

Measuring Seller Style

A Chambr Labs whitepaper on reading selling style from sales conversations

October 2026 Frank Alonso, Founding Engineer · Chambr Inc.

Abstract

Sales teams have described selling styles for decades, through popular sales methodologies and four-quadrant personality models. Most of them rely on questionnaires: how sellers rate themselves, or how customers remember them. Few observe what a seller actually does on a call.

This paper introduces the **Chambr Trade Map**, a method for measuring selling style directly from the conversation. It reads six behavioral choices from each transcript, such as whether the seller tells or asks and whether they lead or follow, and places the call among eight named selling styles. The Trade Map powers **Seller DNA** in Chambr's sales-practice roleplays. It was built from a research sample of 10,648 sales roleplays and tested on 901 scored roleplay calls from 411 sellers practicing against AI-simulated buyers.

Its main properties:

- **Grounded in the transcript.** Every reading is checked against the seller's own words, and verified quotes are shown alongside it. Three dimensions also come with countable figures, such as questions per 100 words.
- **Separated from the scenario.** In our research set, the buyer persona strongly predicted where an uncorrected call landed. After our scenario correction, it no longer did.
- **Outcome-neutral by design.** On the outcome data available to us, no selling style outperforms another once closing behavior is taken into account. The map is meant for coaching conversations, not for ranking.
- **Open about confidence.** Every result shows its certainty as an arc on the ring and a confidence label, so a clear read and a tentative one never look the same.
- **About calls, not people.** Across eight calls, the median seller shows up in five of the eight selling styles. Seller DNA describes each call rather than assigning anyone a type.

This paper sets out the method, the evidence for each claim, and where that evidence stops. Technical detail is in the appendix.

1. Why style frameworks drift into scoreboards

Most style frameworks start out as descriptions and end up as rankings. Someone notices that one type closes more deals, and before long the supposedly neutral map is being used to filter candidates. Questionnaire-based frameworks have a second limit: they capture what sellers believe about their selling, not what they do. Adaptive selling (Spiro & Weitz, 1990) and customer orientation (Saxe & Weitz, 1982), for example, are both measured by self-report. A meta-analysis covering more than 31,000 salespeople found that customer orientation predicts how well sellers rate themselves, but not how their managers rate them or what they actually sell (Franke & Park, 2006). The Trade Map shares no items with these scales.

Three design choices keep the Trade Map from going the same way:

- **Behavior, not self-report.** Every dimension is scored from the transcript and tied to the seller's own words.
- **Style is kept separate from skill.** Any candidate dimension that tracks "this was a good call" is moved to a separate skill layer.
- **Calls, not people.** The product never assigns a seller a type.

We tested every proposed structure against controls that could have disproved it, and dropped whatever did not hold. That is why Chambr does not call the map a circumplex, does not score how widely a seller varies, and does not publish success rates by selling style.

Intended use. Seller DNA is a coaching aid. It is not designed or validated for hiring, promotion, compensation or performance evaluation, and should not be used for those decisions.

2. What a seller sees

After each eligible roleplay, the seller receives a Seller DNA card in five parts:

1. **The nearest selling style**, always with its behavioral subtitle, for example *Qualifier · question-first, buyer-steered*. A confidence label sits above it: "One direction", "Nearest position" or "Nearest position · slight lean".
2. **The ring**, with the position drawn as an arc whose width shows how certain the read is.
3. **Six readings**, one per dimension, each linked to the moment in the transcript behind it.
4. **Context** on the selling style, such as how common it is.
5. **A fixed reminder**: "No position on this map wins more than another. The eight are different choices, not eight rungs."

Every card has the same structure at every confidence level, so a tentative read looks as finished as a clear one. Cards that are shared publicly leave out the transcript quotes.

3. The evidence base

The Trade Map was fitted on **901 scored roleplay calls from 411 sellers**. They are a stratified sample from 5,810 usable roleplays, drawn from a research sample of 10,648 roleplays from Chambr's current and previous platforms.

Stage	Calls	What happens
Research sample	10,648	Roleplays from Chambr's current and previous platforms
With a transcript	10,486	Calls without a stored transcript are dropped
Usable: complete, real, English calls	5,810	Ten quality rules, applied in order. 1,105 sellers, 287 scenarios
Scored for style	901	411 sellers. Stratified by platform, call type and call length

The largest exclusion is calls too short to read, 3,428 in all. The other rules remove internal and test traffic, failed and non-English calls, and deleted records.

Twenty-four sellers have eight scored calls each. We use this group to ask whether a seller has a stable style (Section 8). Most of them sell freight or logistics, so findings from this group should be read with that scope in mind.

Privacy. All analysis used read-only access. Raw transcripts and personal data were never exported from production, and any excerpts are redacted.

4. What we measure: six dimensions

Each dimension runs from -3 to +3. The sign shows a direction and says nothing about quality. Zero is the typical Chambr call, and the anchors are calibrated so that the median roleplay in the research sample scores close to zero on all six.

Dimension	-3 pole	+3 pole	What moves the score	Evidenced by
inquiry	Telling: moves the call forward by explaining and presenting	Asking: questions drive the call	Number and openness of real questions, and whether they open or close a topic	Numbers + quotes
warmth	Task: efficient and impersonal	Person: acknowledges, validates and personalizes	Acknowledgment that responds to what was said, reuse of the buyer's own words	Quotes
assertion	Following: offers options without a view, hedges, defers	Leading: recommends, disagrees, names hard commercial facts	Unprompted recommendations, disagreement that holds under pushback	Quotes
solution_pull	Problem-space: stays in the buyer's world	Solution-space: talks about the offering	Which subject the content covers, and when the offering appears	Numbers + quotes
momentum	Exploring: opens things up and leaves next steps open	Closing: narrows options and secures a specific next action	Next-step requests, commitment tests, persistence after deflection	Numbers + quotes
tempo	Broad: one pass per topic	Deep: stays on a thread through several levels	How often the seller follows up on the buyer's last answer	Quotes

Every level of every scale has written anchor text. For **inquiry**, -3 means zero or one genuine question in the whole call, 0 means about six mixed questions, and +3 means twenty or more, mostly open. The rubric is frozen with every release.

Numbers where they exist, quotes everywhere. Three dimensions have meaningful countable support. The strongest is questions per 100 words, which correlates $r = +0.379$ with **inquiry** and -0.387 with **solution_pull**. A sentence like "you asked 14 questions and followed up on 40% of them" is therefore a fair explanation. **warmth**, **assertion** and **tempo** have little countable support, so we explain them with quotes and never with counts that would suggest more precision than we have.

What we left out, and why. Each exclusion was tested and comes with a reason:

- **Reframing.** It is too rare to place anyone, and the model put 62% of calls at the floor.
- **Explicit structure.** It correlated +0.79 with a call-quality score, so it measures skill rather than style.
- **Adaptation.** It was the least repeatable candidate we tested, and it varied far more within a seller than between sellers. It behaves like a state that changes from call to call, so there is no Adapter position.
- **Listening.** The sales literature measures listening by questionnaire (Ramsey & Sohi, 1997) and has no validated transcript measure, and the turn-taking data that would support one is missing. Out of 172,445 utterance pairs, only 14 overlap.
- **Speaking rate, filler rate and prosody.** These reflect the recording pipeline more than the seller. We measure airtime in words, never in seconds.

5. How we measure

A transcript passes through a fixed pipeline. The output is a direction on the map, with a confidence arc and a confidence level. It is not a score.

The style claim is the angle, read after scenario and halo are removed

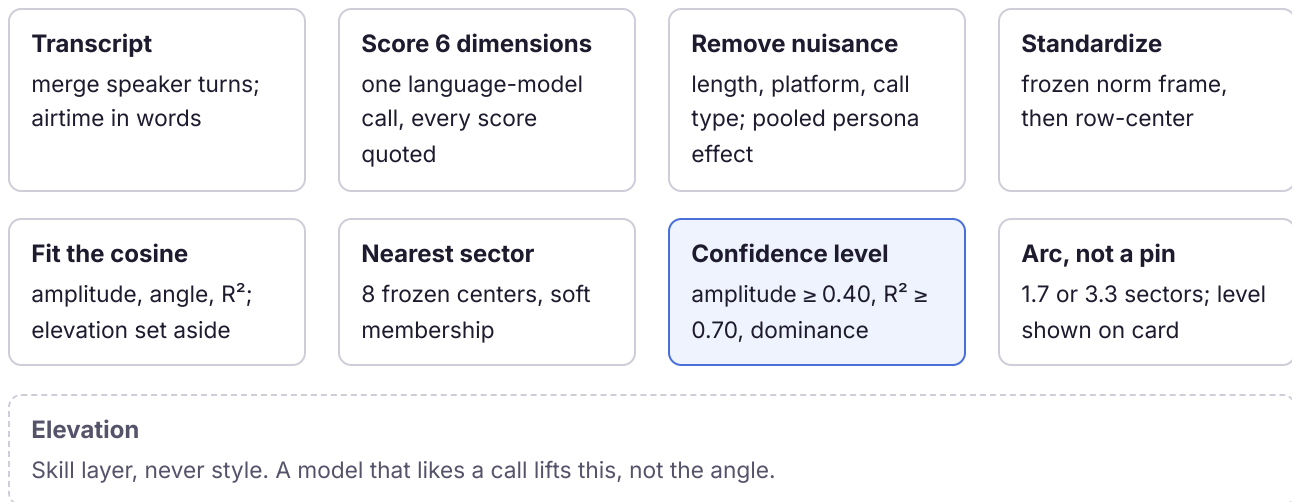


Figure 1. The scoring pipeline, from transcript to confidence level

5.1 Scoring the call

A language model scores all six dimensions in a single pass. Language-model raters are known to anchor when scoring several attributes at once (Stureborg et al., 2024), so the order of the dimensions is shuffled for each call and no dimension is consistently scored first. The model never sees the call's score or outcome, any position name, or any framing about quality.

Each score must cite a verbatim quote from the seller. Quotes are checked against the transcript using a six-word exact match. If a quote fails the check, it is flagged and hidden, and the score is kept.

5.2 Agreement across models

We re-scored the same calls with the same model a second time, and with two other language models. Language-model judges can approach human levels of agreement but carry known biases (Zheng et al., 2023), so we measure agreement rather than assume it. One model scoring a call twice agrees with itself at a median correlation of 0.77, which is the practical ceiling. Different models agree at medians of 0.61 to 0.63, close to that ceiling. `inquiry`, `momentum` and `warmth` agree best. `tempo` is the least stable of the six and is being refined, and `solution_pull` is being monitored. Appendix A gives the figures for each dimension.

Different models also score at slightly different overall levels. For that reason the scoring model is frozen with each release and recorded on every score.

5.3 Separating the seller from the scenario

Raw scores reflect the scenario, call length and call type as well as the seller. If these effects are left in, they dominate any map, so we remove them statistically. Each buyer persona gets its own correction, weighted by how many calls we have seen with that persona (Appendix A).

In the research set the correction worked. Before correction, the buyer persona strongly predicted where a call landed (Cramér's V 0.40 to 0.47). After correction there was no significant association ($p = 0.10$ to 0.96).

In the product, the persona corrections are frozen at release. Scenarios created later, including new custom personas, receive the correction for call length and call type but no persona-specific correction yet. The correction makes a real difference: raw and corrected scores place the same call in the same sector only about half the time (49.1%).

5.4 From six scores to one direction

The corrected scores are standardized against a fixed reference set and reduced to their shape. We then fit a single wave around the ring using the Structural Summary Method (Gurtman, 2009; Zimmermann & Wright, 2017). The fit produces four numbers:

- **Elevation:** how much of everything the seller did.
- **Amplitude:** how strongly the call leans in one direction.
- **Angle:** which way it leans. This is the style.
- **Fit:** how well a single direction describes the call.

This keeps quality out of style. A model that simply likes a call raises every score, the halo effect named by Thorndike (1920). That changes elevation and leaves the angle unchanged. In our data, elevation correlates +0.69 with an independent quality rating, while a realistic quality halo shifts the angle by a median of under 1°. Style is read from the angle alone, and elevation belongs to the skill layer. If a call's profile is flat, no angle is reported.

6. The shape of style

Selling style occupies a **two-dimensional plane**. The six dimensions keep a stable order around it, and calls are spread unevenly across it, with three denser regions. We draw the plane as a ring.

The order holds on new sellers. We derived the order on half of the sellers and tested it on the other half, splitting by seller so that nobody appeared in both halves. On the held-out sellers the order held (Tracey's correspondence index 0.708, $p = 0.017$; Tracey, 1997):

assertion 0° → solution_pull 60° → warmth 120° → inquiry 180° → tempo 240° → momentum 300°

Opposite dimensions are genuine trade-offs: leading against asking, closing against warmth, and product-forward against deep in the buyer's problem.

A plane, not a circle. The interpersonal tradition (Wiggins, 1979) often treats rings like this as circumplexes, in which correlation falls off smoothly with angle. We tested for that against synthetic data with known structure. Our data looks like a two-factor plane, not a true circle (Appendix A), so we describe the ring as a display of the plane rather than as a circumplex.

Calls cluster unevenly. Call angles bunch together far more than noise would produce: $\chi^2 = 68.1$, where noise run through the same pipeline reaches only 13.6 at its 95th percentile. The densest regions sit at 2.6°, 144.4° and 200.6°. They are denser and sparser stretches of a continuum, not three separate types.

7. The Trade Map: eight selling styles

We divide the ring into eight equal 45° sectors, with the first anchored at 2.6°. That angle is the density center with the cleanest profile, which means the data sets the grid rather than an arbitrary zero. Each sector has a name and a fixed behavioral subtitle, and the name is never shown without its subtitle.

Eight selling styles on one plane; three sit on density centers

Outer labels mark the + pole of each dimension; its – pole lies directly opposite. Share of the 901 calls under each name.

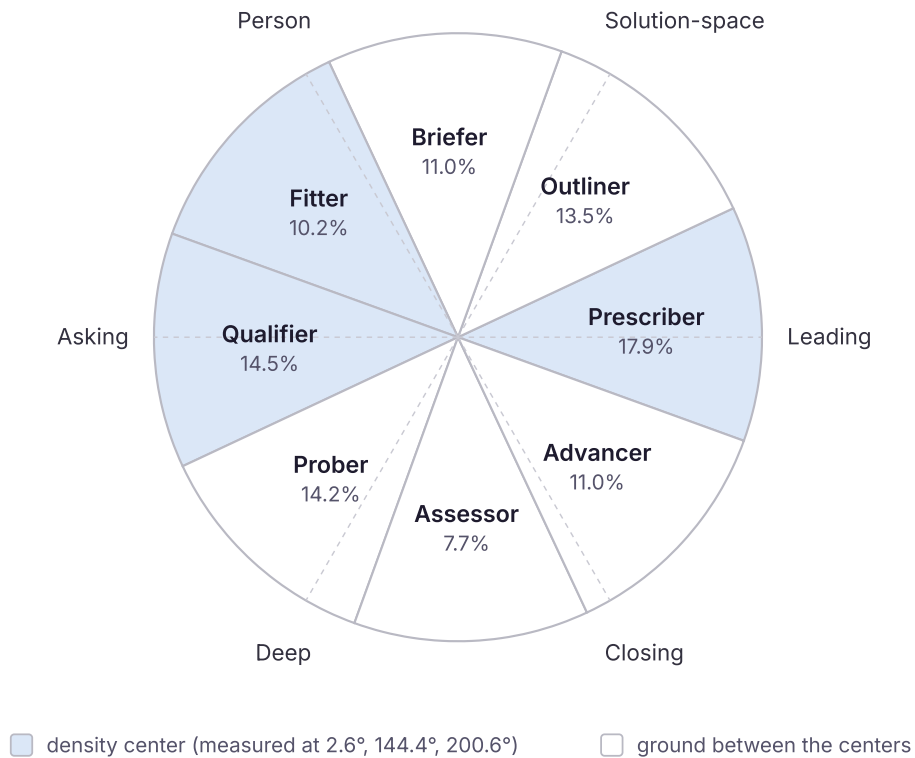


Figure 2. The Trade Map: six dimension poles and eight selling styles

Sector	Selling style · subtitle	Share of calls	Confident read (research rule)
S0	Prescriber · tell-first, next-step-led (density center)	17.9%	46.6%
S1	Outliner · offer-led, broad-pass	13.5%	20.5%
S2	Briefer · person-warm, offer-led	11.0%	9.1%
S3	Fitter · person-first, open-ended (density center)	10.2%	22.8%
S4	Qualifier · question-first, buyer-steered (density center)	14.5%	41.2%
S5	Prober · depth-first, offer-light	14.2%	25.8%
S6	Assessor · depth-first, next-step-led (sparsest arc)	7.7%	15.9%
S7	Advancer · task-first, next-step-led	11.0%	38.4%

Naming is part of the measurement. In research the selling styles were named after working trades (Paver, Frammer, Surveyor and so on). The point was that no trade outranks another. The product uses a plainer sales vocabulary chosen to steer clear of existing framework names and to avoid sounding like a grade: "You're not much of a Prescriber" doesn't land as criticism. Names that imply an outcome, such as Closer or Hunter, were rejected because the data shows no selling style winning.

Why every card carries a name. Some selling styles produce a confident read more often than others. Prescriber passes the research confidence rule on 46.6% of its calls, while Briefer passes it on 9.1%. If only confident reads were named, teams would see far more Prescribers than actually occur. The product therefore names the nearest selling style on every call and shows its confidence alongside.

Blends are the normal case. Membership across the eight selling styles is soft, calculated from the measured angular uncertainty rather than from the sector width. On a single call, the median top weight is 0.37 and the median second weight is 0.29, so most calls are blends of two neighboring styles.

8. What the evidence supports

We tested each claim below in a way that could have disproved it. Chambr shows sellers only the claims that held. The rest are things Chambr deliberately does not say.

Claim	Verdict	Evidence
Selling styles reflect behavior, not the scenario	Holds in the research set	Before correction, the persona predicts position (Cramér's V 0.40–0.47). After correction there is no significant association ($p = 0.10$ – 0.96)
Selling styles describe countable behavior	Holds	Question rate separates the eight sectors ($p = 9e-11$)
Time of day changes style	No effect	Within-seller hour against angle, $p = 0.38$
Sellers facing the same buyer still differ	Preliminary	Against one widely used simulated buyer, sellers spread across the selling styles much as they do overall. The test mixes two closely related personas, so a single-persona re-run is planned
The structure is the same on both platforms	Not established	The dimension order appears on both, but random splits reproduce it 66% of the time. How calls spread around the ring differs ($p = 0.024$)
Some selling styles outperform others	No	Position explains 3.0% of call-score variance ($p = 0.20$). It adds nothing measurable once overall activity and closing behavior are controlled ($p = 0.48$)
A seller has a fixed style	No	Two halves of a seller's own calls land in the same selling style 14.9% of the time, against 12.5% by chance
How widely a seller varies is a trait	No	Split-half $r = -0.12$, 95% CI $[-0.34, +0.14]$

Why no selling style wins. Raw success rates by selling style show a threefold spread. Most of it comes from **momentum**: for a cold call, success means securing a next step, and that is what **momentum** measures. Once **momentum** and overall activity are controlled, no selling style differs significantly from another. The success label itself is a narrow proxy, so testing style against business outcomes such as meetings held is planned work.

Why we validate against outcomes without a language model. Chambr's call score is model-based, and validating one model-based score against another is circular. When we swapped only the source of the outcome label, style–outcome correlations rose from 0.243 to 0.417. That is a 72% inflation from the method alone, a well-known measurement trap when the same kind of tool sits on both sides of a comparison (Campbell & Fiske, 1959). We therefore validate only against outcomes that involve no language model.

Why style belongs to the call. The median seller with eight calls shows up in five of the eight selling styles. After scenario correction, a seller's own calls are no more alike in angle than calls from different sellers ($p = 0.30$). We cannot yet separate genuine call-to-call change from measurement noise. That is exactly why the product describes calls and never assigns a type.

Sellers do show stable tendencies on individual dimensions, but it takes about 11 calls to read them reliably (reliability 0.70, based on intraclass correlations; Shrout & Fleiss, 1979). Seller-level views are therefore built from dimensions, not selling styles.

9. Confidence: what the product may say

Before anything is shown, every call is given a **confidence level**. The research rule uses three checks: how strongly the call leans, how well a single direction fits it, and how dominant the top selling style is.

Confidence level	Condition	Share of 901 calls
Single position	amplitude ≥ 0.40 , $R^2 \geq 0.70$, top weight $> 2 \times$ second	20.0%
Blend	amplitude ≥ 0.40 , $R^2 \geq 0.70$, top weight $\leq 2 \times$ second	9.5%
Dimensions only	amplitude ≥ 0.40 , $R^2 < 0.70$	49.7%
No pronounced style	amplitude < 0.40 , or no angle	20.8%

The research version named a selling style only at the first two levels. The product names the nearest selling style at every level and shows the level as a label, so the uncertainty stays visible without hiding the result.

A position is an arc, never a point. When the same calls are scored twice, the angle moves by a median of 34° . How strongly a call leans controls how much of that movement matters. For the flattest quarter of calls the median move is 64° , and for the most pronounced quarter it is 21° . The arc width reflects this: 1.7 sectors for calls that lean clearly, and 3.3 for those that don't. Measured more strictly, against the frozen reference set and across different models, the arc is wider, at about 2.0 to 2.5 sectors.

Seller-level views are built from dimensions and held back until the evidence supports them.

Each seller-level estimate is pulled toward the population average in proportion to how much evidence there is, a standard technique known as partial pooling (Gelman & Hill, 2007; Appendix A). One call counts for 18% of the estimate, eight calls for 64%, and twenty calls for 81%. No seller-level selling style is ever named, and **tempo** is never used for a seller-level claim. How widely a seller varies is described in words, never scored. Change-over-time claims ("you've

become more of an Advancer") are not made, because the observed movement cannot yet be told apart from noise.

Banned everywhere in the product:

- **Anything that reads as a grade:** "strong", "needs improvement", percentiles, letters, red-to-green scales.
- **Anything that reads as identity:** "You are a Prober".
- **Any outcome-by-style table.** It looks dramatic and it is an artifact (Section 8).

The gate copy explains why without apologizing. A call that doesn't resolve to one direction says so, with six readings and the transcript moment behind each one. It never says "insufficient data". Every confidence level gets the same design weight, because a card that looks less finished would introduce the ranking the copy avoids.

10. What comes next

The dimensions have been checked against three different language models. The next step is a check by people. Planned work, subject to change, includes:

- **Human coding.** Two trained coders scoring 150 calls blind, so that each dimension can be validated against human judgment.
- **Longer histories.** Scoring the full history of sellers with 20 or more calls, to study how style changes over time.
- **Business outcomes.** Linking style to downstream results such as meetings held and opportunities created.
- **A stronger tempo.** Refining its anchors to bring its stability up to the level of the other five.
- **Live calibration.** Re-scoring production calls to refine the scenario correction and the confidence arcs on live data.

Three boundaries on the current evidence are worth keeping in mind:

- **Practice calls with simulated buyers.** All calls are roleplays against AI buyers. For calls on the current platform, we confirmed that the buyer does not mirror the seller's style in 99.3% of calls. Calls on the previous platform were not checked.
- **Text only.** The data contains no tone of voice, pauses or interruptions, so warmth is read from wording alone.
- **Seller and scenario overlap.** Each seller practices only a few scenarios, so correcting for scenario also removes some seller signal. Seller-level effects may be understated.

These limits are why every Seller DNA card shows its confidence, and why no card ranks a seller.

Appendix A. Technical detail

A.1 Agreement and seller stability by dimension

Columns 2 to 4 are correlations (r) between two scorings of the same calls. Seller ICC is the share of a dimension's variation that belongs to the seller rather than to the individual call.

Dimension	Same model, second pass	vs second model	vs third model	Seller ICC [95% CI]
inquiry	0.845	0.785	0.642	0.204 [0.11, 0.30]
momentum	0.813	0.659	0.615	0.259 [0.17, 0.34]
warmth	0.799	0.709	0.622	0.153 [0.04, 0.26]
solution_pull	0.740	0.359	0.563	0.139 [0.00, 0.30]
assertion	0.727	0.537	0.608	0.283 [0.18, 0.37]
tempo	0.504	0.598	0.427	0.031 [0.00, 0.12]

A.2 Scenario correction

Call length, platform and call type are fixed effects. Each buyer persona is a partially pooled random intercept, shrunk toward zero according to how many calls it has:

$$\begin{aligned}
 S &= D \cdot \beta + P \cdot u + e \\
 \hat{u}_j &= n_j \sigma^2_b / (n_j \sigma^2_b + \sigma^2_e) \cdot \text{mean residual of persona } j \\
 S_r &= S - D \cdot \hat{\beta} - P \cdot \hat{u}
 \end{aligned}$$

The variance components are estimated separately for each dimension by REML. The persona's share of variance ranges from 0.15 (tempo) to 0.31 (inquiry).

A.3 Structural summary

Each standardized, row-centered profile is fitted by ordinary least squares:

$$S_i = e + a \cdot \cos(\theta_i - d) \quad \theta = 0^\circ, 60^\circ, 120^\circ, 180^\circ, 240^\circ, 300^\circ$$

Here e is elevation, a is amplitude, d is the angle, and R^2 is the fit. A flat profile, or one with no first harmonic, returns no angle.

A.4 Circumplex test against known structures

Structure	Correspondence index	Cosine-law R^2	Circulant deviation
True circumplex, low noise	+1.000	0.996	0.060
True circumplex, high noise	+1.000	0.984	0.096
Two-factor simple structure (not a circle)	+0.542	0.495	0.664
Pure noise	-0.125	0.007	0.194
Chambr, held-out sellers	+0.708	0.443	0.645

Chambr's cosine-law fit and circulant deviation match the two-factor plane, not the circle.

A.5 Seller-level shrinkage

$$w(n) = n / (n + \lambda) \quad \lambda = (1 - \text{ICC}) / \text{ICC}$$

Calls	1	3	5	8	20
Weight on the seller's own mean	0.18	0.40	0.52	0.64	0.81

A.6 Frozen at each release

- The reference set: per-dimension means and SDs, the correction coefficients, and the persona corrections.
- The rubric and the scoring model, recorded on every stored score.
- The ring angles, and the sector boundaries anchored at 2.6°.
- The confidence thresholds: fit 0.70, amplitude 0.40, median amplitude 0.629, and arc spreads of 19.3° and 38.2°.

References

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. <https://doi.org/10.1037/h0046016>
- Franke, G. R., & Park, J.-E. (2006). Salesperson adaptive selling behavior and customer orientation: a meta-analysis. *Journal of Marketing Research*, 43(4), 693–702. <https://doi.org/10.1509/jmkr.43.4.693>

- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511790942>
- Gurtman, M. B. (2009). Exploring personality with the interpersonal circumplex. *Social and Personality Psychology Compass*, 3(4), 601–619. <https://doi.org/10.1111/j.1751-9004.2009.00172.x>
- Ramsey, R. P., & Sohi, R. S. (1997). Listening to your customers: the impact of perceived salesperson listening behavior on relationship outcomes. *Journal of the Academy of Marketing Science*, 25(2), 127–137. <https://doi.org/10.1007/BF02894348>
- Saxe, R., & Weitz, B. A. (1982). The SOCO scale: a measure of the customer orientation of salespeople. *Journal of Marketing Research*, 19(3), 343–351. <https://doi.org/10.2307/3151568>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Spiro, R. L., & Weitz, B. A. (1990). Adaptive selling: conceptualization, measurement, and nomological validity. *Journal of Marketing Research*, 27(1), 61–69. <https://doi.org/10.2307/3172551>
- Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Large language models are inconsistent and biased evaluators. arXiv preprint. <https://doi.org/10.48550/arXiv.2405.01724>
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1), 25–29. <https://doi.org/10.1037/h0071663>
- Tracey, T. J. G. (1997). RANDALL: a Microsoft FORTRAN program for a randomization test of hypothesized order relations. *Educational and Psychological Measurement*, 57(1), 164–168. <https://doi.org/10.1177/0013164497057001012>
- Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: the interpersonal domain. *Journal of Personality and Social Psychology*, 37(3), 395–412. <https://doi.org/10.1037/0022-3514.37.3.395>
- Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems 36* (Datasets and Benchmarks track). <https://doi.org/10.48550/arXiv.2306.05685>
- Zimmermann, J., & Wright, A. G. C. (2017). Beyond description in interpersonal construct validation: methodological advances in the circumplex Structural Summary Approach. *Assessment*, 24(1), 3–23. <https://doi.org/10.1177/1073191115621795>

Read the findings online and try Jerry, our free AI cold call bot, at chambr.ai/labs/measuring-seller-style. All trademarks belong to their respective owners.