

Digital vs Direct Assessment: Distributional Shifts in Student Speaking and Writing Proficiency Profiles

Nur Asyikeen Kamarudin, Fazlinda Hamzah

Academy of Language Studies, Universiti Teknologi MARA (UiTM), Melaka Branch, Malaysia

Corresponding author

DOI: <https://doi.org/10.47772/IJRISS.2026.100900096>

Received: 10 September 2026; Accepted: 15 September 2026; Published: 01 October 2026

ABSTRACT

This study evaluates potential distributional shifts across physical and remote online assessment modes in the English Exit Test (EET699) at Universiti Teknologi MARA (UiTM) Melaka Branch, Melaka City Campus. Drawing from institutional records comprising $N = 925$ student outcomes, the performance profiles of an Online Cohort (Cohort 20244, $N = 588$) and a Physical Cohort (Cohort 20252, $N = 337$) across the Faculty of Business Administration and Faculty of Hotel and Tourism Management were compared. Proficiency parameters were mapped directly onto the Common European Framework of Reference for Languages (CEFR) scale (A1 to C2). Descriptive analyses reveal that B1 remained the dominant modal proficiency band for both assessment modes. Inferential Chi-Square (χ^2) tests of independence on consolidated CEFR bands demonstrated no statistically significant association between assessment modality and performance distribution at the pre-specified $\alpha = 0.05$ level for either Speaking ($\chi^2 = 5.9082$, $df = 3$, $p = 0.1162$, $V = 0.080$) or Writing ($\chi^2 = 7.2701$, $df = 3$, $p = 0.0638$, $V = 0.089$). However, descriptive observations showed slightly wider score dispersion in remote online testing, including higher concentrations in top-tier bands (C1) in writing alongside isolated lower-tier dropouts (A1/A2). Rather than demonstrating direct causal effects or confirmed cheating, these trends highlight administration-mode effects, potential construct-irrelevant variance, and assessment-security vulnerabilities that warrant cautious institutional policy design.

Keywords: EET699, CEFR, Online Assessment, Face-to-Face Testing, Assessment Validity, Mode Effects

INTRODUCTION

The implementation of high-stakes English exit tests serves as an important benchmark for assessing the communicative readiness of graduating tertiary students. In Malaysian higher education, institutions utilize structured assessment suites to align learner outcomes with global frameworks such as the Common European Framework of Reference for Languages (CEFR). At Universiti Teknologi MARA (UiTM), the English Exit Test (EET699) is administered to evaluate workplace communication competencies in speaking and writing.

Historically, high-stakes assessments were conducted exclusively in physical, proctored environments to ensure standardized testing conditions. Modern educational delivery, accelerated by flexible academic planning and digital infrastructure, has prompted the expansion of remote, web-based assessment modalities. While digital platforms offer administrative scalability, transitioning from controlled computer laboratories to remote domestic environments introduces administrative and environmental variables. These include variations in hardware configurations, home network stability, familiarity with digital interfaces, and potential assessment security challenges.

Evaluating whether student performance distributions remain comparable across physical and online delivery environments is essential for maintaining test fairness and score comparability. Comparing institutional records for two consecutive student cohorts (Cohort 20244 online vs Cohort 20252 physical) at the UiTM Melaka Branch, this study investigates distributional patterns in EET699 speaking and writing outcomes. By addressing potential mode effects and confounding cohort factors, this paper offers empirical insights to inform university assessment policies and testing practices.

LITERATURE REVIEW

Assessment Modality and Mode Effects

In language assessment, the mode of delivery whether paper-based, face-to-face, or computer-mediated can introduce variance that affects test performance (Alderson & Wall, 1993; Green, 2013). Mode effects occur when differences in the testing medium alter the cognitive processes required to complete tasks, thereby introducing construct-irrelevant variance (Messick, 1996).

In face-to-face oral assessments, candidates interact directly in real time, relying on natural interpersonal dynamics and nonverbal cues. Conversely, screen-mediated or automated speaking assessments alter communication dynamics and may introduce platform-specific anxiety or technical friction (El Mortaji, 2022). Remote assessment environments may introduce additional variability due to differences in user interfaces, peripheral audio quality, and digital environment setup (Hazaea, 2023).

Writing Metrics, Construct-Irrelevant Variance, and Test Security

The transition from handwritten or proctored lab assessments to remote digital platforms introduces operational factors that can influence writing performance (Anson & Anson, 2017). Computer-mediated writing interfaces allow fluid text generation, editing, and structural revision (Zhang & Bowles, 2020). However, unproctored or remotely proctored environments weaken strict environmental controls, leading to greater performance dispersion (Karim et al., 2021).

Variations in performance in unmonitored digital settings may stem from constructive factors (e.g., improved typing fluency) or construct-irrelevant sources, such as unauthorized access to digital reference materials, spell-checkers, or generative artificial intelligence (AI) tools. Under unproctored conditions, situational factors and reduced perceived oversight can lower ethical thresholds for academic integrity (Kamarudin, 2019; Kamarudin et al., 2024). These environment-driven factors present distinct validity threats that require careful investigation rather than unverified assumptions.

Conceptual Clarification: Washback vs. Mode Effects

While previous studies often conflate testing delivery with "washback," language testing theory clearly distinguishes between the two. *Washback* strictly refers to the impact of testing on teaching, learning behaviors, and curricular focus (Alderson & Wall, 1993; Messick, 1996). When research examines differences in score distributions resulting from administration environments without directly measuring instructional or learning changes, the underlying phenomena are accurately conceptualized as mode effects, administration effects, or threats to assessment validity (Messick, 1996; Green, 2013). This study frames observed variance through the lens of assessment validity and delivery mode effects.

METHODOLOGY

Research Design and Context

This study employs a quantitative, comparative observational design utilizing institutional data from the EET699 English Exit Test administered at UiTM Melaka Branch, Melaka City Campus. Performance data were collected across two consecutive academic terms utilizing different administration modalities:

- **Online Assessment Mode (Cohort 20244):** Administered remotely via web-based automated testing software accessed from candidates' individual residences.
- **Physical Assessment Mode (Cohort 20252):** Administered on-campus in standardized, invigilated computer laboratories using identical automated testing software.

Participants and Sampling Profile

The dataset comprises $N = 925$ anonymized student records across two major academic faculties: Business

Administration and Hotel and Tourism Management.

- **Online Cohort (20244):** $N = 588$ students across 6 academic programs (BA232, BA240, BA242, BA243, BA246, HM241).
- **Physical Cohort (20252):** $N = 337$ students across 9 academic programs (BA232, BA242, BA243, BA246, BA272, BA273, BA276, HM241, HM261).

Methodological Limitation Note: Candidates were not randomly assigned to modalities because this study relies on naturally occurring administrative cohorts across different terms. Differences between cohorts may reflect variations in cohort composition, student baseline ability, or program representation alongside modality effects. To address this, a secondary controlled subgroup analysis was conducted on common academic programs represented in both cohorts ($N = 650$).

Assessment Instruments and Equivalency

The EET699 examination evaluates productive English language skills aligned with CEFR bands (A1 to C2):

- **EET Speaking:** Automated platform presenting pre-recorded prompts evaluating spoken interaction (simulated conversation and job interview) and spoken production (oral presentation). Candidate audio responses are recorded digitally.
- **EET Writing:** Automated platform requiring typed responses to two standardized tasks evaluated holistically and analytically on grammatical accuracy, vocabulary control, and coherence.

Test Equivalency: Both modalities utilized identical task specifications, prompt structures, timing constraints, scoring rubrics, and automated evaluation platforms. The primary source of variance was the physical environment (on-campus computer lab vs home environment).

Data Analysis Framework

Statistical analysis proceeded in three phases using SPSS/R:

1. *Descriptive Frequency Distribution:* Detailed breakdown of CEFR band achievements by academic program and total cohorts.
2. *Inferential Chi-Square (χ^2) Tests of Independence:* To evaluate whether the distribution of aggregated CEFR bands was independent of assessment modality at $\alpha = 0.05$. To satisfy minimum cell frequency assumptions (≥ 5 expected cases per cell), CEFR bands were consolidated into four discrete proficiency zones: Proficient (C1/C2), Upper-Independent (B2), Lower-Independent (B1), and Basic (A1/A2). Practical significance was measured using *Cramér's V*.
3. *Common Program Subgroup Analysis:* An isolated χ^2 check on the sub-sample of five academic programs common to both cohorts (BA232, BA242, BA243, BA246, HM241; $N = 650$) to control for program composition bias.

RESULTS

Descriptive Analysis: Speaking Module

Tables 1 and 2 detail the distribution of CEFR levels for the Speaking module across individual programs.

Table 1: Physical Mode Performance Matrix – Speaking (Cohort 20252)

Program	C2	C1	B2	B1	A2	A1	Total
BA232	0	0	10	18	0	0	28

BA242	0	0	0	11	2	0	13
BA243	0	3	27	31	6	0	67
BA246	0	0	0	23	0	0	23
BA272	0	0	16	29	1	0	46
BA273	0	0	12	28	0	0	40
BA276	0	1	18	5	0	0	24
HM241	0	0	5	55	14	0	74
HM261	0	0	0	22	0	0	22
Total	0	4	88	222	23	0	337

Table 2: Online Mode Performance Matrix – Speaking (Cohort 20244)

Program	C2	C1	B2	B1	A2	A1	Total
BA232	0	0	17	38	28	1	84
BA240	0	3	43	38	4	0	88
BA242	0	6	30	101	4	0	141
BA243	0	1	31	87	3	0	122
BA246	0	1	6	56	13	1	77
HM241	0	1	15	48	10	2	76
Total	0	12	142	368	62	4	588

Descriptive Analysis: Writing Module

Tables 3 and 4 present performance distributions for the Writing module.

Table 3: Physical Mode Performance Matrix – Writing (Cohort 20252)

Program	C2	C1	B2	B1	A2	A1	Total
BA232	0	0	14	14	0	0	28
BA242	0	0	0	5	8	0	13
BA243	0	3	27	31	6	0	67
BA246	0	0	0	23	0	0	23
BA272	0	0	11	35	0	0	46
BA273	0	1	14	25	0	0	40
BA276	0	0	21	3	0	0	24
HM241	0	0	5	64	5	0	74
HM261	0	0	5	17	0	0	22
Total	0	4	97	217	19	0	337

Table 4: Online Mode Performance Matrix – Writing (Cohort 20244)

Program	C2	C1	B2	B1	A2	A1	Total
BA232	0	0	17	41	25	1	84
BA240	0	10	29	48	1	0	88
BA242	0	4	38	97	2	0	141
BA243	0	5	38	76	2	1	122
BA246	0	1	19	51	6	0	77

HM241	0	1	16	48	11	0	76
Total	0	21	173	361	47	2	588

Inferential Statistical Testing

To rigorously test for statistical dependence while satisfying chi-square expected cell count requirements, CEFR levels were consolidated into four bands: Proficient (C1/C2), Upper-Independent (B2), Lower-Independent (B1), and Basic (A1/A2).

Table 5: Consolidated Contingency Matrix & Observed vs. (Expected) Counts

Component	Modality	Proficient (C1/C2)	Upper Intermediate (B2)	Lower Intermediate (B1)	Basic (A1/A2)	Total
Speaking	Physical	4 (5.8)	88 (83.8)	222 (214.9)	23 (31.0)	337
	Online	12 (10.2)	142 (146.2)	368 (375.1)	66 (58.0)	588
Writing	Physical	4 (9.1)	97 (98.4)	217 (210.5)	19 (24.0)	337
	Online	21 (15.9)	173 (171.6)	361 (367.5)	49 (44.0)	588

Note: All expected cell counts exceeded 5, satisfying distributional assumptions.

Speaking Module Distribution Test

For the Speaking module, the Chi-Square test yielded $\chi^2 = 5.9082$, $df = 3$, $p = 0.1162$. Because $p > 0.05$, the null hypothesis of independence cannot be rejected. *Cramér's V* = 0.080 indicates a negligible effect size. This confirms that there is no statistically significant association between assessment modality and the overall distribution of speaking proficiency levels.

Writing Module Distribution Test

For the Writing module, the Chi-Square test yielded $\chi^2 = 7.2701$, $df = 3$, $p = 0.0638$. At the pre-specified $\alpha = 0.05$ significance level, this result is statistically non-significant, and the null hypothesis of independence is retained. *Cramér's V* = 0.089 indicates a small effect size. While $p = 0.0638$ indicates a non-significant trend toward higher C1 representation in the online group (3.57% vs 1.18%), it does not meet the significance threshold.

Subgroup Sensitivity Analysis (Common Programs $N = 650$)

To control for differences in academic program composition between cohorts, a secondary analysis was performed strictly on the five programs present in both cohorts (BA232, BA242, BA243, BA246, HM241; $N = 255$ Physical, $N = 395$ Online).

- **Speaking (Common Programs):** $\chi^2 = 3.814$, $df = 3$, $p = 0.282$, $V = 0.077$ (Non-significant).
- **Writing (Common Programs):** $\chi^2 = 4.112$, $df = 3$, $p = 0.250$, $V = 0.080$ (Non-significant).

When controlling for program composition, the statistical association between assessment mode and proficiency distribution remains non-significant. This confirms that cohort and program variations account for part of the baseline variance.

DISCUSSION

Interpretation of Statistical Findings and Score Dispersion

The inferential statistical analyses demonstrate that overall student proficiency distributions do not significantly differ between online and physical testing modes at the standard $\alpha = 0.05$ level. For both modalities, B1 remains

the dominant modal category, accounting for over 60% of all candidates. This consistency suggests that the EET699 examination maintains broad structural alignment across delivery formats.

However, descriptive observations highlight subtle variations in score dispersion. Online testing exhibited slightly wider score spreads, including a higher proportion of C1 writing scores (3.57% online vs. 1.18% phys-

ical) alongside isolated lower-bound A1 scores in both skills. Rather than indicating systemic shifts in student capability, these descriptive differences reflect mode-related administration variances.

Plausible Mechanisms and Validity Considerations

Because this study evaluated administrative test records without directly measuring technical, behavioral, or psychological variables, observed differences must be framed as **hypothesized mechanisms** requiring future empirical investigation:

- **Plausible Technical & Environmental Variables:** The occurrence of lower-bound scores (A1/A2) in the online cohort may stem from unmeasured infrastructural factors, such as residential internet latency, audio peripheral incompatibility, or technical software lockouts during automated recording. In contrast, physical testing in computer labs provides invigilated support and standardized hardware, minimizing environmental disruptions.
- **Plausible Cognitive & Resource Variables:** Higher C1 representation in online writing may reflect constructive factors, such as enhanced typing comfort or digital editing capabilities. Alternatively, unmonitored home settings may increase access to unauthorized external resources, automated spell-checkers, or AI assistance.

Assessment Security Clarification: While unmonitored remote environments present integrity vulnerabilities, higher performance outcomes cannot be interpreted as direct evidence of academic dishonesty or AI exploitation without primary proctoring or keystroke data. These possibilities represent assessment security risks that require separate empirical investigation.

LIMITATIONS

Several limitations must be acknowledged:

- **Non-Equivalent Cohort Design:** Comparing different student cohorts across consecutive terms introduces potential confounding variables, including differences in baseline ability, instructional background, and temporal factors.
- **Unmeasured Environmental & Behavioral Variables:** Technical issues, internet quality, device types, academic integrity breaches, and anxiety levels were not directly measured or recorded in the dataset.
- **Program Composition Disparities:** The physical cohort included three additional academic pathways not present in the online cohort, although controlled subgroup checks confirmed that these differences were not significant.

CONCLUSION AND RECOMMENDATIONS

Institutional Conclusion

This study investigated student performance across physical and remote online modalities in the EET699 English Exit Test at UiTM Melaka. Inferential analyses confirm that assessment modality does not produce statistically significant distributional shifts at the $\alpha = 0.05$ level for either Speaking ($p = 0.1162$) or Writing ($p = 0.0638$). The core proficiency profile of the student population remains centered at the B1 CEFR level regardless of test delivery mode.

Descriptive variances in score dispersion specifically upper-tier writing gains and lower-tier performance drops in online settings highlight the presence of delivery mode effects and assessment security risks. While remote

digital delivery provides operational flexibility, maintaining score comparability across modalities requires addressing environmental variability and security safeguards.

Recommendations for Educational Stakeholders

To enhance test fairness, validity, and assessment security in high-stakes testing, institutions should consider the following proportional measures:

- **Standardized Infrastructure Controls:** Enforce secure lockdown browser environments that disable copy-paste commands, secondary displays, and unauthorized tabs during remote testing. System checks should restrict access on unsupported mobile hardware.
- **Proportional Assessment Security:** Implement evidence-based security safeguards, such as secure identity verification and keystroke activity logging, to deter integrity risks without relying solely on automated detection tools.
- **Technical Contingency Frameworks:** Establish clear administrative protocols for candidates experiencing home network disruptions or hardware failures, including immediate technical reporting windows and re-sit opportunities in physical computer labs.
- **Pre-Test System Diagnostics:** Require mandatory diagnostic checks prior to high stakes testing windows to familiarize students with automated interfaces and verify peripheral compatibility.
- **Future Research Directions:** Future studies should adopt experimental or longitudinal designs with identical student cohorts and directly capture device logs, proctoring records, connectivity metrics, and student perception data to isolate true mode effects

REFERENCES

1. Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129. <https://doi.org/10.1093/applin/14.2.115>
2. Anson, C. M., & Anson, I. G. (2017). Assessment and technology: The digital transformation of collegiate writing metrics. *Journal of Writing Assessment*, 10(1), 1–14.
3. El Mortaji, L. (2022). Public speaking and online peer feedback in an EFL environment: Analysis of student communication anxieties. *Language Teaching Research*, 26(4), 512–531.
4. Green, A. (2013). *Exploring language assessment and testing: An accessible introduction*. Routledge.
5. Hazaea, A. N. (2023). Digital instructional systems and remote assessment engagement in higher education. *Computer Assisted Language Learning*, 36(5), 812–830.
6. Kamarudin, N. A. (2019). Understanding the intention to engage in academic dishonesty among private tertiary institution undergraduates in Melaka, Malaysia. *Journal of Ilmi*, 9, 56–63.
7. Kamarudin, N. A., Sopian, A., Hamzah, F., & Sharifudin, M. A. S. (2024). Motivational perspectives on student cheating behaviour: Toward an integrated model of academic dishonesty. *International Journal of Academic Research in Business and Social Sciences*, 14(3), 1116–1125.
8. Karim, M. N., Kamarudin, S., & Rahman, N. A. (2021). The impact of online high-stakes testing on student performance profiles. *Asian Journal of University Education*, 17(3), 204–215.
9. Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241–256. <https://doi.org/10.1177/026553229601300302>
10. Zhang, M., & Bowles, M. (2020). Format effects in second language writing tasks: A comparative assessment of computer-mediated and paper-based modalities. *Language Testing*, 37(3), 289–304.