Test Design under Falsification

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

test and decisions

- 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)

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

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

manipulations/ falsification /cheating, sadly, common

The University of Texas at Austin
Department of Economics

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

baseline setup

- Sender: endowed with 1 or continuum of items
- Sender wants each item to be approved (payoff 1-0)
  - each item is "good" or "bad" $\omega \in \{G, B\}$
  - distributed i.i.d. with $\Pr(\omega = G) = \mu_0$
- Receiver(s) preferences (identical for all receivers)
  - reject $\rightarrow$ 0
  - approve G $\rightarrow$ $g > 0$
  - approve B $\rightarrow$ $-b < 0$
- Receiver approves i iff $\Pr(\omega = G) \geq \hat{\mu}$, where $\hat{\mu} \equiv \frac{b}{g+b}$
  - assume: $\mu_0 < \hat{\mu}$

The University of Texas at Austin
Department of Economics

timing and falsification technology

Time

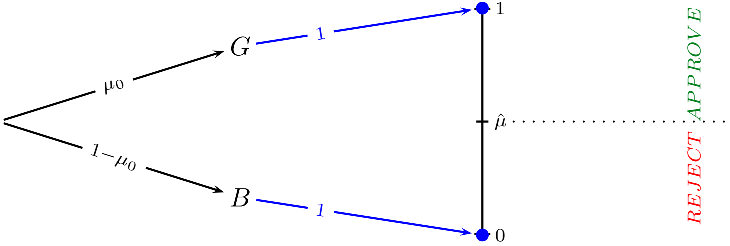
- there is a test
- Sender: chooses **falsification rates** p_B; p_G:
- state(s) realized
- items tested, results revealed
- Receiver(s): based on results, decide whether to approve/reject each item

test–modeled as **Blackwell Experiment:** $H : \Omega \to \Delta(S)$

- maps each state to a distribution over signals: H_B, H_G
- $\Omega = \{B, G\}$; $\mu_s = Pr(\omega = G|s)$; normalization: signals = beliefs; $S = [0, 1]$
- **falsification technology** state B generates signals from H_G–vice versa
- falsification costless or costly
  - install devices that artificially lower emission levels
  - teaching the students to the test
  - inaccurate reporting of asset characteristics

The University of Texas at Austin
Department of Economics

fully informative test receiver-optimal without cheating

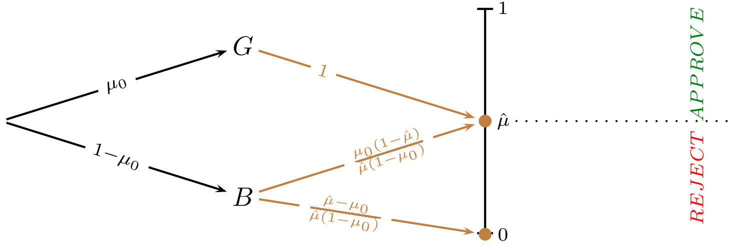


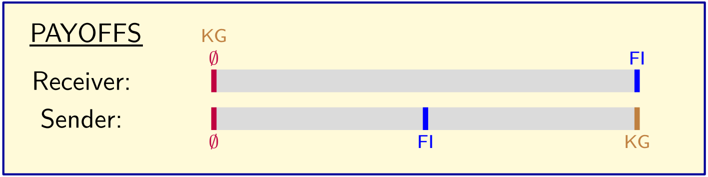
PAYOFFS

Receiver: $\emptyset$ … $\mu_0 g$

Sender: $\emptyset$ … μ_0

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





falsification **endogenously costly** "devalues" signals

- a signal μ has **literal** meaning if $p_B = 0, p_G = 0$
- otherwise μ is "naive" and associated belief $\tilde{\mu}$ satisfies mapping:

$$\mu = \mu_0 \frac{(1-\mu_0)\tilde{\mu} - \mu_0(1-\tilde{\mu})p_G - (1-\mu_0)\tilde{\mu}p_B}{\mu_0(1-\mu_0) - \mu_0(1-\tilde{\mu})p_G - (1-\mu_0)\tilde{\mu}p_B}$$

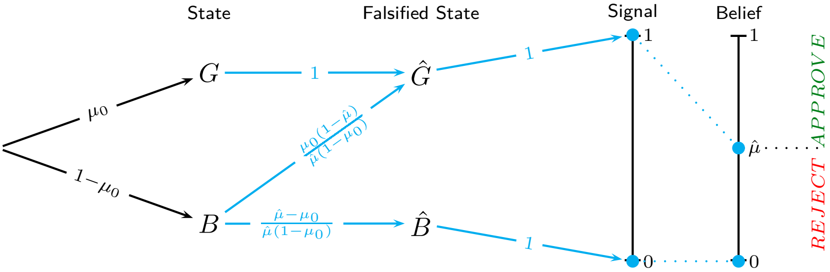
  - if $p_B + p_G \leq 1$ higher μ, associated with higher actual belief $\tilde{\mu}$
  - if $p_B + p_G > 1$ higher μ, associated with lower actual belief $\tilde{\mu}$
- one can show that $p_G = 0$
- **approval threshold:** $\hat{\mu}(p_B)$: signal with belief $\tilde{\mu} = \hat{\mu}$
  - choosing p_B=choosing threshold $\hat{\mu}(p_B)$
- in eqm p_B correctly anticipated–**what about deviations?**

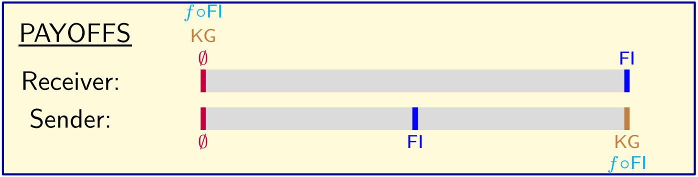
plan

1. perfect/partial observability of p_B
2. unobservable p_B

The University of Texas at Austin
Department of Economics

test + falsification: fully informative test





The University of Texas at Austin
Department of Economics

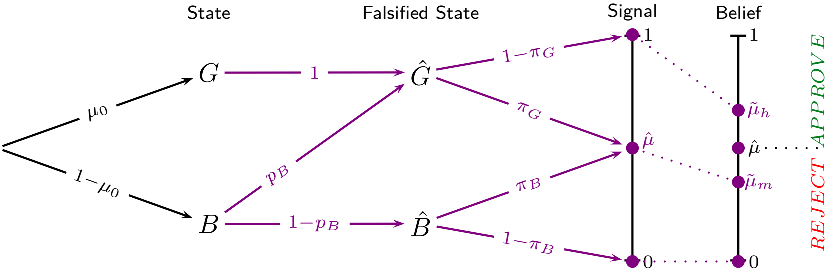
first observation: 2-signal tests

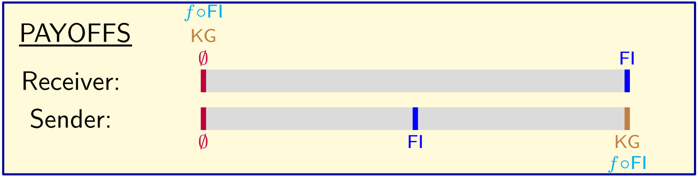
poor performance of two-signal test

any two signal test that would lead to positive probability of approval in the absence of cheating will be faslified and yield 0 to receiver

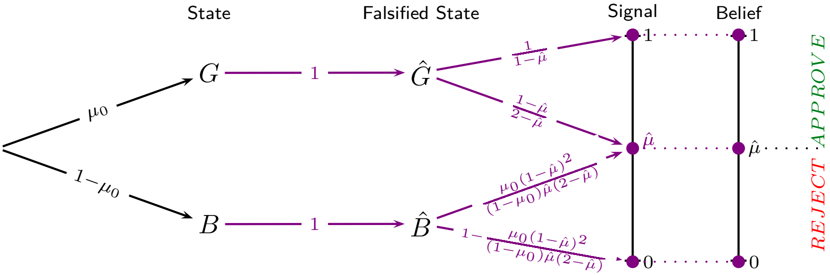
How about adding one extra noisy signal to FI test?

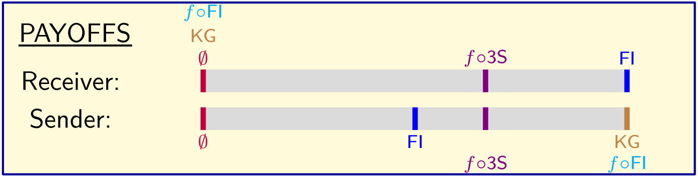
test + falsification: a 3-signal test





test + falsification: a 3-signal test





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

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

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

The University of Texas at Austin
Department of Economics

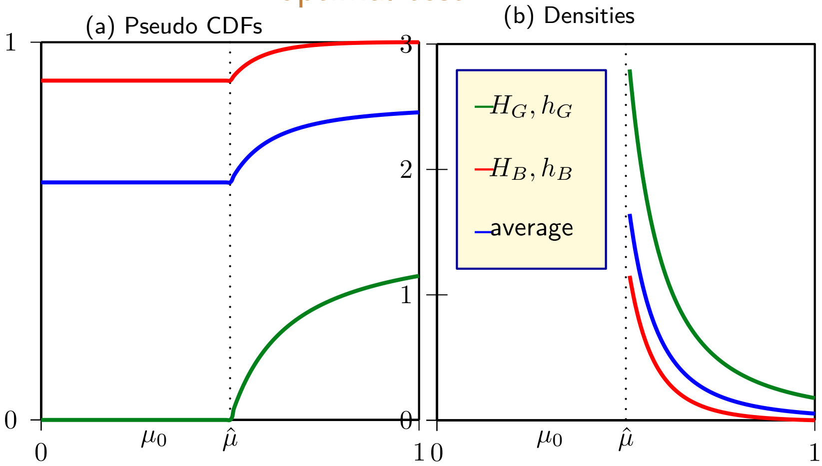
receiver-optimal test with cheating

results in a nutshell

- establish **falsification proofness**–like "revelation principle"
  - intuition: test + optimal cheating = new test → 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 "approve" signals
  - bad type may generate both "approve" or "reject" signals
  - payoffs on Pareto Frontier
  - makes sender **indifferent across all falsification levels** (thresholds)

The University of Texas at Austin
Department of Economics

optimal test



<u>PAYOFFS</u>

Receiver: $f\circ$FI, KG, $\emptyset$ ... $f\circ$3S, $f\circ\mathcal{H}^*$, FI

Sender: $\emptyset$ ... FI, $f\circ\mathcal{H}^*$, $f\circ$3S ... KG, $f\circ$FI

trade-off and clarifications

- Receiver(s) decide after results are in (there is no ex-ante commitment to a signal contingent approval policy)
- Sender is akin to a “constrained” persuader–cannot choose test, but can costlessly falsify state
- **trade-off** faslification can yield “better” test results; more approvals
    - but it can devalue test results: Receiver interprets test results differently

summary

if cheating is fully observable (or endogenously and partially observable) receiver-optimal test enables information provision, even if cheating is costless

The University of Texas at Austin
Department of Economics

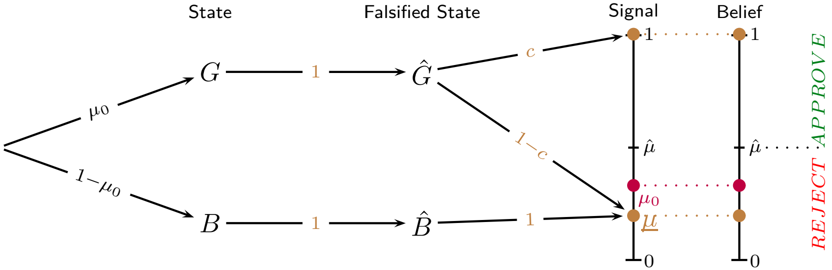
what about unobservable deviations?

suppose that deviations are unobservable/ no inferences are possible

- if cheating is costless, any test that generates higher probability of approval for G, fully falsified
- only possible equilibrium approve G and B equally often–but given prior best to always reject
- when faslification cost is ZERO, not possible to generate any approvals in equilibrium
- explicit costs here help!

The University of Texas at Austin
Department of Economics

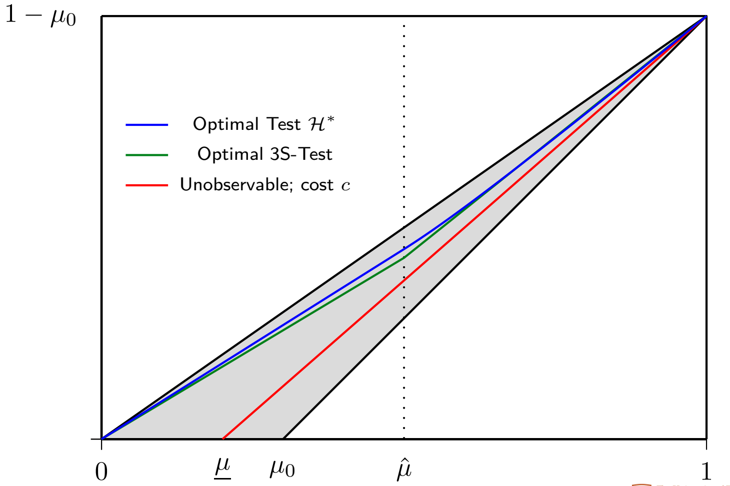
optimal test: costly unobservable deviations



optimal test has a two signals:**pass/fail**, is **faslification-proof**

- Sender, Receiver strictly worse-off, Eq payoffs not Pareto Optimal

optimal test in convex function representation



unobservable deviations: deriving optimal test baseline

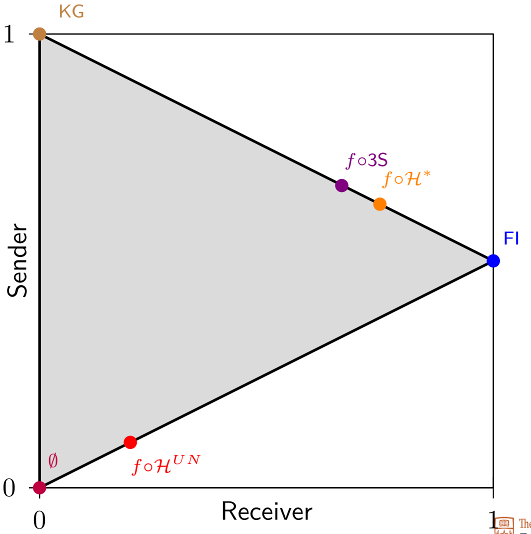
- falsification-proofness (FP) holds for two-states
- formulate a constrained information design problem subject to FP; λ_{BG} multiplier on B not to falsify as G; analogous λ_{GB}
- observation $\lambda_{GB} = 0$; and algebra yield:

$$\inf_{\lambda_{BG} \geq 0} \max_{H \in \Delta([0,1])} \sum_{\mu \in supp(H)} H(\mu) \underbrace{\left\{ (\mu - \hat{\mu})^+ - \lambda_{BG} \mathbb{1}_{\mu \geq \hat{\mu}} \frac{\mu - \mu_0}{\mu_0(1-\mu_0)} \right\}}_{\tilde{v}(\mu, \lambda_{BG})} + \lambda_{BG} c_{BG}$$

- solution:
  - if $c_{BG} < 1$ then $\lambda_{BG} = \mu_0(1 - \hat{\mu})$; constraint binds H **splits the mass between 1 and $\underline{\mu} = \frac{\mu_0(1-c_{BG})}{1-\mu_0 c_{BG}}$**
  - if $c_{BG} \geq 1$, $\lambda_{BG} = 0$, constraint is not binding and **full information feasible thus optimal**

The University of Texas at Austin
Department of Economics

relative performance: payoffs



The University of Texas at Austin
Department of Economics

unobservable deviations: deriving optimal test **Designer-Receiver** conflict

designer threshold $\tilde{\mu} \neq \hat{\mu}$ agent's payoff function is unchanged principal's payoff function becomes $(\mu - \tilde{\mu})\mathbb{1}_{\mu \geq \hat{\mu}}$ still get $\lambda_{GB} = 0$

$$\inf_{\lambda_{BG} \geq 0} \max_{H \in \Delta([0,1])} \sum_{\mu \in (\tau)} \tau(\mu) \left\{ \underbrace{(\mu - \tilde{\mu})\mathbb{1}_{\mu \geq \hat{\mu}} - \lambda_{BG}\mathbb{1}_{\mu \geq \hat{\mu}} \frac{\mu - \mu_0}{\mu_0(1-\mu_0)}}_{\equiv \tilde{v}(\mu, \lambda_{BG})} \right\} + \lambda_{BG} c_{BG}$$

note that:

1. $\tilde{v}(\mu, \lambda_{BG}) = 0$ on $[0, \hat{\mu})$
2. $\tilde{v}(\hat{\mu}, \lambda_{BG}) > 0 \Leftrightarrow \lambda_{BG} < \lambda(\hat{\mu})$
3. $\tilde{v}(1, \lambda_{BG}) > 0 \Leftrightarrow \lambda_{BG} < \lambda(1)$
4. $\lambda(\hat{\mu}) < \lambda(1) \Leftrightarrow \mu_0 < \tilde{\mu}$

where:

$$\lambda(\hat{\mu}) = \frac{(\hat{\mu} - \tilde{\mu})\mu_0(1-\mu_0)}{\hat{\mu} - \mu_0}, \lambda(1) = \mu_0(1 - \tilde{\mu})$$

The University of Texas at Austin
Department of Economics

solution: two cases, and assume $c_{BG} < 1$

case I: $\tilde{\mu} > \mu_0$ $\lambda_{BG} = \lambda(1)$ same solution as in the case with no misalignment
case II: $\tilde{\mu} < \mu_0$ then $\lambda(\hat{\mu}) > \lambda(1) > \lambda^*$ where λ^* equates slope line connecting (0,0) with $(\hat{\mu}, \tilde{v}(\hat{\mu}\lambda_{BG}))$ with that connecting (0,0) with $(1, \tilde{v}(1, \lambda_{BG}))$

value function increasing in λ_{BG} if $c_{BG} > \frac{\hat{\mu}-\mu_0}{\hat{\mu}(1-\mu_0)}$, decreasing otherwise

1. if $c_{BG} < \frac{\hat{\mu}-\mu_0}{\hat{\mu}(1-\mu_0)}$, the solution is the same as in the case of no misalignment
2. $1 > c_{BG} > \frac{\hat{\mu}-\mu_0}{\hat{\mu}(1-\mu_0)}$, minimizing Lagrange multiplier is λ^*, and the optimal splitting (that concavifies the $\tilde{v}(\mu, \lambda_{BG})$ and satisfies constraint with eq) is a split between 0 and
$$\overline{\mu} = \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-c_{BG})}$$
3. $1 \leq c_{BG}$ full info feasible and optimal

- the value λ^* pins down the correct posterior leading to "approve"

The University of Texas at Austin
Department of Economics

literature

information design / Bayesian dersuasion:

- Kamenica and Gentzkow (2011), Gentzkow and Kamenica (2016), Kolotilin (2016)
- with moral hazard: Boleslavsky and Kim (2017), Rodina (2016), Rosar (2017), Roesler and Szentes (2017)

costly state falsification:

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

The University of Texas at Austin
Department of Economics

thank you!