🌻 Comparing groups and testing the differences

22 Sep 2025

Summary

How can we use a **factors table** to summarise a causal map, compare groups (districts, gender, questionnaire sections, etc.), and decide which differences are worth attention?

This extension is essentially a *second-stage transformation*:

1. We start with a **links table** (one row per coded causal claim, with a `source_id` and any source metadata).
2. From that we derive a **factors table** (one row per factor label), by aggregating the links table:
3. counts of citations (how many link rows mention the factor as cause or effect)
4. counts of sources (how many distinct sources mention it)
5. optionally in/out splits (as cause vs as effect), and derived scores like “outcomeness” or influence/importance.
6. This extension then transforms the **factors table again** by adding **group breakdown columns**, and (optionally) a **statistical test** of whether observed group differences are larger than you would expect by chance.

The factors table is already an interpretation layer

The factors table is not “raw data”: it is a summary view derived from links.

- A **factor row** corresponds to a *label* (after whatever upstream label-rewrite transforms you have chosen).
- A **count column** corresponds to a rule for “how evidence accumulates” for that label.

That makes the factors table useful for:

- **Ranking / sorting** factors by frequency (what is most talked about?)
- **Distinguishing roles** (cause-like vs outcome-like) via in/out splits and ratios like outcomeness
- **Comparing contexts** by breaking counts out by source-level group variables

Semantics of the main counts (and why totals are tricky)

Two common evidence units are:

- **Citations**: each coded link row counts as 1 mention.
- **Sources**: each distinct source counts at most 1 (per factor, per group, etc.).

Some important consequences:

- **Citation totals across factors are "factor-mentions", not "number of links".**
  Each causal claim mentions (at least) two factors: a cause and an effect. So if you sum "Citation Count" down the factors table, you are roughly counting *mentions* of factors in links, not distinct causal claims. (A self-loop can be a special case, depending on how you count.)

- **Source totals across factors are not interpretable as "number of sources".**
  The same source contributes to many factors, so summing "Source Count" down the table double-counts sources heavily. It is meaningful only as "total source–factor incidences", not as a population size.

- **In/out splits change the story.**
  "Citation Count: Out" (as a cause) answers "what factors are used as explanations?", while "Citation Count: In" answers "what factors are treated as outcomes?". Sorting by one or the other is a substantive analytic choice.

Totals and normalisations (what they mean)

When you add totals or percentages, you are choosing a baseline:

- **Within-table totals** (e.g. per group column total) reflect overall verbosity or coverage of that group under the current filters and coding density.
- **Percent-of-baseline** views turn absolute counts into *relative prominence*: "what share of this group's factor-mentions does this factor account for?"

These are not cosmetic: they encode a stance on whether you care about absolute volume or relative emphasis.

The core idea: breakdown columns

Let G be a categorical group variable attached to sources (e.g. `district`, `gender`, `section`). For each factor f, we create extra columns that "break out" the factor's frequency by the levels of G.

Two natural "units of analysis" exist:

- **Sources mode**: for each factor and group level, count how many *distinct sources* in that group mention the factor at least once.
- **Citations mode**: for each factor and group level, count how many *link rows* in that group mention the factor (sensitive to sources that make many claims).

These extra columns let you ask questions like:

- "Which factors are disproportionately mentioned by women vs men?"
- "Which outcomes are more prominent in one district than another?"
- "Do different questionnaire sections elicit different causal themes?"

Optional normalisation: percent-of-baseline view

Raw counts can be misleading when groups differ in overall verbosity (e.g. one group has more sources, or makes more links per source).

A complementary view is to show each cell as a **percent of that group's total** across all factors (a "percent-of-baseline" normalisation). This re-expresses the question as:

> "Is this factor unusually prominent *given how much this group mentions factors overall*?"

Optional inference: significance testing

When you choose exactly one grouping variable G, you can attach a statistical test to each factor that asks whether the distribution across group levels departs from expectation.

Intuition (chi-squared style):

> Even if group A and group B differ in total mentions, is factor f *over-represented* in one group relative to that baseline?

For numeric-like groupings (e.g. ordered age bands), an **ordinal trend** interpretation can be more powerful than treating levels as unordered categories.

Why this matters

Group comparisons help you move from "what is mentioned most often?" to "what differs between contexts?"—which is often the analytic point of multi-source causal mapping: you are comparing perspectives, contexts, or sub-populations, not estimating causal effect sizes.