

From Messy Data to Meaningful Insight

Emotion & Appraisal Video Study

Research audit | 13 October 2026 | Exploratory — not a confirmatory publication

Executive overview

A striking 13,264-row psychology study turns out to contain only 434 worker IDs. When the unit of analysis changes, the apparent association changes too.

The dataset appeared ready for a quick emotional-intensity analysis. Instead, I audited its row structure, quality flags, durations, and scoring formulas before interpreting any association.

Dataset provenance and privacy

Uploaded file: full_data(4).csv | Rows: 13,264 | Columns: 127 | Source: Source DOI/URL not verified for supplied file

The uploaded participant-level CSV is deliberately withheld from all public artifacts. This report contains only aggregate, non-row-level results. Source dataset license and original codebook status must be rechecked before public publication.

Phase 1 — Audit

- 13,264 recorded rows, 127 fields; 8.35% literal missing cells.
- 0 whole-row duplicates; structure: 434 unique worker IDs; 3316 tasks (HITs); 188 videos.

Null cells: 140,620 (8.348% of all cells); blank-string cells: 0. Whole-row duplicate count: 0. Numeric columns: 121; other: 6.

Missingness and coding patterns

- CHECK_music_never: 95.27% literal nulls
- CHECK_sleep_per_night: 95.14% literal nulls
- CHECK_ocean_air_balloon: 94.90% literal nulls
- CHECK_speak_day: 94.87% literal nulls
- CHECK_middle_name: 94.82% literal nulls
- intensity_2: 61.96% literal nulls

Observed negative codes: . These are NOT automatically classified as missing. Consult source coding manual.

Observed integrity flags

- quality_flagged: 4
- all_fail: 1226
- time_fail: 352
- duration_over_24h: 24

- max_duration_hours: 312.24
- task_repeats: 9948
- workers: 434

Phase 2 — Analyses and reproducibility

Exploratory associations use Spearman rank correlation with pairwise available cases. Group comparisons use Welch t-tests and pooled-SD Cohen d, but group meanings must be verified. Reliability uses complete cases only and is diagnostic in the absence of confirmed scale keying.

Worker-level aggregated Spearman: $\rho=0.0346$, $n=432$, $p=0.473$; this between-worker analysis shows little overall association. Random-intercept worker mixed model ($n=12,315$, worker groups=432): slope=-2.377, 95% CI [-2.686,-2.067], $p=4.53e-51$. These models answer different questions. IMPORTANT: the random-intercept covariance is singular/at the boundary; MixedLM confidence intervals and p values cannot be treated as reliable evidence. The mixed model also does not model crossed video/HIT effects. Worker-cluster-robust OLS is reported separately as a sensitivity comparator: {'slope': -2.238, 'p': 1.1227190859932084e-16, 'ci_low': -2.767, 'ci_high': -1.709, 'n_rows': 12315, 'n_workers': 432, 'type': 'Pooled marginal OLS, cluster-robust covariance by worker; not a random-effects model'}.

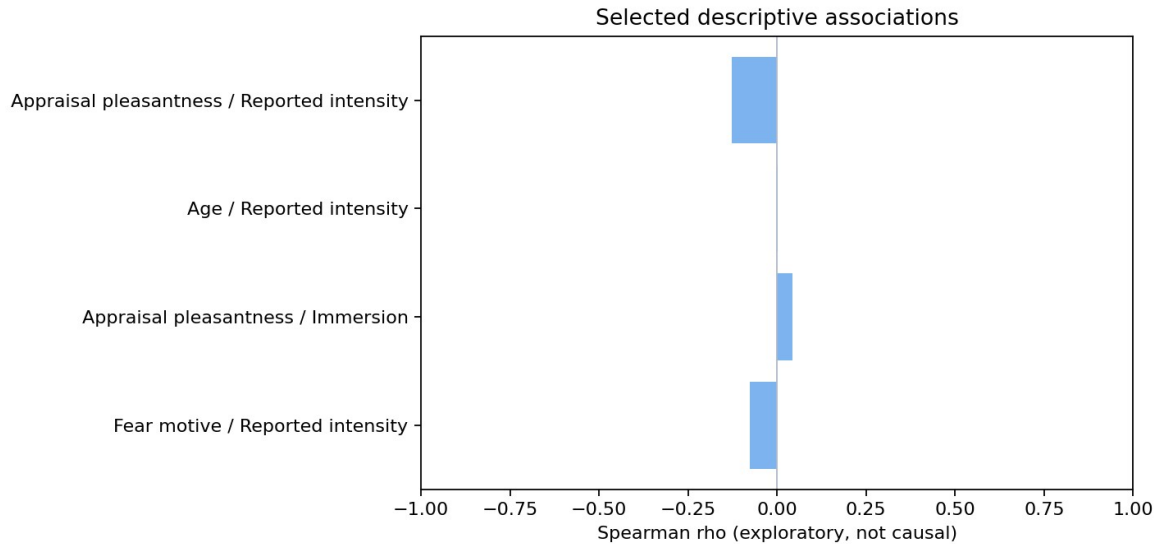
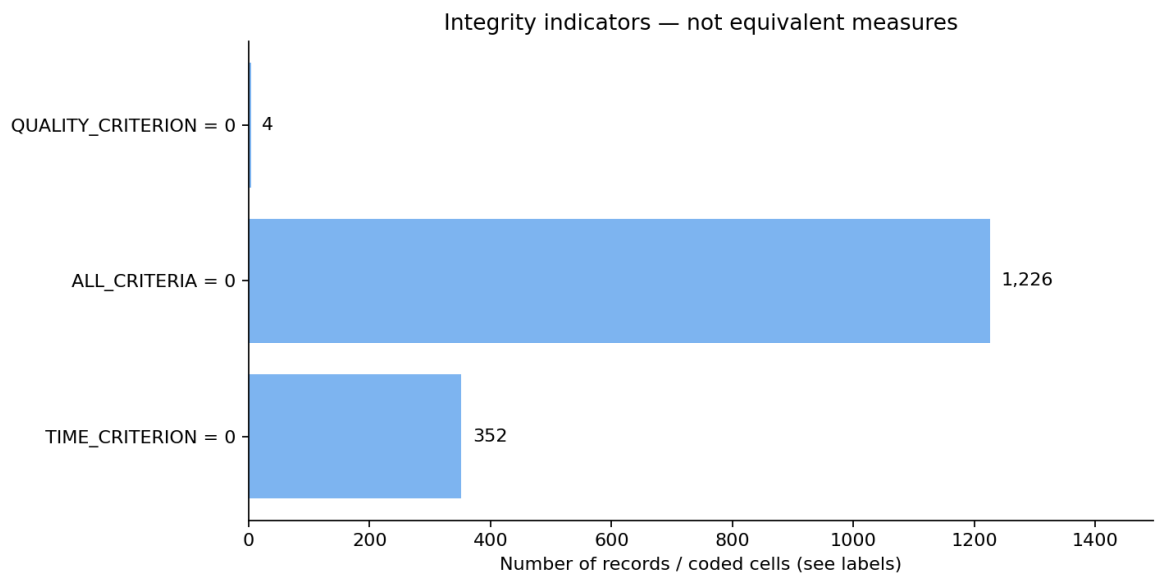
Diagnostic FEAR1/2/3 alpha: 0.844; one-factor omega approximation: 0.854. No validated scoring rule supplied.

Stored named composites vs raw item means were tested as diagnostic comparisons, not official scoring validation:

- {'score': 'FEAR', 'nonmissing_compared': 13260, 'mismatch_gt_0_01': 0, 'max_abs_diff': 0.0}
- {'score': 'ACH', 'nonmissing_compared': 13260, 'mismatch_gt_0_01': 0, 'max_abs_diff': 0.0}
- {'score': 'INT', 'nonmissing_compared': 13260, 'mismatch_gt_0_01': 0, 'max_abs_diff': 0.0}
- {'score': 'AFF', 'nonmissing_compared': 13260, 'mismatch_gt_0_01': 0, 'max_abs_diff': 0.0}

Selected Spearman associations

- Appraisal pleasantness / Reported intensity: $\rho=-0.1272$, $n=12,315$, $p=1.44e-45$ (exploratory)
- Age / Reported intensity: $\rho=+0.0022$, $n=12,315$, $p=0.807$ (exploratory)
- Appraisal pleasantness / Immersion: $\rho=+0.0423$, $n=13,182$, $p=1.17e-06$ (exploratory)
- Fear motive / Reported intensity: $\rho=-0.0775$, $n=12,311$, $p=6.97e-18$ (exploratory)



From findings to action

Records are not people

What we found — 13,264 rows arise from 434 worker IDs, with four entries per HIT.

Why it matters / What to do — Treat workers and HITs as repeated structures rather than 13,264 independent participants.

Quality thresholds disagree

What we found — Only 4 rows fail QUALITY_CRITERION, while 1,226 fail ALL_CRITERIA.

Why it matters / What to do — Review every component flag before calling a dataset clean or unusable; preserve both views.

Aggregation changes the story

What we found — The worker-aggregated Spearman rho is +0.035 (n=432), while a singular worker-intercept model returns -2.377, a numerically unstable estimate requiring caution.

Why it matters / What to do — Report between-worker and conditional-model patterns separately, not as interchangeable effect estimates.

Extreme task durations remain visible

What we found — 24 rows have dur_in_sec above 24 hours; the recorded maximum is 312.24 hours.

Why it matters / What to do — Investigate collection logs and timestamp rules before quality filtering or interpretation.

Data limitations — retained, not silently fixed

- Worker IDs repeat across observations; treating 13,264 rows as independent people is incorrect. HITs and videos also repeat.
- QUALITY_CRITERION and ALL_CRITERIA measure different combinations of checks. Their disagreement is not proof of a computation bug without the original rulebook.
- Duration outliers exceed a day; preserved, not removed or truncated.
- Item scoring and the provenance of coded variables have not been verified. FEAR reliability and component checks are diagnostic.
- Missing responses are nonrandom across survey sections and need instrument/branching documentation.

The mixed model accounts for worker intercepts, not crossed HIT or video dependence. Neither the row-level nor worker-level result demonstrates causation.

Reproducibility and citations

This bundle contains aggregate-results.csv, variable-dictionary.csv, two figures in PNG and SVG, a Python script, and references.bib. Original participant-level files are withheld. See REPRODUCIBILITY.md for exact command lines. This report is exploratory. Original source licenses are not established for the first two uploads.

Core methods: Spearman (1904), Cronbach (1951), McDonald (1999), Welch (1947), Cohen (1988), statsmodels MixedLM, scikit-learn FactorAnalysis. DOI/URL in references.bib.