Supplement: Measuring Proof Burden in Public Bounty Listings: A RentAHuman Case Study

IMAN YECKEHZAARE, Massachusetts Institute of Technology, USA and Honor Education, USA

Additional Key Words and Phrases: proof burden, RentAHuman, bounty markets, crowd work, surveillance, physical-world action, worker privacy, manual audit

Where to find the key checks. *Human Coding and Label Provenance* reports reliability, provenance, and adjudication. *Conceptual Foundations of the Taxonomy* gives the construct grounding, separates theory-derived from authorial choices, and states what this analysis can and cannot show about the scheme. *Proof Burden Score: Computation Detail* defines the score and its alternatives. *Model Specifications*, *Deviations from the Internal Pre-Analysis Plan*, and *Robustness Checks* report the models, deviations, and score-4-or-5 subsets. *Agent-Identified Listing Profile* contains the display-name, composite-sensitivity, and calibration checks. The final sections cover follow-up reachability, co-occurrence, and proposed worker validation.

Reader guide. A *proof requirement* is something a listing asks a worker to submit, show, or make available as evidence of task completion. The *Proof Burden Score (PBS)* is the study's proposed 0–5 screening summary: a 0–4 *base tier* set by the requirements a listing asks for, plus up to three *modifier* points, capped at 5; "elevated" means PBS at least 3 and "severe" means PBS at least 4. A *threshold-positive* listing is one meeting the severe threshold. A *routed* row was sent to the third coder for adjudication; an *adjudicated* label is that coder's final label. A *display-name cluster* groups listings carrying the same normalized public requester display name. OR denotes an odds ratio, CI a confidence interval, and q a multiplicity-adjusted p value. The *primary* population is the 779 content-eligible bounty/task listings; *all-returned* analyses instead retain every one of the 981 objects the public endpoints returned and are labeled sensitivities. These statistical quantities describe this nonprobability snapshot and do not make it representative of a larger population. Throughout this supplement, "agent-identified" and "human-identified" are shorthand for groups formed from platform metadata; they name the same two groups the main paper calls "agent-or-bot-labeled" and "human-labeled", and they do not verify who controlled a listing.

Large-language-model disclosure. Generative AI tools revised portions of the manuscript and supplement and assisted with analytical critique, proposed diagnostic and sensitivity checks, literature discovery, LaTeX and analysis-code development, cross-artifact consistency checking, and build and package verification. The author reviewed every retained contribution, made the final decisions about inclusion and interpretation, and accepts full responsibility for the study and its claims. The tools did not collect listings, assign labels to listing content or study variables, or adjudicate labels.

Supplement and data availability. Only the supplementary PDF accompanies this paper. It contains additional methods, aggregate results, and redacted examples, but excludes listing-level data, raw task text, requester identifiers or hashes, follow-up records, and coder or adjudicator packets. Researchers interested in follow-up work are invited to contact the author to discuss collaboration or access to appropriately minimized, disclosure-reviewed materials. Any sharing would be considered case by case and would remain subject to privacy, legal, institutional, licensing, and platform-policy constraints.

Author's Contact Information: Iman YeckehZaare, Massachusetts Institute of Technology, MIT Center for Collective Intelligence (CCI), Cambridge, Massachusetts, USA and Honor Education, Research, San Francisco, California, USA, oneman@mit.edu, iman@honor.education.

1 Public-Only Collection Protocol

The collection protocol uses only public listing pages and unauthenticated public API endpoints. The study does not create accounts, log in, apply to tasks, contact workers or requesters, join private channels, or collect private messages.

For RentAHuman, broad unauthenticated public queries returned authentication errors when requesting paginated result pages after page 1. The collection script therefore queries public bounty listings by status and by status-category partitions, deduplicates rows by platform identifier, and records the query path used for each raw response. Every recruiting partition returned HTTP 400; one recruiting listing appeared on a broad first page, so recruiting listings beyond that observed row may be absent. For Human Pages, the script queries the public listings endpoint. A public GoHireHumans REST endpoint was checked during collection and returned no jobs. The frozen pre-coding snapshot returned one open Human Pages listing and no GoHireHumans jobs. An auxiliary capture from the same Human Pages endpoint about six hours earlier had returned 15 open listings, so that market's visible inventory changed substantially over that six-hour gap; the analysis frame is the frozen snapshot, to which Human Pages contributes one row. These counts reflect the two smaller markets' small and fast-changing inventories rather than collection gaps. In a recheck on August 12, 2026, during camera-ready preparation, the Human Pages public listings page displayed "No open listings right now" and the GoHireHumans jobs page displayed four postings, each posted by the platform's own operations account and dated July 2026—after the snapshot window, consistent with the collection-day record of zero jobs.

RentAHuman is the empirical center of the study. The public listing fields available in this snapshot support measurement of advertised task requirements: listed price, status, category, free-text descriptions, and sometimes structured evidence or completion criteria. For RentAHuman, the processed-dataset builder joins each detail record to its matching public list record by platform identifier; the list record supplies the visible application count. Human Pages exposes that field directly. It is populated for all 981 listings and positive for 464, supporting the separate diagnostic in Model Specifications. These fields do not support claims about completed work, final payment, rejection, worker choices, or post-completion dispute resolution. The one Human Pages row is retained as a small auxiliary source that enters full-snapshot prevalence estimates and full-sample RQ3 calculations as a singleton; one row cannot support a cross-platform comparison.

The processed snapshot spans 20 RentAHuman public categories. The largest platform-provided category labels are "other," "marketing campaigns," "creative media," "research fieldwork," "hiring," and "tech-dev." Across the harmonized source records, 870 rows are labeled human, 71 RentAHuman rows carry a named agent or bot type, and the Human Pages singleton has a row-level nested `agent/agentId` relation rather than an agent-type field. Those 72 rows form the agent-identified group; 39 rows carry an ambiguous "other" label. The planned content-eligibility screen, applied row by row during the full protocol recode described in *Human Coding and Label Provenance*, leaves 779 content-eligible bounty/task listings as the primary population (all 72 agent-identified rows, 676 human-identified rows, and 31 ambiguous "other" rows); 202 returned records were excluded as service offers (117), records requesting no human action (61), or promotional posts (24), and enter only the all-returned sensitivities. The requester-type fields and agent relations above are too coarse to distinguish who designs, benefits from, or supervises a task. Public status is similarly limited as an outcome measure: the snapshot mixes cancelled, completed, open, assigned, and partially filled listings without longitudinal status history. For joined RentAHuman rows, the processed timestamp and source-URL hash describe the selected detail response, not necessarily the list response that supplied application and spot fields; that field-level list provenance remains only in the withheld raw captures and manifests. The public fields do not reveal whether

completion evidence would later be uploaded through the platform, sent through an external channel, reviewed by a person, reviewed automatically, retained after payment, or disputed by workers.

Manifest rows record source platform, raw response path, source URL hash, UTC collection timestamp, HTTP status, scraper revision field, and collection notes for 2,827 captures. The earliest RentAHuman run also retains 66 detail captures that have no manifest row. A post-hoc integrity index over the raw captures hashes and counts those files as `retained_unmanifested`; it does not assign them invented timestamps, URLs, statuses, or collector metadata. The core RentAHuman manifests record `scraper_commit_hash=NO_COMMIT`, so the exact historical collector revision is unavailable. Raw captures are not intended for public release because task text can contain personal data, direct links, social handles, addresses, phone numbers, or instructions that should not be amplified.

2 Processed Data and Redaction

The controlled internal processed dataset is redacted so routine analysis files expose no task wording. The privately retained CSV and JSON versions contain one row per public listing. Identifiers and source URLs are deterministic SHA-256 hashes of public platform identifiers and URLs, not anonymous study identifiers. The repeated-signature key—which groups listings sharing a title, category, price, and currency—is also an unkeyed digest of normalized title and public metadata. These inputs can be guessed or enumerated, and public quasi-identifiers remain. Creation time alone is unique for 981/981 rows. After removing timestamps, 822/981 remain unique on platform, category, status, listed price format, spots, application count, requester label, and evidence-type text. Rekeying hashes therefore cannot make the microdata suitable for an unrestricted archive, and this supplementary PDF excludes them. Columns named for raw title, description, and requirement fields remain, but every value is a constant placeholder; per-row field presence and length are instead preserved in separate description- and requirement-word-count columns. Full raw text is retained privately and does not accompany this PDF.

Privately retained documentation includes a processed-data schema and a structured source-dataset datasheet. Following the Datasheets for Datasets framework [11], the datasheet documents motivation, composition, collection, labeling, intended and out-of-scope uses, and distribution. The key analytical fields are the 13 coded variables (11 proof-evidence types, recurring monitoring, and physical-world action), Proof Burden Score (PBS), the coder-recorded score components, the content-eligibility decision and structured exclusion reason, rule-signal provenance, proof-signal source, and the legacy `keyword_only_flag` for title/description action-phrase-only rule signals.

3 Human Coding and Label Provenance

All 981 returned rows carry final human-coded labels from the full protocol recode promoted on 2026-08-10. That recode, not the original June 2026 audit, is the source of every label, score, and label-based result in the paper and this supplement; the original audit is retained as a superseded first round and is summarized at the end of this section. Two newly recruited coders, working independently, coded every row under the version-11 codebook, whose criteria were operationalized through a pilot that itself surfaced the original codebook's missing definitions (several flagged by the coders). The codebook was frozen at version 11 before coding; the only later revision (version 12) adds the phone-only base-tier policy described in *Proof Burden Score: Computation Detail*. The recode packets supplied the complete listing context—title, description, requirements, required skills, application-stage link fields, completion/evidence criteria, start instructions, and evidence types, with category and status as context—that the first round's packets had partly omitted. Price, application count, and requester identity were not to determine a proof label, and the codebook required

a completion-linked role rather than a keyword alone; for example, location metadata did not itself establish location proof.

The coders and the third adjudicator accessed raw listing text only under a restricted, local-only protocol, and raw packets were never shared or committed. Coders were blinded to the rule-based labels and to the superseded first-round labels, so their judgments were not anchored to the automated extraction or to the earlier study coding. This blinding applies to rule and prior-study outputs, not to the general study aim. The author additionally asked both coders in writing not to use any LLM or AI assistance under any condition in this project, and both confirmed by reply before coding. That prohibition was delivered in a covering message rather than in the instruction documents themselves; the delivered instruction files therefore differ from the packet-bound instrument versions by one removed clause, a deliberate deviation preserved in a private instrument-deviation record, and the delivered prohibition is broader than the removed clause, not weaker.

Recruitment and compensation. The two recode coders and the blinded adjudicator were doctoral students recruited on Upwork under preset screening filters: based in the United States; English listed as native or bilingual; available at the time of hiring; Top Rated Plus status (Upwork's tier for freelancers with a sustained record of highly rated work, chosen to select for careful attention to requirements); and independent freelancers rather than agencies. We set no time limit on the coding and the coders used that latitude, so we imposed no incentive to trade care for speed; we paid an average of $37 per hour across the three. Hours were billed through Upwork's hourly time tracking with manual time entry disabled; the platform records periodic screenshots during billed time, all three were informed in advance that we would review those records, and we reviewed all of them. We note the reflexivity: this billing arrangement is itself an instance of the workplace monitoring this paper studies. We cite the record only for what it can support: it indicates the hired freelancers performed the coding and adjudication personally, and it is consistent with their written confirmations that no LLM or AI assistance was used.

The recode also applied the internal plan's content-eligibility screen row by row. Each coder recorded a structured eligibility decision and exclusion reason for every row, with disagreement or uncertainty routed to adjudication. After adjudication, 779 listings are content-eligible bounty/task listings and form the primary population; 202 returned records are excluded as service-offer listings (117), records requesting no human action (61), or promotional posts (24; coded `advertisement`). Proof labels were still coded on excluded rows, so all-returned quantities are labeled sensitivities computed from the same coding exercise rather than a separate protocol. A coder-facing ethical retention criterion allowed either coder to request that a row be withheld; no exclusion resulted.

The instructions used context to decide whether a requirement was completion-linked, not to judge how much burden it added. They supplied no field, weight, or decision rule for the portion of a requirement that is incremental to the task a worker had already accepted, and they defined the score as exposure burden rather than task difficulty or pay. The consequence is that identical labels and identical points can describe very different lived demands. Traveling to a place solely to produce location or photo proof for work that is otherwise remote, and filming a summit that a worker was already hiking to, are both coded as physical-world action together with photo or video proof, and both receive the same rubric points. PBS ranks advertised exposure channels; it does not estimate the increment over what the task itself already requires, and no coded field recovers that increment retrospectively. The recode worksheets do record structured location, work-mode, recurrence, bounty-kind, submission, and review context. Those fields narrow this gap but do not close it, because the incremental question concerns the worker's counterfactual rather than the amount of listing text available.

The processed dataset retained the rule-generated provenance value for later provenance checks, but coders did not see the rule labels while coding. Coders recorded each of the 11 proof-evidence variables, recurring monitoring, physical-world action, the Proof Burden Score with its base-tier and modifier components, the eligibility decision, and ambiguity/sensitivity and UNCLEAR flags. Both coders' completed files passed row-order, identifier, schema, and value validation checks. A blinded third adjudicator then reviewed every routed row: any coder disagreement on any recorded field, any ambiguity/sensitivity flag, any UNCLEAR entry, any low-confidence entry, and any difference from the superseded first-round labels all routed a row (the groups overlap). The adjudicator saw the raw listing text and both coders' answers but neither the rule-based labels nor the superseded study labels, and assigned final labels by the version-11 codebook (the step-by-step procedure is in Section 10). Across all 981 returned rows, 873 were routed and 108 carried unflagged exact two-coder consensus; within the primary population the split is 747 routed and 32 consensus (Table 4). Because a difference from the superseded labels was itself a routing trigger, routed does not mean disputed between the two recode coders.

Pre-adjudication reliability. Across all 981 returned rows, the two recode coders recorded the same Proof Burden Score entry for 81.8% of listings (83.0% of the 963 rows where both recorded a numeric score; Table 2), with quadratic weighted $\kappa = .901$. Agreement was 92.2% ($\kappa = .84$) on the proposed elevated threshold (score 3 or higher), 90.2% ($\kappa = .804$) on the proposed severe threshold (score 4 or higher), and 90.3% ($\kappa = .799$) on the capped maximum. Table 2 gives the per-variable agreement shares and κ values; every one of the 13 coded binary variables has $\kappa \geq .76$. Reliability improved on twelve of the thirteen variables relative to the superseded first round and held on the thirteenth (link proof, $\kappa = .94$). Disagreements were most frequent for text proof (82 listings), recurring monitoring (62), and identity proof (61), and rarest for screenshot, video, and photo proof (1, 7, and 13).

Choice of agreement statistic. With two fixed coders and no missing labels, Cohen's κ (binary variables) and quadratic-weighted κ (the ordinal score) are the conventional statistics for this design. The main alternatives make no material difference here. Krippendorff's α reduces in this design, up to a small-sample factor, to Scott's π, which differs from κ only through differences in the two coders' marginal rates; those marginals are close for every variable, and recomputing π from the aggregate reliability file behind Table 2 changes no feature's value by more than 0.002. Prevalence-robust alternatives such as Gwet's AC1 matter most when κ is depressed by rare categories; κ is conservative in that direction, so the reported values understate rather than overstate raw agreement for the rarest features. The table reports raw agreement and κ per variable; the aggregate reliability file behind it adds disagreement counts and each coder's positive counts, so any of these statistics can be recomputed from aggregates alone.

The weakest variables are now recurring monitoring ($\kappa = .76$) and location proof ($\kappa = .79$). Two of the three largest priority groups among threshold-positive listings are anchored by identity proof ($\kappa = .83$) and recurring monitoring ($\kappa = .76$), so the severe profiles still rest partly on the taxonomy's less reliable boundaries, although far less so than in the superseded round. The data do not establish why those boundaries remain harder: feature prevalence, listing complexity, coder differences, and codebook boundaries are all possible contributors. The retained adjudication-note templates identify broad boundary areas but were not systematically open-coded as causes of disagreement. Notes are nonblank for 534 of the 873 routed rows, span 309 distinct templates, and are blank for 339 rows. We therefore report where coders disagreed, not a complete qualitative explanation of why.

Under the promoted labels, 492 primary listings (63.2%) meet the proposed elevated threshold, 438 (56.2%) meet the proposed severe threshold, and 372 (47.8%) are at the capped maximum of 5. The final label set differs from the automated first-pass scores described in the *Proof Extraction Rules* and *Rule-Signal Provenance* sections on 459 of the 981 returned rows (373 of the 779 primary rows).

Because routing was broad—including routing triggered only by a difference from the superseded labels—nearly all final labels are adjudicator-issued: 747 of the 779 primary rows (95.9%) carry third-adjudicated labels. Essentially every threshold-positive label is therefore adjudicator-mediated. Separately, Table 1 shows that 83.8% of severe-threshold-positive listings have high-confidence rule-signal provenance, meaning the first-pass extractor saw a structured or proof-specific signal rather than relying only on title/description action phrases. That provenance supports traceability, not label correctness. Similarly, the main paper's priority-group table reports the inter-coder κ of each group's first-named variable, not a reliability statistic for the compound group.

Table 1. Elevated, severe, and capped-maximum labels in the primary population, with the share backed by high-confidence rule signals. The adjudicated and exact-consensus columns count the promoted recode's provenance: nearly every high-scoring listing was resolved by the blinded adjudicator, consistent with the broad routing rules; the full routed-versus-consensus split is in Table 4.

Definition	Listings	Adjudicated	Unflagged exact-consensus	High-confidence rule signal
Proposed elevated threshold (≥3)	492	473 (96.1%)	19 (3.9%)	416 (84.6%)
Proposed severe threshold (≥4)	438	423 (96.6%)	15 (3.4%)	367 (83.8%)
Capped maximum (score 5)	372	361 (97.0%)	11 (3.0%)	319 (85.8%)

Table 2. Two-coder reliability for all proof and context variables, the score, and thresholds under the promoted protocol recode, before adjudication, on all 981 returned rows. Binary variables use Cohen's κ; the Proof Burden Score row uses quadratic weighted κ and conditions on the rows where both coders recorded a numeric score; threshold rows are binary κ at the stated cutoffs [6, 7].

Variable / threshold	Agreement	κ
Text proof	91.6%	0.83
Link proof	97.8%	0.94
Screenshot proof	99.9%	1.00
Photo proof	98.7%	0.97
Video proof	99.3%	0.98
Identity proof	93.8%	0.83
Account proof	97.3%	0.92
Location proof	96.7%	0.79
Phone proof	98.5%	0.81
Financial proof	98.3%	0.89
Public-post proof	98.3%	0.92
Physical-world action	98.3%	0.96
Recurring monitoring	93.7%	0.76
Proof Burden Score	83.0%	0.90
Proposed elevated threshold (≥3)	92.2%	0.84
Proposed severe threshold (≥4)	90.2%	0.80
Capped maximum (score 5)	90.3%	0.80

Table 3 distinguishes adjudication routing from disagreement about the severe threshold, over all 981 returned rows. Among the 456 all-returned final threshold-positive rows, both coders had already met the threshold for 427, exactly one had for 27, and both coders were below it for 2. The final rate therefore does not arise from converting a large both-coders-below set, although the defining feature profile remains adjudicator-mediated. Restricted to the 779 primary listings, the coders' pre-adjudication scores classify 461 (59.2%) and 456 (58.5%) listings as severe, versus 438 (56.2%) under the final adjudicated labels; these are the coder-level rates cited in the main paper.

Adjudicator decisions. The adjudicator's decisions can be characterized directly (Table 4). On contested cells, the final labels sided with coder 1 and coder 2 almost equally (55% versus 45%), so they do not simply mirror either coder; across all adjudicated cells, the adjudicator matched both coders on 95.4%, coder 1 alone on 2.4%, coder 2 alone on 1.9%, and neither on 0.3%. The adjudicator flagged 356 rows for author review of retention or sensitivity; every one was retained after that review, with final labels unchanged. The adjudicator's separate confidence field is high for 788 of the 873 routed rows, medium for 70, and low for 15 (703, 37, and 7 within the primary population). This field is the adjudicator's ordinal self-rating of how confidently the final coding vector followed the codebook; it is not calibrated against another adjudicator. These figures quantify, rather than remove, the single-adjudicator dependence discussed in the main paper's limitations. Separately, the institutional IRB determined that this public-data study was not human-subjects research; therefore, IRB review was not required. This determination is neither IRB approval nor an exemption.

Five rows whose only base-setting feature was adjacent to phone-only contact were initially marked `UNCLEAR` under the codebook's fail-closed phone-only rule; the three exact phone-only rows among them carry a prospectively author-approved tier-3 base adopted on 2026-08-10 as a new policy, as described in *Proof Burden Score: Computation Detail* and in the *Deviations* section.

Table 3. Both pre-adjudication coder classifications versus the final proposed severe-threshold classification. Counts cover all 981 listings; unrouted rows have identical coder and final classifications by construction.

Final rubric classification	Both coders below	Exactly one coder meets	Both coders meet
Below severe threshold	453	69	3
Meets severe threshold	2	27	427

Table 4. Adjudication routing and the adjudicator-confidence field for the 779 primary rows under the promoted recode. Routing triggers were any coder disagreement; an ambiguity/sensitivity flag, UNCLEAR entry, or low-confidence entry from either coder; or a difference from the superseded first-round labels.

Adjudication route or confidence field (routing over $n = 779$ rows; confidence over the 747 routed)	Rows	Share
Recorded adjudication route/outcome		
Routed to the blinded third adjudicator	747	95.9%
Exact two-coder consensus, no routing	32	4.1%
Adjudicator self-rated confidence		
Self-rated high	703	94.1%
Self-rated medium	37	5.0%
Self-rated low	7	0.9%

The superseded first round. The originally submitted paper's labels came from a June 2026 audit: two research-team coders, a third adjudicator, and packets that blanked most skill context and omitted an application-stage link field. The first round's coders were recruited on Upwork under the same screening filters and comparable pay as the recode personnel described in *Recruitment and compensation* above. That round's pre-adjudication reliability, on the same 981 returned rows, was materially lower (score exact agreement 67.9%, quadratic weighted $\kappa = .745$; severe threshold 82.4%, $\kappa = .628$; identity, location, and recurring monitoring between $\kappa = .50$ and .54), and several of its codebook criteria were never operationally defined. The recode, coded with full context under the piloted version-11 codebook, changed the label or score of 78% of rows and corrected two errors in the originally submitted figures: location proof had been conflated with physical-world action (the two overlapped on 91% of the first round's location positives; the combined physical-or-location share moves only from 37.0% to 38.9% on the all-returned rows), and a commissioned deliverable had been counted as completion proof (it now sets no base tier by itself, and the labels stay descriptive). The first round's artifacts—its label set, reliability and threshold tables, adjudication-note aggregates, and sensitivity outputs—are preserved unchanged under `legacy` names as the historical record and are not renumbered or deleted.

Two first-round records still bear on how this supplement reads. First, that round designated 50 rows for calibration coding, and the retained record cannot establish whether the calibration review occurred as described or which instruction version governed which rows; a private protocol-version map records the surviving byte-exact snapshots, SHA-256 digests, and these limits. Throughout this section, "calibration" refers to that first-round coder-training exercise, not to the statistical calibration of the sign-flip test; the 28 calibration-consensus rows cited in the main paper are the calibration rows whose labels rested on coder consensus without a documented post-discussion recode. Table 5 preserves the first round's key reliability quantities after excluding those 50 rows; it describes the superseded coders, not the recode. Table 6 shows that the promoted headline results barely move when those historically designated rows are excluded. Second, the first round's workflow marked 250 rows with a broad internal sensitivity-or-uncertainty flag (`needs_pi_review`); the name is neither an IRB designation nor evidence of exclusion. Table 7 preserves the disclosure-reviewed aggregate of that round's nine private note templates. The promoted dataset does not carry that flag; the recode's own author-review flag (356 rows, all retained) replaces it.

Table 5. Superseded first round: key two-coder reliability quantities of the original June 2026 audit after excluding its 50 designated calibration rows. Historical record only; the operative labels and reliability are the recode's (Table 2).

Measure	n	Agreement	κ	Disagreements
Proof Burden Score	931	66.7%	0.736	310
Proposed elevated threshold (≥3)	931	86.6%	0.687	125
Proposed severe threshold (≥4)	931	81.6%	0.615	171
Score 5	931	84.3%	0.684	146
Identity proof	931	80.2%	0.483	184
Location proof	931	81.4%	0.517	173
Physical-world action	931	84.4%	0.635	145
Recurring monitoring	931	88.7%	0.522	105

Table 6. Promoted-label headline results after excluding the first round's historically designated calibration rows. The consensus scope removes only the designated calibration rows whose promoted labels rest on unadjudicated exact coder consensus (one listing in the primary population); the 28 calibration-consensus rows cited in the main paper are the superseded first round's consensus-provenance category, defined in the text above. No inferential test is reported because these exclusion scopes were not included in the simulation-based test calibration.

Scope	Listings	Score 4 or 5	Agent composite	Human composite
Full snapshot	779	438/779 (56.2%)	54/72 (75.0%)	374/676 (55.3%)
Exclude all 50 designated calibration rows	735	414/735 (56.3%)	52/68 (76.5%)	354/637 (55.6%)
Exclude unadjudicated calibration-consensus rows	778	437/778 (56.2%)	54/72 (75.0%)	374/675 (55.4%)

Table 7. Superseded first round: disclosure-reviewed aggregate of the nine private adjudication-note templates among that round's 250 internally flagged rows, with that round's severe counts. Historical record only. The first category is disjunctive—sensitive *or* unresolved—and cannot identify a clean unresolved subset. No note text or row-level category mapping appears in this PDF.

Private-note category	Rows	Severe	Agent	Human	Other
Sensitive or unresolved context	60	53	0	56	4
Identity/account/telecom credentials	56	54	2	51	3
Financial or crypto exposure	44	31	5	38	1
Live-animal physical-world task	22	22	5	17	0
Representation/confidential document	19	5	11	7	1
Regulated medical delivery	17	16	2	15	0
Private-home evidence	15	14	0	15	0
Security/recovery-related task	15	5	0	11	4
In-person identity/location exposure	2	2	0	2	0

4 Proof Extraction Rules

The preceding sections describe the current adjudicated labels used throughout the paper; this section and the next document the rule-based first pass, which supplies rule-signal provenance and the rule-versus-audit comparison, not the labels used in the substantive analyses. All rules are deterministic: exact matches on the platform's structured evidence-type values plus hand-written, case-insensitive regular expressions; no machine learning, embedding similarity, or generative AI is involved. The extractor separates 11 proof-evidence variables, recurring monitoring, and physical-world action. Proof-evidence variables are text proof, link proof, screenshot proof, photo proof, video proof, identity proof, account proof, location proof, phone proof, financial proof, and public-post proof. Recurring monitoring captures repeated checks over time. Physical-world action records whether the listing asks the worker to act in a physical setting; it is coded separately because the action may affect exposure without itself being an artifact submitted as proof.

This rule set distinguishes three evidence sources:

Structured evidence labels Platform-provided evidence labels, when present.

Proof-specific fields Evidence criteria, completion criteria, requirements, and start instructions.

Title/description action phrases Title or description phrases that place a proof term within 90 characters of an action verb (submit, provide, send, share, upload, attach, show, include, confirm, verify, record, film, bring, keep, enter, or deliver), in either order.

Category labels, required skills, location metadata, and work-mode labels are not sufficient by themselves to trigger proof-evidence variables. This is a deliberate validity safeguard: a task can involve a camera, account, place, or payment without requiring completion-linked evidence about that element.

5 Rule-Signal Provenance

The generated rule-signal provenance labels are:

High At least one proof signal comes from structured evidence labels or proof-specific fields.

Medium Proof is triggered only by title/description action phrases.

None The rule-based extractor found no proof signal.

Of the 981 returned rows, 681 have high-confidence rule signals, 68 have medium-confidence rule signals, and for 232 the rule-based extractor found no proof signal. The medium-confidence rows are internally flagged by the legacy variable name `keyword_only_flag`, but the flag means that the rule signal came only from title/description action phrases rather than structured or proof-specific fields. They are not discarded because some platforms express proof requirements only in free text; they were therefore included in the human audit.

6 Conceptual Foundations of the Taxonomy

6.1 Why Three Families, and Why They Are Coded Separately

The taxonomy separates three questions about one advertised object—the requirement a requester states as a condition of payment: *what* is disclosed or produced (completion evidence), *for how long* the requester retains a claim on the worker (recurring monitoring), and whether the worker must act in a physical setting (physical-world action). They are coded separately because a listing can vary on one while the others are held fixed. A submitted photo, a weekly check-in, and a trip to a named address are not substitutes: the first fixes a modality, the second a temporal structure, the third a setting. Collapsing them would make a photo taken at a specified place and repeated over a week indistinguishable from one photo taken at home. Distinctness here is a claim about what coders record, not an empirical claim that the families are separable latent factors; we ran no dimensionality analysis, and the co-occurrence lift in Figure 1—how often two features appear together relative to the rate expected if they were independent—is descriptive.

6.2 Relation to Privacy Constructs

Solove's taxonomy of privacy [31] treats privacy problems as a family of distinct activities rather than one quantity. We use that structure as a sensitizing analogy, not a one-to-one mapping. Identity or account proof can create identification risks; phone or financial proof can do so when it contains a persistent identifier. Depending on their content and audience, photos, videos, public posts, screenshots, links, location evidence, or repeated checks can increase accessibility, reveal incidental context, or support aggregation or surveillance. The listings do not show what workers actually submitted, who viewed it, how long it was retained, or whether it was reused. The features therefore mark possible pathways to exposure, not observed privacy harms. Text proof remains a low-exposure reference category, not a claim that text is harmless in every setting.

Contextual integrity [27] supplies the reason burden context is coded at all. The relevant question is not whether information is sensitive in the abstract but whether its flow—sender, recipient, subject, information type, transmission principle—fits the norms of the context. An advertised completion requirement fixes several of those parameters in public text: the recipient (the requester), the information type (the artifact), and the transmission principle (submit this in order to be paid). It does not fix retention, onward sharing, or review, which is why we describe coded features as exposure *requests* rather than transfers. Contextual integrity also implies that one modality is not one thing: a photo of a public storefront and a photo of the worker's residence differ in flow, not in artifact type. Because listing text often

fixes the setting without making the setting the evidence, the setting is recorded as context rather than as a further proof-evidence type.

6.3 Relation to Surveillance and Administrative-Burden Constructs

Surveillance research motivates the temporal separation. A one-time artifact and a request for repeated verification differ: repeated checks can extend a requester's claim on a worker's time or visibility. The listings do not reveal how long such checks continue, whether they continue after another artifact is delivered, or how they are reviewed. Monitoring experiments report reduced willingness to accept tracked work [17]; gig-sector studies report that greater algorithmic surveillance tracks lower trust and higher privacy concern [30, 34]; worker-centered sousveillance treats observation as a design parameter rather than a byproduct [10]; and platform-scale accounts situate task-level monitoring within broader behavioral extraction [38]. Recurring monitoring is therefore its own variable and its own tier-4 base, not a modifier on artifact type.

Administrative-burden research [25] supplies the closest available vocabulary for the demand side, and we use it with an explicit mismatch. It decomposes citizen–state encounters into learning, compliance, and psychological costs. Our coded features are advertised *demands* whose nearest analogue is compliance cost: documentation and action required as a condition of receiving something. Recurrence extends that cost over time, and identity or public-post requirements are the ones most plausibly carrying psychological cost. We measure none of the three. We do not observe how long a worker spent, what they had to learn, or how the requirement felt; and the settings differ in ways that matter, since a benefits applicant typically faces a monopoly provider whereas a worker can decline a listing. We take the learning/compliance/psychological decomposition as a reminder that an advertised requirement is not a measured cost, not as a validated instrument transported to this setting. The representative literature reviewed for this study did not identify a prior application of administrative-burden theory to bounty-listing proof requirements. This limited review cannot establish an exhaustive novelty claim, so we present the connection as an analogy to be tested rather than an inherited, validated construct.

6.4 Relation to Crowd-Work Constructs

Crowd-work research explains why the unit is the requirement rather than the task or the price. Quality-control scholarship treats verification as the requester's problem—redundancy, gold questions, acceptance rules—and proof of completion appears there as a design option whose cost to the worker is not the object of study [13, 14]. Earnings research shows that listed price omits search, screening, waiting, and rejection risk [12, 33], and unclear acceptance criteria make evaluation itself a worker risk [24]. Proof requirements may be another part of that gap: they are stated as conditions of payment but are not separately itemized in the listed price or represented by category alone. A requester may nevertheless have considered them when setting the price; these data cannot show whether or how. Studies of worker disclosure show workers already trade pay against sensitivity and rejection risk [29, 37], and workflow interventions show task design can bound sensitive input [15], which is why we code the requirement as designed rather than the platform as a whole. Field and spatial crowdsourcing supply the physical and locational dimensions [1, 32].

6.5 Why One Framework, and Which Choices Are Ours

The families are combined because payment can depend on several requirements at once, not because those requirements are interchangeable or reflect one hidden trait. PBS is therefore a *formative index*: its selected features define

the total, whereas items in a reflective scale are treated as effects of an underlying trait [9]. Two consequences govern how PBS should be read. First, internal-consistency reasoning does not apply—low correlation among features is not evidence against the index, and high correlation is not evidence for it. Second, dropping an indicator changes the construct rather than sampling it less precisely, so the ablations in Table 13 are sensitivity diagnostics on a definition, not reliability estimates. Combination into one number remains the weakest part of the framework, and the feature labels can be used without it.

Three elements are grounded in prior work: treating exposure as a set of distinct activities rather than one quantity (Solove); separating setting from artifact (contextual integrity); and separating a one-time artifact from a repeated obligation (surveillance research). By contrast, all numerical scoring rules are authorial conventions. The base-tier ordering encodes the author's judgment about potential reach, linkability, persistence, and recurrence. Additivity was preferred to a maximum rule because a listing stacking several channels was judged to differ from one using a single channel at the same tier, and equal +1 steps were preferred to weighted or multiplicative combination because no basis existed for choosing weights. Equal increments are a simple and transparent convention for making stacking visible, not a uniquely neutral choice or a claim that the modifiers matter equally to workers. The cap at 5 is inherited from the historical screening rubric and keeps the score on a short communicable range; it has no external criterion and costs information, placing 47.8% of primary listings at the ceiling, which is why uncapped diagnostics and profiles are reported beside it. The retained materials do not document why the cutpoints fall at 3 and 4 rather than elsewhere; retrospectively, elevated marks listings whose base is already high-tier or whose modifiers have accumulated, and severe marks the top two attainable bands. No worker judgment, outcome, or platform policy was used to place either. Table 8 states this choice by choice.

6.6 What This Analysis Can and Cannot Tell Us About the Taxonomy

The retained plan and codebook show that the taxonomy's 13 variables and base tiers were fixed before human coding. The retained historical materials do not, however, document a contemporaneous literature-to-category derivation. The theoretical mappings above are therefore retrospective conceptual grounding, not evidence of how each category was originally selected. There was no inductive or open-coding pass over this snapshot, and no category was added, merged, or dropped in response to what the data contained. The analysis applies the fixed taxonomy rather than deriving it from these listings; the following limits are structural rather than incidental.

The coding process could assign a value for every one of the 13 variables to every listing. That demonstrates executability on this snapshot, not that every category is appropriate, valid, or exhaustive. The analysis can show how often each feature appears and which combinations dominate the threshold-positive set. Through pre-adjudication agreement it can show which definitions two independent coders applied more or less consistently in this exercise: under the promoted recode, screenshot ($\kappa = 1.00$), video (.98), photo (.97), physical-world action (.96), and link proof (.94) are at or near ceiling, while recurring monitoring (.76), location (.79), and phone proof (.81) are the weakest, and in the superseded first round the location, recurring-monitoring, and identity boundaries fell to $\kappa = .50$–$.54$. That pattern is consistent with some boundaries being harder to apply, but it does not identify a cause: prevalence, listing complexity, or differences between coders could also affect κ. Some variables carrying weight in the threshold-positive profiles are among those with comparatively lower agreement. The author-arranged co-occurrence display also shows stronger within-group lift for a physical-world/location-adjacent group and a digital/account-linked group, but it is not an independent discovery of those groups.

The analysis cannot show that the categories are exhaustive. Coders assigned only from the fixed scheme and no residual or "other" proof category was recorded, so nothing in the design could surface a proof type the codebook omitted. It cannot establish that the three families are empirically separable dimensions: co-occurrence lift is descriptive, no factor or dimensionality analysis was run, and for a formative index such an analysis would not be the right evidence in any case. It cannot validate the tier ordering, the equal-weight modifiers, the additive form, the cap, or the two cutpoints, none of which was estimated from anything. It cannot speak to worker-perceived burden, which requires evidence external to the coding process [8]. And it cannot establish that the taxonomy travels: one platform, one collection window, one nonprobability snapshot.

The appropriate next check on the taxonomy is inductive rather than confirmatory. The completed full-context recode improved reliability but still assigned only from the fixed scheme; a coding pass that admits an open residual category would let the data indicate what that scheme misses. The present design cannot.

7 Proof Burden Score: Computation Detail

The 0–5 rubric and the two thresholds (elevated, score 3 or higher; severe, score 4 or higher) are defined in the main paper; this section gives the scoring guidance coders applied. Under the promoted recode, coders assigned PBS by hand and recorded its components directly: the base tier, each modifier, the uncapped total, and the final capped score. The analyses use the final adjudicated score as the recorded study value. "Final" does not mean externally authoritative or construct-validated, and the score is not determined mechanically by the binary labels. This is a change from the superseded first round, whose coders did not store components, leaving only score-informed post-hoc reconstructions. For a listing, let the *base* be the highest applicable tier under Table 9, including its tier-4 combination rules; the nominal rubric total is

$$\mathrm{PBS}_{\mathrm{nominal}} = \min\big(5,\ \mathrm{base} + m_1 + m_2 + m_3\big),$$

where $m_1 = 1$ if more than one of the eleven proof-evidence types is present, $m_2 = 1$ if the listing involves physical-world action and carries any proof-evidence or recurring-monitoring requirement, and $m_3 = 1$ if the listing requires public posting or account use; each modifier contributes at most one point. Recurring monitoring counts toward the tier-4 base and can satisfy the second conjunct of m_2 when physical-world action is also present, but does not count toward m_1, which counts submitted proof-evidence types only.

The public-post/account modifier m_3 applies whenever the listing requires public posting or account use, *including* when public-post or account proof is also the base tier. The modifier is intentionally retained because public-facing or account-mediated completion requirements carry an additional advertised exposure-relevant dimension beyond the feature's base classification. That dimension is not independently coded: the same coded feature supplies both contributions, so m_3 records a rubric judgment about double weighting rather than a separate observation. In the coder-recorded components, m_3 applies to 200 primary listings. Removing it moves 15 of the 779 primary rows below the severe threshold. This is a sensitivity of the rubric definition, not independent evidence that the audited score or modifier is valid. For example, a listing coded for public-post proof receives the public-post base tier because completion is tied to a public post, and it also receives the modifier because the requirement uses a public or account-mediated channel. The coder-recorded uncapped total $\mathrm{base} + m_1 + m_2 + m_3$ is retained for the score-5 saturation analysis below.

Primary summaries treat PBS as ordinal; the later ordinary-least-squares (OLS) models are explicitly diagnostic linear approximations that treat the numeric score as a predictor or an outcome.

Applying the nominal formula to the coder-recorded base and modifiers reproduces the final adjudicated score for all 779 primary listings, and it classifies 492 as elevated, 438 as severe, and 372 at the cap, matching the audited counts exactly. A separate rule-based reconstruction from the binary labels alone, retained as a comparison, disagrees with the recorded components on 218 rows, almost always by sitting higher, because it cannot see the codebook rule that a commissioned deliverable is not itself completion proof. The verification script checks this row-level crosswalk before manuscript compilation.

Table 8 separates conceptual motivation from evidentiary status. Where the retained historical materials do not document an empirical rationale, the table says so; the explanations are not retrospective validation.

Table 8. Rubric design choices, motivation, and evidentiary status.

Choice	Conceptual motivation	Evidentiary status
Three feature families	Proof-evidence features record requested artifact or attribute modalities; recurrence records temporal persistence; physical-world action records exposure-relevant task context. Privacy harms, contextual integrity, surveillance, administrative-burden, and crowd-work research are sensitizing foundations; Section 6 gives the construct-by-construct mapping and the theory-versus-convention split.	Candidate taxonomy fixed before human coding. Its mapping to the cited theories is retrospective conceptual grounding. Frequencies, reliability, profiles, and co-occurrence describe its application here; they do not establish exhaustiveness or distinct latent dimensions.
Base-tier ordering	Orders requirements by the author's judgment about potential reach, linkability, persistence, and recurrence, from minimal acknowledgement through identity-linked or repeated requirements.	Author heuristic, not derived from a worker ranking, outcome criterion, or statistical model. Retained files do not document an empirical basis for the exact tier boundaries.
Equal +1 modifiers and additivity	Makes stacking, physical-world coupling, and public/account use visible in a simple screening total.	Author convention. Equal increments do not imply equal experienced costs, and base tiers and modifiers reuse some observations. Table 13 reports post-hoc ablations.
Public/account modifier	Treats a public or account-mediated channel as an additional exposure-relevant aspect even when it also sets the base tier.	Intentional double weighting, not an independently coded dimension. The same features can also make the multiple-evidence modifier positive.
Cap at 5	Bounds the historical screening rubric.	Range-limiting convention without an external criterion. It produces a 47.8% ceiling; uncapped component diagnostics and profiles expose the compressed variation.
Cutpoints 3 and 4	Names elevated and severe screening bands for descriptive summaries.	Author-defined labels, not validated clinical, behavioral, or worker-perceived thresholds.
Contextual departures	Allowed coders to apply the rubric to the listing as a whole rather than mechanically summing binary fields.	Recode coders recorded the base tier, each modifier, and the uncapped total directly, so components are human-coded; the superseded first round preserved neither components nor structured override reasons.

Table 9. Base-severity mapping. The base score is the highest applicable tier, including the tier-4 combinations; a minimal text acknowledgement alone yields tier 0 even though the text-proof indicator is set. Tier labels describe the requirement pattern coded for a listing, not its unobserved submission channel.

Tier	Coded variables mapped to this base severity
0	No specified proof, or a minimal text acknowledgement only.
1	Structured text submission; a private link to such a submission.
2	Screenshot proof; photo proof; or another discrete non-text artifact—an image, recording, scan, document, or other file—submitted as evidence about the work; how it arrives (attached, uploaded, or linked) does not change the tier.
3	Video proof; public-post proof; account proof; location proof; financial proof; standalone phone proof under the prospective 2026-08-10 policy (three listings).
4	Identity proof; recurring monitoring; or physical-world action paired with any of: phone proof; public-post/account use; or location proof together with photo, video, or financial proof.

The tiering scheme is a conceptually motivated heuristic; its tier ordering and point values are not empirically estimated (see Planned Worker-Facing Validation). The historical rubric never assigned a base tier to standalone phone proof; that silence was a gap in the materials rather than affirmative evidence of a policy, and the version-11 codebook marks any phone-only case `UNCLEAR` rather than silently assigning a tier. In the recode, five phone-only-adjacent rows were initially marked `UNCLEAR` under that fail-closed rule. The three exact phone-only rows among them carry a tier-3 base under a prospectively author-approved policy adopted on 2026-08-10; the policy is recorded as version 12 of the codebook, is labeled a new policy rather than a recovered historical one, and is listed in the *Deviations* section. The tier-4 rule for physical-world action paired with phone proof is unchanged.

Table 10. Coder-recorded base-plus-modifier points among primary listings assigned the capped maximum score. These illustrate, rather than recover, the cap's compression.

Uncapped points	Score-5 listings	Share of score 5
5	150	40.3%
6	187	50.3%
7	35	9.4%

Table 11. Proof-burden scores by public status among the 779 primary listings. Status was shown at collection; these scores were not re-coded over time (see the ten-day follow-up section).

Public status	Listings	Share of all	Score 4 or 5	Score 5
cancelled	404	51.9%	230 (56.9%)	200 (49.5%)
completed	146	18.7%	66 (45.2%)	57 (39.0%)
open	135	17.3%	92 (68.1%)	70 (51.9%)
assigned	74	9.5%	37 (50.0%)	33 (44.6%)
partially filled	20	2.6%	13 (65.0%)	12 (60.0%)
non-cancelled subtotal	375	48.1%	208 (55.5%)	172 (45.9%)

Because the base tiers and modifier points are heuristic, Table 12 first reports four ordinal-score summaries and then four feature-based alternatives. The ordinal rows distinguish the elevated threshold, severe threshold, cap, and coder-recorded uncapped-total diagnostic above the cap; they are not alternative definitions of one quantity. The feature-set rows ask whether the broad descriptive pattern depends on the capped rubric. Every share uses the 779 primary listings as its denominator. The row labeled "Any of eight selected higher-exposure features" is the union of video, identity, account, location, phone, financial, public-post, and recurring-monitoring requirements—so it includes phone proof even though the historical rubric assigns no standalone-phone base tier, while excluding physical-world action and photo proof. It is a named subset, not a comprehensive exposure index.

Table 12. Ordinal-score summaries and feature-set alternatives. Shares use the 779 primary listings as the denominator.

Definition	Listings	Share of all 779 listings	Interpretation
Proposed elevated threshold (≥3)	492	63.2%	Listings assigned audited PBS 3 or higher under contextual application of the rubric guidance.
Proposed severe threshold (≥4)	438	56.2%	Listings assigned audited PBS 4 or higher; the primary threshold used for profile summaries.
Capped maximum (score 5)	372	47.8%	Listings assigned audited PBS 5; coder-recorded uncapped totals illustrate the cap's compression.
Score 5 with uncapped points >5	222	28.5%	Listings whose coder-recorded uncapped totals exceed 5; this illustrates, rather than recovers, cap compression.
Any of eight selected higher-exposure features	528	67.8%	Deliberately broader than tier 3: video, identity, account, location, phone, financial, public-post, or recurring-monitoring features; the historical rubric assigns no standalone-phone base tier.
Public post or account use	200	25.7%	A coded public-post or account-use requirement that may create a public trace or link completion to an external account.
Identity proof, location proof, phone proof, or financial proof	345	44.3%	Coded completion-linked requirements involving personal identity, place, telecom, or transaction information.
Physical-world action with a coded proof-evidence feature	358	46.0%	Physical-world task context paired with any coded proof-evidence feature.

Table 13 removes each equal-weight modifier in turn. Every row conditions on the recorded base; it tests the overlapping heuristic point assignments, not how coders would have rescored the listings. The final row also treats public-post and account proof as one family for the multiple-evidence modifier while omitting their separate public/account modifier. These are sensitivity diagnostics on the rubric definition rather than alternative labels.

To make the high-score set interpretable, the main paper reports mutually exclusive, priority-assigned summary groups for score-4 and score-5 listings. These are order-dependent summaries, not discovered clusters. The first five priorities are identity-linked proof, recurring monitoring, public-post plus account use, location plus financial proof, and physical-world action plus proof. If none applies, assignment proceeds through video, public-post, account, financial, and location proof, followed by other stacked lower-tier proof reaching the severe threshold. Counts are displayed only after this assignment. The priority order is an author choice, set before any group counts were tabulated: it starts from the most person-linked exposure (identity), continues through ongoing and compound exposures, and ends with single evidence types, so each listing is summarized by its most exposure-laden qualifying feature. The non-exclusive overlap table below shows the same listings without any ordering. "Physical-world action plus proof" requires physical-world action together with photo, video, location, financial, public-post, or account proof, after higher-priority groups

Table 13. Threshold and cap counts under modifier ablations. All variants use the 779 primary listings and the coder-recorded base; no row is an independently coded score.

Ablation variant	Score ≥3	Score ≥4	Score 5	Interpretation
Full nominal crosswalk	492	438	372	Coder-recorded base plus all three author-defined modifiers.
Omit multiple-evidence modifier (m1)	459	394	255	Removes the equal +1 increment for more than one proof-evidence type.
Omit physical-with-proof modifier (m2)	466	417	280	Removes the equal +1 increment for physical-world action paired with proof or monitoring.
Omit public/account modifier (m3)	484	423	350	Removes the equal +1 increment for public posting or account use.
Base tier only	404	310	0	Removes all three additive modifiers; the cap is then unreachable.
Collapse public/account family and omit m3	484	423	350	Counts public-post and account proof once for m1 and removes their separate m3 increment.

have been assigned. To avoid duplicating the main paper's priority-group table here, the supplement instead reports the non-exclusive completion-evidence, recurring-monitoring, and physical-world-action feature overlap among severe-threshold-positive listings, which complements the mutually exclusive groups by showing how often each feature appears across the score-4 and score-5 set. The 438 threshold-positive rows span 154 exact 13-feature profiles; 77 occur once, and the ten most common cover only 155 (35.4%). This heterogeneity is another reason not to treat one capped score as a complete disclosure.

Table 14. Non-exclusive completion-evidence, recurring-monitoring, and physical-world-action feature overlap among threshold-positive score-4 and score-5 listings. The final column, "Full-snapshot count", counts listings carrying the feature among all 779 primary listings, threshold-positive or not. This complements the priority-assigned group table in the main paper.

Feature	Kind	Threshold-positive count	Within-threshold share	Full-snapshot count
Text proof	Evidence	259	59.1%	450
Physical-world action	Context	243	55.5%	378
Photo proof	Evidence	203	46.3%	289
Identity proof	Evidence	199	45.4%	241
Account proof	Evidence	160	36.5%	190
Recurring monitoring	Verification	151	34.5%	151
Video proof	Evidence	142	32.4%	207
Link proof	Evidence	131	29.9%	231
Screenshot proof	Evidence	130	29.7%	158
Public-post proof	Evidence	98	22.4%	123
Financial proof	Evidence	76	17.4%	80
Location proof	Evidence	65	14.8%	83
Phone proof	Evidence	42	9.6%	43

8 Paraphrased Listing Patterns

Table 15 gives redacted paraphrases of observed rows, checked against the adjudicated labels; they are not quotations. They illustrate potential exposure mechanisms rather than realized exposure; submission channels and evidence retention are unobserved.

Table 15. Redacted proof-burden paraphrases and potential exposure mechanisms. Parentheses report the row's final adjudicated Proof Burden Score, not its base tier.

Pattern (audited score)	Paraphrase and potential exposure mechanism
Minimal text acknowledgement (0)	*Observed row.* A listing asks the worker to complete a simple task and provide only a short free-text completion acknowledgement. *Potential mechanism.* This request may add little exposure beyond ordinary completion review.
Public post plus account proof (5)	*Observed row.* A listing asks the worker to promote a service publicly and submit links plus screenshots. *Potential mechanism.* Public posting and screenshots could link the work to a worker account or public reputation.
Physical-world action plus proof (5)	*Observed row.* A listing asks the worker to create in-person promotional media and submit photos plus a short video. *Potential mechanism.* The worker's surroundings, equipment, or production setting could become visible.
Recurring monitoring (5)	*Observed row.* A listing asks for recurring in-person service after an initial visit, with before/after photos and a video walkthrough. *Potential mechanism.* Repeated proof requests could create repeated observation rather than a one-time completion signal. *Scoring note.* Rows of this kind carry a recorded tier-4 base with modifiers whose uncapped total exceeds 5; the reported score is capped at 5.

9 Status-Stratified Descriptive Check

This supplement reports a status-stratified descriptive table (Table 11) because public status is mixed and more than half of rows are cancelled. That table is not an outcome analysis: status is a cross-sectional public field, and the collection does not observe task lifecycle, acceptance, work submission, rejection, payment, or disputes. It is reported only to show that excluding cancelled rows leaves a similar severe-threshold share among the 375 non-cancelled primary listings, of which 208 (55.5%) meet the threshold (Table 23).

10 Coding and Adjudication Procedure

The rule-based extraction is not a substitute for human coding and adjudication. The recode procedure had four stages. First, a pilot operationalized the codebook criteria and surfaced missing definitions before full coding began. Second, two newly recruited coders independently annotated every returned row for content eligibility, all 11 proof-evidence variables, recurring monitoring, physical-world action, the Proof Burden Score with its recorded components, and ambiguity/sensitivity and `UNCLEAR` flags. Third, the coder-results analysis identified every routing trigger: coder disagreements, flags, `UNCLEAR` entries, low-confidence entries, and differences from the superseded first-round labels. Fourth, the blinded third adjudicator assigned final codes for the resulting packet.

First-round packet gaps, resolved by the recode. The next three notes describe packet-construction defects in the superseded round; the adjudicator-dependence note after them concerns the recode's own adjudication.

Skill context. The first round's packet builder read the public field `requiredSkills`, but RentAHuman publishes required skills as `skillsNeeded`; the sole Human Pages object uses the former name. Thus 980 of 981 first-round packet cells were blank even though the instructions directed coders to consult skills, and 773 RentAHuman listings had non-empty skill text in the raw capture that those coders never saw. The recode builder reads both field names, so the recode worksheets carried the full skill context and the promoted labels were produced under the full view.

Application-link context. The first-round packets also omitted RentAHuman's application-stage `requiredLinks` field for 121 listings. Platform documentation defines these as applicant-supplied profile, portfolio, résumé, or custom links rather than completion evidence [28], so their absence removed context needed to distinguish application links from completion proof. The recode worksheets include this field, and the promoted link-proof labels were coded with it; the first round's link results were provisional for this reason and are superseded.

Eligibility coding. The previously blank prospective eligibility packet has now been completed as part of the recode: each coder recorded `eligible_bounty_task` and a structured exclusion reason for every row, with disagreement or uncertainty routed to adjudication. Proof labels were still coded on excluded rows, so the *all-returned* sensitivity—retaining every object returned by the public endpoints—can be compared with the eligibility-filtered primary analysis throughout. The rowwise content screen does not by itself resolve the plan's separate normalized-text duplicate rule (see the *Deviations* section) or replace collection-level public-access provenance.

Adjudicator dependence. Because routing was broad, adjudication provenance is high in both RQ3 groups: 69 of 72 agent-identified listings carry a third-adjudicated label, as do 649 of the 676 eligible human-identified listings. The agent comparison is thus heavily dependent on one adjudicator. As a cross-round check that relies on neither the recode's adjudicator nor its coders, we recomputed each focal gap on the outcome-specific rows where the superseded first round's two coders agreed before any adjudication (an all-returned scope: 72 agent-identified versus 870 human-identified rows). Physical-world action occurs in 63.9% of 61 agent rows versus 23.7% of 742 human rows. For location proof, the corresponding values are 54.0% of 50 versus 18.5% of 726; for photo proof, 60.6% of 66 versus 34.2% of 783; and for recurring monitoring, 37.7% of 69 versus 6.8% of 768. The physical-world/location/monitoring composite (at least one of physical-world action, location proof, or recurring monitoring) occurs in 62.5% of 48 agent rows versus 21.4% of 571 human rows. The denominator changes by outcome. Four of the five gaps—physical-world action, photo, recurring monitoring, and the composite—retain the promoted direction under an entirely different pair of coders, which is the point of the check. The location gap does not: the first round's coders marked location far more often, consistent with the location/physical-world conflation the recode corrected, so first-round location agreement is not corroborating evidence. Agreement-selected rows are nonrandom, so this corroborates direction without establishing the full-sample magnitude.

A full-sample variant of the same cross-round check holds the all-returned 72-versus-870 grouping and its 595 display-name clusters fixed, then substitutes each superseded-round coder's complete labels for all 942 RQ3 rows. First-round coder 1 gives a composite of 47/72 versus 345/870 (nominal cluster-score sign-flip $p = .0575$). First-round coder 2 gives 49/72 versus 328/870 (nominal cluster-score sign-flip $p = .0034$). The promoted labels give 54/72 versus 377/870 on the same scope (nominal cluster-score sign-flip $p = .0010$). We report physical-world action because it is the largest positive focal gap and a composite component, and link proof because its observed gap is negative in all four display-name scopes. A privately retained aggregate contains the complete 13-feature family. Within each first-round coder's family, the physical-world Benjamini–Hochberg (BH) diagnostics are .0711 for coder 1 and .0143 for coder 2; the link-proof BH diagnostics are .0234 and .0143, respectively. These coder-vector calculations were not separately calibrated and are used only to assess label sensitivity across instrument rounds. The composite direction persists, but its strength and the physical-world pattern depend on the label vector. The privately retained aggregate has 28 rows: 14 outcomes (the 13 coded features plus the composite) for each of two coders. The focal values needed to interpret this sensitivity are reported above; no row-level coder data accompany the PDF.

Routing and adjudication steps. The step-by-step procedure was as follows:

(1) The adjudicator reviewed every row where the two recode coders disagreed on the eligibility decision or its structured reason, any proof-evidence variable, recurring monitoring, physical-world action, or the Proof Burden Score or its components.
(2) The adjudicator reviewed every row flagged as ambiguous or sensitive by either coder, every row either coder marked `UNCLEAR` or coded with low confidence, and every row whose consensus differed from the superseded first-round labels, even if the coders agreed with each other.
(3) The adjudicator filled the adjudicated eligibility decision, the 13 adjudicated binary variables (11 proof-evidence variables, recurring monitoring, and physical-world action), and the adjudicated Proof Burden Score with its components, without changing the completed coder files. The written instructions directed the adjudicator to treat neither coder as automatically correct, decide each field from the raw listing text and version-11 codebook definitions, and apply the scoring rubric to the adjudicated labels rather than average or split the two coder scores. On contested cells the final labels sided with coder 1 and coder 2 almost equally (55% versus 45%).
(4) The adjudicator recorded brief redacted adjudication notes for boundary cases and flagged rows for author review of retention or sensitivity (356 rows; all retained after review).
(5) The analysis recomputed descriptive tables, threshold counts, severe-threshold profiles, and diagnostic models from the adjudicated label set.

The privately retained adjudicator-facing packet omits both the rule-based labels and the superseded first-round labels to avoid anchoring the adjudicator to automated extraction or the earlier study coding. It contains the 873 routed rows, with raw public listing text, both recode coders' answers, adjudicated-label fields, disagreement indicators, and a notes column. The packet must remain local-only and must not be committed or shared publicly.

The private workflow has two branches. One branch can regenerate stratified audit-sample packets for authorized future pilots, but doing so requires local raw listing text that is not distributed. The full-coding branch produces aggregate reliability results and a private routing packet for internal comparison. A separate step removes rule labels from the adjudicator-facing packet, and the final build combines adjudicated rows with unflagged exact-consensus rows to create the controlled, no-raw-text analysis dataset.

11 Model Specifications

How to read this section. A regression coefficient (β) records the estimated direction and size of an association after the listed controls; it does not establish causation. OLS denotes ordinary least squares, MLE a maximum-likelihood estimate, and logistic models describe odds. An observation's *leverage* measures how far its own outcome determines its own fitted value; a leverage of one means the fit reproduces that observation exactly. HC3 adjusts standard errors for unequal residual variation while still treating listings as independent; display-name-clustered standard errors instead allow listings sharing an apparent requester name to be related. A model that does not converge has not produced a trustworthy finite estimate. Throughout, $p < .05$ is only a conventional reference threshold, and the text states when multiplicity, clustering, or model-selection concerns make it unsuitable for inference.

The listed-price, PBS, adjusted-application, and RQ3 category-adjusted diagnostics use the collapsed-variable definitions below; the planned status model instead uses raw categories. Category controls collapse categories with fewer than 20 rows into a low-frequency level. Requester type (used as a control in the price and adjusted application models)

collapses values other than human and agentic_or_other—including the platform and unknown labels—into an unknown_or_platform level. This diagnostic control uses the processed broad requester-type field; it is distinct from the RQ3 grouping, which compares 71 RentAHuman rows carrying named agent-or-bot labels plus the Human Pages row identified through its nested agent relation against 676 eligible human-labeled rows and excludes 31 ambiguous "other" rows (all-returned versions use 870 human-labeled rows). Price and adjusted application models exclude nonpositive listed prices and prices above 10,000 listed units; in this snapshot, that upper filter removes one extreme placeholder-like USD 999,999 fixed-price listing. Because currencies and price formats are mixed, these models remain diagnostic only.

Listed price is treated as a public listing field, not as wage or realized compensation. The snapshot mixes currencies and fixed/hourly formats, so the model estimates are reported only as diagnostic associations. The table below gives the measurement context for those checks.

Table 16. Listed-price measurement context by currency and price type. "Share" is each cell's share of the 779 primary listings; median listed price is in native listed units (a fixed total or an hourly rate, not converted across currencies); the final column is the share meeting the proposed severe threshold within the cell.

Currency	Price type	Listings	Share	Median listed price	Proposed severe threshold (score ≥4)
USD	fixed	676	86.8%	25	54.6%
USD	hourly	67	8.6%	35	74.6%
EUR	fixed	28	3.6%	45	53.6%
EUR	hourly	5	0.6%	34	60.0%
USDC	fixed	2	0.3%	255	0.0%
ETH	fixed	1	0.1%	3	100.0%

Reported log-price models use log(1 + listed price) after the price filters above. Application counts enter as one unadjusted descriptive diagnostic, three principal adjusted specifications, and seven generated sensitivity specifications; they are not treated as a competition outcome.

Listed price model. The outcome is log(1+listed price). The focal predictor is Proof Burden Score; in this diagnostic numeric-score model, coefficients are reported in log points per score point. Controls are collapsed category, requester type, price type, listed currency, and description length. Table 17 reports the complete generated family of planned price checks and added rule-provenance diagnostics under both HC3 and display-name-clustered covariance, with HC3 shown only where it is defined. Eight of the ten designs contain a singleton currency level and hence an observation with leverage one; because HC3 divides by $(1 - h_{ii})^2$ [22], their finite-looking software values are invalid and the generated table masks them. This includes the planned primary HC3 test, which therefore could not be completed as specified. In that $n = 778$ primary filtered sample, the coefficient is .0886 log points in log(1 + listed price) per score point. Exponentiation describes 1 + listed price, not listed price itself. With covariance clustered by a hash of each lowercased, trimmed public display name, the point estimate is unchanged but its 95% interval is $[-.0044, .1815]$ log points and includes zero ($p = .062$; 468 clusters). This finite clustered contrast conditions on the sparse currency controls and is only a diagnostic. Among the ten clustered specifications, four are nominally below .05 (the binary elevated and severe scores, the keyword-screen exclusion, and the one-per-signature subset, $p = .026$–$.046$), but no specification survives the within-family BH transformation (smallest $q = .088$ across the ten), the fixed-only and hourly-only subsets are null ($p = .056$ and .769; $n = 706$ and 72), and the primary ordinal specification itself is not

below .05. The nominal crossings therefore do not support a compensation claim, and the association is treated as diagnostic and non-robust.

Table 17. Complete generated listed-price sensitivity family, combining planned checks with added rule-provenance diagnostics. Coefficients are on the log(1 + listed price) scale. "Clustered" uses public display-name clusters. The BH transformation across the ten clustered p values is reported in the text.

Score/sample definition	Coefficient	HC3 p (if defined)	Clustered p	Clustered 95% CI	N
Ordinal score; primary filtered sample	+0.089	–	0.062	[-0.004, +0.182]	778
Binary elevated score (3–5)	+0.404	–	0.026	[+0.049, +0.758]	778
Binary severe score (4–5)	+0.339	–	0.044	[+0.008, +0.670]	778
Ordinal score; fixed-price listings only	+0.098	–	0.056	[-0.003, +0.199]	706
Ordinal score; hourly listings only	-0.031	0.820	0.769	[-0.241, +0.178]	72
Ordinal score; excluding unknown/platform requester controls	+0.082	–	0.109	[-0.018, +0.183]	707
Ordinal score; excluding keyword-screen risk flags	+0.099	–	0.044	[+0.002, +0.195]	745
Ordinal score; excluding title/description-only rule signals	+0.086	0.036	0.085	[-0.012, +0.185]	711
Ordinal score; high-confidence rule-signal rows	+0.104	–	0.056	[-0.002, +0.211]	598
Ordinal score; one row per exact title/category/price/currency signature	+0.094	–	0.046	[+0.002, +0.187]	738

Applications/competition models. RentAHuman detail records omit the list-only `applicationCount` and `spotsAvailable` fields. The internal builder therefore joins those fields from the corresponding list record by platform identifier and leaves genuinely unobserved values missing; substituting zero and one would collapse the observed variation. The controlled processed data and internal model outputs implement this rule. The processed row does not retain a second timestamp or URL hash for the list response, so it cannot quantify list/detail timing drift independently of the withheld raw manifests.

Among the returned rows, 464 of 981 listings (47.3%) show at least one visible application. RentAHuman listings carry 66,984 visible applications in total (median 0, third quartile 13, maximum 8,919), and 257 listings advertise more than one spot; these are collection-time platform-metadata facts and span the full returned snapshot. We report these fields descriptively rather than treating visible application count as a competition outcome.

The unadjusted diagnostic regresses log(1 + applications) on Proof Burden Score over the 778 primary RentAHuman rows. It gives $\beta = -0.153$ per score point (display-name-cluster-robust 95% CI $[-0.259, -0.047]$, $p = .005$, 468 clusters). This is an unadjusted association in a nonprobability snapshot; it conditions on nothing and is not treated as a competition finding.

The adjusted applications-per-spot specification (the main paper's applications per position) is Model 2 in the internal plan and generated model record. It regresses log(1 + applications/available spots) on score, log listed price, collapsed category, requester type, price type, and log(available spots); rows require positive available spots. The privately retained generated output contains both HC3 and clustered inference for all specifications. The planned coefficient is −0.112 (HC3 $p = .003$; display-name-clustered $p = .021$). Adding observed listing age—`created_at_age_days`, complete for all returned rows and named as a robustness check in the internal plan—attenuates it to −0.045 (HC3

$p = .214$; display-name-clustered $p = .329$). Listed prices are in four native units. Adding currency indicators to the age-adjusted specification shifts the intercepts while retaining one log-price slope across incomparable units; the coefficient is -0.048 (display-name-clustered $p = .271$). Its HC3 covariance is undefined because at least one currency control contains an observation with leverage one; the generated model record therefore masks that HC3 interval and p value. Clustered p values throughout use the statsmodels default cluster covariance with its small-sample correction enabled. The unadjusted diagnostic uses 778 primary RentAHuman rows and 468 clusters; the three adjusted OLS models use $n = 777$ and 467 clusters because the plan specifies the filtered `log_price_model`. Every value in this paragraph is generated and automatically cross-checked before manuscript compilation.

The negative association therefore depends on whether listing age is controlled: the planned specification is nominally below .05 under clustering, but adding age attenuates it to null, with or without currency indicators. Because exposure time is confounded with visible application counts, the age-unadjusted crossings are not interpreted. Nothing here supports a competition claim in either direction.

Table 18. Complete adjusted applications-per-spot OLS family. Coefficients use the score/sample definition in the first column; "clustered" uses public display-name clusters. The one-row-per-signature check uses exact title/category/price/currency groups, not normalized task text.

Score/sample definition	Coefficient	HC3 p	Clustered p	N
Ordinal score; planned controls	-0.112	0.003	0.021	777
Ordinal score; + listing age	-0.045	0.214	0.329	777
Ordinal score; + listing age and currency	-0.048	–	0.271	777
Binary elevated score (3–5)	-0.343	0.024	0.069	777
Binary threshold indicator (score 4–5)	-0.272	0.049	0.067	777
Ordinal score; fixed-price only	-0.108	0.006	0.040	705
Ordinal score; hourly only	-0.274	0.085	0.052	72
Ordinal score; excluding keyword-screen risk flags	-0.113	0.004	0.030	744
Ordinal score; excluding unknown/platform requesters	-0.098	0.013	0.048	706
Ordinal score; one row per exact signature	-0.122	0.002	0.010	737

The seven generated score/subset sensitivities in Table 18 give clustered p values between .010 and .069; like the planned specification, none of them controls listing age, so their nominal crossings inherit the same exposure-time confound and are not interpreted. The exact-signature reduction retains 737 rows; it is a partial duplicate stress test, not the plan's normalized-text deduplication rule.

The plan also called for a count model. Counts are extremely overdispersed (empirical variance/mean about 2,301 in the primary filtered sample, far above the ratio of one a Poisson model assumes). The raw-category Poisson does not converge, so its Pearson statistic is only an optimizer-iterate diagnostic; the empirical ratio already triggers a count-model assessment.

The literal planned negative-binomial (NB2) model regresses application count on PBS, log listed price, raw category, and requester type, with a log-spots offset, on $n = 778$ rows (468 clusters). Although its design is full rank, no finite MLE exists: all 11 `creative` rows and the sole `home-services` row have zero applications, so the corresponding intercepts/contrasts can diverge. The generated model record therefore masks every raw-category coefficient, interval, and p value. Winsorizing—replacing counts above the stated percentile with the value at that percentile—at the planned 95th and 99th percentiles leaves those zero cells unchanged and does not repair identification.

Two explicitly post-hoc diagnostics converge. Collapsing categories gives $\beta = -.319$ (MLE $p < .001$; clustered $p < .001$; fitted overdispersion $\alpha = 7.97$). Dropping the two all-zero categories (12 rows) gives $\beta = -.261$ (MLE $p < .001$; clustered $p = .003$; $\alpha = 7.55$). Both are negative and nominally below .05, but they do not turn the unestimable planned model into a completed confirmatory test: neither controls listing age, alternative models for heavy zero inflation were not attempted, and visible application counts combine exposure time with unobserved ranking and promotion. No valid diagnostic supports a competition claim.

Status disposition. The plan's named outcome literally includes assigned and partially filled listings (94 primary positives), not completed listings. Its raw-category logit has a full-rank design but fails with a singular matrix under separated rare-category outcome cells, and the collapsed-category version fails the same way. The historical builder had instead broadened the named field to a non-open/non-cancelled status definition (240 positives). We retain that result only as a post-hoc diagnostic: its collapsed-category coefficient is −0.096 (listing-level $p = .042$, treating listings as independent; display-name-clustered $p = .085$, 469 clusters). None supports an allocation or lifecycle claim. Both status definitions, formulae, design-matrix ranks, and fit dispositions are retained in the private analytical record.

Requester-type and RQ3-group diagnostics. The plan's Model 4 uses the broad three-level requester control on all 779 primary rows. Its preferred proportional-odds model—an ordered logit, which fits one coefficient per predictor across all score cutpoints—converges. Relative to human, the clustered log-odds coefficients are −.333 for `agentic_or_other` (95% CI [−1.495, .829], $p = .575$) and +.112 for `unknown_or_platform` ($p = .802$). Diagnostic OLS gives the same null conclusion: −.288 score points for `agentic_or_other` (clustered 95% CI [−1.264, .688], $p = .563$) and −.010 for `unknown_or_platform` ($p = .975$).

A separate post-hoc OLS uses the RQ3 grouping: 71 RentAHuman rows carrying named agent-or-bot labels plus the Human Pages row identified through its nested agent relation versus 676 eligible human-labeled rows, excluding 31 ambiguous "other" rows ($n = 748$). Its agent-identified coefficient is −.091 score points (clustered 95% CI [−.799, .617], $p = .800$; 462 clusters). All three models control collapsed category, price type, physical-world action, and the legacy title/description-only rule-signal flag. Because physical-world action is part of PBS, this is a component-adjusted residual contrast, not an overall PBS contrast. Omitting physical-world action and the legacy rule-signal control gives +.222 score points (clustered 95% CI [−.559, 1.002], $p = .578$); with only 20 display-name clusters containing an agent-identified listing, fixed-seed wild-cluster restricted bootstrap-t sensitivities give $p = .674$ under Rademacher weights and .670 under Webb weights, two standard choices of resampling weight distribution. Under the promoted labels, neither the residual nor the overall-PBS specification indicates a requester-type difference in the contextual score, consistent with the main paper's null severe-rate comparison; the group difference that persists is in the feature composite, not the score.

One severe-rate sensitivity deserves explicit statement. On the all-returned population (981 records), the unadjusted severe-rate comparison reaches nominal significance: 58.3% (42/72) of agent-or-bot-labeled versus 45.4% (395/870) of human-labeled records score 4 or higher (two-proportion $p = .034$), and the category-pooled Mantel–Haenszel interval marginally excludes 1 ([1.03, 3.33]). The display-name cluster-based sign-flip test retains the null ($p = .41$), as does every severe-rate test on the primary population (58.3% versus 55.9%; two-proportion $p = .69$, sign-flip $p = .83$). We therefore report the primary-population null as the operative conclusion and this all-returned crossing as a labeled sensitivity rather than a finding. The complete generated results are retained privately, and the values needed for this comparison appear above.

Physical-world/location/monitoring logistic (RQ3 category adjustment). The outcome is the physical-world/location/monitoring composite (physical-world action, location proof, or recurring monitoring); the predictor

is the same agent-identified indicator, with collapsed-category controls ($n = 748$). Under the promoted labels no category completely predicts the outcome, so the maximum-likelihood (ML) logistic converges (in the superseded round, a fully separated category had forced its masking). The category-adjusted descriptive summary is the Mantel–Haenszel odds ratio [23] pooled across informative category strata (Table 19), which does not rely on ML convergence: 3.9 (95% CI 1.9–8.2, $p < .001$); the converged ML fit gives a near-identical estimate. Here an informative stratum contains both requester labels and enough outcome variation to contribute to the pooled comparison. The pooled estimate is category-sensitive: 30 of the 72 agent-identified listings are research fieldwork and all 30 require the composite (as do 30 of the 53 eligible human research-fieldwork listings), and omitting that category moves the pooled odds ratio to 1.9 (95% CI 0.8–4.3), whose interval includes 1. The unadjusted Breslow–Day test [4] provides evidence against odds-ratio homogeneity across the informative strata in the full primary sample ($p = .044$) but not once research fieldwork is omitted ($p = .53$); with sparse exploratory strata this is an asymptotic diagnostic rather than a categorical verdict, so full-sample pooling rests on a homogeneity assumption for which these data provide contrary evidence. The direction of the contrast does not depend on that one category, but its magnitude and the strength of the evidence do: with research fieldwork omitted, the category-adjusted evidence is no longer nominally below .05. Like the rest of RQ3, this is exploratory and descriptive.

Table 19. Category-adjusted odds of requiring the physical-world/location/monitoring composite (physical-world action, location proof, or recurring monitoring), agent-identified versus human-identified listings in the primary population. Both the Mantel–Haenszel pooling and the converged ML fit are exploratory descriptive summaries, not stable common-effect estimates; see the fieldwork sensitivity and heterogeneity test.

Estimator	OR	95% CI	p	N
Mantel–Haenszel category pooling; listing-level inference	3.9	[1.9, 8.2]	<0.001	748
ML logistic, category-adjusted; listing-level covariance	3.7	[1.8, 7.4]	<0.001	748

12 Deviations from the Internal Pre-Analysis Plan

The plan was internal and not externally registered. Table 20 consolidates material changes between that plan and the reported study; Table 21 records the disposition of the plan's robustness checks.

Table 20. Core design deviations from the internal pre-analysis plan.

Planned item	Implemented and reason	Status and consequence
Requester type predicting overall PBS.	The exact planned three-level Model 4 is now reported under ordered-logit and OLS specifications. The paper separately adds 13 feature contrasts, a three-feature composite, a nominal within-scope Benjamini–Hochberg (BH) transformation, category-pooled estimation, and display-name-clustered sensitivities because feature profiles are more interpretable than one contextual score.	The planned model is null. The RQ3 grouping and feature analyses are data-informed, post-hoc, and exploratory. Reported p values do not adjust for selecting the three-feature composite; confirmation requires a new snapshot. Multiplicity, category composition, and dependence are addressed by separate diagnostics.

Continued on next page

Table 20 continued.

Planned item	Implemented and reason	Status and consequence
Base-plus-modifier computation with stored coder components.	The first round did not store coder-recorded components. The recode coders recorded the base tier, each modifier, the uncapped total, and the final score directly, closing the gap.	Resolved by the recode. The nominal crosswalk from the recorded components reproduces all 779 primary scores; a label-only rule reconstruction, kept as a comparison, differs on 218 rows.
Deduplicate by platform ID and normalized task text.	Query overlaps were deduplicated by platform ID at collection. The normalized-text duplicate rule was not applied in the original coding; it has now been resolved retrospectively with an earliest-posted retention rule: within each cross-row duplicate group, the earliest-posted eligible row is kept. Primary analyses still report the full eligible set, with the deduplicated set as a labeled sensitivity.	Resolved retrospectively with the plan's normalized-text rule (lowercased, punctuation-stripped, whitespace-collapsed title and description): 30 duplicate groups, 42 eligible rows dropped, leaving 737 listings with a severe share of 56.3% (less than half a point from 56.2%). The separate exact title/category/price/currency signature check (68 rows in 28 groups within the primary population; one-per-signature severe share 56.6%) is retained for the listed-price and applications OLS families; applying both rules together leaves 729 listings at 56.4%.
Content-eligibility screen for human-action bounty/task listings.	Not applied in the original coding: all unique returned objects were processed, and no row-level eligibility decision was logged. The recode applied the screen row by row with two-coder eligibility fields, structured exclusion reasons, and adjudication of disagreements.	Resolved by the recode. 779 listings are content-eligible and form the primary population; 202 are excluded (117 service offers, 61 not-human-action records, 24 promotional posts). Excluded rows keep proof labels and enter only the labeled all-returned sensitivities.
No base tier for standalone phone proof.	The historical rubric assigned no base tier to standalone phone proof, and the version-11 codebook fail-closes such rows as `UNCLEAR`. During the recode, five phone-only-adjacent rows were initially `UNCLEAR`; the author prospectively approved a tier-3 base for the exact phone-only case on 2026-08-10, recorded as codebook version 12.	New prospective policy, not a recovered historical one. It governs the three exact phone-only rows; it does not alter any other tier and is disclosed here rather than presented as part of the frozen version-11 instrument.
Labels from the original audit.	The originally submitted labels came from the first-round audit, whose codebook left several criteria without operational definitions. A full protocol recode (two new independent coders, blinded third adjudicator, complete listing context, piloted version-11 codebook) replaced every label and score and was promoted on 2026-08-10.	Post-submission correction. 78% of rows changed label or score. Two first-round errors are corrected: location evidence had been conflated with physical-world action (91% overlap among first-round location positives), and a commissioned deliverable had been counted as completion proof (it now sets no base tier by itself, and the labels stay descriptive). First-round artifacts are preserved under `legacy` names.

Continued on next page

Table 20 continued.

Planned item	Implemented and reason	Status and consequence
Pay, applications, status, and allocation as primary outcomes.	Feature and score prevalence became primary. Listed price is the outcome in pay diagnostics; PBS is a predictor there and in application models, and an outcome only in requester-type diagnostics. Applications enter descriptively with one unadjusted diagnostic, three adjusted specifications, and the planned raw-category negative-binomial count model; several age-unadjusted specifications are nominally below .05, but the age-controlled specifications are null and none is treated as conclusive. The literal planned status model is unidentified. The historical builder had broadened it to a non-open/non-cancelled status definition; that post-hoc diagnostic is also inconclusive. The original pivot cited near-zero application counts, but that was a dataset-construction artifact, now corrected (464/981 listings show visible applications); price is mixed-format and non-robust, and status is cross-sectional. The planned "applicant-to-dollar" ratio is neither analyzed nor retained: the denominator mixes four currencies, fixed totals, and hourly rates, so it is not a comparable dollar quantity.	Post-plan outcome pivot. The paper makes no compensation, competition, allocation, or lifecycle claims.

Table 21. Disposition of robustness checks in the internal pre-analysis plan.

Status	Planned checks	Disposition
Implemented for listed-price models	Elevated/severe binary scores; missing/extreme-price filters; fixed- and hourly-price subsets; unclear-requester and keyword-screen exclusions; one row per exact signature.	Table 17 reports clustered inference for every generated check and HC3 where defined; the rule-provenance rows were added post-pivot. Eight HC3 rows are masked because singleton currency levels create leverage-one observations. Four clustered specifications are nominally below .05, but none survives the within-family BH transformation (smallest $q = .088$ across the ten).
Implemented for application models	Binary scores; fixed/hourly, requester, keyword-screen, and exact-signature subsets; task-age control.	Table 18 reports every OLS check; currency control was added post hoc. Several age-unadjusted checks are nominally below .05 when clustered, but the age-controlled specifications are null, and the exposure-time confound is uncontrolled elsewhere. Count-model checks are reported separately.
Attempted for application counts	Poisson, negative-binomial, and 95th/99th-percentile winsorization.	The raw-category likelihood has no finite MLE because two categories have all-zero outcomes; winsorization cannot repair it. The pipeline masks those fits and reports collapsed-category and 12-row-exclusion diagnostics.
Implemented post hoc, descriptively only	Alternative-score prevalence thresholds and a no-modifier ablation.	Threshold prevalences and the base-tier-only ablation of the coder-recorded components are reported. The latter removes all three modifiers but is not an independently coded alternative score and does not reproduce every contextual judgment.
Replaced or added after the pivot	No exact plan counterparts.	Feature-set alternatives, rule-signal-confidence subsets, non-cancelled and pre-adjudication-consensus summaries, a nominal within-scope BH transformation, category-pooled and display-name-clustered sensitivities, and the ten-day reachability check were added. The follow-up is exploratory and does not establish a lifecycle or trend.
Implemented for Models 1–4	Evidence-count-only alternative.	The generated count sums the 11 final audited proof-evidence binaries and excludes recurring monitoring, physical-world action, and the stale stored count. It replaces PBS as predictor in Models 1–3. Because PBS is Model 4's outcome, that check instead uses the count as an explicitly labeled alternative ordered outcome.

Table 22. Generated evidence-count-only sensitivity. The predictor for Models 1–3 is the row sum of the 11 final audited proof-evidence binaries. Model 4 instead uses that count as an explicitly labeled alternative ordered outcome, so its coefficients are requester-type log-odds contrasts. Dashes mark inference withheld because HC3 is undefined or no finite likelihood optimum exists.

Specification	Estimate	HC3/MLE p	Clustered p	N
Model 1: listed price (OLS)	+0.066	–	0.236	778
Model 2: applications per spot (OLS)	-0.096	0.017	0.087	777
Model 2: application count (NB2)	–	–	–	778
Model 3: planned status, raw category (logit)	–	–	–	779
Model 3: planned status, collapsed category (logit)	–	–	–	779
Model 4 alternative outcome: agentic/other vs human	-0.583	0.086	0.220	779
Model 4 alternative outcome: unknown/platform vs human	-1.131	<0.001	0.058	779

The recomputed count (Table 22) ranges from 0 to 8. The legacy stored `evidence_count` differs from it in 621 of 981 rows and is never used. The listed-price coefficient is +.066 (clustered $p = .236$; HC3 undefined). For applications per spot it is −.096 (HC3 $p = .017$; clustered $p = .087$), so the association that crosses .05 when each listing is treated as an independent observation does not survive display-name clustering. The application-count NB2 with its planned log-spots offset remains unidentified because the same two raw categories have all-zero outcomes. The planned status logit remains unidentified under both raw and collapsed categories. Both Model 4 alternative-outcome contrasts are null when clustered ($p = .220$ and .058). The top count level contains a single row, so the ordered-outcome fit is a sparse-tail diagnostic, not validation of ordinal spacing. These checks do not change the paper's diagnostic conclusions.

13 Robustness Checks

Table 23 reports shares meeting the proposed severe threshold with 95% listing-level Wilson reference intervals that assume independent listings [36]. For the full snapshot, this interval is 52.7–59.7%. A separate intercept-only logistic interval clusters covariance on 468 RentAHuman display-name hashes plus one Human Pages singleton; it is 51.5–60.9%. Exactly: the model is $\text{logit}\,\Pr(\text{severe}_i = 1) = \beta_0$, fit by maximum likelihood over the 779 eligible listings, where severe_i indicates that listing i's final Proof Burden Score is 4 or 5. The only parameter is the intercept β_0; there are no covariates. The covariance is cluster-robust with each display-name cluster as one group, and the interval is the inverse-logit of $\beta_0 \pm 1.96$ cluster-robust standard errors. Neither interval addresses nonprobability coverage, temporal variation, coding uncertainty, or construct uncertainty. The table's "Excluding keyword-screen risk flags" row drops listings marked by the extractor's `possible_scam_flag`, a keyword screen for specified language related to fraud or sensitive payments. It is not a human fraud determination, a coder flag, or an adjudication-workflow flag. The one-row-per-signature row is a repeated-signature sensitivity based on title, category, exact price, and currency. It is not the plan's normalized-text duplicate rule, whose retrospective resolution (earliest-posted retention; 30 groups; 42 eligible rows; severe share 56.3%) is recorded in the *Deviations* section.

Two rows and two flag-based exclusions in this section are preserved records of the superseded first round rather than recode quantities. The first round's workflow flag marked 250 of the 981 returned rows; its note-template aggregate distinguishes 60 rows carrying a disjunctive "sensitive or unresolved context" template from 190 rows whose eight templates name specified sensitive contexts. Under that round's labels, excluding the ambiguous-template subset left 574/921 severe-threshold-positive (62.3%), close to that round's full-sample share; those values document the superseded round and are not comparable with the promoted 56.2%. The promoted dataset carries no rows with that flag, so the "internally flagged" and "excluding all internally flagged rows" entries in Table 23 are empty and identical to the full snapshot, respectively. The corresponding recode-era stress test is the author-review flag (356 rows, all retained), which did not change any final label.

Superseded-round flag exclusions (historical record). Under the first round's labels, the 60-row ambiguous-template exclusion gave a composite of 51/72 (70.8%) for agent versus 336/814 (41.3%) for human listings (display-name-clustered OR 3.45, normal-Wald reference interval 2.23–5.34, nominal sign-flip $p = .0024$). These historical values showed that the first round's composite direction did not depend on its flagged rows; the promoted labels reproduce the same direction on the full primary population (54/72 agent versus 374/676 human; OR 2.42, reference interval 1.40–4.18, sign-flip $p = .0067$; identical to the full-comparison row of Table 26 because the flag exclusion is now empty). Neither exercise resolves coding uncertainty.

Across the non-cancelled, one-per-signature, fixed-price, keyword-screen-excluded, and title/description-only-excluded subsets, the score-4-or-5 share is 54.5–56.6%. Hourly listings are higher (73.6%, on only 72 listings, which

Table 23. Proposed severe-threshold share by analysis subset, with 95% listing-level Wilson reference intervals. The ambiguous 60-row note-template row and the pre-adjudication two-coder score-consensus row are preserved records of the superseded first round (its flags, coders, and labels) and are computed on that round's all-returned 981-record scope, so their n can exceed the 779 primary listings; the internally flagged rows of that round do not exist in the promoted dataset, so those two flag rows are empty and identical to the full snapshot, respectively.

Subset	n	Threshold-positive	Share	95% CI
Full audited snapshot	779	438	56.2%	[52.7, 59.7]
Excluding ambiguous 60-row note-template subset	921	574	62.3%	[59.1, 65.4]
Internally flagged rows only	0	–	–	–
Excluding all internally flagged rows	779	438	56.2%	[52.7, 59.7]
Non-cancelled listings	375	208	55.5%	[50.4, 60.4]
One row per title/category/price/currency signature	739	418	56.6%	[53.0, 60.1]
Fixed-price listings	707	385	54.5%	[50.8, 58.1]
Hourly listings	72	53	73.6%	[62.4, 82.4]
Excluding keyword-screen risk flags	746	412	55.2%	[51.6, 58.8]
High-confidence rule-signal provenance	598	367	61.4%	[57.4, 65.2]
Excluding title/description-only rule-signal rows	712	395	55.5%	[51.8, 59.1]
Two-coder score-consensus subset (pre-adjudication)	666	445	66.8%	[63.2, 70.3]

may reflect category mix; no category-standardized hourly analysis is reported), as is the high-confidence rule-signal subset (61.4%), an expected selection effect because those first-pass signals came from structured or proof-specific fields.

Primary label-based analyses use the promoted recode label set. The pre-adjudication consensus row in Table 23 uses the superseded first round's two-coder score consensus and is retained as a cross-round record, not a recode subset. Analyses of the labeling process—inter-coder reliability, coder-specific prevalence, and the rule-versus-audit comparison—necessarily use coder labels or rule outputs.

Two further checks require data or labels beyond this snapshot: finer audited subtypes of the saturated score-5 set, and platform-specific models, which we do not fit because all but one row comes from a single platform.

14 Agent-Identified Listing Profile

This section asks whether the 72 listings carrying a platform agent/bot label or relation differ from the 676 eligible human-labeled listings (all-returned variants use 870). The comparison is exploratory: repeated listings can share an apparent requester, only a small number of display-name clusters contain agent-identified listings, and the agent designation does not verify autonomous operation. We therefore report the observed feature shares first and then show how the result changes when dependence, mixed-label display names, category composition, and test calibration are considered.

14.1 Feature Diagnostics and Multiplicity Calculations

The main paper plots the observed completion-evidence, recurring-monitoring, and physical-world-action shares for agent-identified and human-identified listings; the full table is below. "Diff" is the agent-identified share minus the human-identified share. Its Newcombe interval, two-proportion p, and Benjamini–Hochberg (BH) q [2, 26] are listing-level references that ignore display-name dependence. Every calculation conditions on the current adjudicated label vector and omits coder/adjudicator uncertainty.

The clustered diagnostic is intended to avoid treating repeated listings under the same public display name as independent evidence. Under a hypothetical world with no group difference, it randomly multiplies each whole cluster's score contribution by +1 or −1 and asks how often the resulting contrast is at least as extreme as the observed one. The technical name used below is a fixed-denominator restricted cluster-score Rademacher sign-flip procedure. A mixed-label display-name cluster is a hashed public display name containing both agent- and human-labeled listings. The four scopes are the full comparison, removal of the mixed cluster with the most agent listings, removal of all mixed clusters, and the latter removal plus exclusion of Human Pages. All 13 calculations remain defined in every scope, although several agent-positive cells become very sparse. Table 24 reports nominal within-scope BH transformations and joint single-step max-$|T|$ values, a familywise adjustment that compares each feature's statistic with the largest standardized contrast seen across the family in resampling. Together with the listing-level BH q values in Table 25, these are the three multiple-testing adjustments reported for the 13 feature comparisons. Privately retained generated outputs additionally contain the raw sign-flip and Benjamini–Yekutieli (BY) values [3]. The calibration below shows severely inflated false-positive rates for several sparse outcomes. Consequently, none provides validated feature-level false-discovery-rate (FDR) or familywise-error control.

Table 24. Nominal display-name-cluster sign-flip diagnostics for all 13 coded features in all four scopes. "Diff" is the agent-minus-human share in the full sample. Each cell gives the within-scope BH transformation followed by the joint single-step max-$|T|$ value. These are descriptive because feature-family error control is not validated.

Feature	Diff (full)	Diagnostic BH q / max-$\|T\|$ p			
		Full	Drop agent-heaviest hash	Drop all mixed	Drop all mixed, RentAHuman only
Physical-world action	+30.3%	0.0052/0.0058	0.0234/0.0134	0.0048/0.0136	0.0013/0.0050
Photo proof	+21.9%	0.3539/0.7410	0.7832/0.9969	0.8681/1.0000	0.9230/1.0000
Recurring monitoring	+21.6%	0.8117/0.9998	0.5236/0.6968	0.8681/1.0000	0.9619/1.0000
Financial proof	+4.4%	0.5696/1.0000	0.7832/0.9902	0.8681/1.0000	0.9230/1.0000
Phone proof	-4.5%	0.4683/0.9628	0.8115/0.9999	0.9683/1.0000	0.9858/1.0000
Location proof	-8.9%	0.2636/0.5148	0.5236/0.8824	0.6402/0.9896	0.7819/0.9965
Public-post proof	-9.0%	0.4830/0.9894	0.9223/1.0000	0.6402/0.9698	0.7819/0.9921
Text proof	-9.4%	0.4683/0.9958	0.8121/1.0000	0.8681/1.0000	0.9230/1.0000
Account proof	-14.9%	0.4683/0.9623	0.8121/1.0000	0.8681/1.0000	0.9230/1.0000
Screenshot proof	-16.8%	0.2622/0.4922	0.5236/0.8957	0.1488/0.3504	0.2018/0.4483
Video proof	-17.9%	0.4658/0.8450	0.8115/0.9999	0.0045/0.0020	0.0056/0.0033
Identity proof	-19.0%	0.4683/0.9455	0.8115/1.0000	0.6402/0.9979	0.8437/1.0000
Link proof	-27.0%	0.0188/0.2150	0.0474/0.0479	0.0045/0.0013	0.0013/0.0025

14.2 Feature Results

Four raw gaps are positive: physical-world action, photo proof, recurring monitoring, and financial proof. The other nine, including location proof, are negative; the superseded first round's positive location gap does not survive the recode's separation of location evidence from physical-world action. Table 24 reports the clustered nominal diagnostics; Table 25 reports the listing-level references. Because calibration does not support feature-family inference, we do not classify any feature as significant.

The observed contrasts are sensitive to the mixed-label display-name structure. The mixed cluster containing the most agent-identified listings contains 36 agent and 13 human rows; it is not the largest mixed cluster overall. Its agent rows supply 25 positives for each of physical-world action, photo proof, and recurring monitoring, and none for location proof. Removing the whole cluster leaves 36 agent and 663 human listings and reverses recurring monitoring.

Removing all 7 mixed clusters leaves 20 agent and 591 human listings. Removing the auxiliary Human Pages row as well leaves only 12 agent-containing clusters and changes the nominal physical-world BH value from .0048 to .0013. These threshold crossings have no inferential interpretation.

Link proof and video proof both have negative observed gaps in all four scopes. They differ in the nominal within-scope BH value: link proof's is below .05 in all four scopes, whereas video proof's is below .05 only in the two scopes that remove all mixed-label display-name clusters. Their nominal BH and max-$|T|$ values remain in Table 24; raw sign-flip and BY values are retained privately. Unlike the superseded first round, the recode packets included the application-stage `requiredLinks` context, so the link labels here distinguish application material from completion evidence as well as the listing text allows.

The agent group is small ($n = 72$), and identification uses platform-provided or self-reported source labels or relations, so this is a descriptive characterization, not a verified-autonomy or platform-general claim.

Table 25. Completion-evidence, recurring-monitoring, and physical-world-action feature prevalence in agent-identified ($n = 72$) versus human-identified ($n = 676$) eligible listings. CI, p, and BH q are listing-level references that ignore display-name dependence and are not used for feature-level inference.

Feature	Agent-identified	Human-identified	Diff (pp)	95% CI	p	q
Physical-world action	75.0%	44.7%	+30.3	[+19, +40]	<0.001	<0.001
Photo proof	56.9%	35.1%	+21.9	[+10, +33]	<0.001	<0.001
Recurring monitoring	38.9%	17.3%	+21.6	[+11, +33]	<0.001	<0.001
Financial proof	13.9%	9.5%	+4.4	[-2, +14]	0.232	0.232
Phone proof	1.4%	5.9%	-4.5	[-7, +2]	0.109	0.128
Location proof	2.8%	11.7%	-8.9	[-12, -2]	0.021	0.030
Public-post proof	8.3%	17.3%	-9.0	[-14, +0]	0.051	0.066
Text proof	48.6%	58.0%	-9.4	[-21, +3]	0.126	0.137
Account proof	11.1%	26.0%	-14.9	[-21, -5]	0.005	0.008
Screenshot proof	5.6%	22.3%	-16.8	[-21, -8]	<0.001	0.002
Video proof	11.1%	29.0%	-17.9	[-24, -8]	0.001	0.002
Identity proof	13.9%	32.8%	-19.0	[-26, -9]	<0.001	0.002
Link proof	5.6%	32.5%	-27.0	[-32, -18]	<0.001	<0.001

14.3 Display-Name Structure, Composite, and Calibration

Listings are the observations, but rows sharing a lowercased, trimmed public requester display name may be dependent, so we also cluster by its hash (Table 28). A privately retained mapping assigns each RentAHuman listing a deterministic, unkeyed, namespaced SHA-256 hash of its normalized display name; neither the mapping nor the hashes accompany this PDF. These pseudonymous keys are not verified accounts or anonymous identifiers: spelling variants can split one requester, shared names can combine requesters, and low-entropy names may be dictionary-linkable. The unmatched Human Pages row remains a singleton.

The full comparison contains 72 agent-identified and 676 human-identified listings in 462 clusters: 461 RentAHuman hashes plus one singleton. Within labels, these form 20 agent and 449 human hash–label cells; 7 display-name clusters carry both labels. The agent-heaviest mixed-label cluster contributes 36 agent-identified and 13 human-identified listings.

Table 26. Post-hoc composite contrast across the four display-name scopes. "Reference interval" is a normal-Wald interval; it is descriptive rather than a calibrated 95% confidence interval.

Scope	OR	Reference interval	Sign-flip p
Full comparison	2.42	1.40–4.18	.0067
Remove agent-heaviest mixed cluster	3.36	1.20–9.40	.0169
Remove all mixed clusters	7.37	2.23–24.38	.0022
RentAHuman only; remove all mixed clusters	14.73	3.24–66.89	.0003

The physical-world/location/monitoring composite is a data-informed post-hoc summary added after the internal plan; its p values do not adjust for that selection. At listing level, it occurs in 75.0% versus 55.3%; averaging within hash–label cells gives 81.4% versus 55.2%. Table 26 puts the four reported sensitivity scopes in one place. An OR above 1 means that the composite is more common among agent-identified listings; the sign-flip p values are the score-based diagnostic evaluated below. The intervals are normal-Wald references and are not asserted to have 95% coverage because Wald calibration is poor.

The exact-signature duplicate stress tests also preserve the composite pattern: keeping one row per signature gives 53/71 agent versus 355/644 human positives (OR 2.40, sign-flip $p = .0061$; uncalibrated nominal diagnostic), while dropping every member of a repeated-signature group gives OR 2.29 (sign-flip $p = .0097$; uncalibrated nominal diagnostic). The calibration study below does not cover either exact-signature scope. These tests use exact title/category/price/currency signatures, not the plan's normalized-text rule.

Removing the agent-heaviest mixed-label cluster reverses recurring monitoring: 3/36 (8.3%) versus 110/663 (16.6%). Only 20 of 462 clusters contain an agent-identified listing; the agent-heaviest-removed, all-mixed-removed, and RentAHuman-only samples contain 19, 13, and 12 treated clusters, respectively. Few treated clusters and unequal cluster sizes are both known to degrade cluster-robust inference [20, 21], so rather than assume adequate calibration, we evaluated the candidate procedures by simulation. The private calibration program uses a reproducible fixed seed. The null data-generating process has no group difference and reuses the observed cluster assignment and fixed agent-label indicator, preserving treated-cluster scarcity and size imbalance. Cluster random intercepts are normal on the logit scale. The intercept is re-solved at every standard deviation to hold the marginal event rate at the observed value, and each design uses a fixed, seeded set of 999 sign flips.

We do not estimate a single dependence parameter. Because most of the 462 clusters are singletons (the median cluster in both groups contains one listing), a within/between moment approach considered during development depends heavily on how singleton Bernoulli variation is handled; it was neither prospectively specified nor independently validated. We therefore sweep normal-logit random-intercept standard deviations (SDs) 0, 0.5, 1, 2, and 4. We also simulate beta-binomial cluster probabilities at intracluster correlations (ICCs) .15 and .30, providing a non-normal dependence sensitivity. These seven conditions include independent listings and substantial within-cluster heterogeneity, but they do not bound every plausible data-generating process.

The calibration covers all 56 core cells: 13 features and the composite in each of four scopes. It also retains the composite scope defined by excluding rows carrying the superseded round's internal flag—an exclusion that is now empty, so that scope coincides with the full comparison—plus three severe-rate scopes, for 60 designs total. Each design uses 1,000 datasets under each of the seven dependence conditions and a fixed set of 999 sign flips.

Across the 56 core cells, sign-flip rejection rates at nominal .05 range from .035 to .415. The four composite cells range from .037 to .077. In contrast, phone cells in the all-mixed-removed and RentAHuman-only scopes range from

.304 to .415, financial cells in those scopes range from .181 to .311, location cells from .115 to .199, and recurring monitoring reaches .103, as does public-post proof. Those samples retain only 19 or 20 agent listings across 12 or 13 treated clusters; under the pooled null, empty agent-positive cells are common and the Bernoulli score contributions are strongly asymmetric. The Rademacher sign-flip approximation is therefore not uniformly calibrated for the feature family.

The procedure draws on wild-cluster bootstrap methods and score perturbations for M-estimators [5, 16, 19] but is a bespoke fixed-denominator restricted cluster-score Rademacher sign-flip test, not any cited procedure; the simulation supplies design-specific evidence rather than a transferred validity theorem. Reported composite values use 9999 draws. The feature-family shared-sign max-$|T|$ calculation uses 99,999 draws, but no outer joint-family simulation validates it. BY also requires valid marginal p values. Accordingly, feature-level BH, BY, and max-$|T|$ values are nominal diagnostics, not tests with demonstrated error control. Table 27 summarizes the simulation. With 1,000 datasets, Monte Carlo SE is about .007 near .05 and at most .016. The table reports ranges for the named design families rather than every individual grid cell; cell-level outputs are retained privately.

Table 27. Empirical rejection rates at nominal .05 across five normal-logit random-intercept SDs and two beta-binomial ICCs. Every design preserves its observed cluster assignment and fixed agent-label indicator and targets its observed prevalence under no group difference. "Wald all" divides rejections by all datasets; "Wald finite" conditions on successful fits, and "unavailable" reports failed fits. Feature-family rows range over all 13 outcomes. The sign-flip column is marginal, not the joint max-$|T|$ calculation.

Comparison	Clusters (treated)	Wald all	Wald finite	Wald unavailable	Sign-flip all
Composite, full	462 (20)	0.118–0.237	0.118–0.237	0.000–0.000	0.053–0.077
Composite, drop agent-heaviest	461 (19)	0.060–0.088	0.060–0.088	0.000–0.000	0.048–0.056
Composite, drop all mixed	455 (13)	0.068–0.100	0.068–0.100	0.000–0.000	0.041–0.063
Composite, drop all mixed; RentAHuman only	454 (12)	0.065–0.103	0.065–0.103	0.000–0.001	0.037–0.061
All 13 features, full	462 (20)	0.094–0.273	0.094–0.275	0.000–0.190	0.043–0.223
All 13 features, drop agent-heaviest	461 (19)	0.029–0.096	0.030–0.096	0.000–0.279	0.035–0.288
All 13 features, drop all mixed	455 (13)	0.034–0.098	0.037–0.098	0.000–0.415	0.042–0.415
All 13 features, drop all mixed; RentAHuman only	454 (12)	0.038–0.112	0.039–0.112	0.000–0.413	0.038–0.413
Phone, drop all mixed	455 (13)	0.044–0.058	0.068–0.094	0.304–0.415	0.304–0.415
Phone, drop all mixed; RentAHuman only	454 (12)	0.048–0.064	0.072–0.102	0.329–0.413	0.330–0.413
Financial, drop all mixed	455 (13)	0.043–0.061	0.058–0.078	0.178–0.265	0.181–0.267
Financial, drop all mixed; RentAHuman only	454 (12)	0.049–0.062	0.063–0.083	0.207–0.311	0.208–0.311
Recurring monitoring, drop all mixed; RentAHuman only	454 (12)	0.038–0.057	0.040–0.059	0.026–0.091	0.060–0.103
Composite, legacy flags excluded	462 (20)	0.130–0.220	0.130–0.220	0.000–0.000	0.050–0.069
Severe rate, full	462 (20)	0.118–0.232	0.118–0.232	0.000–0.000	0.048–0.078
Severe rate, drop agent-heaviest	461 (19)	0.058–0.091	0.058–0.091	0.000–0.000	0.045–0.061
Severe rate, drop all mixed	455 (13)	0.072–0.094	0.072–0.094	0.000–0.000	0.044–0.056

1000 datasets per outcome and dependence condition. "Wald all" uses all datasets; "Wald finite" conditions on finite fits; "unavailable" is the non-finite share. Entries are min–max over the seven dependence conditions. Every dataset yielded a sign-flip result.

The composite remains positive and is nominally below .05 in all four scopes. Its sign-flip rejection rates are near, but not uniformly equal to, .05 on the studied grid. Its post-hoc selection remains unadjusted, and the evidence is weaker than the Wald numbers imply.

14.4 Category-and-Cluster Diagnostics

Table 25 reports listing-level references, and the feature subsection reports nominal clustered diagnostics. Table 26 reports the separately studied composite sensitivity. BH numerically transforms the 13 nominal feature values without

established FDR control. The Mantel–Haenszel estimate (Table 19) adjusts the composite for category without clustering, while the cluster-robust estimates address display-name dependence without category adjustment. A model addressing both category composition and display-name clustering is estimable on all 748 RQ3 rows, since no category is separated under the promoted labels. There, the category-adjusted logistic with display-name-clustered covariance gives OR 3.7 (normal-Wald reference interval 1.8–7.5, $p = 4.25 \times 10^{-4}$; 462 clusters). Omitting research fieldwork as well gives OR 1.8 (normal-Wald reference interval 0.7–4.5, $p = .222$; 443 clusters). Like the Mantel–Haenszel pooling, both estimates are exploratory, and their reference intervals inherit the few-treated-cluster caveat above.

Table 28. Display-name-cluster structure for the full 72-versus-676 comparison (461 RentAHuman display-name hashes plus the Human Pages singleton; 462 clusters). The 7 mixed-label display-name clusters occur in both within-label columns. Platform labels are retained; indented rows give the stated exclusions, including the agent-heaviest cluster (36 agent- and 13 human-identified listings).

Quantity	Agent-identified	Human-identified
Hash–label cells (a mixed hash appears in both columns)	20	449
Listings	72	676
Physical-world/location/monitoring composite, listing level	75.0%	55.3%
Composite, unweighted mean across hash–label cells	81.4%	55.2%
Composite OR, unadjusted for category; display-name-clustered		2.42 [1.40, 4.18]†, sign-flip $p = .0067$
all mixed-label hashes removed		7.37 [2.23, 24.38]†, sign-flip $p = .0022$
agent-heaviest mixed-label hash removed		3.36 [1.20, 9.40]†, sign-flip $p = .0169$
all mixed-label hashes and Human Pages removed (RentAHuman only)		14.73 [3.24, 66.89]†, sign-flip $p = .0003$
Severe-rate OR, unadjusted for category; display-name-clustered		1.10 [0.56, 2.19]†, sign-flip $p = .83$
agent-heaviest mixed-label hash removed		0.70 [0.27, 1.84]†, sign-flip $p = .49$
all mixed-label hashes removed		0.44 [0.18, 1.07]†, sign-flip $p = .15$

†Brackets are normal-Wald reference intervals, not asserted 95% coverage intervals; p values use the bespoke restricted cluster-score Rademacher sign-flip test. No aligned interval is constructed.

15 Ten-Day Follow-Up Snapshot (No New Proof Labels)

On June 10, 2026—ten days after collection began on May 31—we reran the documented public-query protocol with the same endpoints, status and category partitions, and politeness settings, and compared the result with the June 1 snapshot by bounty identifier (Table 29). This comparison uses only platform metadata—public reachability, status, agent-type label, category, and creation timestamp—so it requires no new proof labels, and we make no proof-burden claims about follow-up listings. Aggregate tables are generated by the private analysis pipeline; the follow-up raw captures stay local, like all raw captures in this study. Because the original manifest does not identify a collector commit, exact code identity across the two collections cannot be verified.

The comparison has three descriptive findings. First, the public query surface changed over ten days: only 476 of the 980 June 1 RentAHuman listings (48.6%) were still reachable through the same queries. The follow-up also returned 67 listings not seen in the June 1 snapshot. For this classification, we define the end of the initial collection at its documented completion timestamp, `2026-06-01T02:32:58Z`. Of the 67 first-seen listings, 23 were created after that cutoff; the other 44 were older listings newly returned by the queries, showing that return through the sampled query surface varies under deep pagination. Absence is therefore "no longer publicly reachable through these queries," not necessarily deletion.

Second, absence from the follow-up concentrates in cancelled listings: 81.9% of June 1 cancelled listings were no longer reachable, versus 7.7% of completed ones, while only 4 of the 476 still-reachable listings changed status (0.8%).

The cancelled share of any single snapshot therefore reflects one point in a changing public query surface, which reinforces the main paper's caution against reading status shares as a task lifecycle.

Third, newly created agent-identified listings were rare: just 1 of the 23 newly created listings (4.3%) carries an agent label; 18 carry a human label and 4 carry an ambiguous "other" label. Agent-identified June 1 listings also remained reachable at a higher rate (69 of 71, 97.2%) than human-identified listings (376 of 870, 43.2%). The 97.2% agent-identified reachability rate exceeds a 70.4% status-composition benchmark, obtained by applying the full June 1 RentAHuman sample's status-specific reachability rates to the agent-identified group's status distribution. This is a short-window reachability check; it does not establish growth in new agent-identified requester activity or a temporal trend. The snapshot's 7.3% (72/981) agent-identified share spans all 981 rows, including Human Pages; the newly created and reachability comparisons here use RentAHuman rows only.

Table 29. June 1 versus June 10 RentAHuman public snapshots, matched on bounty identifier. The comparison uses no new human-coded proof labels, only platform metadata and platform-provided status and agent-type labels; "absent" means no longer reachable through the same public queries.

Quantity	Value
June 1 snapshot listings	980
June 10 follow-up listings	543
June 1 listings still returned	476 (48.6%)
June 1 listings absent from follow-up	504
Still-reachable listings with a status change	4 (0.8%)
First seen in follow-up	67
created after collection completed	23
older but newly reachable	44
Agent-identified share among June 1 listings	7.2%
Agent-identified share among all June 10 follow-up listings	13.1%
Agent-identified share among newly created listings	4.3%
newly created agent-identified / human / other	1 / 18 / 4
Still reachable: agent-identified listings	69 of 71 (97.2%)
Still reachable: human-identified listings	376 of 870 (43.2%)
Agent-identified reachability expected from status-specific rates	70.4%

Because reachability is not a task-lifecycle state, this is only a descriptive metadata check. Every June 1 RentAHuman listing carries a follow-up indicator built only from platform metadata—whether it remained publicly reachable on June 10—so proof burden and early public reachability can be compared descriptively (Table 30). The indicator-to-listing mapping and the raw June 1 and June 10 captures are retained privately and do not accompany this PDF. The aggregate results needed to interpret the check are reported below, but an outside reader cannot independently rebuild the indicator from the PDF alone. The status-stratified percentages in Table 30 predate the promoted recode and classify listings by the superseded first round's severe indicator; they are retained as that round's record and were not regenerated, so they should be read as a superseded-label diagnostic. Under that classification, cancelled listings had nearly identical observed reachability shares by severe-threshold group (18.0% versus 18.2%), and among non-cancelled status strata, point estimates were lower for severe-threshold-positive listings—76.9% versus 98.0% still

reachable among assigned listings, and 69.4% versus 86.5% among open listings—but some strata are too small for a stable comparison (the non-severe partially filled cell holds three listings). The logistic model—fit in the same pre-recode run, so its severe-threshold indicator is likewise the superseded round's—uses reachability as the outcome and the proposed severe-threshold indicator as the predictor, adjusts for status and an agent-label indicator, and applies RQ3's label inclusion rule within RentAHuman: 941 named-agent/bot or human-labeled listings, with the 39 ambiguous "other" rows excluded. The point estimate for threshold-positive listings is lower (OR 0.60, 95% CI 0.41–0.89, $p = .0098$). Clustering covariance by 594 public display-name hashes leaves the OR unchanged but widens its 95% CI to 0.30–1.19, which includes 1 ($p = .145$). Among human-identified listings alone, a status-adjusted model—recomputed under the promoted labels on the 676 eligible human-labeled listings—gives a similar OR 0.79 (unclustered $p = .250$; display-name-cluster-robust $p = .48$). We read these as lower descriptive point estimates, not evidence of a stable association or mechanism: absence can reflect filling, withdrawal, hiding, moderation, or query-surface variation.

Ten days on, the auxiliary platforms were unchanged: Human Pages still exposed a single public listing and the GoHireHumans endpoint still returned no jobs, so the follow-up remains centered on RentAHuman.

Table 30. Early public reachability by proposed severe-threshold status, within June 1 listing-status strata (all 980 RentAHuman listings). The stratified percentages classify listings by the superseded first round's severe indicator and are retained as that round's record. The logistic rows use the 941 named-agent/bot or human-labeled listings, excluding 39 ambiguous "other" rows; reachability is the outcome and the severe-threshold indicator is the predictor, with status and agent-label adjustment. Descriptive; "still reachable" means returned on June 10 through the same public queries.

Status	Severe (≥4) % returned June 10	Non-severe % returned June 10
open	69.4% (n=108)	86.5% (n=37)
assigned	76.9% (n=52)	98.0% (n=50)
partially filled	76.5% (n=17)	100.0% (n=3)
completed	90.6% (n=106)	94.7% (n=76)
cancelled	18.0% (n=344)	18.2% (n=187)
Agent/human-label logit; severe predictor	OR 0.60, 95% CI [0.41, 0.89], $p = .0098$	
same model; display-name clustered	OR 0.60, 95% CI [0.30, 1.19], $p = .145$	

16 Proof and Context Feature Co-occurrence

Figure 1 reports pairwise co-occurrence *lift*: how often two features appear together compared with the frequency expected under independence. Formally, lift is the observed joint prevalence divided by the product of the two features' individual prevalences, so a value above 1 means a pair co-occurs more than this independence baseline. This is a descriptive view, not a formal clustering or permutation test. The author ordered rows and columns to place two theory-informed feature groups next to each other; no algorithm discovered those groups. The computed lift values can therefore describe whether the chosen grouping is consistent with the observed co-occurrences, but the visual blocks are not independent evidence for the grouping. Individual cells involving rare features rest on few joint occurrences, and we attach no uncertainty to them. In the author-specified *physical-world/location-adjacent* group are physical-world action, location proof, recurring monitoring, photo proof, and financial proof. This is broader than the main paper's three-feature measure (this supplement's physical-world/location/monitoring composite); within it, location proof and physical-world action have lift 2.0. The author-specified *digital/account-linked* group includes account proof,

public-post proof, identity proof, screenshot proof, and link proof. Account proof and public-post proof have lift 3.8, screenshot proof and public-post proof 3.1, and identity proof and public-post proof 2.8. In this snapshot, pairs crossing the two chosen groups usually co-occur less often than the independence baseline predicts: most cross-group cells have lift below 1. Most of the exceptions pair recurring monitoring or financial proof with a digital/account-linked feature and sit barely above 1; the largest is location proof×identity proof at lift 1.9. Read descriptively, these values are consistent with one author-specified group centered on workers' physical activity, places, or surroundings and another centered on accounts and account-linked artifacts. Only public-post proof necessarily denotes public posting, so we do not read the whole pattern as reputational. This descriptive view helps interpret why the threshold-group and agent-identified comparison summaries separate physical-world/location-adjacent features from digital/account-linked features.

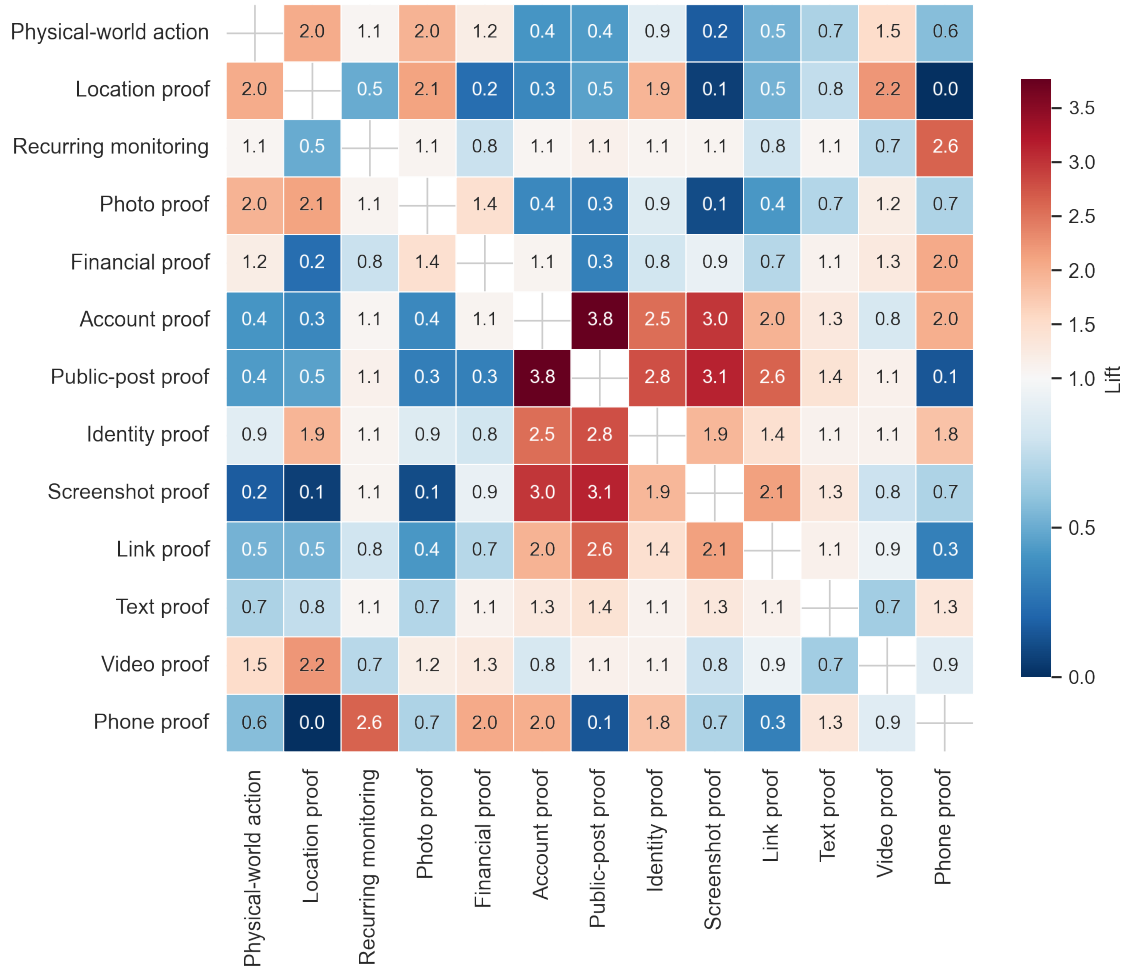

Fig. 1. Pairwise completion-evidence, recurring-monitoring, and physical-world-action feature co-occurrence lift (1 = independence baseline; diagonal omitted). Red (>1) marks pairs that co-occur more than the baseline, blue (<1) less than the baseline. The feature order is author-specified rather than algorithmically derived. Lift is relatively high within two author-specified regions—physical-world/location-adjacent (upper-left) and digital/account-linked (center)—but no formal clustering test is used.

17 Intended Use and Non-Uses

At present the taxonomy and PBS form a candidate manual coding framework: a protocol applied by trained coders to public listing text to describe what completion evidence a corpus of listings advertises. Calling it a candidate framework emphasizes that this study tests its use on one snapshot but does not validate it as a general-purpose instrument. Three further uses are commonly expected of such a framework, and the evidence assembled here supports none of them.

Worker-facing disclosure. The 13 feature labels and the profile summaries are plausible disclosure candidates, but no worker has evaluated either. PBS is a screening total whose base tiers, equal-weight modifier points, cap, and

cutpoints are author conventions rather than estimated quantities; it should not be shown to workers as a burden rating, used to rank or filter tasks, or used to set or justify pay.

Platform moderation. Nothing here supports removal, penalty, or requester sanction. Threshold-positive listings are not shown to be harmful: the score records advertised exposure-relevant requirements, not experienced burden or realized harm, and the ten-day follow-up yields only an imprecise reachability contrast whose display-name-clustered interval includes 1.

Automated extraction at scale. The rule-based first pass is not adequate on its own. Its score differs from the promoted adjudicated score on 373 of the 779 primary rows (47.9%; 459 of all 981 returned rows). Per-feature recall against the promoted labels, computed from the counts in the main paper's rule-versus-audit table, ranges from 10.4% for identity proof and 13.9% for recurring monitoring to 94.8% for link proof; precision ranges from 29.0% for location proof and 40.8% for phone proof to 96.7% for screenshot proof. The two lowest-recall variables anchor two of the three largest threshold-positive groups, so a rule-only pipeline would systematically miss the profiles of greatest interest while over-flagging location and phone proof.

What the single score adds. A platform could simply publish the required proof types, and for a worker judging one task that list is in several respects the better instrument: it is directly inspectable, imposes no ordering, and preserves the 154 distinct feature profiles that PBS compresses into scores 4 and 5. The score's value, if any, is corpus-level: it supplies one ordinal summary for describing distributions, comparing subgroups, and entering diagnostic models, which a 13-dimensional profile cannot do compactly. That convenience is not evidence that the ordering matches worker-perceived burden, and the three-arm comparison described next is meant partly to test whether the score adds anything workers can use.

18 Planned Worker-Facing Validation

The rubric's base tiers and modifier points are a transparent heuristic, not estimates of workers' experience. Privately retained worker-facing planning drafts are unregistered, unpowered, and unpiloted; they are not an approved or field-ready study and do not accompany this PDF. A central next study should compare a feature checklist, PBS, and both together while holding the underlying task constant. In worker-neutral language and redacted examples, the current prototypes provide four starting components: multidimensional ratings of all 13 coded variables; a balanced best–worst exercise estimating relative utilities without a common rating scale [18]; paired vignettes; and willingness-to-apply and minimum-additional-pay questions. They do not yet implement that three-arm disclosure comparison.

The paired vignettes compare: an online answer plus a note with an in-person store visit plus the same note (`physical_with_proof`); one photo with a photo, video, and written report (`multiple_evidence`); and a private screenshot with a public post from the worker's own account plus a submitted link (`public_or_account`). Only the physical-with-proof pair approximates a single-modifier contrast: it keeps the text-acknowledgement requirement approximately fixed while adding physical-world action (m_2). The multiple-evidence pair changes both evidence count and the highest base feature (photo alone versus a set that includes video), so it does not isolate m_1. The public-or-account pair changes the base feature and tier along with the m_3 trigger, so it does not isolate m_3. Those two pairs must be redesigned or prospectively treated as combined contrasts. The sample size, estimands, multiplicity plan, exclusions, and analysis would likewise need to be justified and locked before fielding.

The best–worst descriptions do not map one-to-one onto the PBS base tiers and modifier points: standalone phone lacks a historical base tier, physical-world action is context, and text/link tiers depend on context. Worker judgments could inform a prospectively specified index or sensitivity analysis, but cannot simply replace the rubric's point values

and recompute PBS. The design would prespecify attention, response-pattern, timing, and free-text checks and apply exclusions before scoring; such controls matter because prior work reports nontrivial language-model use in an MTurk text-summarization task [35]. Fielding would require prospective institutional ethics review or determination, consent as applicable, and fair compensation. Recruitment should span workers experienced in digital and physical gig work and collect open-ended accounts of why the same requirement may affect workers differently.

With the independent full-context recode complete, a separate extraction study should compare transparent rules with context-aware classifiers—fine-tuned transformer encoders and prompted or instruction-tuned language models—first against the completed 981-row human annotations as a benchmark, then on later and cross-platform held-out listings. Benchmarking against the existing annotations bounds what such a system could be credited with: they are one team's labels, the weakest features still sit at $\kappa = .76–.81$, and one adjudicator supplied final labels for all 873 routed rows, so agreement with them is a reproduction check rather than evidence of construct validity. The study should report per-feature calibration and subgroup errors, preserve an abstention option, and route uncertain, disagreement-prone, or high-burden cases to human review, restricting any automated path to higher-agreement features such as link, screenshot, photo, and video proof ($\kappa \geq .94$). Neither the current rules nor a future classifier should automate worker-facing scores, ranking, payment, moderation, or enforcement without external construct and use validation.

19 Closing Interpretive Cautions

Two cautions govern this supplement. First, the labels operationalize advertised exposure-relevant completion requirements, not worker experience. Because the function served by each requirement was not coded, the same positive label can refer to a submitted work product, an access channel, infrastructure needed for the task, or a condition extending over time. Agreement and provenance do not make PBS an estimate of marginal verification burden. Worker-facing validation remains necessary. Second, unkeyed hashes and distinctive public metadata can relink rows to live listings that may name third parties. Raw captures therefore stay private, examples are paraphrased, row-level processed data remain nonpublic, and any future sharing must use aggregate or appropriately minimized materials unless a separate disclosure-control review supports more.

References

[1] Elena Agapie, Jaime Teevan, and Andrés Monroy-Hernández. 2015. Crowdsourcing in the Field: A Case Study Using Local Crowds for Event Reporting. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 3, 1 (2015), 2–11. https://doi.org/10.1609/hcomp.v3i1.13235

[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[3] Yoav Benjamini and Daniel Yekutieli. 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* 29, 4 (2001), 1165–1188. https://doi.org/10.1214/aos/1013699998

[4] Norman E. Breslow and Nicholas E. Day. 1980. *Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies.* Number 32 in IARC Scientific Publications. International Agency for Research on Cancer, Lyon, France.

[5] A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics* 90, 3 (2008), 414–427. https://doi.org/10.1162/rest.90.3.414

[6] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000104

[7] Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70, 4 (1968), 213–220. https://doi.org/10.1037/h0026256

[8] Lee J. Cronbach and Paul E. Meehl. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin* 52, 4 (1955), 281–302. https://doi.org/10.1037/h0040957

[9] Adamantios Diamantopoulos and Heidi M. Winklhofer. 2001. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research* 38, 2 (2001), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

[10] Kimberly Do, Maya De Los Santos, Michael Muller, and Saiph Savage. 2024. Designing Gig Worker Sousveillance Tools. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 384, 19 pages. https://doi.org/10.1145/3613904.3642614

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[12] Kotaro Hara, Abi Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 449, 14 pages. https://doi.org/10.1145/3173574.3174023

[13] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. 2013. Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms. *Mathematical and Computer Modelling* 57, 11–12 (2013), 2918–2932. https://doi.org/10.1016/j.mcm.2012.01.006

[14] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. Association for Computing Machinery, New York, NY, USA, 64–67. https://doi.org/10.1145/1837885.1837906

[15] Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey P. Bigham, Jaime Teevan, Ece Kamar, and Walter S. Lasecki. 2017. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5, 1 (2017), 89–98. https://doi.org/10.1609/hcomp.v5i1.13314

[16] Patrick Kline and Andres Santos. 2012. A Score Based Approach to Wild Bootstrap Inference. *Journal of Econometric Methods* 1, 1 (2012), 23–41. https://doi.org/10.1515/2156-6674.1006

[17] Chen Liang, Jing Peng, Yili Hong, and Bin Gu. 2023. The Hidden Costs and Benefits of Monitoring in the Gig Economy. *Information Systems Research* 34, 1 (2023), 297–318. https://doi.org/10.1287/isre.2022.1130

[18] Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9781107337855

[19] James G. MacKinnon, Morten Ørregaard Nielsen, and Matthew D. Webb. 2025. Cluster-Robust Jackknife and Bootstrap Inference for Logistic Regression Models. *Econometric Reviews* (2025), 1–29. https://doi.org/10.1080/07474938.2025.2515161

[20] James G. MacKinnon and Matthew D. Webb. 2017. Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics* 32, 2 (2017), 233–254. https://doi.org/10.1002/jae.2508

[21] James G. MacKinnon and Matthew D. Webb. 2018. The Wild Bootstrap for Few (Treated) Clusters. *The Econometrics Journal* 21, 2 (2018), 114–135. https://doi.org/10.1111/ectj.12107

[22] James G. MacKinnon and Halbert White. 1985. Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics* 29, 3 (1985), 305–325. https://doi.org/10.1016/0304-4076(85)90158-7

[23] Nathan Mantel and William Haenszel. 1959. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute* 22, 4 (1959), 719–748. https://doi.org/10.1093/jnci/22.4.719

[24] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2271–2282. https://doi.org/10.1145/2858036.2858539

[25] Donald Moynihan, Pamela Herd, and Hope Harvey. 2015. Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions. *Journal of Public Administration Research and Theory* 25, 1 (2015), 43–69. https://doi.org/10.1093/jopart/muu009

[26] Robert G. Newcombe. 1998. Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods. *Statistics in Medicine* 17, 8 (1998), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

[27] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (2004), 119–157.

[28] RentAHuman. 2026. How to Hire a Human: Bounties Documentation. https://rentahuman.ai/docs/bounties Accessed July 30, 2026.

[29] Shruti Sannon and Dan Cosley. 2019. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 282, 12 pages. https://doi.org/10.1145/3290605.3300512

[30] Shruti Sannon, Billie Sun, and Dan Cosley. 2022. Privacy, Surveillance, and Power in the Gig Economy. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 619, 15 pages. https://doi.org/10.1145/3491102.3502083

[31] Daniel J. Solove. 2006. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154, 3 (2006), 477–560.

[32] Hien To, Gabriel Ghinita, and Cyrus Shahabi. 2014. A Framework for Protecting Worker Location Privacy in Spatial Crowdsourcing. *Proceedings of the VLDB Endowment* 7, 10 (2014), 919–930. https://doi.org/10.14778/2732951.2732966

[33] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the Invisible Labor in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, Article 319 (2021), 26 pages. https://doi.org/10.1145/3476060

[34] Ward van Zoonen, Monika E. von Bonsdorff, and Beatrice I. J. M. van der Heijden. 2026. Algorithmic Surveillance and Workers' Compliance: The Role of Trust, Privacy Concerns, and Fairness in Online Crowdwork. *Human Relations* 79, 7 (2026), 795–824. https://doi.org/10.1177/00187267251379698

[35] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv preprint arXiv:2306.07899. https://doi.org/10.48550/arXiv.2306.07899

[36] Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212. https://doi.org/10.1080/01621459.1927.10502953

[37] Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017. "Our Privacy Needs to Be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, Article 113 (2017), 22 pages. https://doi.org/10.1145/3134748

[38] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* PublicAffairs, New York, NY, USA.