

Measuring Public Speaking Outcomes: Overview and Framework

- Outcomes span cognitive, affective, behavioral, and audience-level results
- Separate real outcomes from attendance, satisfaction surveys, and anecdotes
- Four-level logic: reaction, learning, behavior, results — with clear limits
- Core vocabulary: formative vs. summative; criterion- vs. norm-referenced; reliability vs. validity
- Goal: a usable measurement framework for coaches, trainers, educators, evaluators



Defining Outcome Domains and Competency Models

1

Map observable domains:
content and organization,
delivery and vocalics,
language and clarity

2

Translate competence
frameworks into low-inference
behavioral indicators, not
trait labels

3

Separate speaker-internal
outcomes (skill, self-efficacy)
from audience-internal
outcomes (comprehension,
retention, action intent)

4

Choose domains in scope by
program level and audience type:
classroom, workplace, community,
conference

5

Build a lightweight competency
map linking each program goal
to measurable indicators



From Competency Statements to Observable Indicators

- Trained observers must see or hear the indicator in one viewing
- Anchor behaviors: eye contact, pacing, transitions, evidence quality, Q&A handling
- Align rubric language to the specific speaking context and audience
- Decide criterion-referenced or norm-referenced scoring before rating begins
- Every indicator must trace back to an intended outcome, not an incidental behavior



Validity, Reliability, and Rater Bias in Speaking Assessment

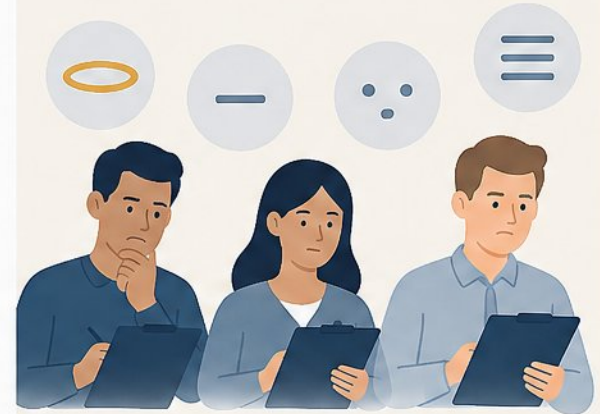
- **Validity:** ratings reflect true speaking ability, not rater habits
- **Reliability:** multiple raters agree on the same performance
- **Common biases:** halo effect, leniency/severity, central tendency, order effects
- Accent, dialect, gender, or major bias contaminates ratings
- High inter-rater reliability does not guarantee valid judgments
- **Construct-irrelevant variance:** anxiety, topic familiarity, room acoustics, recording quality

Valid and Reliable Judgments



Clear signal for fairer judgments

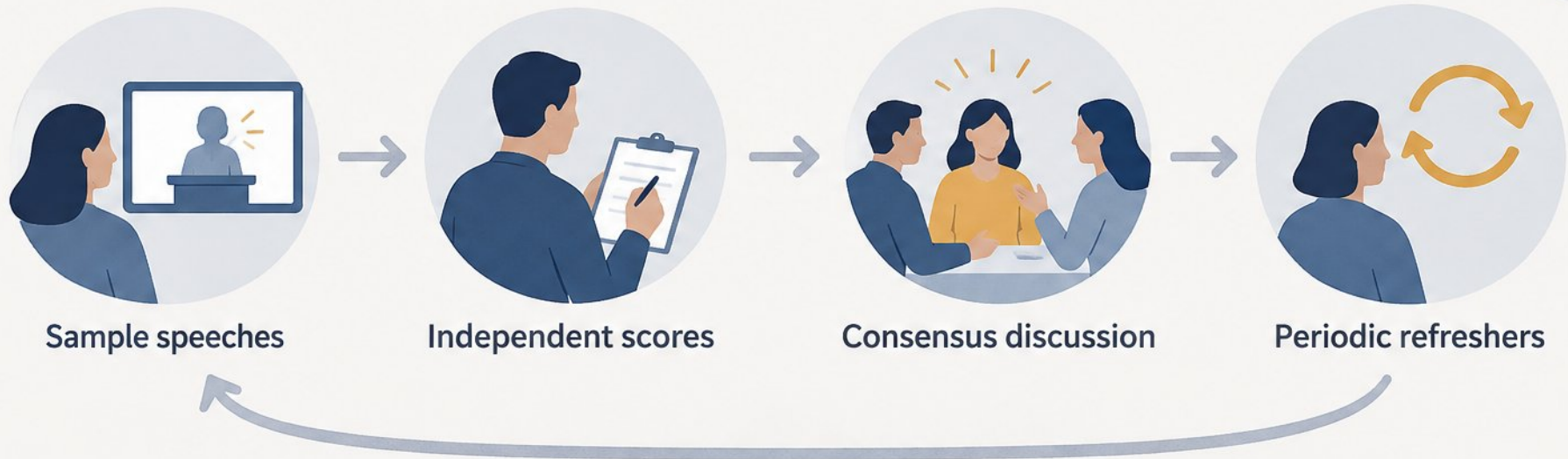
Biased Judgment Patterns



More noise can distort ratings



Rater Training, Calibration, and Guardrails



Short calibration cycles: shared sample speeches, independent scores, consensus discussion



Double scoring and blind review as practical reliability safeguards



Training effects fade; schedule periodic refreshers and ongoing monitoring



Track severity, leniency, and bias during and after each rating cycle



Document rater decisions so results can be audited and defended

Rubrics, Rating Scales, and Observation Protocols

- Analytic rubrics for coaching feedback; holistic rubrics for fast program-level judgments
- Behaviorally anchored scales describe each performance level; checklists and counts flag specific behaviors
- Write descriptors naming observable actions to reduce interpretive drift between raters
- Observation protocols: live versus recorded, with time-sampling and event coding options
- Pilot instruments on a small set of speeches before full deployment



Direct Measures: Performance Tasks and Audience Response



Performance tasks and simulations that mirror real speaking contexts



Immediate audience checks: comprehension, recall, clarity, attitude, action intent



Behavioral data: timing, fillers, gestures, slide dependence, Q&A responsiveness



Audience response systems for live comprehension and engagement signals



Triangulate direct measures with observer ratings; no single source suffices



Using Audience Response and Comprehension Checks Well



Design short post-talk quizzes targeting recall, inference, and misconceptions



Read response distributions to spot divided or collectively wrong understanding



Add very short answers to reveal reasoning beyond fixed options



Interpret engagement, confidence, and comprehension as separate measures

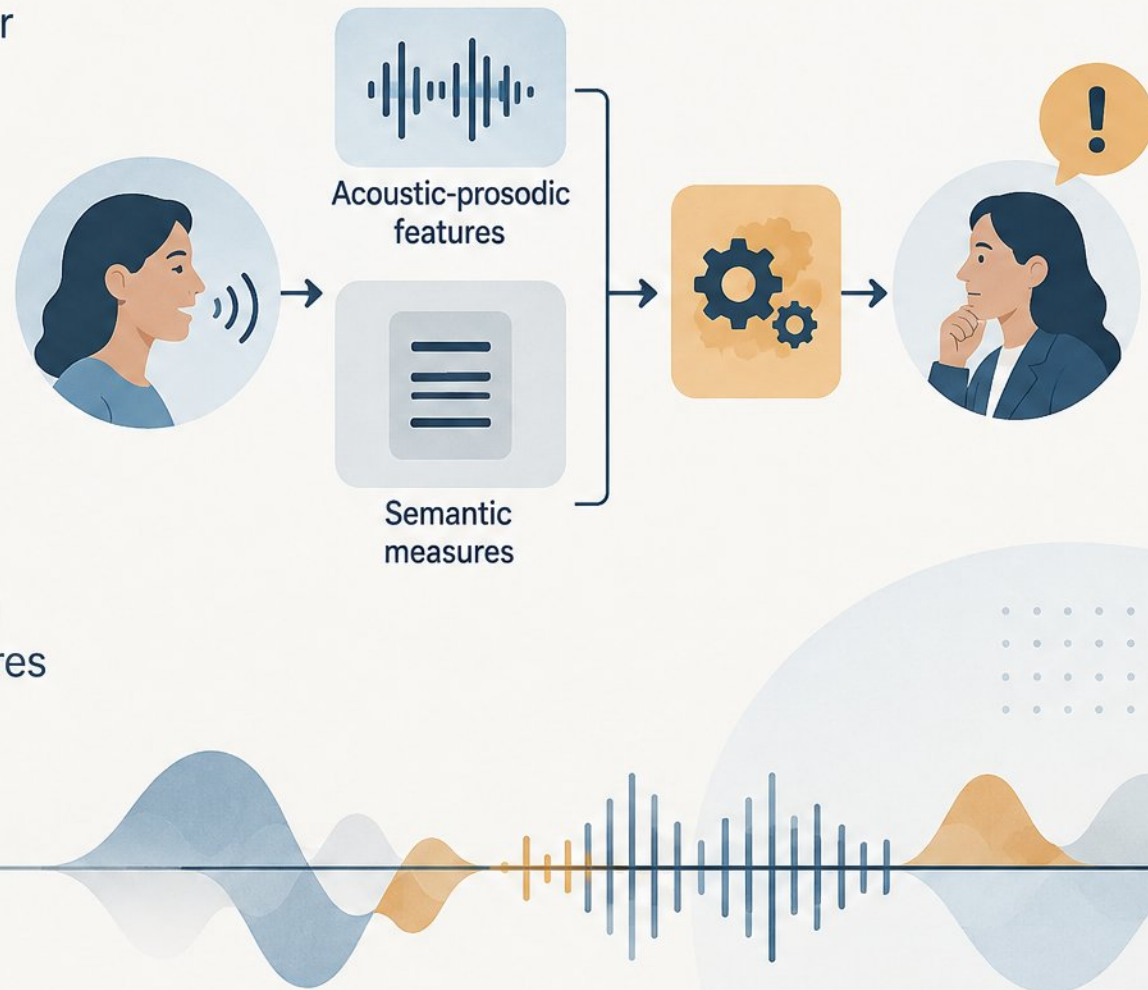


Keep polling ethical: consent, anonymity, low-pressure participation



Technology-Assisted Measurement: Automation and Its Limits

- Automated scoring is strongest on delivery ($r \approx .78$), weakest on grammar and vocabulary ($r \approx .50$)
- Human-machine correlations vary by construct: match the tool to the intended outcome
- Hybrid human-machine composites can match or exceed human holistic reliability
- Acoustic-prosodic features show high test-retest reliability; semantic measures are less stable
- Flag fluent but off-topic content and abnormal speech for human review



Indirect Measures: Self-Report, Feedback, and Perceived Competence



Use self-report for speaking anxiety, self-efficacy, and perceived improvement



Validated tools: PRPSA-18, SSPS (positive/negative self-statements), speaking self-efficacy scales



Design brief peer and audience feedback forms with clear, low-inference items



Watch for social desirability, acquiescence, and recall bias in self-report data



Treat gaps between self-rated and observer-rated performance with caution



Designing an Evaluation Plan: Baselines, Growth, and Comparison



Choose a design: pre-post, repeated measures, cohort comparison, or single case



Set baselines and realistic growth targets for program length and starting points



Decide sampling, timing, and dosage across the program cycle



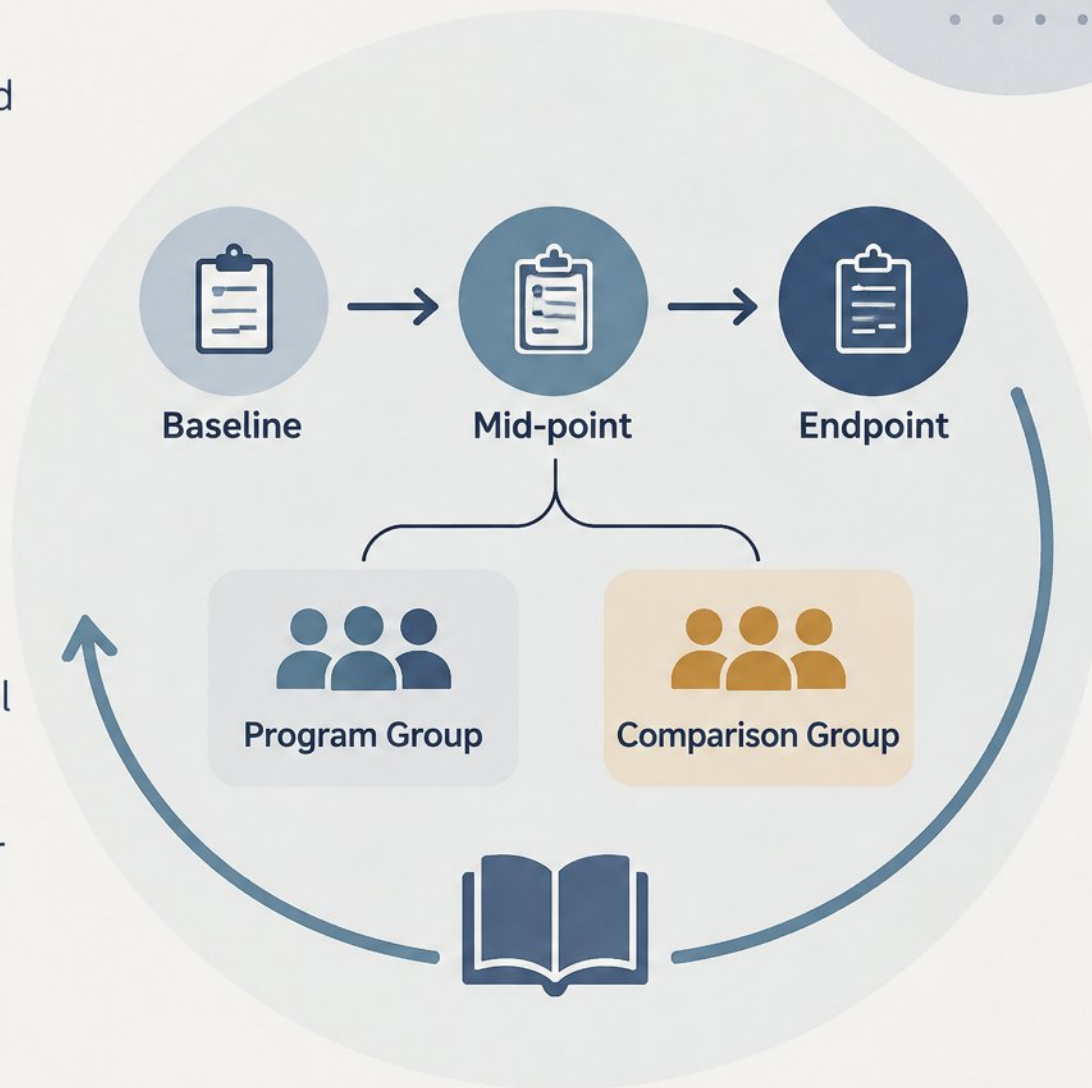
Use comparison groups where feasible; interpret non-experimental results cautiously



Align the plan with resources, rater capacity, and ethical review

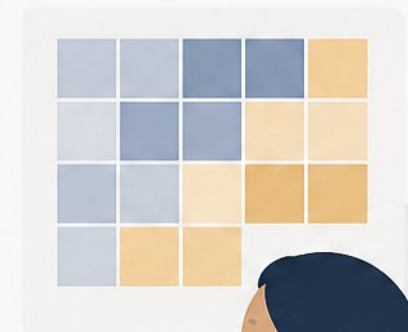
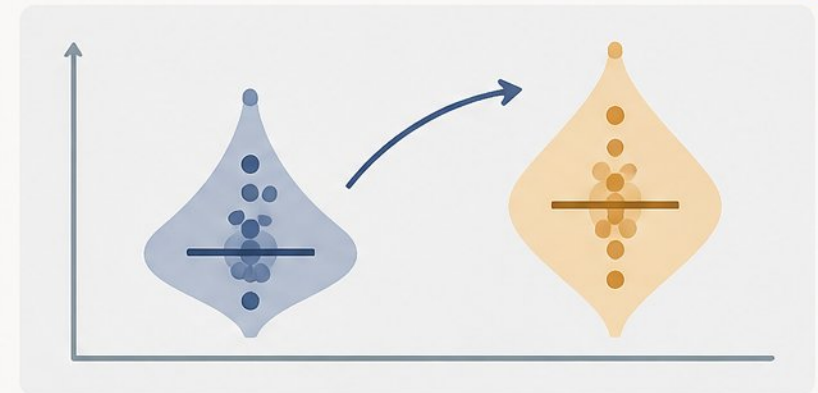


Obtain informed consent and IRB approval when required



Analyzing and Interpreting Speaking Outcome Data

- Start with descriptives: means, distributions, and change scores
- Report effect sizes with confidence intervals, not just p-values
- Judge practical significance against similar speaking studies and program goals
- Sample size shapes effects: small samples often inflate magnitudes (Cheung & Slavin, 2016)
- Address missing data, rater disagreement, and outliers transparently
- Visualize growth: before/after charts and rubric heatmaps
- Avoid over-interpreting small samples, p-hacking, and causal claims from correlation



Reporting Results and Closing the Improvement Loop



Tailor detail to learners, instructors, administrators, and funders



Turn findings into actionable feedback and targeted coaching decisions



Report ethically: protect anonymity, prevent score misuse, state uncertainty



Communicate limitations, methods, and scoring honestly and transparently



Schedule review cycles to revisit instruments, targets, and evaluation questions



Capstone: Building and Defending a Measurement Framework

- Assemble one complete framework for your program: outcomes, measures, timeline, raters, analysis
- Connect every measure to an intended outcome; state validity, fairness, and feasibility rationale
- Anticipate stakeholder questions: reliability evidence, rater effects, cost, and scoring safeguards
- Use double scoring, calibration cycles, and documented rater decisions to defend results
- Peer-review with a critique checklist, then revise instruments and targets before full deployment



PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/3dc8d156/public