Research Design Memo:
Same Information, Different Source —
Does Cheap Talk Outperform Verified Disclosure in Bargaining?

Pre-registration draft

February 2026

Status: Ready for game implementation via `/create-2-player-game`

1 Research Question

Does the source of partial information about the bargaining zone — game-verified disclosure vs. the opponent's unverified claim — have opposite effects on deal rates, even when the information content is identical?

We test a *source-channel reversal*: game-verified disclosure of the opponent's approximate cost activates exploitative anchoring (the informed buyer pushes toward the seller's limit), *reducing* deal rates relative to a no-information baseline. Conversely, the same information conveyed as the opponent's voluntary, unverified claim activates reciprocal trust (the opponent "chose to share," creating cooperative obligation), *increasing* deal rates. If confirmed, this implies that the behavioral channel (reciprocity vs. exploitation) dominates the informational channel (content accuracy) — less credible information can produce better outcomes.

2 Causal Model

Variable Definitions

Testable Implications

1. **Source-channel reversal:** Deal rate ranks: opponent-claimed > no-info (baseline) > game-disclosed. Same content, opposite directions by source.

2. **Anchoring mechanism:** First offers in the game-disclosed condition are significantly closer to the seller's true cost ($50) than in the other two conditions, indicating exploitative anchoring.

3. **Reciprocity mechanism:** Concession speed (rounds to deal, conditional on dealing) is fastest in the opponent-claimed condition, consistent with cooperative obligation from voluntary disclosure.

4. **Mediation:** Controlling for first-offer anchoring attenuates the game-disclosed → lower deal rate path. Controlling for concession speed attenuates the opponent-claimed → higher deal rate path.

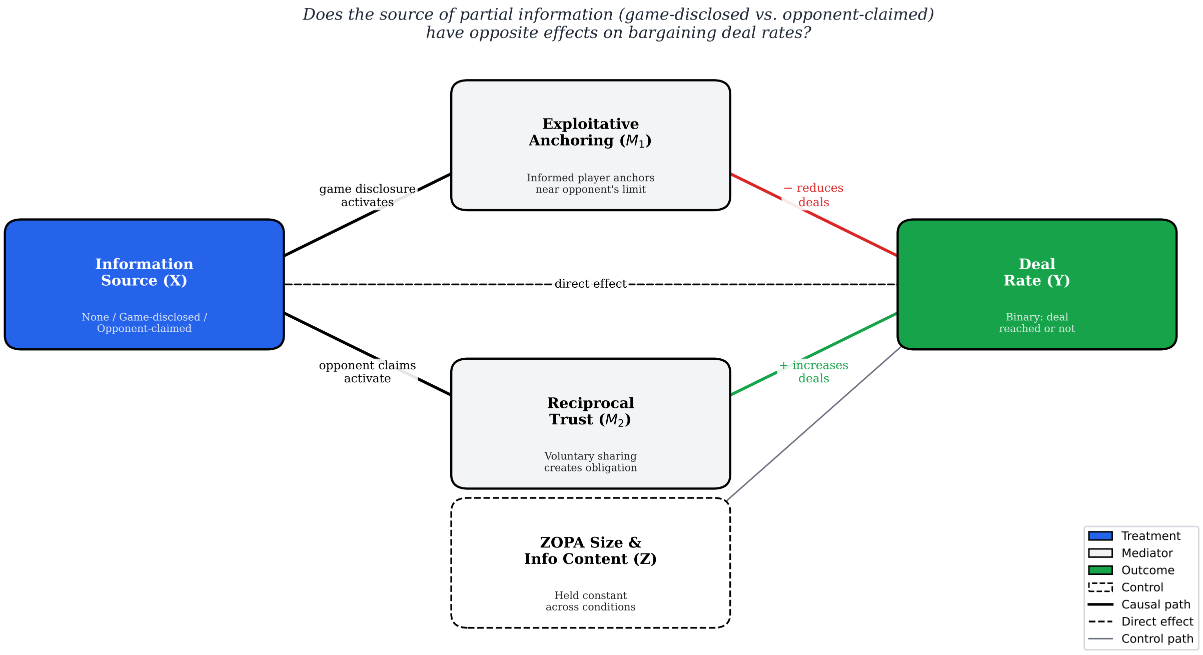

Does the source of partial information (game-disclosed vs. opponent-claimed)
have opposite effects on bargaining deal rates?
Exploitative Anchoring (M_1)
Informed player anchors near opponent's limit
game disclosure activates
− reduces deals
Information Source (X)
None / Game-disclosed / Opponent-claimed
direct effect
Deal Rate (Y)
Binary: deal reached or not
opponent claims activate
+ increases deals
Reciprocal Trust (M_2)
Voluntary sharing creates obligation
ZOPA Size & Info Content (Z)
Held constant across conditions
Treatment
Mediator
Outcome
Control
Causal path
Direct effect
Control path

Figure 1: Pearlean causal DAG. The treatment (blue) is information source: none, game-verified, or opponent-claimed. Two competing mediators (gray): game disclosure activates exploitative anchoring (M_1, red arrow, negative effect), while opponent claims activate reciprocal trust (M_2, green arrow, positive effect). Outcome (green) is deal rate. ZOPA size and information content (dashed border) are held constant. The key prediction: identical information content produces opposite effects depending on source.

Variable	**Type**	**Operationalization**	**Game Measurement**
X	Treatment	Information source: (1) none, (2) game-verified disclosure, (3) opponent's unverified claim	Three game variants; only the information channel differs
M_1	Mediator	Exploitative anchoring: buyer's first offer relative to seller's true cost	Anchoring index: $(offer_1 - V_s)/ZOPA$; lower = more exploitative
M_2	Mediator	Reciprocal trust: speed of concession across rounds	Rounds to deal (conditional on dealing); concession \$/round
Y	Outcome	Deal reached (binary); surplus captured (continuous)	Transcript: accept/reject; final price if deal
Z	Control	ZOPA size (\$30); info content ("\$45–\$55" range); AI strategy; payoffs	Identical across all three conditions

Identification Strategy

- **Randomized:** Information source (3-level between-subjects) via game variant assignment.

- **Held constant:** Buyer valuation ($80), seller cost ($50), ZOPA ($30), information content ("$45–$55"), round count (6), AI strategy (anchored concession, identical `player.md`), payoff formula. The AI does *not* change behavior based on condition.

- **Content equivalence:** In both informed conditions, the buyer learns the same range ("$45–$55"). The only difference is *who says it*: the game rules (Condition B) or the AI opponent (Condition C). This isolates the source channel.

- **Confounds ruled out:** Random assignment eliminates selection. Identical ZOPA eliminates difficulty differences. Identical AI strategy eliminates behavioral adaptation. Content equivalence isolates source from content.

- **Limitations:** (1) AI's "voluntary" disclosure in Condition C is scripted, not truly spontaneous. (2) Beliefs are unobservable — only offers reveal trust/exploitation. (3) Human–AI bargaining may differ from human–human. (4) Three conditions require ∼50+ sessions/cell for adequate power.

3 Experimental Design

Condition A: No Information (Baseline)

The buyer knows their own valuation ($80) and that the seller has some cost, but receives no information about the seller's cost. No chat phase. Standard alternating offers for 6 rounds.

Folder: `games/bargain_source_none/`

Condition B: Game-Disclosed (Verified)

Identical to Condition A, except the game rules include: *"The seller's cost is approximately $45–$55."* This is stated as a game fact (verified, certain). No chat phase. The buyer can use this information when formulating offers.

Folder: `games/bargain_source_verified/`

Condition C: Opponent-Claimed (Unverified)

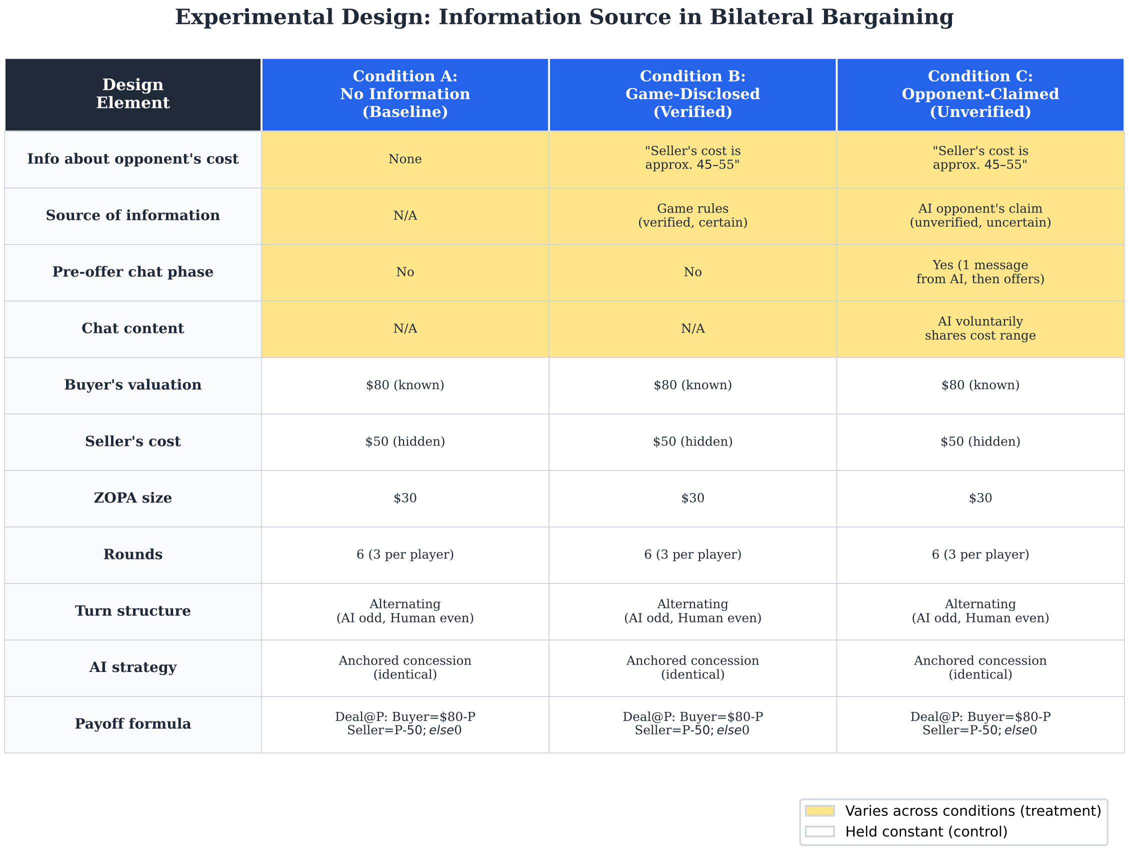
Identical to Condition A in terms of game rules (no information about seller's cost). However, before offers begin, the AI opponent sends a single message: *"I want to be upfront with you — my cost is around $45–$55."* The human then proceeds to standard alternating offers. The information content is identical to Condition B; only the source and credibility differ.

Folder: `games/bargain_source_claimed/`

Outcome Measures

4 Analysis Plan

Primary analysis: Logistic regression of deal rate on treatment condition (2 dummy variables: game-disclosed and opponent-claimed vs. no-info baseline). Test: (1) game-disclosed coefficient is negative (fewer deals than baseline), (2) opponent-claimed coefficient is positive (more deals than baseline), (3) the two coefficients differ in sign (the reversal).



Experimental Design: Information Source in Bilateral Bargaining

Design Element	Condition A: No Information (Baseline)	Condition B: Game-Disclosed (Verified)	Condition C: Opponent-Claimed (Unverified)
Info about opponent's cost	None	"Seller's cost is approx. 45–55"	"Seller's cost is approx. 45–55"
Source of information	N/A	Game rules (verified, certain)	AI opponent's claim (unverified, uncertain)
Pre-offer chat phase	No	No	Yes (1 message from AI, then offers)
Chat content	N/A	N/A	AI voluntarily shares cost range
Buyer's valuation	$80 (known)	$80 (known)	$80 (known)
Seller's cost	$50 (hidden)	$50 (hidden)	$50 (hidden)
ZOPA size	$30	$30	$30
Rounds	6 (3 per player)	6 (3 per player)	6 (3 per player)
Turn structure	Alternating (AI odd, Human even)	Alternating (AI odd, Human even)	Alternating (AI odd, Human even)
AI strategy	Anchored concession (identical)	Anchored concession (identical)	Anchored concession (identical)
Payoff formula	Deal@P: Buyer=$80-P Seller=P-50; *else*0	Deal@P: Buyer=$80-P Seller=P-50; *else*0	Deal@P: Buyer=$80-P Seller=P-50; *else*0

🟨 Varies across conditions (treatment)
□ Held constant (control)

Figure 2: Design matrix for the three-condition experiment. Yellow cells indicate elements that **differ** across conditions (the treatment manipulation); white cells indicate elements **held constant** (controls). The only differences are: (1) whether the buyer receives information about the seller's cost, (2) the source of that information (game rules vs. opponent claim), and (3) whether a pre-offer chat phase exists. All other game mechanics, payoffs, and AI behavior are identical.

Measure	**Type**	**Operationalization**
Deal rate	Primary	Binary: did the pair reach agreement within 6 rounds?
First-offer anchoring	Mechanism	Buyer's first offer as fraction of ZOPA: $(offer_1 - 50)/30$
Concession speed	Mechanism	Rounds to deal (conditional); $/round concession rate
Surplus captured	Secondary	Price minus seller cost (seller surplus); buyer value minus price (buyer surplus)
Offer variance	Exploratory	Std. deviation of human offers across rounds

Secondary analyses:

- **Mechanism — anchoring:** OLS regression of first-offer anchoring index on treatment. Expect game-disclosed $<$ baseline $\approx$ opponent-claimed.

- **Mechanism — concession:** OLS regression of concession rate on treatment (conditional on $\geq$2 rounds). Expect opponent-claimed > baseline > game-disclosed.
- **Mediation:** Causal mediation analysis (Baron & Kenny or Imai et al. 2010). Path 1: game-disclosed → anchoring → lower deal rate. Path 2: opponent-claimed → concession speed → higher deal rate.
- **Surplus (conditional on deal):** Among deals, compare price and surplus split across conditions. Expect game-disclosed deals (when they happen) to favor the buyer more (exploitative offers that happened to succeed).

Power considerations: With 3 conditions and a primary binary outcome, detecting a 15–20 percentage-point difference in deal rates (e.g., 60% vs. 45% vs. 75%) at $\alpha = 0.05$, $\beta = 0.80$ requires approximately 60–80 sessions per condition (∼200 total). With 6 rounds per session, each session provides multiple offers for mechanism analyses.

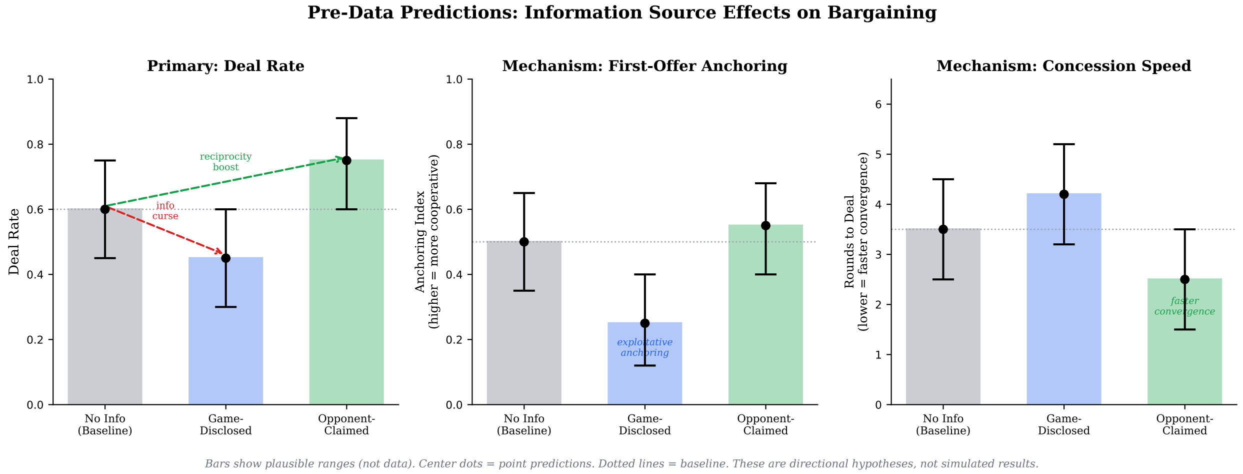


Figure 3: Pre-data predictions across three conditions. **Left:** Deal rate — game-disclosed (blue) is predicted to fall below baseline (gray), while opponent-claimed (green) is predicted to rise above it (the source-channel reversal). **Center:** First-offer anchoring — game-disclosed buyers anchor exploitatively (lower index), while baseline and opponent-claimed buyers anchor cooperatively. **Right:** Rounds to deal — opponent-claimed pairs converge faster (fewer rounds), consistent with reciprocal trust. Bars represent plausible ranges, not simulated data. Center dots are point predictions.

5 Game Implementations

Condition	Game Folder	Key Difference
A (No Info)	`games/bargain_source_none/`	No cost info, no chat
B (Game-Disclosed)	`games/bargain_source_verified/`	Game rules state cost range, no chat
C (Opponent-Claimed)	`games/bargain_source_claimed/`	AI claims cost range in pre-offer chat

Each folder contains four files (`config.toml`, `manager.md`, `player.md`, `sim_human.md`). The `player.md` (AI strategy) is *identical* across all three conditions. The `manager.md` and `config.toml` differ only in the information display rules and chat phase presence.

6 Limitations

- **Scripted disclosure.** In Condition C, the AI's "voluntary" claim is scripted into the game flow, not a genuine strategic choice. The reciprocity mechanism assumes the human perceives it as voluntary — if they recognize it as scripted, the effect may attenuate.
- **Unobservable beliefs.** We cannot measure whether the buyer *trusts* the AI's claim or *believes* the game's disclosure. We infer trust from offer behavior, which is indirect.
- **AI opponent.** Players know they face an AI, which may dampen both exploitation ("it's just a program") and reciprocity ("I don't owe a machine"). Effects may be larger with human opponents.
- **Content equivalence imperfect.** The game-disclosed range is presented as a rule ("The seller's cost is approximately \$45–\$55"); the opponent-claimed range is a conversational statement ("My cost is around \$45–\$55"). Framing differences beyond source are hard to eliminate entirely.
- **Single ZOPA size.** We test with ZOPA = \$30. The reversal may depend on ZOPA size — exploitative anchoring matters less when surplus is large. A follow-up experiment varying ZOPA would test this boundary condition.
- **Three conditions, one treatment dimension.** We cannot decompose the opponent-claimed effect into "information" vs. "social act of sharing." A fourth condition — opponent shares irrelevant information — could isolate the social channel but increases sample requirements.