INTERNAL RELIABILITY OF AUTOMATED LARGE LANGUAGE MODEL SPEAKING ASSESSMENT: A FOUR-RUN FEASIBILITY STUDY OF SPEAKAUDITOR.COM

Hendi Pratama[1], Izzati Gemi Seinsiani[1], Kushardiyanti Novinda[1], Stephanie Priyanto Putri[2]
[1] *Faculty of Languages and Arts, Universitas Negeri Semarang, Indonesia*
[2] *Independent Researcher, PT Transformasi Pendidikan Indonesia, Indonesia*
Corresponding author: hendipratama@mail.unnes.ac.id

Abstract

Automated speaking assessment increasingly relies on generative language models, yet the measurement stability of such systems is rarely reported, and instruments developed elsewhere are routinely applied to speakers of English in the Expanding Circle without evidence that they behave consistently. This feasibility study examined the internal reliability of SpeakAuditor, an assessment platform scoring public speaking on nine dimensions and twenty-seven sub-variables through a pipeline combining Automatic Speech Recognition with Large Language Model scoring. One hundred English-language speech videos produced by Indonesian speakers were scored four times on a single day under identical pipeline conditions. Agreement was analysed with repeated-measures Analysis of Variance, the Intraclass Correlation Coefficient, the Standard Error of Measurement, Fleiss Kappa and weighted Cohen Kappa. The first run completed without technical fault yet differed systematically from the later three by 2.65 points, whereas those three were statistically indistinguishable from one another. A label-based explanation was tested and rejected, leaving provider-side non-stationarity as the most plausible remaining account. Reliability was therefore estimated on the three homogeneous runs, yielding an Intraclass Correlation Coefficient of 0.835 for a single run and 0.938 for the mean of three, with a Standard Error of Measurement of 3.26 points. Both sets are reported side by side: excluding the first run raises the single-measure coefficient by 0.009 while slightly worsening four other indices, so the exclusion cannot be read as improving the result. Dimension reliability ranged from 0.867 for visually anchored confidence to 0.422 for humour, and level classification proved unstable because the reporting band is narrower than the minimal detectable change. Generative scoring appears stable enough for developmental use when scores are aggregated, and transparent reliability reporting is argued to be a precondition for deploying such systems on speakers already judged against imported norms.

Keywords: automated speaking assessment; large language model; test-retest reliability; intraclass correlation coefficient; decolonial assessment ethics

INTRODUCTION

Public speaking is among the most consequential and least reliably assessed competences in higher education. Evaluating it demands trained judges, sustained attention across long performances, and simultaneous attention to content, structure, delivery, voice, and bodily conduct. In Indonesian universities, where a single lecturer may face several hundred student presentations in one semester, that combination is rarely available. Speaking is therefore assessed impressionistically or not at all, and detailed formative feedback becomes a privilege of students who can afford private coaching.

Systems combining Automatic Speech Recognition (ASR) with Large Language Model (LLM) scoring appear to dissolve this constraint. They read a video, transcribe it,

extract acoustic and visual features, and return a scored report in minutes at negligible marginal cost. The appeal for institutions in the Global South is obvious, and so is the danger. An assessment instrument encodes a definition of competent speech, and most commercially available systems were trained, calibrated, and validated on speakers whose English belongs to Kachru's (1985) Inner Circle. Applying such instruments to Indonesian speakers without evidence risks converting a colonial linguistic hierarchy into an apparently neutral number. Koenecke et al. (2020) showed that widely deployed ASR systems produce substantially higher error rates for speakers of African American English, which demonstrates how technical infrastructure can carry sociolinguistic bias without explicit intent.

A prior question comes first, and it is the one this article addresses. Before asking whether an automated instrument measures the right construct, one must ask whether it measures anything consistently. Generative language models are stochastic. The same input submitted twice may yield different output, and vendors update deployed models without notice. An instrument returning different scores for the same performance on consecutive mornings cannot be defended on any ground.

Published evidence is thin. Research on language models as evaluators has concentrated on agreement with human judgement rather than agreement with themselves, and where self-consistency is reported it is usually for short verdicts on text rather than for multidimensional rubrics applied to multimodal performance. We located no published test-retest study for automated public speaking assessment combining ASR, acoustic analysis, vision, and generative scoring. The gap matters because such pipelines chain several stochastic components, each contributing variance that a composite score conceals.

This article reports a four-run internal reliability study of SpeakAuditor, an assessment platform developed in Indonesia. One hundred English-language speech videos were scored four times on a single day under fixed conditions. Four questions guided the analysis. How stable is the composite score across repeated runs? Which rubric dimensions are stable, and does the pattern follow the interpretive demand each dimension makes? How stable is the level classification derived from the composite? And

what does the observed behaviour imply for responsible deployment in Expanding Circle contexts?

The contribution is threefold. We provide what appears to be the first published multi-run reliability profile for a multimodal LLM speaking assessment pipeline. We show that repeated observation can separate ordinary scoring noise from a step change originating outside the researcher's control, a distinction invisible in the two-run designs that dominate the area. And we argue that disclosing the weaknesses of one's own instrument belongs to the ethics of assessment rather than only to its statistics.

LITERATURE REVIEW

Rubric-based public speaking assessment has a mature instrumentation tradition. The Public Speaking Competence Rubric of Schreiber et al. (2012) remains the reference instrument, comprising eleven descriptive items covering topic selection, organisation, supporting material, language, vocal expression, and nonverbal behaviour, with reported factor structure and predictive validity. Its influence on computational work has been direct, since studies of automated speech quality generally adopt it as the human benchmark against which machine features are validated. The same tradition documents that trained human raters agree imperfectly with one another, which establishes the realistic ceiling against which any automated system should be judged.

Work on language models as evaluators has expanded rapidly. Zheng et al. (2023) established the paradigm of using a strong model as judge and reported substantial agreement with human preference, and Bavaresco et al. (2025) extended the evidence across twenty evaluation tasks while recommending task-specific validation before deployment. A parallel and less comfortable literature documents internal instability. Systematic evaluations of judge models report that same-verdict rates on repeated presentation of identical input fall as sampling temperature rises, and that prompt phrasing, option order, and response length shift verdicts independently of content. Two implications follow. Self-consistency and human agreement are distinct properties, and a system may possess either without the other. Because deployed models are updated by their providers, moreover, the reliability of an assessment pipeline is only partly under the control of whoever built it.

Reliability theory supplies the appropriate instruments. Shrout and Fleiss (1979) formalised the Intraclass Correlation Coefficient (ICC) family, and Koo and Li (2016) provide the standard guidance on selecting among its ten forms and interpreting results through the lower bound of the confidence interval rather than the point estimate, treating values below 0.50 as poor, 0.50 to 0.75 as moderate, 0.75 to 0.90 as good, and above 0.90 as excellent. Two further quantities matter in practice. The Standard Error of Measurement (SEM) expresses run-to-run noise in the units of the scale, and the Minimal Detectable Change (MDC) derived from it states how large a difference must be before it can be distinguished from noise. Where results are reported as ordinal bands, Fleiss (1971) Kappa and weighted Cohen (1960) Kappa quantify classification agreement.

The decolonial dimension is usually treated separately, but it converges here. Scholarship on native-speakerism has long argued that treating Inner Circle usage as the implicit standard positions all other speakers as deficient (Phillipson, 1992; Canagarajah, 1999; Cook, 1999; Holliday, 2006). Automated assessment intensifies the problem twice over. It conceals the norm inside a model whose training data and calibration are inaccessible, and it delivers the resulting judgement with a numerical authority that human raters do not command. A student who receives sixty-two out of one hundred rarely asks who decided what one hundred means.

The proposition guiding this study follows from that convergence. If an automated instrument is to be used on speakers already disadvantaged by standard language ideology, its developers owe users a clear statement of how much of the reported score is signal and how much is noise, expressed in the units the user actually sees. Read this way, reliability reporting becomes a mechanism of accountability that returns interpretive agency to the assessed. This study accordingly treats internal reliability as its object rather than its preface.

RESEARCH METHOD

Design

The study used a quantitative repeated-measures design analogous to test-retest reliability. Because the stimulus is fixed video and the rater is a software pipeline, the interval between administrations affects only the state of the scoring system, never the

state of the performer. Variation observed between runs is therefore attributable to the instrument, which makes the design a direct measurement of instrument-side error.

Instrument

SpeakAuditor (speakauditor.com) scores public speaking on nine weighted dimensions: Quality of Content (16), Quality of Argument (16), Delivery (10), Prosody (11), Confidence Visual (11), Humor (7), Politeness (9), Emotional Appeal (10), and Linguistic Quality (10). Each dimension decomposes into three sub-variables, giving twenty-seven scored elements. Descriptors are anchored in established frameworks, among them audience design for content, the Toulmin model for argument, prosodic theory for voice, gesture and nonverbal immediacy for visual confidence, benign violation theory for humour, face theory for politeness, and complexity, accuracy and fluency for linguistic quality. The weighted composite runs from 0 to 100 and maps onto a ten-tier level matrix in which levels from the third upward occupy bands of ten points.

The pipeline chains four components. Speech is transcribed by a commercial ASR service. Acoustic parameters including speech rate, phonation ratio, fundamental frequency statistics, intensity, and harmonics-to-noise ratio are extracted with *Parselmouth* libary. Sampled video frames are scored for the three visual sub-variables by a vision-capable model. Transcript, acoustic profile, and visual scores are then passed to a generative model *Deepseek Flash V4* that returns sub-variable scores with written justifications. Scoring temperature was fixed at 0.2. No model, temperature, weight, or prompt was altered at any point during the study.

The scoring behaviour of the generative model is governed by prompt engineering rather than by fine-tuning. No parameter update was performed on the underlying language model. Instead, the rubric, the twenty-seven sub-variable descriptors, the anchoring theoretical frameworks, and the scoring procedure for each dimension are encoded in a structured Markdown specification file, which is injected as system context into every scoring call made by the Python pipeline. The generative model therefore performs the assessment task in a zero-shot manner with respect to model weights, conditioned at inference time on an externally maintained rubric rather than on labelled training examples. This design keeps the scoring logic auditable and independent of the underlying model, since revising a descriptor means editing the specification file rather

than retraining a model, and it also means that the reliability reported here is a property of a fixed prompt-model pairing rather than of a calibrated scoring model, a distinction developed further below.

SpeakAuditor.com was initially designed to evaluate general English public speaking skills, such as for job entry assessments, promotion tests, or training outcome evaluations. However, because its designers specialize in Applied Linguistics and TEFL, it meticulously measures key linguistic and pragmatic dimensions, including prosody and pronunciation, grammatical accuracy, sociopragmatics, and vocabulary usage. Consequently, SpeakAuditor.com can also serve as a diagnostic tool in general English-speaking classes to track students' speaking development over time. Currently, this pipeline is also being piloted to support IELTS speaking instruction and factual speech classes at one of the authors' universities.

Corpus

The dataset comprises one hundred English-language speech videos recorded by Indonesian speakers across three genres: casual speech (19), cultural speech (56), and tourism speech (25). The cultural speech subset consists of undergraduate speaking examinations in a cross-cultural understanding course. All speakers are first-language Indonesian users of English, which scopes the reliability claim to Expanding Circle performance rather than to native-speaker norms. Videos were held read-only and were byte-identical across all four runs.

Procedure

The full corpus was processed four times on 28 July 2026, at 09:33, 11:41, 13:21, and 18:29, referred to hereafter as Run 1 to Run 4. Every run used the same notebook, the same model endpoints, the same parameters, and the same input files. All one hundred videos completed successfully in every run and were detected as English in every run, so the analysis rests on a complete four-observation matrix with no imputation and no substitution.

One procedural detail requires disclosure because it later became analytically relevant. The reporting metadata field recording the requesting organisation carried different values across runs, and the sequence was deliberately arranged so that the first and fourth runs shared a value while the second and third shared another. This

arrangement was introduced to test whether that field influenced scoring, and its outcome is reported below.

Analysis

Analysis proceeded in five stages. Run-level descriptive statistics were computed first. Systematic difference was then tested by paired comparison and by one-way repeated-measures Analysis of Variance with partial eta squared as the effect size. Reliability was estimated using two-way random effects ICC for absolute agreement, reported as single-measure ICC(2,1) and average-measure ICC(2,k) with ninety-five per cent confidence intervals, alongside consistency forms for comparison. The Standard Error of Measurement was derived from the residual mean square, and the Minimal Detectable Change at the ninety-five per cent level was computed for a single run and for run averages. Level classification agreement was assessed with Fleiss Kappa, weighted Cohen Kappa for pairs, and a point-biserial correlation between level stability and distance from the nearest band boundary. Every reliability estimate was computed twice, once on all four runs and once on the reduced set, so that the effect of any exclusion is visible rather than assumed. Computation was performed in Python using NumPy, pandas, and SciPy.

RESULTS AND DISCUSSION

What four observations revealed about the first run

Table 1 reports composite descriptives by run. Run 1 sits roughly two and a half points above the other three, which are almost identical to one another.

Table 1. Composite score by run (n = 100)

Run	Time	Mean	SD	Median	Minimum	Maximum
Run 1	09:33	60.49	8.81	59.8	40.5	86.9
Run 2	11:41	57.48	8.24	56.9	38.3	80.7
Run 3	13:21	58.00	7.95	58.0	32.8	76.7
Run 4	18:29	58.03	7.87	58.7	40.3	87.3

The difference is systematic. Run 1 exceeds the mean of the later three by 2.65 points, with a ninety-five per cent confidence interval of 1.94 to 3.36 and a standardised effect of 0.74 ($p < 0.001$). Repeated-measures Analysis of Variance across Runs 2, 3, and 4 returns $F(2, 198) = 0.904$, $p = 0.407$, with partial eta squared of 0.009. Those three runs cannot be told apart, and the effect size is negligible.

Table 2 shows the pattern holding without exception across all six pairwise comparisons.

Table 2. Pairwise agreement on the composite score

Pair	Bias	p	Mean abs. diff.	Max diff.	Pearson r	ICC(2,1)
Run 1 vs Run 2	-3.01	<0.001	4.44	20.0	0.841	0.792
Run 1 vs Run 3	-2.49	<0.001	3.80	13.6	0.887	0.846
Run 1 vs Run 4	-2.45	<0.001	4.14	15.5	0.861	0.821
Run 2 vs Run 3	+0.52	0.209	3.19	11.7	0.871	0.870
Run 2 vs Run 4	+0.55	0.252	3.69	14.8	0.824	0.822
Run 3 vs Run 4	+0.03	0.946	3.83	13.2	0.809	0.810

Our initial explanation was wrong, and reporting the error is more useful than concealing it. With only three runs available we had attributed the shift to the organisation metadata field, since the value attached to Run 1 differed from that attached to Runs 2 and 3. Run 4 was executed specifically to test that hypothesis, carrying the same field value as Run 1. The hypothesis predicted a return to approximately 60.5. Run 4 landed at 58.03, statistically identical to Run 3, and the comparison between the two runs sharing a field value shows a difference of 2.45 points at $p < 0.001$. The metadata explanation was therefore eliminated by the data it predicted.

What remains is inference by elimination, and its limits should be stated plainly. Nothing under our control changed between runs: identical files, identical notebook, identical parameters, identical endpoints. The disturbance is confined to a single boundary in time, between 09:33 and 11:41, after which the system is stable for the rest

of the day. We cannot inspect the provider's serving infrastructure and therefore cannot demonstrate that the deployed model changed. Non-stationarity on the provider side is the most parsimonious account consistent with the evidence, and readers should treat it as the best available explanation rather than a verified cause.

Run 1 was not a failed run. All one hundred videos completed, language detection agreed with the other runs, and transcription output was effectively identical. Excluding it from the reliability estimates is a decision about analytic assumptions, not the reporting of a technical fault, and the distinction governs how the remaining coefficients should be read. A test-retest estimate assumes that repeated administrations differ only by random error; Run 1 violates that assumption, and an ICC computed across a set containing it averages over a shift while appearing to describe noise.

One participant illustrates the magnitude at stake. Composite scores across the four runs were 69.8, 49.8, 60.6, and 54.3, a range of twenty points and a level path from the fifth tier down to the third and back up to the fourth. This is the extreme case rather than the norm, since the mean four-run range was 7.14 points, but it shows what a single administration can produce for an individual student.

Composite reliability, and what the exclusion actually costs

Because exclusion decisions invite suspicion, Table 3 reports both sets side by side.

Table 3. Reliability of the composite score, with and without Run 1

Statistic	Runs 2 to 4	All four runs	Effect of exclusion
ICC(2,1), absolute agreement	0.835	0.826	+0.009
Lower bound of 95% CI	0.779	0.800	-0.021
ICC(2,k), absolute agreement	0.938	0.950	-0.012
ICC(3,1), consistency	0.835	0.847	-0.012
Standard Error of Measurement	3.26	3.21	+0.05 worse
MDC95, single run	9.04	8.91	+0.13 worse
MDC95, mean of runs	5.22	4.45	+0.77 worse
Fleiss Kappa, levels	0.454	0.456	-0.002

Scores and thresholds in points on a 0 to 100 scale. Average-measure indices are not strictly comparable across the two sets because they rest on three and four observations respectively.

Removing Run 1 raises the single-measure coefficient by nine thousandths and worsens five of the eight indices reported, including the lower confidence bound that Koo and Li (2016) treat as the basis for interpretation. The exclusion cannot be read as an attempt to improve the result. It is justified by the assumption it protects, not by the number it produces, and we would draw the same substantive conclusions from either set.

Taking the reduced set as the primary estimate, a single run reaches a lower bound of 0.779 and qualifies as good, while the mean of three reaches 0.914 and qualifies as excellent. Absolute-agreement and consistency coefficients coincide to four decimal places within this set. The two indices diverge only in the presence of systematic bias between raters, so their coincidence independently confirms, without reference to the Analysis of Variance, that no residual drift remains.

The Standard Error of Measurement carries the practical message. Run-to-run noise is 3.26 points on a hundred-point scale, and a single administration cannot distinguish two performances differing by less than roughly nine points. Averaging three runs lowers that threshold to 5.22 points. Within-participant dispersion across the three runs was modest: mean standard deviation 2.81 points, mean coefficient of variation 4.93 per cent, mean range 5.35 points, with half the sample varying by under five points and eighty-nine of one hundred by under ten. These figures compare favourably with agreement typically reported between trained human raters of public speaking, which supports the feasibility claim while leaving validity entirely open.

Dimension-level reliability

Table 4 disaggregates reliability by dimension, again reporting both sets.

Table 4. Reliability by dimension

Dimension	Weight	Mean	ICC(2,1) Runs 2-4	95% CI	ICC(2,1) all four	SEM
D1 Quality of Content	16	54.56	0.661	0.563 to 0.743	0.666	8.17

Dimension	Weight	Mean	ICC(2,1) Runs 2-4	95% CI	ICC(2,1) all four	SEM
D2 Quality of Argument	16	45.25	0.765	0.697 to 0.830	0.737	7.34
D3 Delivery	10	39.40	0.786	0.716 to 0.842	0.776	6.17
D4 Prosody	11	67.95	0.572	0.461 to 0.670	0.613	6.87
D5 Confidence Visual	11	73.20	0.867	0.819 to 0.903	0.855	3.69
D6 Humor	7	43.56	0.422	0.298 to 0.539	0.412	8.19
D7 Politeness	9	86.44	0.493	0.378 to 0.606	0.546	7.25
D8 Emotional Appeal	10	62.60	0.762	0.689 to 0.825	0.738	8.92
D9 Linguistic Quality	10	53.15	0.673	0.581 to 0.755	0.682	5.63

No dimension changes category under exclusion; every difference between the two columns falls inside the confidence interval.

The ordering is informative. Confidence Visual leads at 0.867, scored from sampled video frames against observable postural and gestural cues. Delivery, Quality of Argument, and Emotional Appeal follow in the good range. Humor at 0.422 and Politeness at 0.493 trail. Reliability broadly tracks the interpretive latitude each descriptor leaves open: where the rubric points at something countable or directly observable, repeated scoring converges, and where it asks the model to decide whether an utterance was funny or appropriately deferential, repeated scoring diverges.

Two dimensions need separate comment. Prosody is anomalous because its inputs are numerical acoustic measurements, yet its reliability sits at 0.572. The explanation lies

in sampling rather than judgement: the pipeline selects an analysis window from longer recordings, and different windows yield different acoustic profiles. This is a correctable design decision, and it shows that instability in an automated pipeline is not always attributable to the generative component.

Politeness shows a pronounced ceiling effect, with a mean of 86.44 and forty-seven per cent of scores at or above ninety. Its lower bands require active impoliteness, which is essentially unavailable in a prepared monologue, so the effective scoring range is compressed and the coefficient deflated by restricted variance rather than erratic judgement. Its ICC is the one that moves most under exclusion, from 0.546 to 0.493, which is itself a symptom of low variance rather than of anything Run 1 contributed.

Removing the two weakest dimensions would raise the headline figure slightly. We rejected that option. Humor and Politeness together carry sixteen of one hundred weight points and contribute a minority of composite error variance, so the gain is small. More importantly, they encode the pragmatic competences that distinguish culturally situated speaking from grammatically correct speaking. Deleting them because they are hard to measure would narrow the construct toward the formal, norm-referenced conception of good English that the decolonial critique identifies as the problem. Reporting their instability openly and constraining their interpretation is the defensible response.

Level classification and the band-width problem

Reliability at the level of reported bands is substantially weaker than at the level of scores.

Table 5. Level classification agreement, Runs 2 to 4

Statistic	Value
Identical level across three runs	49 of 100
Level range of one band	47 of 100
Level range of two bands	4 of 100
Fleiss Kappa across three runs	0.454
Weighted Kappa, quadratic, run pairs	0.680 to 0.784
Point-biserial r, stability by distance to boundary	0.564 ($p < 0.001$)

Fewer than half the participants received the same level in all three runs. The apparent tension with an ICC of 0.835 dissolves once the geometry of the reporting scale is examined. Levels occupy bands of ten points while the Minimal Detectable Change for a single run is 9.04 points, so the band is narrower than the noise it must resolve and misclassification near boundaries becomes a mathematical certainty independent of scoring quality. The point-biserial correlation of 0.564 confirms the mechanism directly: participants whose mean score sat 3.45 points from a boundary were stable, those at 1.87 points were not. Almost all disagreement stayed within adjacent bands, with only four cases spanning two.

The implication generalises beyond this instrument. Reporting granularity should be derived from measurement precision rather than chosen for presentational convenience. A system that reports fine-grained ordinal levels without publishing its Standard Error of Measurement presents noise as classification. For an assessment used on students that is not cosmetic: a learner told they are at level four rather than level five has been given a distinction the instrument cannot support.

Implications for responsible deployment

Three conclusions follow for practice. Aggregation rather than tuning is the route to stability. Prior experimentation on this pipeline, including lower temperature settings, alternative scoring models, and revised weighting schemes, failed to reduce drift, whereas averaging three runs raised reliability from good to excellent and cut the detectable-change threshold nearly in half. Stochastic variation appears to be a property of generative scoring rather than a defect awaiting an engineering fix.

Reported output should also match measured precision. Composite scores are defensible for developmental feedback, dimension scores only where reliability is adequate, and fine-grained level labels not at all from a single administration. Reliability studies of pipelines built on hosted models should further be designed with at least three post-baseline observations and should publish the timing of every run, since a two-run design cannot separate ordinary noise from a step change and will silently misattribute the latter to the former.

None of this establishes that the instrument measures the right thing. It establishes only that it measures something consistently, and the distinction is where the decolonial argument bites. A stable instrument calibrated to Inner Circle norms would reproduce those norms with great precision, and its stability would make the resulting judgement harder to contest rather than easier. Reliability reporting supplies the footing on which contestation becomes possible. A user who knows that a score carries an error of 3.26 points, that humour scoring is barely half reliable, and that level labels are unstable near boundaries can treat the report as evidence to be interrogated rather than a verdict to be accepted. Publishing the weaknesses of one's own instrument redistributes interpretive authority, and it is the minimum owed to speakers whose English is already assessed against a standard they did not set.

CONCLUSIONS

Repeated scoring of one hundred English speech videos across four runs shows that a pipeline combining Automatic Speech Recognition with Large Language Model scoring achieves good single-run and excellent three-run reliability at the composite level, with Intraclass Correlation Coefficients of 0.835 and 0.938 respectively and a Standard Error of Measurement of 3.26 points. Dimension reliability varies systematically with interpretive demand, from 0.867 for visually anchored confidence to 0.422 for humour. Level classification is unstable because the reporting band is narrower than the minimal detectable change, a structural limitation rather than a computational one.

The methodological contribution of the four-run design lies in separating two sources of variation that a two-run study conflates. Random scoring noise and a step change in a hosted model produce similar disagreement between any two administrations and entirely different patterns across four. We also report a rejected hypothesis: the metadata explanation we initially favoured was eliminated by the fourth run, and the remaining account, provider-side non-stationarity, is an inference by elimination that we cannot verify directly.

This is a feasibility investigation, and its limits are firm. It establishes internal consistency, not validity. Whether the scores correspond to defensible judgements of speaking competence remains untested, and a concurrent validity study with multiple human examiners over the same corpus is the necessary next stage, followed by expert

appraisal of the narrative feedback the system produces. A related limit concerns the scoring mechanism itself. Because the generative model is guided by a rubric specification rather than fine-tuned or calibrated against human-rated speech samples, the reliability reported here describes the consistency of a prompt-model pairing rather than the accuracy of a trained scoring model, and calibration of the rubric against expert human ratings remains a task for future work alongside the validity study proposed above. Until that evidence exists, output from this and comparable systems should serve reflective and developmental purposes and not selection or summative certification. The wider claim is independent of the particular instrument: for automated assessment deployed on speakers of English outside the Inner Circle, transparent reliability reporting is a condition of legitimacy.

REFERENCES

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., ... Testoni, A. (2025). LLMs instead of human judges? A large-scale empirical study across 20 NLP evaluation tasks. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Canagarajah, A. S. (1999). Interrogating the "native speaker fallacy": Non-linguistic roots, non-pedagogical results. In G. Braine (Ed.), *Non-native educators in English language teaching* (pp. 77–92). Lawrence Erlbaum.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33*(2), 185–209. https://doi.org/10.2307/3587717

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Holliday, A. (2006). Native-speakerism. *ELT Journal, 60*(4), 385–387. https://doi.org/10.1093/elt/ccl030

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge University Press.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

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. https://doi.org/10.1016/j.jcm.2016.02.012

Phillipson, R. (1992). *Linguistic imperialism*. Oxford University Press.

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competence rubric. *Communication Education, 61*(3), 205–233. https://doi.org/10.1080/03634523.2012.670709

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

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems, 36*. https://doi.org/10.48550/arXiv.2306.05685