STATISTICAL FOUNDATIONS OF AI 2026 · UC SANTA BARBARA

Code availability is not enough: Schemas for documenting and extending data analysis decisions

H. Sherry Zhang (hsherryzhang@utexas.edu) · Roger D. Peng | Department of Statistics and Data Science, The University of Texas at Austin

tines website

paper repo

tines

Where are the decisions?

A data analysis is a series of decisions, from cleaning the data to specifying the model; readers need them to **judge an analysis and explore alternatives**. Yet they are **scattered and buried**, as in this Human Development Index (HDI) example:



manuscript pages illustrative

1. Rescale each indicator to $[0, 1]$ by **min–max**.
2. **Average** the two schooling indicators into one education index.
3. Combine the three dimension indices by their **geometric mean**.
   WHY Penalises uneven development: a low score in one dimension cannot be offset by a high one, unlike the arithmetic mean.

Zoom into the code: decisions are **interleaved with implementation details** (reading, renaming, type conversion) in the same syntax, so a decision does not *look* like one.

```
hdi ← read_csv("data/hdi_inputs_2022.csv") |>                implementation
  janitor::clean_names() |>                                  implementation
  mutate(across(le:gni_pc, as.numeric)) |>                   implementation
  mutate(across(le:gni_pc, scales::rescale),                 decision: min–max scaling
         edu = (eys + mys) / 2,                              decision: average
         hdi = (le * edu * gni_pc)^(1/3)) |>                 decision: geometric mean
  arrange(desc(hdi))                                         implementation
```

Working out exactly what was done falls on the reader, who must piece the decisions together without knowing where, or *whether*, each one is recorded.

Take-home

1. **Decisions deserve their own record.** A schema states each step's objective, decision and rationale in one self-contained place, instead of scattering them across the manuscript and the code.
2. **Humans decide; LLMs write the code.** Given the decisions, generating and validating code is close to mechanical.
   Case 2: code from extracted schemas reproduces most of 25 teams.
3. **Any reader can ask "what if?"** An alternative file turns it into a multiverse of runnable analyses.
   Case 1: another analyst's refit matches the reported AIC.
4. **An automatic reproducibility check.** When code from an extracted schema misses the reported result, the gap shows where a report leaves a decision unstated.
   Case 2: 3 of the 6 gaps trace to the report.

REFERENCES

Zhang & Peng (2026). Code availability is not enough: schemas for documenting and extending data analysis. *In preparation.*
Yee, Frigau & Ma (2025). Heaping and seeping, GAITD regression and doubly constrained reduced-rank vector generalized linear models in smoking studies. *Ann. Appl. Stat.* 19(4), 3045–3070.
Silberzahn et al. (2018). Many analysts, one data set. *Adv. Methods Pract. Psychol. Sci.* 1(3), 337–356.

A schema, and its alternatives

A schema is a YAML file that records the analysis as a sequence of steps, each with an **objective**, a **decision** and its **rationale** in plain sentences. The HDI schema, with the three decisions in one place (numbered as on the left):

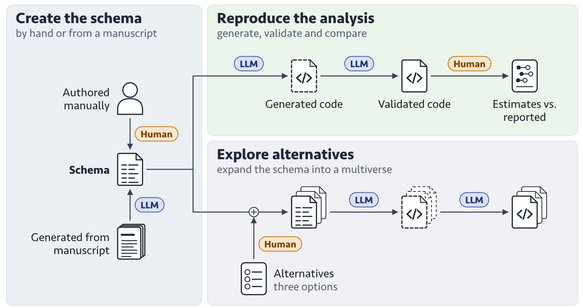
```
   meta:                                                     Human
     type: schema                                            fills in a template
   nodes:
1  - id: step-scaling
     objective: how to put the variables on one scale?
     decision: apply min-max scaling to each variable
2  - id: step-education
     objective: how to combine the education variables?
     decision: take the average
3  - id: step-combine
     objective: how to combine the three dimensions?
     decision: use the geometric mean
     rationale: penalizes uneven development
```

An **alternative file** lists other decisions for the same objectives: `branch: single` gives one new analysis, `branch: multi` runs every combination.

```
   meta:                                                     Human
     type: alternatives                                      another analyst's what-ifs
     branch: multi
   nodes:
2  - overrides: step-education
     alternatives:
     - decision: take the geometric mean
     - decision: take the harmonic mean (2ab / (a + b))
3  - overrides: step-combine
     alternatives:
     - decision: use the arithmetic mean (pre-2010 HDI)
```

With the original decisions: (2 + 1) × (1 + 1) = **6 analyses**.

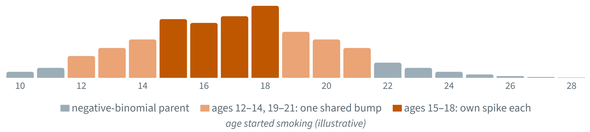
Workflow: humans decide, LLMs translate



Generate, then validate. The LLM writes a script from the schema. A validation agent runs it in a sandbox; on failure, it receives the error message and line number and returns a revised script, until the script runs or a retry limit is reached. Every analysis from an alternative file runs through this loop.

Case 1 Extending a published analysis

Data and model (Yee, Frigau & Ma, 2025, *Ann. Appl. Stat.*). In the NHANES 2017–2020 smoking questionnaire, respondents report the age they started smoking (`SIA`), which is **heaped**: certain ages are reported far more often than a count distribution predicts. GAITD regression modifies a negative-binomial parent at chosen *special* ages, by inflation, deflation, alteration or truncation. **Which ages are special, and which operation each gets, are modeling decisions**, guided by a spikeplot of `SIA`. The authors chose:



STEP 1 **Build the schema (3 of 14 steps) and its code**

```
- id: special_value_high_risk                               Human
  decision: give ages 15–18 each its own extra              fills in a template from the paper
    probability (nonparametric inflation)
  rationale: peak teenage uptake; the largest,
    most persistent spikes in the spikeplots
- id: special_value_mid_risk
  decision: give ages 12–14 and 19–21 one shared
    extra mass, shaped like the parent
  rationale: early and later smokers around the
    peak show a moderate, secondary elevation
- id: covariate_selection
  decision: include race, gender, marital, educ
  rationale: standard adjustments in smoking studies
```

```
# Step 8: special_value_mid_risk                           LLM
i_mix_spec ← c(12, 13, 14, 19, 20, 21)                     from the schema
# Step 9: special_value_high_risk
i_mlm_spec ← c(15, 16, 17, 18)
# Step 13: covariate_selection
cov_fml ← SIA_corrected ~ race2 + gender + marital + educ + ...
# Step 14: fit_final_model
fitted_model ← vglm(cov_fml, data = df_ever_smokers,
  family = gaitdnbinomial(i.mix = i_mix_spec, i.mlm = i_mlm_spec, ...))
```

STEP 2 **Another analyst asks "what if?"**
Could ages 19–21, where uptake declines, be **altered** instead?

```
meta: {type: alternatives, branch: single}                 Human
nodes:                                                     another analyst's what-if
- overrides: special_value_mid_risk
  alternatives:
  - decision: inflate ages 12–14 only
    rationale: 19–21 show declining uptake, a
      different pattern from 12–14
- overrides: alteration_operator_decision
  alternatives:
  - decision: alter ages 19–21 (A.mix)
- overrides: covariate_selection
  alternatives:
  - decision: intercept-only model
    rationale: compare the two age structures
      on equal footing
```

STEP 3 **Code for the alternative**

```
# Step 8: special_value_mid_risk                           LLM
i_mix_spec ← c(12, 13, 14)                                 from schema + alternative file
# Step 10: alteration_operator_decision
a_mix_spec ← c(19, 20, 21)
# Step 13: covariate_selection
covariate_formula ← SIA_corrected ~ 1
```

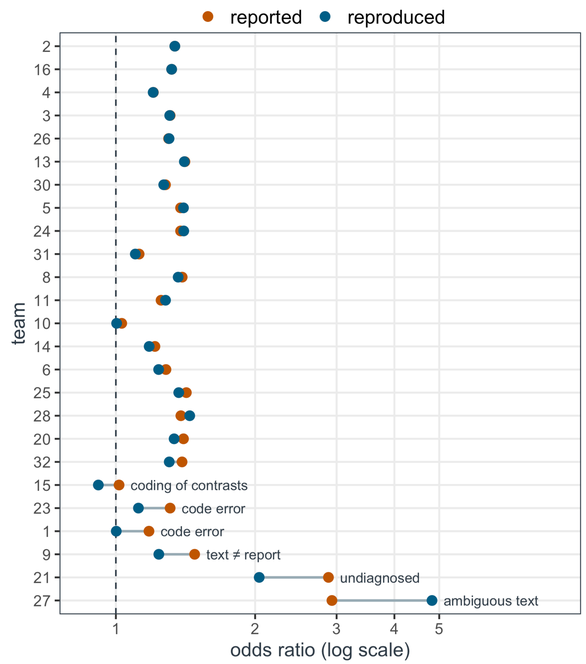
Anyone can reproduce and extend it. The alternative matches the supplement's AIC of 19,340.4, with no access to the authors' code.

Case 2 Many analysts, one dataset

Data and question (Silberzahn et al., 2018). 29 teams got the **same data**, player–referee pairs from top European soccer leagues with skin-tone ratings, red cards and covariates, and the **same question**: do referees give more red cards to dark-skin-toned players? **Teams made different decisions**, in preprocessing, model family (linear, logistic, Poisson, Bayesian) and random effects, and each wrote a methodology report with an **estimated odds ratio**.

STEP 1 **Extract a schema from each team's report**
Claude Opus 4.5 reads each team's methodology report and records every decision it describes as a schema step.

STEP 2 **Generate code and compare odds ratios**
If the schema is faithful, our code's odds ratio matches the team's (25 teams; 4 too costly to run). Overlapping dots mean reproduced; a long bar is a gap, labelled with its cause.



STEP 3 **Diagnose the six gaps**

- REPORT **Team 27.** "Player position and league were included" could mean main effects or interactions.
- REPORT **Team 9.** The text names referee country as a covariate; the full report uses player country.
- REPORT **Team 15.** The contrast coding is unstated; `options(contrasts = c("contr.sum", "contr.poly"))` recovers the reported value exactly.
- CODE **Teams 1, 23.** The schema is right; the code silently mis-reshapes.
- OPEN **Team 21.** Neither the schema nor the code accounts for the gap.

An automatic reproducibility check. Where it fails, a decision is under-stated; no LLM can read the author's mind. Once the author confirms the schema, anyone can test alternatives.