

AGENCY 150 | WHITE PAPER

From Conversations to Measures

How Codebooks Turn Interview Data into Strategic Insight

October 2026

Lamar Pierce & Natasha Swift

Agency 150, Inc.

Executive Summary

Agency 150 turns open-ended employee interviews into scores that executives can compare, audit, and act on. The device that makes this possible is the codebook: a versioned instrument that defines, for each construct, what counts as evidence, how that evidence is scored on an anchored 0-to-4 scale, and how scores roll up into dimension measures and profiles. Vega, Agency 150's AI-led interview platform, makes it economical to interview an entire workforce in depth. The codebook makes the resulting transcripts commensurable across studies, segments, and waves.

This paper describes how that translation works and why it holds up. It covers four topics:

- **The architecture.** Each codebook splits the work into three layers. Hard-coded signal terms recognize candidate evidence; a language model judges that evidence against behavioral anchors and must cite a verbatim quote for every score; deterministic code does all arithmetic. Humans author, calibrate, and audit the instrument.
- **A worked instrument.** The Workplace Sentiment Codebook scores Valence, Activation, and Dominance at the passage level and tags what each feeling is about. In a three-pass calibration on 25 interviews, Valence reached weighted κ of .95 to .97 and Activation 1.00; only 1 of 119 ratified passages showed a score spread of two points or more.
- **The interview designer.** A codebook is also a specification of the evidence an interview must produce. Vega's interview designer uses it to build protocols that elicit every scored construct. The stakes are large: with the same AI readiness codebook, a general future-of-work protocol yielded a full index for 1 of 102 participants, while a purpose-built protocol yielded 59 of 211.

- **An application.** In post-merger integration (PMI), codebook measures of the experienced organization are joined to deterministic measures of the administered organization: org charts, compensation, and integration decisions. The gap between the two layers is the diagnostic.

The common thread is traceability. Every number in a Vega readout links to the coded passages that produced it, and every coded passage links to a sentence an employee said.

1. The Problem: Decisions Need Knowledge That Sits Everywhere

A strategic decision rarely touches one activity. Following Porter, a firm's strategy is a system of complementary activities, so a choice such as outsourcing a delivery fleet ripples through logistics, talent, finance, and customer service at once. The knowledge needed to judge those ripples is dispersed across the people who run each activity. Hayek made the general point about economies; Chandler, Arrow, Williamson, and Garicano made it about firms. Much of that knowledge is tacit and local, and it is summarized, filtered, and reframed as it moves up a hierarchy.

Firms have historically reached this knowledge with two instruments, each shaped by cost.

- **Consulting interviews** go deep but narrow. A typical engagement might interview two dozen people, weighted toward senior leaders, over several months, and produce one deck.
- **Engagement surveys** go wide but shallow. They are closed-ended because open text was too expensive to analyze, annual because fielding was expensive, and standardized at the item level because that was the only affordable way to make results comparable.

AI-led interviewing removes the cost constraint. Vega can interview every member of a population, ask follow-up questions, and finish in days. That exposes a second

constraint. Unstructured qualitative output does not compound. Ten thousand transcripts cannot be compared across departments, tracked across waves, or audited by a skeptical executive unless something imposes structure on them. Summaries produced by a language model, however fluent, cannot be checked against their evidence, and they shift each time the model or the prompt changes.

For AI interviews to inform strategic decisions, the analysis must meet three requirements:

1. **Science-based.** Every measure should apply a construct with a validated research lineage, so that a score means the same thing in every study.
2. **Dynamic.** The system should support new segments, new waves, and firm-specific constructs without breaking comparability with earlier work.
3. **Verifiable.** Any figure in a readout should be traceable to the passages that produced it, and those passages should be checkable in seconds.

The codebook is the instrument that meets all three.

2. What a Codebook Is

An Agency 150 codebook is a measurement instrument for interview text. Methodologically, it sits in the tradition of directed content analysis (Hsieh & Shannon, 2005), where categories come from established theory and text is examined for evidence of them. Agency 150 extends that tradition by anchoring each scored code to a behavioral scale, which places the method in the quantitizing lineage: the disciplined conversion of qualitative data into numbers with every judgment step made explicit (Sandelowski, Voils, & Knafl, 2009).

From construct to measure

Each codebook is built as a short chain:

1. A **source construct** from peer-reviewed research, such as perceived usefulness (Davis) or psychological empowerment (Spreitzer).
2. An **indicator**, the code that applies that construct to interview evidence. Each indicator carries a definition, inclusion and exclusion criteria, evidence-scope notes, signal terms, behavioral anchors for scores 0 through 4, counter-examples, and a library of reference-scored excerpts.
3. A **dimension**, the mean of its indicators, valid only when enough indicators are scored, rescaled to 0 to 100.
4. A **composite index** or an interpretive **profile**, defined by the codebook's aggregation rules.

Two disciplines run through every instrument. The first is that codebooks are **protocol-agnostic**: scoring keys on what the respondent expresses, so interviews conducted with different question wording still produce commensurable scores. The second is **NA discipline**: when an interview never elicits evidence for an indicator, the indicator is coded NA, and NA is never zero. A low score requires expressed evidence of the low pole. Silence is reported as a coverage gap.

The platform instruments

Codebook	Dimensions	Structure	Headline output
Employee AI Readiness v2.5	Motivation, Organizational Support, Capabilities	15 core indicators plus a 6-indicator AI-Stance module reported separately	AI Readiness Index with bands
Organization al Culture	Involvement, Consistency,	12 indicators	Culture tension profile; no composite

Codebook	Dimensions	Structure	Headline output
	Adaptability, Mission (Denison model)		
Workplace Sentiment v1.0	Valence, Activation, Dominance	3 passage-level indicators plus target tags	Affective profile; no composite

Codebooks come in three tiers. **Platform** codebooks such as these apply validated frameworks across clients. **Client** codebooks preserve a framework or prior wave the organization already uses. **Study-level discovery** derives new codes inductively when the data should speak first; those codes enter production only after hardening into full instruments.

Where the codebook lives

A codebook exists in two synchronized forms. The authored document is what researchers and clients read. The **code pack** is its compiled, machine-readable form, defined by a JSON schema and validated before use. Because Vega's scoring tools are generated from the same schema, the instrument a person reads and the instrument the machine executes cannot drift apart.

Both forms sit inside the Vega ontology, a typed object model that links every part of a study. An *Initiative* is the strategic question. A *Study* runs a versioned *Protocol*, made of *Probes*, through *InterviewSessions* that produce *Responses*. A *Codebook* defines *Codes*; each *CodeApplication* links a code to the *Responses* that evidence it; and each *Measure* is a deterministic aggregate over applications. This chain is what makes a readout auditable: any *Measure* can be walked back to the sentences that produced it.

3. From Transcript to Score

Vega applies every codebook through the same three-stage pipeline. The governing rule, enforced by the code pack schema, is that **phrases never map to scores**. Words point to where evidence may be; scores come only from judgment against anchors; and numbers above the passage level come only from code.

Stage 1: recognition. Each code carries a dictionary of signal terms: literal phrases with pole hints, patterns, and routing hints to neighboring codes. Matching is hard-coded, fast, and fully auditable, and it guarantees that signature expressions such as “no training” or “done to us” are never missed. Flagged text can be rejected at Stage 2, and unflagged text can still be scored there.

Stage 2: judgment. Vega’s language model reads each flagged span in the context of the respondent’s full set of answers and scores it against the behavioral anchors. It returns strict JSON for each application: the score, a confidence rating, a verbatim supporting quote, and a one-sentence rationale. The model may score evidence the lexical layer missed. It may never score without a verbatim quote. The reference-scored excerpt library, grown from calibration and audit, is where phrase-to-number knowledge actually lives.

Stage 3: arithmetic. Application records feed deterministic code that computes dimension scores, indices, profiles, and coverage. No model computes a mean, and re-running the aggregation returns identical figures.

All three stages inherit the Agency 150 Qualitative Coding Standards, which govern segmentation, the evidence standard, and the bounded license for latent interpretation. The Standards apply identically to machine and human coders.

Code does the arithmetic; people own the instrument

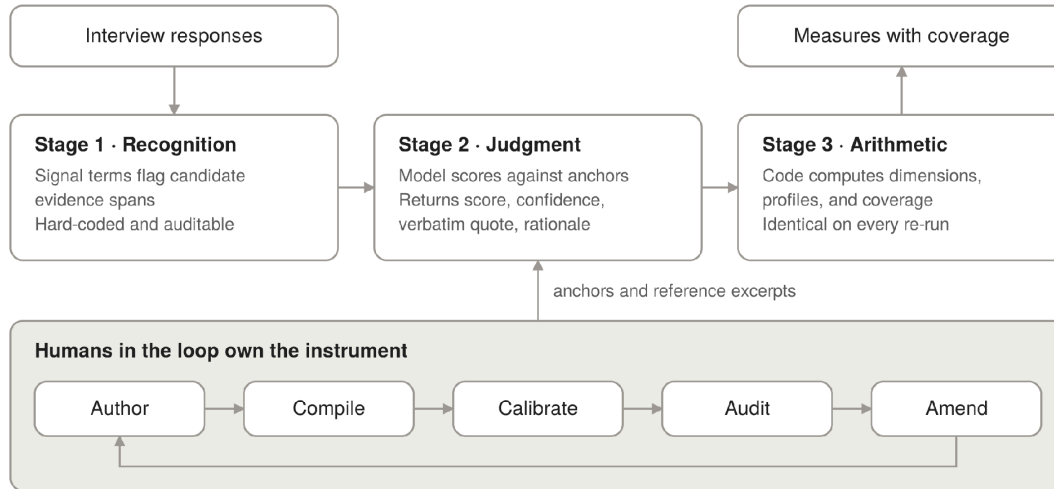


Figure 1. The Vega scoring pipeline and the human cycle that owns it. The model judges only within anchors that people wrote and calibrated, and every amendment produces a new codebook version.

Humans in the loop

The pipeline automates coding. People own the instrument. Each codebook moves through a cycle of authoring, compiling into a code pack, calibration, production coding, audit, and amendment.

- **Calibration.** Before production use on a new population or protocol, Vega and an independent coder (or two human coders) double-code 15 to 25 interviews. Agreement is computed per code; the default acceptance thresholds are quadratic-weighted κ of at least .70 and NA-agreement of at least .80. Codes that fall short are re-anchored. Agreed passages become reference-scored excerpts, so the calibration set becomes part of the instrument.
- **Audit.** In production, a human reviews a random sample of applications plus every low-confidence application. Because each score carries its quote, a

check takes seconds, and confirmed applications are flagged as human-verified in every downstream count.

- **Versioning.** Any amendment increments the codebook version and triggers recalibration. A wave always re-applies one fixed version, so change over time is measured with an unchanged instrument.

The AI readiness instrument illustrates the resulting reliability. In its audit, 25 double-coded interviews produced 525 paired judgments with pooled quadratic-weighted κ of .87 and 85% exact agreement.

Adequacy is measured per cell

Vega studies are segmented by organizational unit, role, and tenure band, and results are reported by segment. A study-level judgment of saturation is the wrong grain for that design: pooled adequacy can look fine while every subgroup is thin, and heterogeneity depresses the effective prevalence of subgroup-specific content. Agency 150 therefore reports two per-cell quantities with every aggregate.

Coverage (the NA rate per indicator and the number of scored applications) shows whether the protocol elicited enough evidence. The engagement's **confidentiality tier** (a minimum cell size of 1, 3, 5, 8, or 11, disclosed to respondents before the interview) sets the floor below which no result or quote is shown.

4. A Worked Example: The Workplace Sentiment Codebook

The sentiment codebook shows the architecture end to end. It scores how employees feel on three dimensions, records what each feeling is about, and reports profiles instead of a single sentiment number.

Why three dimensions

A single positive-to-negative score hides the distinctions clients act on. An energized, angry workforce and a quiet, resigned one are both negative. The first is

mobilized; the second has given up. They call for opposite interventions. The codebook therefore scores three dimensions from the dimensional tradition in affect research:

- **Valence (VAL)**, from strongly negative to strongly positive. Source: Russell's (1980) circumplex and Warr's (1990) job-related affective well-being.
- **Activation (ACT)**, from depleted to highly energized, independent of direction. Source: circumplex arousal, and the vigor and exhaustion poles of work engagement (Schaufeli et al., 2002).
- **Dominance (DOM)**, from helpless to in command of one's situation, independent of valence. Source: Mehrabian and Russell's (1974) PAD model and psychological empowerment (Spreitzer, 1995).

Affective Events Theory (Weiss & Cropanzano, 1996) supplies the unit of analysis. Sentiment arises in reaction to events, so it is scored where it is expressed. Instead of one judgment per indicator per participant, the codebook scores each **evaluative passage**: a span expressing one feeling about one object. Each application also carries a **target tag** (WORK, ORG, CHANGE, or OTHER), because attitudes toward one's own job and toward one's employer are empirically distinct (Judge & Kammeyer-Mueller, 2012). A respondent can love the work and distrust the company.

The scale and the rules

Score	Valence	Activation	Dominance
0	Strongly negative	Depleted, numb, exhausted	Helpless; no control
1	Negative	Low energy, flat	Little control; mostly done-to

Score	Valence	Activation	Dominance
2	Neutral or mixed	Ordinary energy	Partial control
3	Positive	Energized	Largely in control
4	Strongly positive	Highly activated, intense	Strong agency; shaping events

Valence is scored in every application. Activation is scored only when the passage carries energy cues, and Dominance only when it speaks to control, voice, or their absence. A mild opinion evidences valence and nothing else, and the coder leaves ACT unapplied there. Coders score the expression, tag the object of the feeling instead of the topic of the question, and split passages that evaluate two objects.

Three passages, scored

The codebook's worked examples show what the extra dimensions buy.

Passage (synthetic)	Target	VAL	ACT	DOM
R1: "I'm past caring. Decisions get made somewhere above us and we just absorb them."	ORG	1	0	0
R2: "This reorg is a disaster and I'm not going to sit quietly. I've booked time with the VP, and half the team is behind me."	CHANG E	1	4	4
R3: "It's a pleasant place. Things get decided for us and they usually turn out fine, so I just go along with it."	ORG	3	1	1

The first two respondents share a valence score. A one-dimensional measure would call them identical, yet one is resigned and the other is organizing. The third is positive but passive, which a valence-only measure would misread as readiness for a change program that needs discretionary effort.

Aggregation and the affective profile

Code computes everything above the passage. A participant's dimension score is the mean of that dimension's applications, rescaled to 0 to 100. Valence requires at least three applications; Activation and Dominance require at least two and always report their n. Per-target scores apply the same rule within each tag, which is where a split such as WORK 100 and ORG 12.5 on valence becomes visible. Participants with valid Valence and Activation are placed in one of four quadrants at the 50-point midline, and Dominance qualifies the placement. The codebook deliberately computes no composite sentiment index, since collapsing the three dimensions would discard the information the instrument exists to preserve.

Same valence, different states: activation and dominance separate them

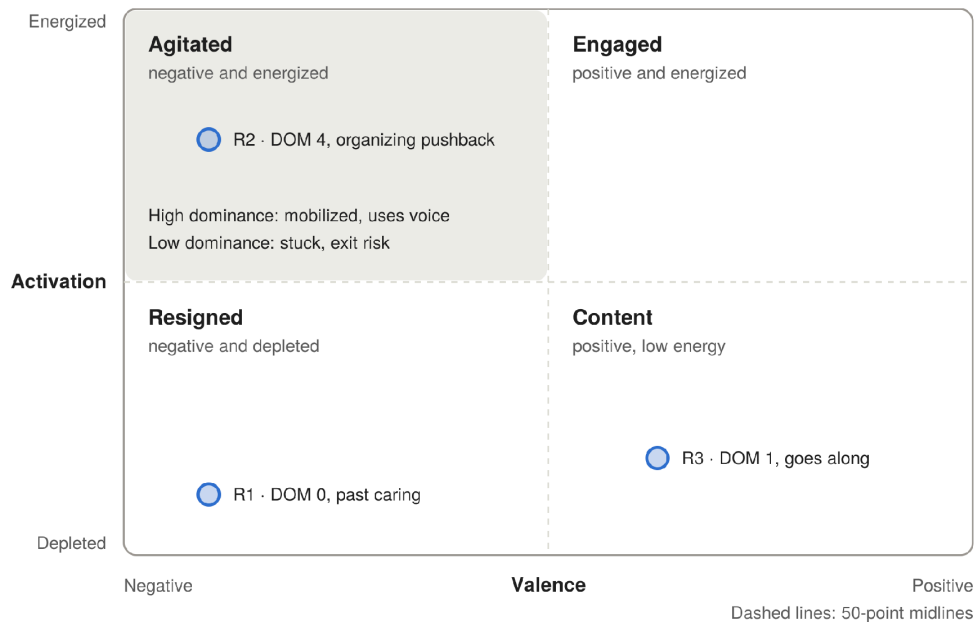


Figure 2. The affective profile: four quadrants, with Dominance as a qualifier. The three example respondents land in different quadrants even though R1 and R2 share a valence score; dots show quadrant placement only.

Calibration results

Three independent coder passes, blind to one another, segmented and scored the same 25 interviews.

Agreement measure	Range across pairs	Threshold
Unitization (same passages found)	.91 to .93	.80
Target tag, matched passages	.95 to .97	.80
Valence, weighted κ	.95 to .97	.70
Activation, weighted κ where both applied	1.00	.70

Agreement measure	Range across pairs	Threshold
Dominance, weighted κ where both applied	.79 to .92	.70

Of 119 majority-ratified passages, one showed a spread of two or more points: a two-object passage that one pass split into separate CHANGE and WORK applications. It now sits in the excerpt library as a unitization exemplar.

The substantive results carried a design lesson. Valence was reportable for 24 of 25 participants, with a median of 56. The affective profile, however, could classify only 7 of 25, because Activation was evidenced in roughly one application in five and rarely reached its floor. The instrument worked; the interview did not ask for feelings directly. Section 6 returns to this finding, because it is exactly the kind of gap the interview designer exists to close.

5. From Scores to Strategic Insight

Scores become insight when they are compared, segmented, joined to other data, and read against the decision at hand. Because every Vega study reduces interviews to the same dimensions on the same 0-to-100 scale, results support four kinds of reading that transcripts alone cannot.

Segment contrasts. Measures are computed per organizational unit, role, tenure band, or any other attribute the ontology records, subject to the confidentiality tier. Two units with identical valence may differ completely in activation and dominance. A unit that is agitated with high dominance is mobilized and likely to use voice. A unit that is agitated with low dominance is frustrated and stuck, and it carries more exit risk. The remedies differ: the first calls for engagement with the people pushing back, the second for restoring a credible channel of influence.

Object contrasts. Target tags separate feelings about the work, the organization, and the change under study. A workforce that is positive about its work and negative about its employer has a leadership and trust problem. Investment in job redesign would do little for it.

Cross-instrument reading. Instruments pair naturally. The readiness codebook says what employees intend and are able to do; the sentiment codebook says how the situation feels while they do it. High readiness motivation combined with low dominance about the change, for instance, suggests a workforce that wants to adopt but believes the rollout is being done to it.

Change over time. A wave re-applies a fixed codebook version to the same population, so movement between waves reflects the organization. Instrument drift is ruled out by construction.

Rolling up without losing honesty

Sub-population studies roll up into a total-population profile, weighted by how well each source study evidenced each dimension. Coverage, application counts, and validity status travel with every aggregate. A department whose interviews rarely touched organizational support contributes little to the organization-wide support score, and the readout says so.

Every claim has a quote

A finding in a Vega readout is a query over stored objects. An executive who doubts that a unit feels helpless about a reorganization can open the Measure, see the applications behind it, and read the verbatim passages, each with its score, confidence, and rationale. Quotes obey the confidentiality tier: a quote plus its demographic labels may never resolve to a cell below the minimum size, so labels are dropped one at a time until it complies. Confidentiality is enforced by the

rendering system and disclosed to respondents before the interview, which makes it a structural property of every deliverable.

The ontology learns

Each client engagement runs on a firm-specific extension of a shared root ontology. The extension holds the firm’s units, roles, and any constructs particular to it. The root holds the platform codebooks and the cross-client structure. When an engagement produces something general, such as a refined evidence-scope note, a calibrated new indicator, or a probe that reliably elicits a hard construct, it becomes a candidate for promotion to the root. Promotion requires science sign-off, and only a consented, de-identified mapping ever leaves a firm’s extension. Each engagement improves the next one without moving client data.

6. How Codebooks Shape Vega’s Interview Designer

A codebook scores evidence, and it can only score evidence the interview produced. Read from the other direction, the codebook is a specification of what an interview must elicit. Vega’s interview designer uses it that way: it starts from the constructs a study needs to measure and works backward to the questions and probes that will produce scoreable evidence for each one.

The cost of ignoring the codebook

The clearest evidence comes from applying one instrument to two protocols. The AI readiness codebook was run, unchanged, over interviews from a general future-of-work protocol and over interviews from a protocol built for AI readiness.

	General protocol	Purpose-built protocol
Participants coded	102	211

	General protocol	Purpose-built protocol
Full readiness index computed	1	59
Participants with a valid Organizational Support score	4	64

The gap does not come from the population or the model. The general protocol never asked about tools, training, access, or self-rated skill, so respondents rarely volunteered that evidence, and the support and capability indicators came back NA. Most of the shortfall could have been closed with three to five targeted probes. In the readiness codebook's words, protocol design is part of the measurement system.

What the designer takes from the codebook

Codebook element	How the interview designer uses it
Source constructs and indicators	Sets the measurement targets. Each Probe in the ontology records the construct it is meant to elicit, so a protocol's coverage of the codebook can be checked line by line before fielding.
Validity floors	Sets how much evidence each dimension needs per respondent. Sentiment Valence needs three applications and Activation two; readiness dimensions need three of five indicators. Probes are budgeted to reach those floors.
Anchors and evidence-scope notes	Define what evidence looks like at each pole. Probes are written to invite descriptions of behavior and events that can be scored, such as asking what happened the last time a respondent raised a concern.

Codebook element	How the interview designer uses it
Signal terms and reference excerpts	Show the register in which evidence usually arrives. They help the designer write probes that open the right territory without supplying the words the coder will later look for.
Target tags and referents	Ensure the protocol asks about each object of feeling separately: the work, the organization, the change, and in a merger each legacy firm.
Coverage history	Shows which probes have actually produced evidence in prior studies, so weak probes are revised and strong ones are reused.

Because scoring keys on expressed content, the designer is free to adapt wording to a population, an industry, or a language without breaking comparability. A probe for hospital nurses and a probe for software engineers can be phrased differently and still feed the same indicator. Question wording is treated as context for interpretation and never as evidence, which also protects against leading questions: a probe that plants the answer produces a passage the coder cannot credit to the respondent.

Closing the sentiment gap

The sentiment pilot shows the loop working. Coverage analytics showed that Activation was evidenced in about one application in five and Dominance in about one in seven, leaving 18 of 25 participants without an affective profile. Three responses were possible: add a probe, accept sparse profiles, or lower the Activation floor. The recommendation was to add a single probe that elicits affect directly, such as asking how the change leaves the respondent feeling at the end of a week, and to leave the floor where it is. Lowering a validity floor silently would make the readout look complete while making it less trustworthy. Adding the probe makes the evidence exist.

Coverage review before fielding

In every engagement, a protocol review stage checks elicitation coverage across all of a codebook's dimensions before interviews begin. For the culture instrument, that review typically adds Mission and Adaptability probes, which general protocols under-elicited. For longitudinal work, coverage requirements change by wave, and the designer builds each wave's protocol against them. Because Probe is a stored object linked to the Responses it produced, the coverage question can be asked of past studies with a query, and the answer improves the next protocol version.

7. Application: Post-Merger Integration

A merger produces two organizations that can be measured. The **administered organization** is what the org chart, the compensation file, and the decision-rights matrix say exists. The **experienced organization** is what employees describe when asked, and it is what the codebooks measure. In steady state, the two mostly align. Integration is the period when they diverge, and the gap between them determines whether integration succeeds or fails.

The design keeps the two layers separate. Coders never score structure, and structural variables never enter the codebooks. We compute structure deterministically from administrative data and join it to the psychological measures at the participant and unit level. The analysis lives in the joins.

The finding sits in the gap between the two organizations

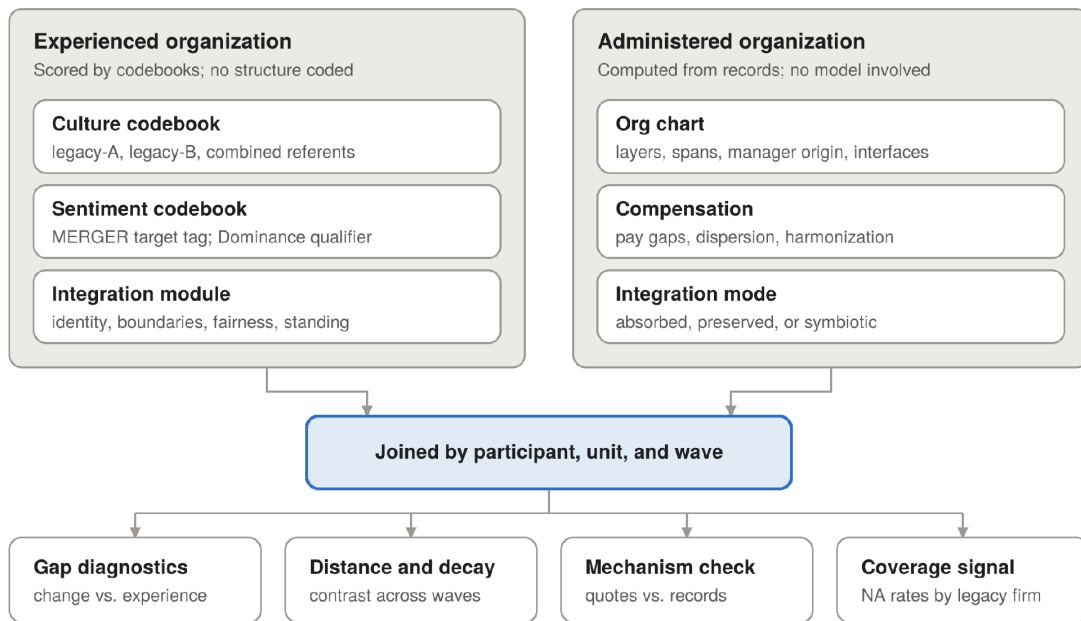


Figure 3. PMI measurement architecture: two layers, one join, four analyses. Each layer is measured on its own terms; the four analyses read the two layers against each other.

Two examples show why. First, suppose empowerment declines in an acquired unit. If that unit's decision rights were formally centralized at close, the decline is the expected result of a deliberate choice. If its autonomy was formally preserved, the same decline signals a failure in how integration was carried out. The psychological readings are identical, and only the join distinguishes the diagnoses. Second, compensation inequity across legacy firms tends to surface in interviews as cynicism, negative sentiment toward the organization, and fairness complaints well before it reaches turnover files. Because every score carries a quote, a passage about counterparts keeping a richer bonus plan can be checked against the compensation data. A grievance that matches a real plan difference and one that does not are both findings, with different remedies.

Layer 1: the codebooks in merger mode

The existing instruments apply with small, engagement-scoped extensions.

Culture, with legacy referents. The culture codebook already records the referent of each application. For mergers, the referent vocabulary adds *legacy-A*, *legacy-B*, and *combined*. Merger respondents use these frames unprompted, and one passage often evidences the same indicator at two referents in opposite directions: decisions that used to be made locally and now go to the acquirer's committee is high empowerment remembered and low empowerment now. Scoring both yields each legacy culture as remembered, the combined culture as experienced, and the contrast within each respondent. That within-person contrast is a better measure of cultural distance than the difference between two separately sampled survey populations. Remembered culture is scored as such and never blended into current-state scores; how favorably each side remembers its past is itself a measure.

Sentiment, with a MERGER target. The sentiment codebook gains a fourth substantive target tag, MERGER, for feelings about the integration process, the deal, and the counterpart firm. Feelings about the combined company as an employer stay ORG, and feelings about one's changed job stay WORK. This separates integration-process sentiment from general morale, which blur together in merger settings and need different responses. Dominance does especially useful work here. An agitated MERGER profile with high dominance is organized opposition, which can be engaged. The same profile with low dominance is resigned helplessness, which predicts exit.

A candidate integration module. Five merger-specific indicators are specified at candidate level, to be reported alongside the culture dimensions and never folded into them, because they describe states the integration induces.

Code	Indicator	What it captures	Source construct
PMI1	Identity continuity	Whether the combined firm feels like a continuation of one's organization or its loss	Post-merger identification (van Knippenberg et al., 2002)
PMI2	Intergroup boundary salience	Whether daily work runs on "us versus them" by legacy firm	Social identity in mergers
PMI3	Integration procedural fairness	Whether integration decisions were made and communicated fairly	Procedural justice (Colquitt, 2001)
PMI4	Acculturation preference	Congruence between the mode of combination wanted and the mode perceived	Nahavandi & Malekzadeh (1988)
PMI5	Standing change	Gain or loss of status, voice, and prospects	Elsass & Veiga (1994); Nickerson & Zenger (2008)

Boundary salience is the module's leading indicator. Salience that fails to decay across waves is the classic sign of sociocultural integration failure.

Layer 2: the structural data

Structural measures are computed from HRIS extracts, org chart exports, compensation files, and integration program documents. No language model touches them. They fall into three families:

- **Org chart:** layers and spans per unit; the share of managers from each legacy firm at each level; whether each respondent's manager changed and to

whom; and, for each pair of units that must now coordinate, whether they were merged, linked, or left separate.

- **Compensation:** pay differences across legacy lines for comparable roles, pay dispersion within newly merged peer groups, differences in bonus design, and the harmonization schedule. A merger creates new and visible comparison sets, and the evidence suggests that the perceived legitimacy of a pay differential, as much as its size, drives the behavioral response (Shaw, 2014).
- **Integration mode:** whether each acquired unit was absorbed, preserved, or combined symbiotically (Haspeslagh & Jemison, 1991), validated against the chart. A unit labeled as preserved whose reporting lines were rewired has in practice been absorbed, and the mislabel is itself a finding.

These measures attach to existing ontology types. Participants gain legacy affiliation and manager-change fields from the HRIS; organizational units gain legacy composition, integration mode, and structural changes per wave; and a new interface object records how each required coordination link was treated. Legacy affiliation always comes from administrative records and is never inferred from transcripts.

Waves and the interview designer

PMI measurement runs in three waves: a baseline as close to close as possible (separately in each legacy firm where deal confidentiality permits), an early-integration wave at roughly 100 days, and a consolidation wave at 12 to 18 months. Waves are aligned to the integration milestones in the structural layer, so change between waves can be read against what was done between them.

The interview designer builds each wave's protocol against the codebooks' coverage needs, which shift as integration proceeds. The baseline needs mission and identity content. The 100-day wave needs coordination, agreement, and fairness content,

because that is where the action moves. Every wave needs probes that invite respondents to speak about each legacy firm and the combined firm separately, so the referent extension has evidence to score.

What the analysis delivers

- **Gap diagnostics.** For each dimension, compare what the structural layer says changed with what employees experience. Experienced decline without structural change points to how integration was carried out. Structural change without experienced decline marks a change that was absorbed well and is worth understanding and repeating.
- **Distance and its decay.** Every merger shows cultural distance at the first wave. The question is whether the within-respondent contrast and boundary salience decline, and whether the decline differs by integration mode, interface treatment, and pay harmonization. Persistent boundaries in preserved units are expected; in absorbed units they signal failure.
- **Mechanism corroboration.** Fairness and standing-change passages that name concrete differentials are checked against compensation and chart data. Corroborated and uncorroborated grievances call for different interventions, and only this architecture can tell them apart.
- **Coverage as signal.** If acquired-firm respondents produce no mission content at the second wave under the same protocol that elicits it from acquirer staff, the combined firm's direction may not have reached them.

Confidentiality in mergers

Legacy affiliation is often the most identifying attribute in an acquisition. A forty-person target inside a five-thousand-person acquirer makes every target-side cell small. Three rules follow. The confidentiality tier is chosen with the smaller population in mind, and the participant notice names legacy affiliation as a reporting

dimension. Quotes labeled by legacy firm and role are subject to the same cell-size rule as any other breakdown. And because legacy-A and legacy-B partition the population, suppression is applied to the partition, since suppressing one side while publishing the other and the total would reveal the suppressed cell.

Sequencing

The extensions are ordered by cost and readiness: first the legacy referents in the culture code pack and the MERGER tag in the sentiment code pack, both small changes; then the structural join on the ontology; and last the integration module, calibrated on interview data from a first live engagement. Merger interviews are emotionally charged and full of irony, so the module should not support production reporting until a calibration study with human ratification is complete.

8. Design principles and limits

Principles

The architecture rests on a short list of commitments, each enforced in the schema, the Coding Standards, or the rendering system.

1. **Phrases never map to scores.** Signal terms locate evidence. Judgment against anchors assigns scores.
2. **Evidence is expressed content.** Interpretation uses the transcript and its interview context, with latent reading allowed only within bounded rules.
3. **NA is never zero.** Absence of evidence is reported as a coverage gap and never averaged in as a low score.
4. **Every score has a quote.** No application exists without the verbatim passage that supports it.
5. **Arithmetic belongs to code.** Language models extract and judge; deterministic code aggregates.

6. **One version per wave.** Change over time is measured with an unchanged instrument.
7. **Coverage travels with every number.** Aggregates report how much evidence stands behind them.
8. **Trust is architectural.** Confidentiality tiers are disclosed in advance and enforced at render time, and identity is kept outside the analytical layer.

Limits

The system has known limits, and Agency 150 reports them alongside results.

- **Stability is not validity.** Agreement among independent passes of the same model shows that the instrument is applied consistently. Passes that share a model's biases can still agree and be systematically wrong. Before scaled client use of a new instrument, a human-scored subsample or a panel of different models should anchor validity.
- **Natural samples are mostly easy.** In the sentiment pilot, a largely engaged technology workforce produced few depleted or furious passages, so some pole exemplars remain synthetic. Future calibration should oversample hard cases such as irony, mixed feelings, and reported speech, and should draw on more varied populations.
- **Respondents may perform for the interviewer.** Social desirability and platform effects can shift levels. Where a wave changes modality or interviewer framing, within-wave contrasts are more trustworthy than level shifts.
- **Borrowed text has a different register.** External corpora such as online employer reviews are written, retrospective, and addressed to an audience, while Vega interviews are conversational and probe-elicited. Such corpora can help with broad sentiment vocabulary and pole exemplars. They fit poorly for constructs that require probing, such as access to AI training.

Each limit has a remedy inside the same loop of calibration, audit, and amendment, which is why the instrument is versioned and why its excerpt library grows with every engagement.

9. Conclusion

AI interviewing makes it affordable to ask an entire organization what it knows. Codebooks make the answers usable. They turn conversation into evidence with a clear construct lineage, anchored scores, recorded quotes, and honest coverage, and they do it the same way in every study, so results compare across units, accumulate across waves, and join to the structural data executives already hold.

The same instrument does double duty. Read forward, it scores what employees said. Read backward, it tells Vega's interview designer what employees need to be asked. The sentiment codebook shows both directions: a three-dimensional measure that separates the resigned from the mobilized, and a coverage finding that turned directly into a better probe. Post-merger integration shows where the approach leads: two organizations measured side by side, the administered and the experienced, with the gap between them as the most useful reading a leadership team can get during an integration.

References

- Colquitt, J. A. (2001). On the dimensionality of organizational justice. *Journal of Applied Psychology*, 86(3), 386–400.
- Elsass, P. M., & Veiga, J. F. (1994). Acculturation in acquired organizations. *Human Relations*, 47(4), 431–453.
- Haspeslagh, P. C., & Jemison, D. B. (1991). *Managing acquisitions*. Free Press.

- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.
- Judge, T. A., & Kammeyer-Mueller, J. D. (2012). Job attitudes. *Annual Review of Psychology*, 63, 341–367.
- Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. MIT Press.
- Nahavandi, A., & Malekzadeh, A. R. (1988). Acculturation in mergers and acquisitions. *Academy of Management Review*, 13(1), 79–90.
- Nickerson, J. A., & Zenger, T. R. (2008). Envy, comparison costs, and the economic theory of the firm. *Strategic Management Journal*, 29(13), 1429–1449.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.
- Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research*, 3(3), 208–222.
- Schaufeli, W. B., Salanova, M., González-Romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout. *Journal of Happiness Studies*, 3(1), 71–92.
- Shaw, J. D. (2014). Pay dispersion. *Annual Review of Organizational Psychology and Organizational Behavior*, 1, 521–544.
- Spreitzer, G. M. (1995). Psychological empowerment in the workplace. *Academy of Management Journal*, 38(5), 1442–1465.
- van Knippenberg, D., van Knippenberg, B., Monden, L., & de Lima, F. (2002). Organizational identification after a merger. *British Journal of Social Psychology*, 41(2), 233–252.

- Warr, P. (1990). The measurement of well-being and other aspects of mental health. *Journal of Occupational Psychology*, 63(3), 193–210.
- Weiss, H. M., & Cropanzano, R. (1996). Affective events theory. *Research in Organizational Behavior*, 18, 1–74.