset of S and $s \in S$ a state of the world, then $\phi(T|s)$ denotes the probability that the true state s, or source, is falsified as a target state in T
- **truth-telling strategy** Markov kernel δ mapping each state s to the Dirac measure δ_s on S
- prior π and falsification strategy ϕ define joint probability measure denoted $\phi\pi$ on $S \times S$
- falsification costless or costly
  - install devices that artificially lower emission levels
  - teaching the students to the test
  - inaccurate reporting of asset characteristics
  - psychological lying costs
- falsification cost $c(t|s)$ cost of falsifying source state s as target state t
- cost of falsification strategy ϕ is $C(\phi) = \int_{S\times S} c \, d\phi\pi$

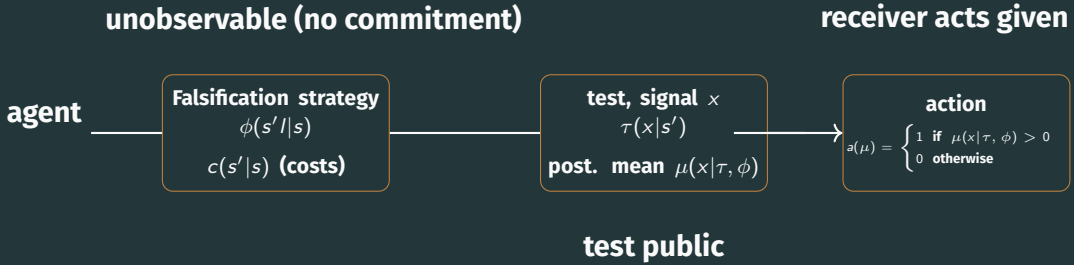
posterior beliefs, actions and resulting payoffs

- prior, falsification strategy and test define a joint distribution over $X \times S$ denoted by $\tau\phi\pi$
- posterior belief given x is $\tau\phi\pi_x \in \Delta S$
- $\mu(x|\tau,\phi) = \int_S s\, d\tau\phi\pi_x(s)$: expected state according to $\tau\phi\pi$
- receiver approves whenever $\mu(x|\tau,\phi) \geq 0$
- signal approval set of the receiver $\bar{X}(\tau,\phi) = \{x : \mu(x|\tau,\phi) \geq 0\}$

$$A(\tau,\phi) = \int_{\bar{X}(\tau,\phi)\times S} d\tau\phi\pi \qquad \text{ex ante probability of approval}$$

$$U(\tau,\phi) = A(\tau,\phi) - C(\phi), \qquad \text{agent's payoff}$$

$$V(\tau,\phi) = \int_{\bar{X}(\tau,\phi)\times S} \mu(x|\tau,\phi) d\tau\phi\pi(x,s) \qquad \text{receiver's payoff}$$


unobservable (no commitment)
receiver acts given x
agent
Falsification strategy
$\phi(s'|s)$
$c(s'|s)$ (costs)
test, signal x
$\tau(x|s')$
post. mean $\mu(x|\tau,\phi)$
action
$a(\mu) = \begin{cases} 1 & \text{if } \mu(x|\tau,\phi) > 0 \\ 0 & \text{otherwise} \end{cases}$
test public


unobservable (no commitment)
receiver acts given x
agent
Falsification strategy
$\phi(s'|s)$
$c(s'|s)$ (costs)
test, signal x
$\tau(x|s')$
post. mean $\mu(x|\tau,\phi)$
action
$a(\mu) = \begin{cases} 1 & \text{if } \mu(x|\tau,\phi) > 0 \\ 0 & \text{otherwise} \end{cases}$
observable (commitment) in paper
test public

Committed versus non-committed falsification

beliefs with observable (committed) falsification the meaning of x 'reacts' to actual ϕ

beliefs with unobservable (non-committed) falsification meaning depends on τ; equilibrium falsification ϕ^E

- with **commitment** agent is a "constrained" persuader: instead of choosing *any* experiment, he can only induce information structures consistent with τ
  - **signals** $\neq$ **action recommendations**
  - $\rightarrow$ need **continuum** "pass" signals even **binary** state
  - challenge 2: entire information structure & approval thresholds change with falsification

Committed versus non-committed falsification

beliefs with observable (committed) falsification the meaning of x 'reacts' to actual ϕ

beliefs with unobservable (non-committed) falsification meaning depends on τ; equilibrium falsification ϕ^E

- with **commitment** agent is a "constrained" persuader: instead of choosing *any* experiment, he can only induce information structures consistent with τ
  - **signals $\neq$ action recommendations**
  - $\rightarrow$ need **continuum** "pass" signals even **binary** state
  - challenge 2: entire information structure & approval thresholds change with falsification
  - **in Perez-Richet and Skreta (2018)**

Committed versus non-committed falsification

beliefs with observable (committed) falsification the meaning of x 'reacts' to actual ϕ

beliefs with unobservable (non-committed) falsification meaning depends on τ; equilibrium falsification ϕ^E

- with **commitment** agent is a "constrained" persuader: instead of choosing *any* experiment, he can only induce information structures consistent with τ
  - **signals $\neq$ action recommendations**
  - $\rightarrow$ need **continuum** "pass" signals even **binary** state
  - challenge 2: entire information structure & approval thresholds change with falsification
  - **in Perez-Richet and Skreta (2018)**

Committed versus non-committed falsification

beliefs with observable (committed) falsification the meaning of x 'reacts' to actual ϕ

beliefs with unobservable (non-committed) falsification meaning depends on τ; equilibrium falsification ϕ^E

- with **commitment** agent is a "constrained" persuader: instead of choosing *any* experiment, he can only induce information structures consistent with τ
  - **signals $\neq$ action recommendations**
  - $\rightarrow$ need **continuum** "pass" signals even **binary** state
  - challenge 2: entire information structure & approval thresholds change with falsification
  - **in Perez-Richet and Skreta (2018)**
- **NEW unobservable falsification**
  - akin to mechanism design without transfers
  - here **signals =action recommendations**
  - if falsification costless: WLOG no falsification (a.k.a "truth-telling") best response
  - but without costs no test works....
  - characterisation of optimal test: **involves falsification!**
  - derivation of falsification proof test

overview of results

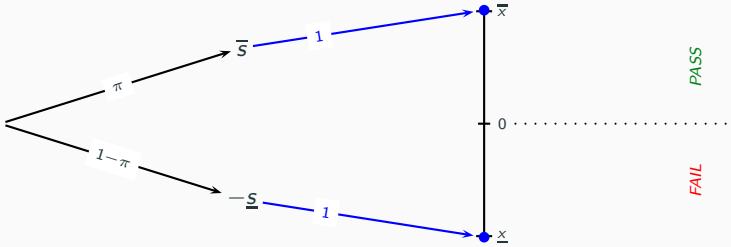
- general framework to study manipulations
- **mechanism design with costly reports; no transfers**
- issues with revelation principle
- **optimum involves lying–and lying is essential**
- optimal falsification-proof test strictly worse
  - constrained infinite dimensional program
  - usual relaxed program not ususefull
  - non-local IC bind
  - and continuum of binding IC
  - novel characterization via auxiliary problem/dual of optimal transportation problem

Warm-up Binary state

baseline setup

- agent: endowed with 1 or continuum of items
- agent wants each item to be passed (payoff 1-0)
  - each $S = \{-\underline{s}, -\underline{s}\}$
  - distributed i.i.d. with $\Pr(s = \overline{s}) = \pi_0$;
- Receiver(s) preferences (identical for all receivers)
  - fail $\rightarrow$ 0
  - pass $\overline{s}$ $\rightarrow$ $\overline{s} > 0$
  - pass $-\underline{s}$ $\rightarrow$ $-\underline{s} < 0$
- Receiver pass item i iff $\Pr(s = \overline{s}) \geq 0$,
- test $\overline{\tau}$ and $\underline{\tau}$
- falsification state $-\underline{s}$ generates signals from $\overline{\tau}$: ϕ
  (WLOG ignore 'downwards' falsification)

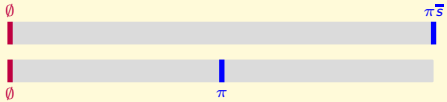
fully informative test receiver-optimal without cheating



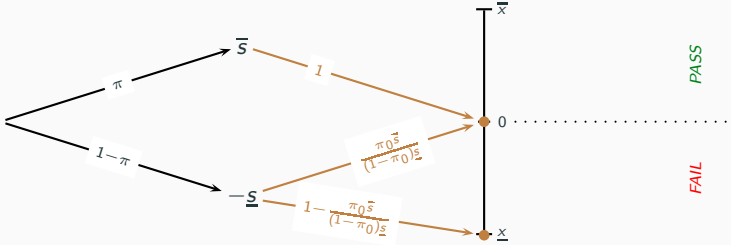
PAYOFFS

Receiver:

agent:



agent-optimal a.k.a. Kamenica-Gentzkow test



PAYOFFS

	KG		
Receiver:	∅		FI
agent:	∅	FI	KG

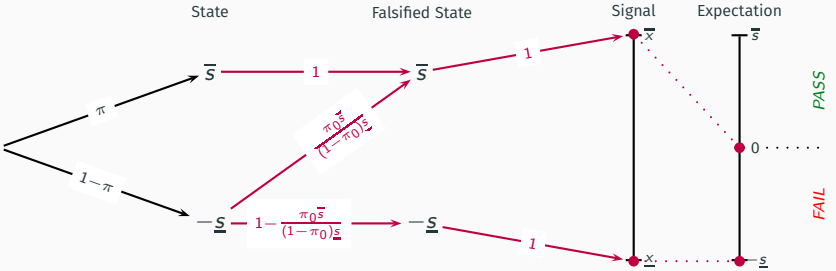
The University of Texas at Austin
Department of Economics

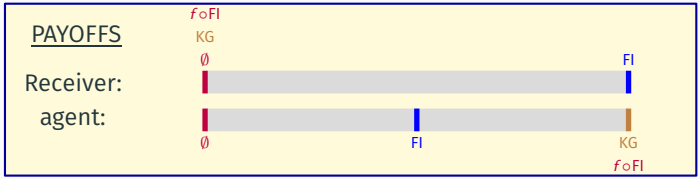
falsification of two-signal tests

Suppose there is a fully revealing two-signal test: $X = \{\underline{x}, \overline{x}\}$

- suppose ϕ observable–**endogenously costly** "devalues" signals
  - signal $\overline{x}$ yields pass if: $\pi_0\overline{s} - (1-\pi_0)\phi\underline{s} > 0$
  - $\pi_0 + \phi(1-\pi_0)(1-c)\ \mathbb{1}\left(\phi \leq \frac{\pi_0\overline{s}}{(1-\pi_0)\underline{s}}\right)$
  - if $1-c>0$ optimal ϕ is $\phi = \frac{\pi_0\overline{s}}{(1-\pi_0)\underline{s}}$
  - setting $\phi = \frac{\pi_0\overline{s}}{(1-\pi_0)\underline{s}}$ and "approve" after $\overline{x}$ is an equilibrium if ϕ observable
  - agent achieves optimum!
- no equilibrium with pos. prob of "approve" if ϕ unobservable (if that was the case agent chooses $\phi = 1$, but then receiver never approves
- both benefit when falsification observable/detectable
- this talk: **what can be done in unobservable case when falsification is costly**

fully informative test + observable falsification



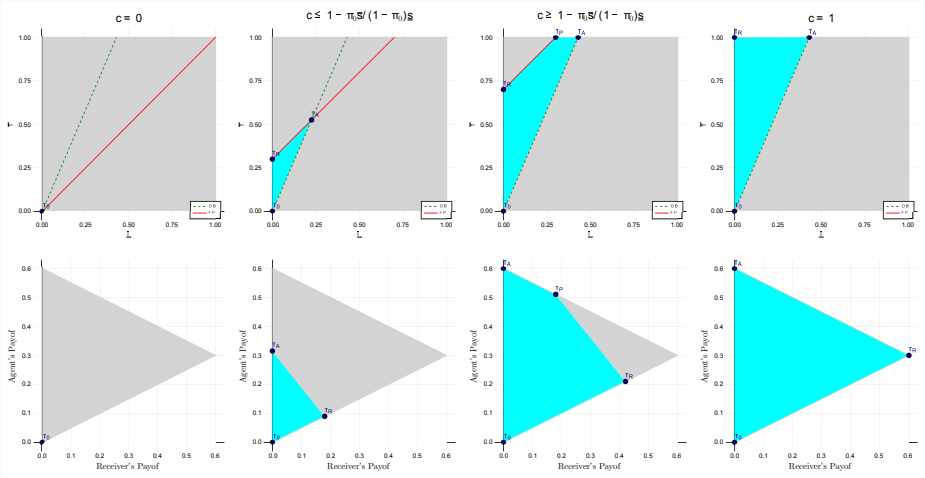


Undetectable falsification: two states

Binary State: Set of feasible tests given by:

- WLOG test as an approval probability $\overline{\tau}, \underline{\tau}$
- "Falsification proofness"/informativeness condition: $\overline{\tau} - \underline{\tau} \leq c$
- otherwise $-\underline{s}$ falsifies as $\overline{s}$; $\rightarrow$ no information
- Obedience Constraint: $\overline{\tau}\pi_0\overline{s} - \underline{\tau}(1-\pi_0)\underline{s} \geq 0$

Binary State



Undetectable falsification: general state space

signals: action recommendations

Let ϕ be an equilibrium falsification strategy under τ. Then the test τ' with binary signal space $X' = \{\text{Pass}, \text{Fail}\}$ defined by

$$\tau'(\text{Pass}|s) = \tau(\bar{X}(\tau, \phi)|s)$$

is such that ϕ is an equilibrium under τ' is equivalent in terms of payoffs and approvals. So:

- we redefine **tests as measurable functions** $\tau : S \to [0, 1]$
- nominal passing probability $\tau(s)$: probability test recommends passing s
- falsification induces the " true" passing probability that can differ from nominal

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s)$$

$$\text{s.t.} \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi\pi(t,s) \geq \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi'\pi(t,s), \quad \forall \phi' \text{ ex-ante optimal falsification}$$

$$\int_{S\times S} s\tau(t)d\phi\pi(t,s) \geq 0 \text{ receiver obedience}$$

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s)$$

$$\text{s.t.} \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi\pi(t,s) \geq \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi'\pi(t,s), \quad \forall \phi' \text{ ex-ante optimal falsification}$$

$$\int_{S\times S} s\tau(t)d\phi\pi(t,s) \geq 0 \text{ receiver obedience}$$

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s)$$

$$\text{s.t.} \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi\pi(t,s) \geq \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi'\pi(t,s), \quad \forall \phi' \text{ ex-ante optimal falsification}$$

$$\int_{S\times S} s\tau(t)d\phi\pi(t,s) \geq 0 \text{ receiver obedience}$$

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s)$$

$$\text{s.t.} \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi\pi(t,s) \geq \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi'\pi(t,s), \quad \forall \phi' \text{ ex-ante optimal falsification}$$

$$\int_{S\times S} s\tau(t)d\phi\pi(t,s) \geq 0 \text{ receiver obedience}$$

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s)$$

$$\text{s.t.} \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi\pi(t,s) \geq \int_{S\times S} \{\tau(t) - c(t|s)\} d\phi'\pi(t,s), \quad \forall \phi' \text{ ex-ante optimal falsification}$$

$$\int_{S\times S} s\tau(t)d\phi\pi(t,s) \geq 0 \text{ receiver obedience redundant}$$

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s)$$

$$\text{s.t.} \int_{S\times S} \big\{\tau(t) - c(t|s)\big\} d\phi\pi(t,s) \geq \int_{S\times S} \big\{\tau(t) - c(t|s)\big\} d\phi'\pi(t,s), \quad \forall \phi' \text{ ex-ante optimal falsification}$$

$$\int_{S\times S} s\tau(t)d\phi\pi(t,s) \geq 0 \text{ receiver obedience redundant}$$

ex ante optimal falsification (EOF) is equivalent interim (IOF): ϕ_s puts probability 1 on set of (interim) optimal falsification targets

program

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s) \qquad \text{(P)}$$

$$\text{s.t. } \phi\left(\Phi(s;\tau)|s\right) = 1, \quad \forall s \in S \qquad \text{(IOF)}$$

$$\text{where } \Phi(s;\tau) = \operatorname{argmax}_t \ \tau(t) - c(t|s) \qquad \text{(optimal falsification targets)}$$

assumptions on falsifications costs

cost: given $c(t|s)$ where $c : S \times S \to \mathbb{R}_+$ is measurable, and continuous in t.

- $c(s|s) = 0$.
- Monotonicity (MON): If $c \neq 0$, $c(t|s)$ increasing in t and decreasing in s if $s < t$, ...
- Triangular inequality (TRI): $c(t|m) + c(m|s) \geq c(t|s)$.

Let

$$s_0 = \max\left\{s' \in S : \mathbb{E}_\pi(s|s \geq s') \leq 0\right\}.$$

In particular, if π has no atom at s_0, then $\mathbb{E}_\pi(s|s \geq s_0) = 0$.

further simplifications

- if the cost function satisfies (TRI), we can restrict attention to tests that are falsification proof for negative states
- we can also restrict attention to tests when positive states don't have incentive to falsify as negative

optimal class of tests

We now consider a class of tests defined by two parameters: the highest nominal passing probability $p \in [0, 1]$, and the cutoff state $\hat{s} \in S^+$ above which nominal probabilities are set to p:

$$\tau_{p,\hat{s}}(s) = \begin{cases} p \text{ for } s \geq \hat{s} \\ \left[p - c(\hat{s}|s)\right]^+ \text{ for } s < \hat{s} \end{cases}$$

here $\check{s}(p, \hat{s}) = \min\{s \leq \hat{s} : c(\hat{s}|s) \leq p\}$ and such that $\check{s}(p, \hat{s}) \leq 0$.

properties

(i) $\tau_{p,\hat{s}}$ is continuous on S, strictly increasing on $\left(\check{s}, \overline{s}\right]$ and constant and equal to 0 below $\check{s}(p, \hat{s})$ and and constant and equal to p above $\hat{s}$

(ii) if the cost function satisfies (TRI), then **truth-telling** optimal for every $s \in S$ ($s \in \Phi(s; \tau_{p,\hat{s}})$).

(iii) for every $s \in \left(\check{s}(p, \hat{s}), \hat{s}\right]$, falsify to $\hat{s}$ ALSO optimal $\hat{s} \in \Phi(s; \tau_{p,\hat{s}})$

(iv) receiver-preferred falsification $s \in \left(0, \hat{s}\right)$ falsify to $\hat{s}$

Theorem

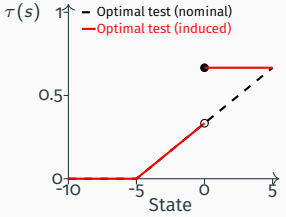
Suppose the cost function satisfies (TRI). Then $(\tau_{p^*,s^*}, \phi_{p^*,s^*})$ *maximizes* (P)*, where*

$$p^* = \min\left\{c(\overline{s}|s_0), 1\right\},$$

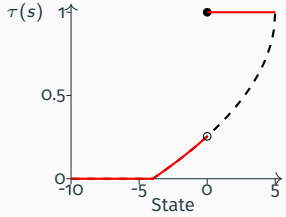
and

$$s^* = \max\left\{s \in S : c(s|0) \leq 1\right\}.$$

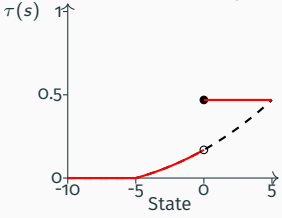
Furthermore, the receiver gets her first-best payoff if and only if $c(\overline{s}|0) \geq 1$*. However, the pair of resulting payoffs* (U^*, V^*) *never lies on the Pareto frontier.*



(a) $c(t|s) = \alpha(t-s)$, $\alpha = \frac{1}{15}$



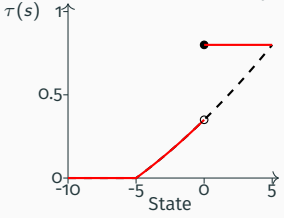
(b) $c(t|s) = \alpha\sqrt{t-s}$, $\alpha = \frac{1}{3}$



(c)

$c(t|s) = \alpha e^{2\beta s}\left\{(t-s) + \beta(t-s)^2\right\}$,

$\alpha = \frac{1}{20}$ and $\beta = \frac{1}{30}$



(d) $c(t|s) = \alpha\left\{(t-s) - \beta(t-s)^2\right\}$,

$\alpha = \frac{1}{10}$ and $\beta = \frac{1}{50}$

Proof

Step 1:

- use (TRI) to show that we can transform any test so that negative states are truthful while improving both receiver and agent payoffs.
- **IDEA: give negative states their falsification payoff**

Step 2: we can further transform any test so that nonnegative states do not falsify as negative states while improving both receiver and agent payoffs.

Step 3: we can replace any such transformed test by a test of the form $\tau_{p,\hat{s}}$ and increase the receiver's payoff.

Step 4: optimize on p and $\hat{s}$

optimal falsification proof test

$$\sup_{\tau} \int_S s\tau(s) dF_\pi(s) \tag{FPProg}$$

$$\text{s.t. } \tau(t) - \tau(s) \leq c(t|s), \quad \forall s, t \in S \tag{FPIC}$$

properties of FPIC tests

Let τ be a test that satisfies (FPIC) then:

1. τ is continuous
2. there exists a K-Lipschitz and nondecreasing test function $\hat{\tau}$ that also satisfies (FPIC) and makes the receiver better off
3. $\implies$ test increasing and K-Lipschitz, differentiable a.e. derivative τ' bounded in $[0, K]$

These properties result to **envelope characterization** $\tau(s) = \underline{\tau} + \int_{-\underline{s}}^{s} \tau'(z) dz$
Can choose scaler $\underline{\tau} \in [0, 1]$ and the function $\{\tau'(s)\}_{s \in S}$
Need: Regularity (REG): $c(t|s)$ is continuously differentiable in t on $[s, \overline{s}]$ and in s on $[-\underline{s}, t]$, and there exists $K > 0$ such that, for every $t > s$,

$$c(t|s) \leq K(t - s).$$

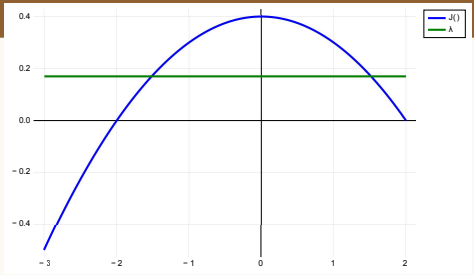
an auxiliary function

Let $J : S \to \mathbb{R}$ be:

$$J(z) = \int_z^{\overline{s}} s dF_\pi(s)$$

properties of $J(z)$

1. $J(z) < 0$ for $z < s_0$
2. $J(z) \geq 0$ for $z \geq s_0$
3. continuous
4. increasing on S^-
5. decreasing and S^+
6. $\implies$ single-peaked at 0



$\pi = Uniform([-3, 2])$

$$\underline{\tau}\mu_0 + \int_S \tau'(z)J(z)dz \qquad \text{(reformulated objective function of (FPProg))}$$

$$\text{s.t.}\underline{\tau} + \int_S \tau'(z)dz \leq 1 \qquad \text{(probability bound)}$$

$$\int_s^t \tau'(z)dz \leq c(t|s), \text{ for all } s < t \qquad \text{(FPIC)}$$

Reducing $\underline{\tau}$ increases the objective function as $\mu_0 < 0$, relaxes the probability constraint, and has no effect on the incentive constraints, implying that it is optimal to set $\underline{\tau} = 0$

We ignore FPIC and solve relaxed program, where we treat the probability constraint with the Lagrangian method. Let $\lambda \geq 0$

$$\mathcal{L}(\tau', \lambda) = \int_S \tau'(z) J(z) dz + \lambda \left(1 - \int_S \tau'(z) dz\right) = \int_S \tau'(z) \big(J(z) - \lambda\big) dz + \lambda$$

Maximize $\mathcal{L}(\tau', \lambda)$ where $\tau' : S \to [0, K]$ is feasible if, for every $s < t$, $\int_s^t \tau'(z) dz \leq c(t|s)$, and $\int_{s_0}^{\overline{s}} \tau'(z) dz \leq 1$.

Any solution must satisfy $\tau'(s) = 0$ for almost every s such that $J(s) < \lambda$, that is, by continuity and single-peakedness of J, outside of an interval $[s_*, s^*]$ such that $J(s_*) = J(s^*) = \lambda$.

Lemma (Lagrangian sufficiency theorem)

Suppose that there exists $\hat{\lambda} \geq 0$*, and a feasible* $\hat{\tau}'$ *such that:*

(a) $\hat{\lambda} = 0$ *or* $\int_S \tau'(z)dz = 1$*;*

(b) *For every feasible* τ'*,* $\mathcal{L}(\hat{\tau}', \hat{\lambda}) \geq \mathcal{L}(\tau', \hat{\lambda})$*.*

Then there exists an interval $[s_*, s^*]$ *such that:*

(i) $s_0 \leq s_* \leq 0 \leq s^* \leq \overline{s}$ *and* $J(s_*) = J(s^*) = \hat{\lambda}$*;*

(ii) $\hat{\tau}'(s) = 0$ *for every* $s \notin [s_*, s^*]$*;*

(iii) *The test* $\hat{\tau}(s) = \left(\int_{s_*}^{s^* \wedge s} \hat{\tau}'(z)dz \right) \mathbb{1}(s \geq s_*)$ *is a falsification-proof receiver optimal test.*

matching function

Matching function help us guess optimal Lagrange multiplier: Choose

$$s_* = \min\{s \in [s_0, 0] : c(m(s)|s) \leq 1\},$$

Then: $\lambda^* = J(s_*)$ Note that $s_* = s_0$ whenever $c(\overline{s}|s_0) \leq 1$

- matching function $m : [s_0, 0] \to [0, \overline{s}]$ is
  - decreasing
  - implicitly defined by $J(s_*) = J(m(s^*))$
  - or equivalently by $\int_{s_*}^{m(s^*)} s dF_\pi(s) = 0$
  - s_0 is matched with $m(s_0) = \overline{s}$
- each choice of $s_* \in [s_0, 0]$ uniquely pins down $s^* = m(s_*)$

next step: program reformulation

Instead of solving the Lagrangian problem, we go back to the original program. Focus on tests τ that are constant outside of $[s_*, s^*]$, and $\tau(s_*) = 0$.

Also relax the program by getting rid of the constraint that $\tau(s^*) \leq 1$, and only keeping the incentive constraints for pairs (s, t) such that $s_* \leq s \leq 0 \leq t < s^*$

Change variables and let $y = -s \in Y = [0, -s_*]$ and $z = t \in Z = [0, s^*]$. Finally, we let $\rho : Y \to \mathbb{R}$, and $\psi : Z \to \mathbb{R}$ be the functions defined by $\rho(y) = \tau(-y) = \tau(s)$, and $\psi(z) = \tau(z) = \tau(t)$. With these notations, the remaining incentive constraints become

$$\psi(z) - \rho(y) \leq c(z| - y), \quad \forall (y, z) \in Y \times Z.$$

And, up to multiplication by the constant $\mu^* = \int_0^{s^*} s dF_\pi(s)$, the objective function of the program becomes

$$\int_Z \psi(z) dQ(z) - \int_Y \rho(y) dP(y),$$

where $Q(z) = \frac{1}{\mu^*} \int_0^z x dF_\pi(x)$, and $P(y) = \frac{1}{\mu^*} \int_0^y x dF_\pi(-x)$ define atomless cumulative distribution functions on, respectively, Z and Y.

new relaxed and reformulated program:

$$\sup_{\rho,\psi} \int_Z \psi(z) dQ(z) - \int_Y \rho(y) dP(y)$$
$$\text{s.t. } \psi(z) - \rho(y) \leq c(z|-y), \quad \forall (y,z) \in Y \times Z,$$

is dual of the following well known Monge-Kantorovich optimal transport problem

$$\inf_{\varphi \in \mathcal{M}(P,Q)} \int_{Z \times Y} c(z|-y) d\varphi(z,y),$$

where $\mathcal{M}(P,Q)$ is the set of joint distributions on $Z \times Y$ with marginals Q on Z, and P on Y.

Assume: Upward increasing differences (UID): $c(t'|s') - c(t|s') \geq c(t'|s) - c(t|s)$ for $s < s' \leq t < t'$
$\implies$ transportation cost function of this problem, $c(z|-y)$ is submodular, well known solution for both problems

Let c_t and c_s we denote the partial derivatives of the cost function.

$$\tau^*(s) = \begin{cases} -\int_{s_*}^{s} c_s\big(m(x)|x\big)\,dx \text{ for } s \in [s_*, 0] \\ c(s^*|s_*) - \int_{s}^{s^*} c_t\big(x|m^{-1}(x)dx\big) \text{ for } s \in (0, s^*] \\ 1 \text{ for } s > s^* \end{cases}$$

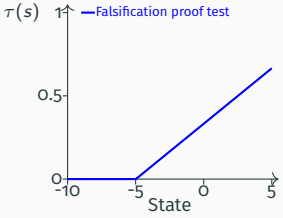
The following theorem shows that τ^* solves our initial problem.

Theorem

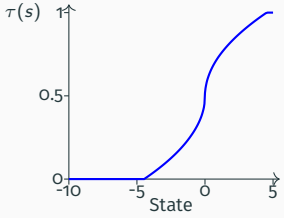
The test τ^* solves (FPProg) *and is therefore a receiver-optimal falsification-proof test. The corresponding receiver's payoff is given by*

$$U(\tau^*, \delta) = \int_{s_*}^{0} -sc\big(m(s)|s\big)\,dF_\pi(s) = \int_{0}^{s^*} tc\big(t|m^{-1}(t)\big)\,dF_\pi(t).$$

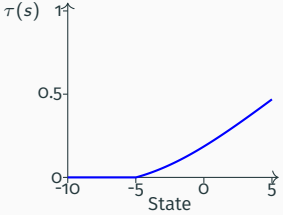
Furthermore, the outcome (τ^*, δ) is Pareto inefficient.



(e) $c(t|s) = \alpha(t-s), \alpha = \frac{1}{15}$

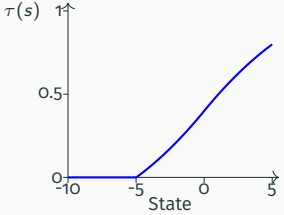


(f) $c(t|s) = \alpha\sqrt{t-s}\ \alpha = \frac{1}{3}$

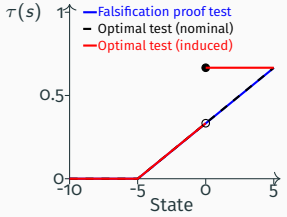


(g)

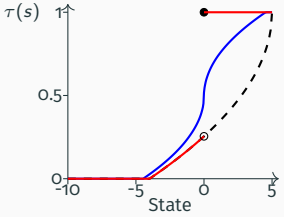
$c(t|s) = \alpha e^{2\beta s}\left\{(t-s) + \beta(t-s)^2\right\}$,
$\alpha = \frac{1}{20}$ and $\beta = \frac{1}{30}$



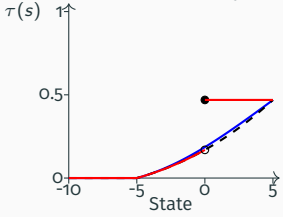
(h) $c(t|s) = \alpha\left\{(t-s) - \beta(t-s)^2\right\}$,
$\alpha = \frac{1}{10}$ and $\beta = \frac{1}{50}$



(i) $c(t|s) = \alpha(t-s)$, $\alpha = \frac{1}{15}$

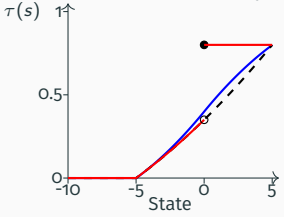


(j) $c(t|s) = \alpha\sqrt{t-s}$, $\alpha = \frac{1}{3}$



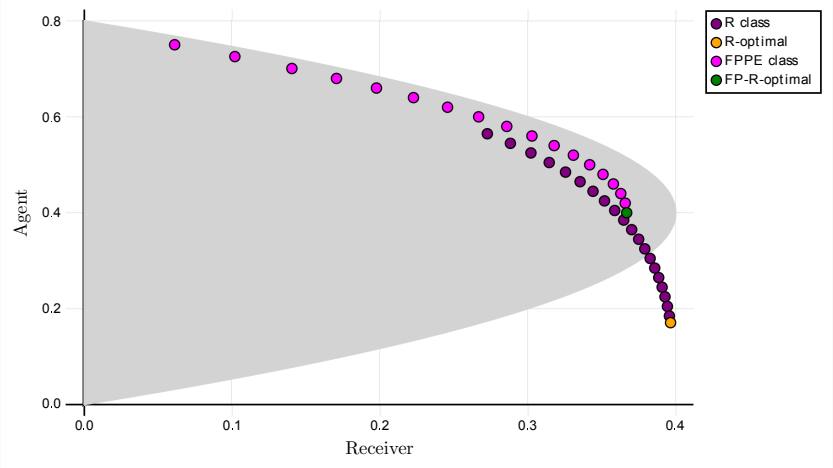
(k)

$c(t|s) = \alpha e^{2\beta s}\left\{(t-s) + \beta(t-s)^2\right\}$,
$\alpha = \frac{1}{20}$ and $\beta = \frac{1}{30}$



(l) $c(t|s) = \alpha\left\{(t-s) - \beta(t-s)^2\right\}$,
$\alpha = \frac{1}{10}$ and $\beta = \frac{1}{50}$

Payoff plots



$$c(t|s) = \frac{1.33|t-s|}{1+|t-s|} \qquad \pi = Uniform([-3,2])$$

literature

information design / Bayesian persuasion:

- Kamenica and Gentzkow (2011), Gentzkow and Kamenica (2016), Kolotilin (2016)
- with manipulations: Frankel and Kartik (2019), Guo and Shmaya (2019), Nguyen and Tan (2020)

mechanism design with costly reporting:

- Kephart and Conitzer (2016), Deneckere and Severinov (2017), Severinov and Tam (2019)

mechanism design without transfers:

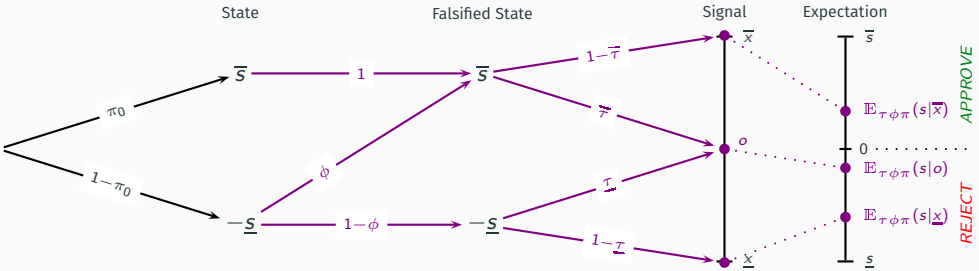
- Amador and Bagwell (2013), Amadon, Werning, Angeletos (2006), Ben-Porath, Dekel and Lipman (2014)

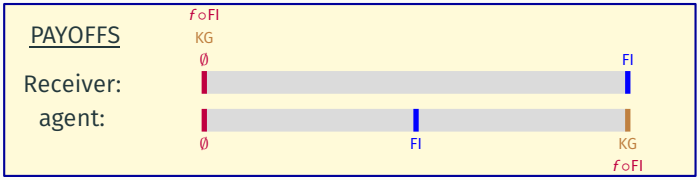
costly state falsification:

- mechanism design: Lacker and Weinberg (1989), Landier and Plantin (2016)
- testing: Cunningham and Moreno de Barreda (2015)

Observable falsification in Perez-Richet and Skreta (2018)

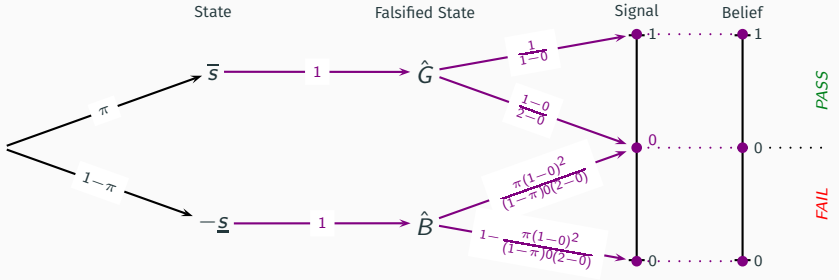
test + observable falsification: a 3-signal test

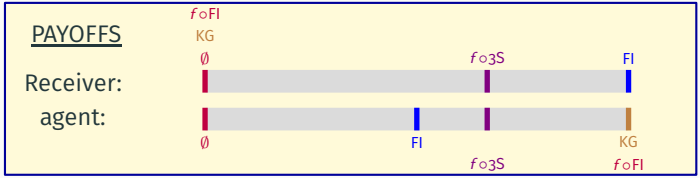




The University of Texas at Austin
Department of Economics

test + falsification: a 3-signal test for observable case





The University of Texas at Austin
Department of Economics

second result (observation)

adding an extra (noisy) signal helps!

the 3-signal test contains a simple practical insight: introducing a “noisy" (pooling) grade that is associated with approval in the absence of falsification, can make falsification so costly that it prevents it, rendering this noisy test much better than the (manipulated) fully informative test

next

second result (observation)

adding an extra (noisy) signal helps!

the 3-signal test contains a simple practical insight: introducing a “noisy" (pooling) grade that is associated with approval in the absence of falsification, can make falsification so costly that it prevents it, rendering this noisy test much better than the (manipulated) fully informative test

next

- is the three signal test optimal?
- how many signals do we need?
- is optimal test falsification-proof?
- how can we tractably find it?

receiver-optimal test with observable falsification

results in a nutshell

- any test feasible with unobservable fals, feasible with observable
- with 2 states/ establish **falsification proofness**–like “revelation principle"
  - intuition: test + optimal cheating = new test $\rightarrow$ offer new test
  - no incentive to cheat in new test–otherwise cheating not optimal in old test
  - argument can fail with certain costs/more than two states
- formulate tractable program derive optimum
- optimal test is rich: signals $\neq$ recommendations
  - one failing signal
  - a **continuum** of passing signals
  - **clustering** of signals above the approval threshold
  - good type only generates “pass” signals
  - bad type may generate both “pass” or “fail” signals
  - payoffs on Pareto Frontier
  - makes agent **indifferent across all falsification levels** (thresholds)

Both the receiver and the agent STRICTLY benefit if falsification is observable

It is useful to publish the empirical distribution of grades: simple yet important practical insight: tests can gain credibility if the principal publishes the empirical distribution of test results. doing so enables fraud detection

Test credibility benefits everyone ex-ante

thank you!

The University of Texas at Austin
Department of Economics