PRACTICE GROUND BY VIRBELA

From Leadership Theory to Observable Behavior

The empirical foundations and validation agenda for AI-powered leadership simulations

WHITEPAPER THESIS

The research base strongly supports the constructs, behavioral measurement format, and simulation method. The validity of AI scoring in open-ended leadership conversations must be established through Virbela's own reliability, fairness, and criterion-related validation program.

Version 1.0 | July 2026

Prepared from Virbela's Empirical Foundations v1.2

Executive summary

Leadership is expressed in moments of action: giving difficult feedback, holding standards under resistance, coaching a struggling employee, or navigating conflict without losing trust. Yet much leadership development still measures participation, knowledge, or self-reported confidence rather than what a leader actually does in the conversation.

Practice Ground is designed around a different premise. Leaders enter structured, voice-based scenarios, respond naturally to an adaptive AI counterpart, and receive feedback tied to observable behavior. This approach sits at the intersection of four empirical traditions: leadership competency research, behaviorally anchored rating scales, simulation-based assessment, and deliberate practice.

THE CENTRAL CONCLUSION

Practice Ground operates in a well-established empirical neighborhood, but it does not automatically inherit the validity coefficients of assessment centers, structured interviews, role plays, or situational judgment tests.

What the published evidence supports

- The six-competency model has recognizable foundations in industrial-organizational psychology and leadership research.
- Behaviorally Anchored Rating Scales have a six-decade history as a way to translate abstract constructs into observable performance standards.
- Structured observation in controlled scenarios can support useful workplace inferences when the assessment is well designed and validated.
- Repeated practice with specific feedback creates conditions that can support skill development.

What Virbela must establish

- Agreement between AI scoring and calibrated expert human scoring.
- Score consistency across equivalent scenarios, repeated use, and system updates.
- Fairness across demographic, linguistic, cultural, and accessibility-relevant groups.
- Construct validity, including whether scores reflect intended leadership behavior rather than scenario-specific or language-style artifacts.
- Relationships between simulation performance, learning transfer, and consequential workplace outcomes.

1. Why leadership development needs behavioral evidence

Courses can transfer concepts. Workshops can create insight. Neither guarantees that a manager can hold two standards at once when a valued employee resists accountability. The gap between knowing and doing becomes visible only when the learner must choose words, sequence the conversation, respond to resistance, and move toward a clear commitment.

Simulation makes that behavior observable under controlled conditions. It does not recreate the workplace perfectly, but it creates a repeatable setting in which consequential behaviors can be elicited, recorded, examined, and practiced again.

A lineage that predates generative AI

Practice Ground extends work that began with Virbela's GMAC-funded global business simulation at the University of California San Diego. In 2015, Howland and colleagues described that virtual environment as a proof of concept for a virtual assessment center, connecting immersive technology to the established assessment-center tradition. The current product changes the medium and scoring architecture, but retains the core logic: put people into realistic situations and examine what they do.

DESIGN PRINCIPLE

The unit of development is not the course. It is the consequential moment, the observable behavior within it, and the next rep.

2. The six-competency model

The model organizes leadership behavior into six competencies. They are not equally mature in the published literature, and the boundaries between them are not perfectly clean. That is normal in leadership measurement, where communication, interpersonal behavior, coaching, and accountability are inherently connected.

Competency	Working focus	Relative empirical foundation
Communication	Clear, audience-aware, adaptive exchange	Substantial, with overlap and cultural variability
Interpersonal Effectiveness	Trust, empathy, collaboration, and relationship navigation	Substantial but broad and multidimensional
Coaching &	Feedback, inquiry, goal support, and growth	Moderate and distributed across

Competency	Working focus	Relative empirical foundation
Development		adjacent literatures
Judgment	Interpreting conditions and making sound decisions	Strong
Integrity & Accountability	Standards, ownership, follow-through, and ethical conduct	Strong
Learning Agility	Learning from experience and adapting behavior	Thinnest as a unified construct; supported through adjacent constructs

The model should therefore be used as a practical behavior framework, not as six perfectly independent latent traits. Reports are most defensible when they describe behavior observed in the scenario context rather than claiming to reveal a context-free essence of the person.

3. Why behaviorally anchored measurement

Behaviorally Anchored Rating Scales, first formalized by Smith and Kendall in 1963, replace vague labels with descriptions of observable performance at different levels. Instead of asking whether a leader demonstrated 'good communication,' a behavioral rubric asks whether the leader named the issue, acknowledged the other person's perspective, held the standard under resistance, and secured a specific commitment.

What BARS contributes

- A shared scoring language anchored in observable behavior.
- Greater transparency about what a score means.
- A practical bridge from assessment to coaching.
- A foundation for calibrating human and automated raters.

What BARS does not solve

- A well-written rubric does not prove that scores are reliable or valid.
- Behavioral examples can become overly literal prototypes, causing raters to reward expected phrasing rather than equivalent behavior.
- Scenario content can dominate scores, creating exercise effects.
- Competency overlap limits claims that each dimension is fully distinct.

IMPORTANT BOUNDARY

The strongest current use is developmental: describe behavior, support reflection, and guide practice. Higher-stakes selection or promotion use requires

a substantially stronger job-specific validation case.

4. The case for simulation-based assessment

Assessment centers, role-play exercises, situational judgment tests, and structured behavioral interviews all support the broader proposition that structured responses to job-relevant situations can yield useful information about workplace behavior. Meta-analytic findings across these traditions are meaningful, but they belong to those methods and study conditions.

Research tradition	What transfers as a design principle	What does not transfer automatically
Assessment centers	Multiple observations, structured exercises, trained scoring, data integration	Published validity coefficients
Role-play exercises	Behavior elicited through interpersonal interaction	Equivalence between human and AI counterparts
Situational judgment tests	Job-relevant scenarios can sample judgment and likely behavior	Validity of open-ended voice responses and LLM scoring
Structured interviews	Standardization and behavior-focused evaluation improve inference quality	Interview validity estimates for a different scoring process

5. AI scoring: the central open question

Automated scoring has a substantial history in structured and constructed-response assessment. Large-language-model scoring of open-ended leadership conversations is different. The response is interactive, the AI counterpart affects what behavior is elicited, and the scorer must distinguish the intended construct from verbal fluency, cultural style, scenario familiarity, and prototype matching.

The correct first validation target is inter-rater reliability and agreement: when calibrated organizational psychologists and the AI score the same transcript against the same rubric, how closely do their ratings align? Strong agreement would establish scoring consistency with expert judgment. It would not, by itself, establish construct validity, fairness, learning transfer, or prediction of workplace outcomes.

A three-system architecture

Practice Ground separates three jobs that are often collapsed into one model interaction. This separation is intended to reduce role conflict, improve traceability, and let each system operate with the context appropriate to its task.

System	Primary job	Output
1. Roleplay	Elicit natural behavior through an adaptive, voice-based conversation	Conversation and transcript
2. Assessment	Evaluate the completed transcript against scenario-specific behavioral criteria	Structured assessment, evidence, and development priorities
3. Coaching	Use the full transcript, structured assessment, and coaching prompt to guide reflection and rehearsal	Dialogic coaching, meaning-making, and the next rep

The third system is more than a feedback screen. The coach can examine the whole interaction, ground its questions in the independent assessment, invite the learner to interpret what happened, and help translate evidence into a stronger next move. Public competitor materials commonly describe roleplay followed by automated feedback or an AI coach. Some vendors explicitly describe a separate transcript-analysis step. Based on currently available public descriptions, the exact sequence of independent assessment followed by a transcript-and-assessment-informed coaching conversation is not yet a standard, clearly documented category pattern. This should be positioned as a distinctive architecture, not claimed as unique without deeper product verification.

WHY THE SEPARATION MATTERS

The roleplay system should stay in character. The assessment system should judge evidence. The coaching system should help the learner make sense of that evidence and act on it. Separating these functions makes the logic easier to test, improve, and explain.

Evidence question	Recommended study	What success would establish
Does AI score like calibrated experts?	Blind AI-human inter-rater study with adjudication	Scoring consistency and error patterns
Are scores stable?	Test-retest and alternate-form studies	Reliability across time and equivalent scenarios
Is the system fair?	Subgroup, language, and transcript-production analyses	Detection and mitigation of differential error
Does it measure the intended constructs?	Multitrait-multimethod and experimental criterion studies	Convergent and discriminant validity
Does it matter at work?	Longitudinal studies with workplace outcomes	Criterion-related validity and transfer

CURRENT SCORING SCOPE

The simulations are voice-based, while the current scoring layer evaluates transcripts. Vocal tone, volume, pacing, facial expression, gesture, posture, and eye contact are not currently part of the scored construct.

6. Fairness and the transcript-production layer

Fairness does not begin with the scoring model. It begins with whether speech is transcribed accurately and whether the scenario elicits comparable opportunities to demonstrate the target behavior. Accent, speech differences, language proficiency, audio quality, interruption behavior, and culturally variable communication norms can all affect the transcript presented to the scorer.

- Evaluate transcription error by subgroup and relevant speech characteristics.
- Separate communication style from criterion-relevant behavior wherever possible.
- Test culturally alternative but behaviorally equivalent ways of demonstrating a criterion.
- Provide accessible alternatives and document the construct implications of accommodations.
- Monitor model and transcription changes as measurement-system changes, not merely software updates.

7. From one score to deliberate practice

Practice Ground is designed for development and repeated use. The deliberate-practice literature supports structured tasks near the edge of capability, immediate feedback, attention to weak components, and opportunities to repeat with adjustment. However, effects in professional and educational settings are more modest and variable than popular accounts often imply.

Improvement on the same scenario may reflect familiarity with that scenario rather than growth in the underlying competency. Developmental use therefore requires alternate forms that exercise the same criteria at comparable difficulty, stability when no intervention occurs, and sensitivity when meaningful learning occurs.

DEFENSIBLE DEVELOPMENT CLAIM

Simulation with structured behavioral feedback creates conditions that can support skill development. The magnitude of development depends on engagement, feedback quality, prior capability, repetition, and alignment with the workplace context.

8. Construct-validity engineering

Scenario-based assessment has a known construct-validity problem: performance can vary more by exercise than by the named competency. A second problem appears when behavioral examples are treated as exhaustive. A learner may demonstrate the intended behavior through an unexpected path and be scored too low, while another may mimic canonical language without demonstrating the deeper construct.

Virbela's response is a criterion-authoring discipline that distinguishes the underlying construct, multiple acceptable paths of execution, and observable evidence. This is paired with a staged four-layer validation architecture.

1. A structural validator checks criterion completeness and internal consistency in the production pipeline.
2. An AI review layer examines criteria for ambiguity, contamination, and over-literal prototypes.
3. Synthetic learners test whether diverse strategies, including non-prototypical valid strategies, are scored appropriately.
4. Sampled expert human review provides ongoing calibration, error discovery, and drift detection.

As of Empirical Foundations v1.2, the structural validator is operational. The AI-review and synthetic-learner layers are in staged implementation. The sampled human-review cadence begins when the bundle library reaches sufficient scale. External claims should reflect that implementation status.

9. A cumulative validation roadmap

Validation is not a single study or badge. It is an accumulating argument that the interpretations and uses of scores are justified. The sequence below prioritizes the evidence needed for responsible developmental use before expanding toward more consequential claims.

Stage	Priority	Decision enabled
1. Scoring reliability	AI-human agreement, disagreement taxonomy, rater calibration	Whether automated feedback is sufficiently consistent for development
2. Technical stability	Alternate forms, retest effects, model-version regression testing	Whether score change can be interpreted over time
3. Fairness	Subgroup analyses, transcription effects, accessibility and	Where restrictions, adjustments, or

Stage	Priority	Decision enabled
	cultural calibration	warnings are required
4. Construct validity	Convergent, discriminant, and method-effect studies	How precisely competency-level scores can be interpreted
5. Transfer and outcomes	Longitudinal workplace criteria and intervention studies	Whether practice predicts or produces meaningful workplace change

Use restrictions during validation

- Position scores as developmental evidence, not a complete judgment of leadership capability.
- Do not use scores as the sole basis for hiring, promotion, discipline, or termination.
- Communicate the current transcript-only scoring scope.
- Preserve versioning, provenance, and auditability for scoring-system changes.
- Distinguish evidence about the broader method from evidence about Practice Ground's specific implementation.

10. Claims discipline

Commercial pressure tends to compress nuanced evidence into broad claims. The following register preserves the distinction between a research-grounded design and a validated operational system.

Permitted language	Avoid
Built on research from behavioral observation, BARS, and simulation-based assessment.	Scientifically proven AI assessment.
Designed to provide behavior-based developmental feedback.	Objectively measures leadership.
AI scoring is being validated against calibrated expert human ratings.	AI scores are equivalent to expert assessors.
Practice creates structured opportunities to develop leadership behavior.	A few simulations transform leadership performance.
The competency model draws on established leadership and I-O psychology research.	All six competencies are equally validated and distinct.

Conclusion

Practice Ground is not empirically defensible because it uses AI. It is defensible to the extent that it applies established behavioral-measurement principles, exposes its assumptions, restricts its claims, and produces evidence for the novel parts of the system.

The published literature provides a strong foundation for the constructs, measurement format, and simulation method. Virbela's criterion-authoring discipline is a substantive response to known threats. The remaining work is empirical: demonstrate scoring reliability, characterize fairness and method effects, test construct interpretations, and connect simulation performance to development and workplace outcomes.

THE STANDARD

Strong foundations do not eliminate the need for validation. They make rigorous validation possible.

Selected references

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. Personnel Psychology, 56(1), 125-153.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. Personnel Psychology, 50(3), 655-702.

Christian, M. S., Edwards, B. D., & Bryant, M. S. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. Personnel Psychology, 63(1), 83-117.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72(3), 493-511.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. International Journal of Selection and Assessment, 15(4), 405-411.

Hickman, L., Herde, C. N., Lievens, F., & Tay, L. (2023). Automated analyses of text responses: Emerging tools for personnel selection. Personnel Psychology.

Howland, A. C., Rembisz, R., Wang-Jones, T. S., Heise, S. R., & Brown, S. (2015). Developing a virtual assessment center. Consulting Psychology Journal: Practice and Research, 67(2), 110-126.

International Test Commission & Association of Test Publishers. (2022). Guidelines for technology-based assessment.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. Journal of Applied Psychology, 89(2), 377-385.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review. Personnel Psychology, 67(1), 241-293.

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads. Research in Personnel and Human Resources Management, 28, 99-152.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. Psychological Science, 25(8), 1608-1618.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50(9), 741-749.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology. Psychological Bulletin, 124(2), 262-274.

SIOP. (2023). Considerations and recommendations for the validation and use of AI-based assessments for employee selection.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47(2), 149-155.

About Practice Ground

Practice Ground by Virbela is an AI-powered leadership simulation platform designed to help leaders rehearse consequential workplace conversations, receive behavior-based feedback, and practice stronger responses before the real moment arrives.

Learn more: practiceground.ai | practice.virbela.com

Method note: This public whitepaper is a compressed derivative of Virbela's Empirical Foundations of the Competency Model v1.2, which contains 180 references and detailed construct-by-construct analysis. It is not a validation report and does not present original empirical results from Practice Ground.